CAMILA CAMPOS MANTELLO

"MAPEAMENTO GENÉTICO MOLECULAR EM *HEVEA BRASILIENSIS*"

"*GENETIC LINKAGE MAPPING IN HEVEA BRASILIENSIS*"

CAMPINAS

2014

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Biologia

CAMILA CAMPOS MANTELLO

"MAPEAMENTO GENÉTICO MOLECULAR EM *HEVEA BRASILIENSIS*"

"GENETIC LINKAGE MAPPING IN HEVEA BRASILIENSIS"

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Genética e Biologia Molecular na área de Genética Vegetal e Melhoramento

Thesis presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Genetics and Molecular Biology, in the area of Plant Genetics and Genetic Breeding

Orientadora/ Supervisor: Profa. Dra. Anete Pereira de Souza

Co-orientador/ Co-supervisor: Dr. Antonio Augusto Franco Garcia

Este exemplar corresponde à versão final da tese defendida pela aluna Camila Campos Mantello e orientada pela Profa. Dra. Anete Pereira de Souza.

Profa. Dra. Anete Pereira de Souza

CAMPINAS

2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

M318m

Mantello, Camila Campos, 1985-
Mapeamento genético molecular em *Hevea brasiliensis* / Camila Campos Mantello. – Campinas, SP : [s.n.], 2014.

Orientador: Anete Pereira de Souza.
Coorientador: Antonio Augusto Franco Garcia.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Seringueira. 2. Mapeamento cromossômico. 3. Transcriptoma. 4. RNA-seq. 5. Marcadores moleculares. I. Souza, Anete Pereira de,1962-. II. Garcia, Antonio Augusto Franco. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Genetic linkage mapping in *Hevea brasiliensis*
Palavras-chave em inglês:
Rubber plants
Chromosome mapping
Transcriptome
RNA-seq
Molecular markers
Área de concentração: Genética Vegetal e Melhoramento
Titulação: Doutora em Genética e Biologia Molecular
Banca examinadora:
Anete Pereira de Souza [Orientador]
Gabriel Rodrigues Alves Margarido
Rafael Vasconcelos Ribeiro
Adonias Castro Virgens Filho
Mirian Perez Maluf
Data de defesa: 03-07-2014
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, **3** de **julho** de **2014**

BANCA EXAMINADORA

Profa. Dra. Anete Pereira De Souza (orientadora)

Assinatura

Prof. Dr. Gabriel Rodrigues Alves Margarido

Assinatura

Prof. Dr. Rafael Vasconcellos Ribeiro

Assinatura

Dr. Adonias Castro Virgens Filho

Assinatura

Dra. Miriam Perez Maluf

Assinatura

Dra. Maria Marta Pastina

Assinatura

Prof. Dr. Marcelo Menossi Teixeira

Assinatura

Prof. Dr. Marceo Carnier Dornelas

Assinatura

Resumo

Aproximadamente 2.500 espécies são conhecidas por produzirem borracha natural e a seringueira, [*Hevea brasiliensis* (Willd. ex Adr. de Juss.) Muell-Arg.], espécie nativa da Amazônia e pertencente ao gênero *Hevea*, é a maior fonte de borracha natural do mundo. A borracha natural é matéria prima para fabricação de mais de 40.000 produtos tendo importância fundamental na indústria de pneus. Apesar de a região Amazônica oferecer condições climáticas adequadas para seu crescimento e desenvolvimento, esta área também possui condições favoráveis à ocorrência do mal-das-folhas (*Microcyclus ulei* P. Henn v. Arx), doença também conhecida como mal sulamericano das folhas (*South American Leaf Bligth* – SALB). Dessa maneira, a heveicultura se expandiu para áreas de escape que propiciam novas condições de estresse, limitando o seu crescimento e a produção de látex. O melhoramento genético vem buscando cultivares adaptados a estas regiões de escape, porém o ciclo de melhoramento da seringueira é longo e não permite o rápido desenvolvimento de novos cultivares. O desenvolvimento de ferramentas na biologia molecular permite o melhor entendimento da espécie e pode diminuir o tempo gasto nos ensaios de campo. O presente trabalho desenvolveu novos marcadores microssatélites (SSRs) a partir de bibliotecas enriquecidas em microssatélites. A caracterização destes marcadores mostrou a alta variabilidade alélica dentro de *H. brasiliensis* e o teste de transferibilidade dos SSRs em outras seis espécies do gênero *Hevea* mostrou alelos exclusivos para as mesmas e taxas de amplificação superior a 80%. Com o objetivo desenvolver novos marcadores SSRs e *single nucleotide polymorphisms* (SNPs) em larga escala, foi sequenciado na plataforma Illumina GAIIx o transcriptoma de painel de dois cultivares importantes para a heveicultura (GT1 e PR255). A montagem e a caracterização do

transcriptoma permitiu o melhor entendimento da dinâmica do transcriptoma em *H. brasiliensis* e identificou novos transcritos para a espécie. As sequências do transcriptoma foram submetidas à busca de SSRs e SNPs. No transcriptoma, foram identificados 1.709 novas sequências contendo SSRs para seringueira, a uma frequência de um SSR a cada 2,8 kb. Já a busca de SNPs identificou 404.114 SNPs com frequência de um SNP a cada 125 pb. Através da anotação no *Kyoto Encyclopedia of Genes and Genomes* (KEGG), foram identificadas sequências anotadas a todas as enzimas referentes às duas vias de síntese de látex (mevalonato - MVA e C-metileritritol 4-fosfato -MEP). Apesar de as vias MVA e MEP serem muitos estudadas, esta foi a primeira vez que SNPs foram identificados e validados. Os marcadores SSRs e SNPs foram então mapeados em uma população segregante F1. O mapa genético obtido contém 383 marcadores mapeados em 20 grupos de ligação. Neste trabalho foram desenvolvidos 52 SSRs e 51 SNPs do total de marcadores mapeados. Como o número de grupos de ligação esperado é 18 (2n=36), conclui-se que o mapa genético obtido mostra que ainda há uma cobertura incompleta do genoma. Devido à alta frequência de SNPs no genoma, o desenvolvimento de novos marcadores poderá saturar o mapa de forma homogênea, permitindo o agrupamento dos marcadores nos 18 grupos de ligação esperados.

Abstract

Approximately 2.500 species are known to produce natural rubber. *Hevea brasiliensis* (Willd. ex Adr. de Juss.) Muell-Arg. also known as rubber tree is a species native to the Amazon rainforest and is the largest source of natural rubber in the world. Natural rubber has been used in more than 40,000 products and has great importance in the tire industry. Although the Amazon rainforest offers optimal conditions for growth and rubber yields due to its warm and humid climate, this region also provides optimal conditions for the fungus *Microcyclus ulei* P. Henn v. Arx which causes the South American Leaf Blight (SALB) disease. Thus, rubber tree plantations have expanded to escape areas that provides new stress conditions limiting their growth and latex production. The rubber tree breeding is trying to create a new cultivar that is resistant to these new conditions but the rubber tree cycle breeding is long and does not allow a rapid cultivar development. Thus, molecular biology techniques could provide a greater knowledge of *H. brasiliensis* genetic and could optimize field evaluation and, thus, reduce the time and area required for experiments. The present work developed new microsatellites (SSRs) markers for rubber tree from genomic enriched libraries. The new microsatellites were characterized and demonstrated a high allelic variability within *H. brasiliensis* genotypes. The transferability rate in other six species of the genus *Hevea* was greater than 80%. To develop new SSRs and single nucleotide polymorphisms (SNPs) markers, the panel transcriptome from two important cultivars (GT1 and PR255) was sequenced in Illumina GAIIx platform. The transcriptome obtained allowed a better knowledge about *H. brasiliensis* transcriptome and identified new transcripts for rubber tree public database. The sequences were submitted to a SSR and SNP search. The SSR frequency was one SSR each 2.8 kb and it was identified 1,709 new SSRs in new sequences for

rubber tree database. A total of 404,114 putative SNPs were detected with a frequency of one SNP every 125 bp. Through Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation, it was identified contigs corresponding to all the enzymes of mevalonate (MVA) and -C-methyl-D-erythritol 4-phosphate (MEP). Despite MVA and MEP pathways are being very well studied, since they are directly involved to rubber biosynthesis, this is the first time that molecular markers have been developed for such important pathways. The SSRs and SNPs developed were mapped in a full-sib population. The genetic linkage map has 383 molecular markers distributed in 20 linkage groups. This project contributed with 52 SSRs and 51 SNPs of the total mapped markers. Although the expected number of linkage groups are 18 (2n=36), the new genetic linkage map still has an incomplete coverage of the genome. Due to the high frequency of the SNPs in the genome, the development of new markers can saturate this map homogeneously.

Sumário

Agradecimentos xiii

Organização da tese xv

Introdução 1

Revisão bibliográfica 5

Gênero *Hevea* 7

Contexto da cultura da seringueira 9

Marcadores Moleculares 10

RNA-seq 13

Mapeamento genético 15

Objetivos 19

Objetivo Geral 21

Objetivos Especificos 21

Capítulo I 23

"Microsatellite marker development for the rubber tree (*Hevea brasiliensis)*: characterization and cross-amplification in *Hevea* wild species."

Capítulo II 43

"De novo assembly and transcriptome analysis of the rubber tree *(Hevea brasiliensis*) and SNP marker development for rubber biosynthesis pathways"

Capítulo III 81

"Saturação de um mapa-genético molecular em seringueira (*Hevea brasiliensis*) com marcadores SNPs."

Considerações Finais 99

Conclusões 107

Perspectivas 111

Literatura Citada 115

Anexo I 123

Anexo II 139

AGRADECIMENTOS

Aos meus pais Marcos e Cecilia e ao meu irmão Guilherme por todas as conversas, por estarem sempre ao lado e me ajudarem a superar todos os obstáculos que a vida impõe. Vocês são um exemplo de vida para mim e não tenho palavras para agradecer e demonstrar todo o carinho que sinto por vocês.

À minha avó Anna, por seu exemplo de superação e aos meus tios e padrinhos Zezé e Alair por sempre estarem ao meu lado.

Ao meu namorado Felipe, que ao longo destes 3 anos me ajudou muito e esteve sempre ao meu lado nos momentos em que mais precisei. Muito Obrigada!

À Professora Anete Pereira de Souza, por me apresentar o fascinante mundo da genética de plantas. Sua alegria e interesse por novas descobertas é contagiante. Obrigada por abrir as portas de seu laboratório em 2007 durante a minha graduação e por acreditar no meu potencial durante todos estes anos. Foram inúmeros os ensinamentos e sua luta pela pesquisa e ética é um grande exemplo para mim.

Ao Professor A. Augusto Franco Garcia, por ter me ensinado a genética estatística, que apesar da sua complexidade sempre foi passada de forma clara e didática. É uma honra tê-lo como meu co-orientador.

Ao Professor Paulo Gonçalves, por confiar e acreditar no meu trabalho. Todo seu conhecimento e paixão pela seringueira é uma inspiração para mim.

Ao Professor Renato Vicentini, por me ajudar a desvendar o mundo da bioinformática. Seus ensinamentos foram fundamentais para o trabalho.

À Lívia, que me acompanha desde a minha iniciação científica e com quem eu compartilho todos os momentos do laboratório e em campo. Obrigada pela paciência, compreensão e pelo companheirismo no trabalho. Ao longo destes 7 anos aprendi muito com você e se hoje eu encerro mais uma etapa, saiba que você teve papel fundamental para isto acontecer!

À Carla, que com seu jeito divertido de ser, compartilhou comigo suas experiências em biologia molecular e me ensinou muito sobre extrações de RNA e sequências expressas em seringueira. Muito Obrigada!

Ao Benicio, que trabalha com cana-de-açúcar, mas que também “puxei” um pouco para o mundo da seringueira e me ajudou com as ferramentas de bioinformática e com novas ideias.

À Melina, pela amizade cultivada ao longo dos anos, pela troca de experiências na bancada e por disponibilizar seu tempo para me ensinar a usar o Sequenom e finalizar meu doutorado.

Ao Rodrigo Gazaffi e ao Marcelo Mollinari, por me ajudarem com as análises dos marcadores e com o mapa genético e por esclarecer minhas dúvidas quando precisei.

Aos meus colegas e amigos do LAGM que me proporcionaram um ambiente de trabalho tão agradável e momentos de muitas risadas em todos estes anos. Todos vocês tiveram contribuição significativa para meu conhecimento e crescimento.

Ao Juver, que me ajudou tantas vezes no laboratório e sempre se preocupou comigo como se fosse um pai.

Aos técnicos do laboratório Danilo, Aline e Pathy por serem tão prestativos comigo.

Aos membros da pré-banca Dr. Vicent Le Guen e a Dra. Mirian Perez Maluf pela disponibilidade e ajuda na avaliação dos resultados do doutorado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa concedida e pelo apoio financeiro.

À Michelin e ao CIRAD por desenvolver e manter a população de mapeamento estudada.

Organização da tese

O presente trabalho está apresentado no formato de capítulos e as diferentes etapas desenvolvidas são apresentadas no formato de artigos científicos.

Inicialmente, será apresentada uma introdução geral e uma revisão bibliográfica que compreende os seguites tópicos: (1) Gênero *Hevea*, (2) Contexto da Cultura da Seringueira, (3) Marcadores Moleculares, (4) RNA-seq e (5) Mapeamento Genético, seguido dos objetivos gerais e específicos do trabalho.

O capítulo I apresenta o artigo "Microsatellite marker development for the rubber tree (*Hevea brasiliensis)*: characterization and cross-amplification in *Hevea* wild species." que foi publicado na revista *BMC Research Notes* (5: 329,2012). Este capítulo relata o desenvolvimento e caracterização de marcadores microssatélites para *H. brasiliensis* e a transferibilidade dos mesmos para outras espécies do gênero *Hevea*.

No capítulo II apresentamos a montagem do transcriptoma de seringueira a partir do método *de novo*. Este trabalho teve como objetivo montar e caracterizar o transcriptoma de painel, por ser o tecido onde se localizam os vasos latíceferos que são responsáveis por produzir e estocar látex, bem como contribuir para o conhecimento do conjunto de genes e sua expressão em seringueira. Pela primeira vez foram identificados e validados marcadores SNPs em genes envolvidos nas vias de síntese de látex. Os resultados obtidos foram relatados no artigo intitulado "De novo assembly and transcriptome analysis of the rubber tree *(Hevea brasiliensis*) and SNP marker development for rubber biosynthesis pathways" publicado na revista *Plos One* (9 (7):e102665).

O capítulo III ainda não submetido à publicação, mas também apresentado no formato de artigo, apresenta o mapa genético-molecular da população de mapeamento PB217xPR255 saturado com marcadores SNPs desenvolvidos a partir de sequências ESTs e do transcriptoma apresentado no capítulo II e microssatélites.

Os resultados obtidos e apresentados nos capítulos I, II, III serão discutidos nas Considerações Finais, seguida das Conclusões e Perspectivas.

Em anexo apresenta-se o mapa genético que foi publicado na revista *Plos One* (8: e61238.2013) no artigo intitulado “QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (*Hevea bradiliensis*) Evaluated in a Sub-Tropical Climate” para o qual o trabalho desenvolvido durante esta tese contribuiu com 52 marcadores SSRs apresentados no capítulo I.

Introdução

A seringueira [*Hevea brasiliensis* (Willd. ex Adr. de Juss.) Muell-Arg.] espécie nativa da Amazônia, pertencente ao gênero *Hevea* e à família Euphorbiaceae, sendo ela a principal fonte de borracha natural do mundo. A borracha é matéria-prima de grande importância em diversos setores industriais devido à sua qualidade superior ao produto similar sintético (CORNISH, 2001).

Apesar de o Brasil ser centro de origem e diversidade da seringueira, o país responde por apenas 1,5% da produção mundial, de modo que no ano de 2013 precisou importar 53,2% da borracha natural consumida no país (IRSG 2014).

A bacia Amazônica oferece condições ótimas para o cultivo da seringueira, porém a baixa produção da borracha natural no Brasil em regiões de clima quente e úmido se deve ao desenvolvimento do fungo *Microcyclus ulei*, causador do mal-das-folhas, principal doença da cultura, diminuindo bruscamente a produção do látex e limitando a produção da borracha nestas regiões (PUSHPARAJAH, 2001). Dessa forma, a heveicultura no Brasil tem se expandido para áreas de escape como, por exemplo, a região do planalto sul. As áreas de escape propiciam condições de estresse como baixas temperaturas, altitude elevada, vento e outras pragas. Devido a estes estresses existentes, o melhoramento genético vem buscando clones adaptados às regiões de escape por propriciarem novas condições de cultivo (PUSHPARAJAH, 1983).

Como o ciclo de melhoramento genético da seringueira leva de 20 a 30 anos para se concretizar (GONÇALVES et al., 1999) existe a necessidade de desenvolver novas técnicas de avaliação precoce, que possibilitem diminuir e otimizar as avaliações para essa cultura. As ferramentas desenvolvidas pela biologia molecular são de grande importância para o melhoramento genético da cultura. Elas podem facilitar o trabalho e diminuir o tempo gasto nas avaliações de novos clones, possibilitando a expansão do cultivo da seringueira em áreas de escape, e assim, aumentar a produtividade da seringueira no Brasil

Devido ao longo ciclo de melhoramento, a construção de mapas genéticos e o mapeamento de *Quantitative Trait Loci* (QTLs) buscam associar marcas genéticas a características fenotípicas. Desta forma, os mapa genéticos podem auxiliar na seleção de genótipos superiores, reduzindo o tempo dos programas de melhoramento da seringueira. Dessa forma, este trabalho teve como objetivo desenvolver marcadores moleculares do tipo

microssatélite e *single nucleotide polymorphisms* (SNPs) a partir de sequências gênicas e genômicas para serem utilizados na saturação de um mapa genético. O desenvolvimento destas ferramentas também proporcionaram novos conhecimentos genéticos e genômicos sobre a seringueira.

O presente trabalho teve como colaboradores o Centro de Seringueira do Instituto Agronômico de Campinas (IAC), a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), o Departamento de Genética da ESALQ/USP, a Unidade de Pesquisa Amélioration Génétique et Adaptation des Plantes do CIRAD em Montpellier-França e a empresa Michelin Ltda do Brasil.

Revisão Bibliográfica

Gênero *Hevea*

O gênero *Hevea* é membro da família Euphorbiaceae que inclui outros importantes gêneros de culturas tropicais, tais como *Ricinus* (mamona) e *Manihot* (mandioca). De acordo com Bouychou (1963), a primeira descrição de uma espécie pertencente ao gênero *Hevea* data de 1749 pelo botânico Fusée Aublet que descreveu o gênero *Hevea*, e dentro deste, a espécie *Hevea guianensis* (Aublet) .

Em 1824, Jussieu publicou o binômio *Siphonia brasiliensis*, sob autoria de Willdenow. Como o nome *Siphonia* não estava de acordo com o Código Internacional de Nomenclatura, o gênero *Hevea* foi restabelecido em 1865 por Müller, que propôs para a seringueira a combinação conhecida hoje como *Hevea brasiliensis* (PIRES; SECCO; GOMES, 2002). Os mais recentes estudos taxonômicos foram conduzidos por Baldwin (1947), Seibert (1947), Pires (1973) e Schultes (1977). Baldwin (1947), baseado no levantamento acurado de "Heveas" nativas da Amazônia, combinando as observações citológicas, chegou à conclusão que o gênero possuía nove espécies. No Peru, Seibert (1947) reconheceu 8 espécies. Os estudos de Schultes (1977) levaram-no a reconhecer nove espécies e quatro variedades, e finalmente Pires (1973) reconheceu 11 espécies com a inclusão no grupo da mais nova espécie, a *H. carmagoana*, encontrada na Ilha de Marajó.

As 11 espécies que compões os gênero são: *H. benthamiana* (Muell- Arg.), *H. brasiliensis*, *H. carmagoana* (Pires), *H. camporum* (Ducke), *H. guianensis* (Aublet), *H. microphylla*, *H. nitida* (Mart. ex-Muell-Arg.), *H. pauciflora* (Spruce ex-Benth), *H. rigidifolia* [(Spruce ex-Benth) Muell-Arg.,] *H. spruceana* [Benth. Muell.-Arg. Sin. *H.discolor* (Spruce ex-Benth.) Muell.-Arg.] e *H.paludosa* (Ule) (PRIYADARSHAN; GONÇALVES, 2003), dentre as quais destaca-se a *H. brasiliensis* por apresentar maior capacidade produtiva de látex.

Na região da bacia Amazônica, sete espécies são encontradas no alto Rio Negro, sendo esta região considerada o centro de origem do gênero. Já outras cinco espécies incluindo *H. brasiliensis* são encontradas mais ao sul da bacia Amazônica no alto Rio Madeira, que se localiza fora do centro de origem (SAHA; PRIYADARSHAN, 2012). O gênero *Hevea* ocorre no Brasil, Bolívia, Peru, Equador, Venezuela, Guiana Francesa, Suriname e Guiana. De todas as espécies

que compõem o gênero, apenas *H. microphylla* (Ule) não ocorre no Brasil (PRIYADARSHAN; GONÇALVES, 2003).

O gênero é pertence a classe das dicotiledôneas e possuem como características serem monoicas com flores unissexuadas, apétalas, pequenas, amarelas e dispostas em racemo. As folhas são longamente pecioladas e repartidas em três folíolos. O fruto é uma cápsula trilocular grande que geralmente apresenta três sementes (PAIVA, 1992).

Com a exceção de um clone triploide de *H. guianensis* (2n=54) e da existência de um genótipo de *H. pauciflora* com 18 cromossomos (BALDWIN, 1947), todas as espécies possuem 36 cromossomos (2n=36). Apesar de o gênero *Hevea* se comportar como um organismo diplóide, acredita-se que sua origem seja anfidiplóide (2n=36; x=9) e se estabilizou durante o curso da evolução (ONG, 1975). Estudos de hibridização *in situ* revelaram dois locos distintos 18S-25S de rDNA e um loco 5S rDNA sugerindo a perda de 1 loco 5S rDNA durante a evolução, sugerindo que o gênero tenha origem alotetraplóide (LEITCH et al., 1998).

O gênero apresenta muita variabilidade morfológica, com uma grande amplitude de ambientes ecológicos, variando de florestas altas a florestas arbustivas. As espécies *H. guianensis* e a *H. brasiliensis* são as mais altas do gênero, podendo atingir até 50m, enquanto *H. camargoana* e *H. camporum* são relatadas como árvores de porte baixo e arbustivo (SCHULTES, 1970). Já as espécies *H. spruceana* e *H. microphylla* estão localizadas em igapós ou pântanos muito encharcados, possuem troncos ventricosos na base e são vulgarmente designadas de “seringa” barriguda. Estas espécies são as únicas que não possuem deiscência explosiva e látex em todas as partes da planta (SCHULTES, 1970).

Apesar das diferenças fenotípicas, todas as espécies podem ser cruzadas interespecificamente, sendo considerado um complexo de espécies devido à falta de barreira reprodutiva (SAHA; PRIYADARSHAN, 2012).

Dentre as 11 espécies do gênero, a *H. brasiliensis* desperta maior interesse econômico por apresentar maior capacidade produtiva e látex de alta qualidade; *H. benthamiana*, *H. pauciflora* também são espécies interessantes para os programas de melhoramento por apresentarem resistência ao fungo *Microcyclus ulei* (GONÇALVES, 2008) causador do mal-das-folhas,

principal doença que acomete a cultura no Brasil e uma ameaça permanente para plantações da Ásia e África.

Contexto da Cultura da Seringueira

Não se sabe exatamente quando a borracha foi descoberta. Os primeiros registros sobre a borracha datam da viagem de Cristóvão Colombo à América, quando seu uso parecia ser bastante difundido entre os nativos do Continente (POLHAMUS, 1962).

Em 1743, La Condamine escreveu (“Relation abrégée d'un voyage fait dans l'intérieur de l'Amérique Méridionale”) a respeito de uma resina de uma árvore que os nativos de Quito e das beiras do *rio Marañón* usavam para fazer garrafas, calçados e bolsas. Estes objetos chamavam atenção pela sua impermeabilidade e elasticidade. Atualmente, a borracha natural é utilizada para a produção de mais de 40.000 produtos, incluindo mais de 400 dispositivos médicos, além de ter importância fundamental na indústria de pneus (CORNISH, 2001).

Sabe-se que aproximadamente 2.500 espécies são conhecidas por sintetizar borracha, mas poucas espécies como *Hevea brasiliensis*, *Parthenium argentatu* e *Taraxacum koksaghyz* produzem borracha natural de alta qualidade com peso molecular superior a 1 milhão de daltons. A composição química da borracha consiste no alto peso molecular de cis-polisoprene (SAKDAPIPANICH, 2007) formado pela condensação sequencial de isopentenyl difosfato (IPP) (CHOW et al., 2007). Devido a esta composição química, ela é e possuidora de elasticidade, resistência ao desgaste, plasticidade, impermeabilidade para líquidos e gases e propriedades isolantes de eletricidade e não pode ser integralmente substituída pela borracha sintética (CORNISH, 2001).

Dentre todas as espécies que produzem látex de boa qualidade, a *H. brasiliensis* é a maior fonte fornecedora de borracha natural, sendo cultivada em aproximadamente 11,3 milhões de hectares em todo o mundo. O Sudoeste Asiático é responsável por 92% da produção, seguido da África com 6% e América Latina com 2% (SAHA; PRIYADARSHAN, 2012).

De acordo com dados do IRSG (2014), em 2013, a produção mundial de borracha natural atingiu 12,04 milhões de toneladas, das quais o Brasil contribuiu com apenas 173 mil toneladas,

cerca de 1,5% do total produzido no mundo, sendo São Paulo o maior estado produtor com 53,2%. Nesse mesmo ano o consumo nacional foi de 395 mil toneladas, ou seja, importou-se 56% da borracha consumida no país.

Apesar da bacia Amazônica oferecer condições ótimas para o crescimento e desenvolvimento da cultura, nesta região também é área de ocorrência do fungo *Microcyclus ulei* causador do SALB (*South American Leaf Bligth*) ou mal-das-folhas que devastou plantações na região norte do Brasil como aconteceu nas plantações de Henry Ford no início do século XX e até hoje é considerado uma ameaça permanente para a indústria da borracha (PUSHPARAJAH, 2001). Por isso, a heveicultura tem se expandido para áreas de escape como o norte da Índia, Vietnã, sul da China e o planalto sul do Brasil (RAJ et al., 2005) Além das novas condições para o desenvolvimento da cultura, essas novas áreas propiciam condições de estresse como baixas temperaturas, altitude elevada e vento, que limitam a produção de borracha (PRIYADARSHAN; GONÇALVES, 2003).

Desta forma o melhoramento genético em seringueira vem buscando clones que são bons produtores e tolerantes às condições de estresse impostas por estas áreas de escape (PRIYADARSHAN; GONÇALVES, 2003). Por ser uma árvore perene, seu ciclo de vida longo dificulta avanços no melhoramento genético. São necessários de 20 a 30 anos para se obter a avaliação total em testes de campo de novos genótipos oriundos de cruzamentos. Dessa forma, técnicas de biologia molecular podem possibilitar a diminuição e otimização das avaliações para essa cultura.

Marcadores Moleculares

Estudos na variação das sequências de DNA são de suma importância para estudos genéticos. Neste contexto, marcadores moleculares são ferramentas altamente poderosas para auxiliar programas de melhoramento, uma vez que podem ser usados na discriminação de acessos duplicados em bancos de germoplasma, em estudos de diversidade genética do material pertencente a estes bancos, na caracterização de variedades comerciais, na construção de mapas genéticos e na identificação de genes de interesse econômico, em estudos de mapeamento

associativo e na seleção assistida por marcadores moleculares (SAM) (VARSHNEY; GRANER; SORRELLS, 2005).

Existe uma variedade de marcadores moleculares disponíveis na literatura, como *Restriction Fragment Length Polymorphisms* (RFLPs), *Random Amplification of Polymorphic DNAs* (RAPDs), *Amplified Fragment Length Polymorphisms* (AFLPs), *Simple Sequence Repeats* (SSRs) ou microssatélites e *Single Nucleotide Polymorphisms* (SNPs) sendo os dois últimos amplamente empregados em programas de melhoramento devido ao caráter codominante e alta reprodutibilidade (VARSHNEY; GRANER; SORRELLS, 2005).

Os marcadores microssatélites são combinações de 1 a 6 pares de bases repetidas em *tandem* e a natureza de sua variação é o número de repetições das unidades repetidas que o compõem por processos de inserção ou deleção ocasionada por *slippage* no processo de replicação (LEVINSON; GUTMAN, 1987). Os SSRs apresentam como características principais: (1) alta reprodutibilidade via reações em cadeia da polimerase (*PCR-Polymerase Chain Reaction*), (2) padrão de herança Mendeliana do tipo codominante, (3) abundância no genoma e alto nível de polimorfismo.

Podemos subdividir os microssatélites em duas categorias: (1) microssatélites genômicos e (2) microssatélites de sequências expressas (EST-SSRs). Os EST-SSRs apresentam menor polimorfismo comparado aos microssatélites genômicos por estarem em regiões gênicas e sofrer maior pressão de seleção. Estes marcadores possuem grande interesse nos programas de melhoramento, justamente por se encontrarem dentro de genes e possivelmente estarem relacionados à variação fenotípica de uma característica de interesse (VARSHNEY; GRANER; SORRELLS, 2005). Os marcadores microssatélites também podem ser testados quanto a sua transferibilidade em espécies do mesmo gênero e entre gêneros, a fim de facilitar e elucidar mecanismos envolvidos na divergência populacional e especiação.

Os marcadores SNPs são variações genéticas de uma única base e são comumente encontradas no genoma. A partir de dados de transcriptoma, a frequência de SNPs em *Eucalyptus grandis* foi de 1 SNP a cada 195 pb (NOVAES et al., 2008), em maçã foi de 1 SNP a cada 149 pb (CHAGNÉ et al., 2012) e em uva 1 SNP a cada 117 pb (LIJAVETZKY et al., 2007).

Na literatura, os SNPs são classificados como marcadores bi-alélicos, ou seja, apenas duas das quatro possíveis variantes são normalmente encontradas em pelo menos 1% da população (WANG et al., 1998). Os SNPs localizados em sítios regulatórios de genes podem afetar as taxas de transcrição e mudar a expressão das proteínas correspondentes a estes sítios, enquanto SNPs encontrados em regiões codificadoras podem ser divididos em duas classes: SNP sinônimo e SNP não-sinônimo.

O SNP é considerado sinônimo quando a mudança da base não afeta a sequência do aminoácido e portanto, não afeta a estrutura da proteína, porém podem modificar a estrutura e a estabilidade do RNA mensageiro (KWOK, 1999). Já o SNP classificado como não-sinônimo indica alteração no aminoácido da proteína e pode consequentemente alterar sua estrutura e funcionalidade. Estes marcadores se baseiam na alteração de uma única base nitrogenada no DNA em uma região específica do genoma. As mutações mais recorrentes são as do tipo transição, em que há troca de purina por outra purina (A↔G) ou de uma pirimidina por outra pirimidina (C↔T). As transversões são menos frequentes e acontecem quando há troca de uma purina por uma pirimidina, ou vice-versa (C↔G, C↔ A, T↔ G ou T↔ A).

Low et al. (1996) descreveram pela primeira vez marcadores microssatélites desenvolvidos em seringueira através de uma busca em banco de dados de sequências gênicas. A primeira construção de bibliotecas enriquecidas em SSRs para *H. brasiliensis* foi descrita por Atan et al. (1996) e posteriormente outros trabalhos utilizando esta metodologia foram publicados (ROY; NAZEER; SAHA, 2004; SOUZA et al., 2009; LE GUEN et al., 2011b; MANTELLO et al., 2012)

O desenvolvimento de marcadores microssatélites é fundamental para o entendimento da diversidade genética e fluxo gênico em seringueira, além de auxiliar programas de melhoramento genético na construção e saturação de mapas-genéticos e identifição de marcadores ligados a locos responsáveis por características de interesse econômico. Diversos trabalhos foram publicados em seringueira com o desenvolvimento e marcadores microssatélites genômicos (ROY; NAZEER; SAHA, 2004; LE GUEN et al., 2011; MANTELLO et al., 2012;) e EST-SSRs (FENG et al., 2008; LI et al., 2012; TRIWITAYAKORN et al., 2011), que foram utilizados para mapeamento genético (LE GUEN et al., 2011a; LESPINASSE et al., 2000a;

TRIWITAYAKORN et al., 2011; SOUZA et al., 2013;), mapeamento de QTLs (LE GUEN et al., 2011a; LESPINASSE et al., 2000b; SOUZA et al., 2013) e estudos populacionais (LE GUEN et al., 2009). Apesar dos marcadores SNPs apresentarem maior frequência quando comparados aos microssatélites, este marcador foi pouco explorado em seringueira. Até o momento, um total de 225 SNPs foram descritos para seringueira (POOTAKHAM et al., 2011; SALGADO et al., 2014; SILVA et al., 2014).

RNA-seq

Nos últimos anos, o sequenciamento pelo método de Sanger tem sido substituído por sequenciadores de nova geração (NGS) para obtenção de transcriptomas, genomas e epigenomas que necessitam de grande volume de dados para suas análises (HAMILTON; BUELL, 2012). O rápido avanço das plataformas NGS associado ao barateamento dos custos desta tecnologia tem permitido sequenciamentos em larga escala com alta qualidade das bases, proporcionando uma alta relação custo-benefício (WEI et al., 2011). Dentre as diversas plataformas disponíveis no mercado, os sequenciadores Roche/454 e Illumina têm sido os mais citados na literatura e utilizam a metodologia de sequenciamento por síntese (EGAN; SCHLUETER; SPOONER, 2012).

A plataforma Roche 454 utiliza a técnica de pirosequenciamento para detectar as bases, enquanto a plataforma Illumina utiliza nucleotídeos terminadores marcados com diferentes fluoróforos, assim como o sequenciamento em Sanger (MOROZOVA; MARRA, 2008). A diferença entre as plataformas Roche/454 e Illumina não se restringe apenas em como as bases são detectadas, mas também na quantidade de dados gerados por sequenciamento e tamanho dos fragmentos sequenciados. A tecnologia do Roche/454 sequencia fragmentos maiores, variando de 200 até 800 pb e produz aproximadamente 700Mb enquanto a plataforma Illumina sequencia fragmentos menores (36-150 pb), porém com maior profundidade, produzindo até 540-600 Gb de dados (CARVALHO; SILVA, 2010).

As plataformas NGS permitem rápida obtenção de dados, porém as sequências obtidas nestas plataformas NGS são pequenas. Dessa maneira, faz-se necessário montar e reconstruir

sequências que possam dar informações a serem interpretadas (MARTIN; WANG, 2011). A montagem de genomas ou transcriptomas a partir de pequenas sequências possui muitos desafios na bioinformática, por isso ao longos dos últimos anos, novos algoritmos e novas ferramentas de bioinformática foram desenvolvidas e/ou aprimoradas.

Apesar do barateamento do custo de sequenciamento em plataformas NGS, a montagem de genomas é extremamente complexa, principalmente porque informações como tamanho do genoma, ploidia da espécie e a porcentagem de regiões repetitivas são de extrema importância e determinam a complexidade destas montagens (WANG; GERSTEIN; SNYDER, 2009). Recentemente, foi publicado um *draft* do genoma de seringueira utilizando diferentes plataformas NGS (Roche/454, Illumina e SOLiD) onde estão disponíveis em banco de dados públicos 1.1 Gb dos 2.1 Gb do genoma haplóide estimado. Como aproximadamente 78% do genoma da seringueira possui sequências repetitivas de DNA, a montagem do seu genoma é extremamente complexa. Os autores acreditam ser possível estruturar o genoma em cromossomos associando o sequenciamento em plataformas NGS, *Bacterial Artificial Chromossomes* (BACs) e mapas genéticos bem saturados como foi feito para a montagem do genoma da cevada, que possui um genoma de 5.1 Gb com 84% de sequências repetitivas (RAHMAN et al., 2013).

Montar transcriptomas utilizando sequências provenientes de sequenciadores NGS têm sido amplamente difundido na literatura por serem menos complexas quando comparadas a montagens de genomas. Entender a dinâmica do transcriptoma é essencial para entender e interpretar elementos funcionais do genoma, revelando os constituintes moleculares de células e tecidos (WANG; GERSTEIN; SNYDER, 2009). O objetivo chave na obtenção do transcriptoma é catalogar transcritos, descobrir *splicing* alternativo, quantificar mudanças na expressão gênica de cada transcrito durante um estágio de desenvolvimento específico e/ou sob diferentes condições fisiológicas, além de oferecer resolução de base única que detecta diferenças de uma única base para subsequente identificação e desenvolvimento de marcadores SNPs em regiões expressas (GANAL; ALTMANN; RÖDER, 2009; LANGMEAD B, TRAPNELL C, POP M, 2009; WANG; GERSTEIN; SNYDER, 2009)

A montagem de transcriptomas sem ter um genoma de referência pode ser feita pela estratégia de montagem *de novo* (WARD; PONNALA; WEBER, 2012). Dentre os métodos computacionais disponíveis para o *de novo*, o grafo de Bruijn é o método mais difundido na literatura. No grafo, cada *read* é dividido em um comprimento fixo de *k* nucleotídeos (*k-mer*) onde *k*-1 deve se sobrepor a *k* de forma a montar os transcritos (GRABHERR et al., 2011; WARD; PONNALA; WEBER, 2012). Programas de montagem como o Trinity (GRABHERR et al., 2011), Velvet (ZERBINO; BIRNEY, 2008), Oases (SCHULZ et al., 2012) e CLCBio são amplamente utilizados na literatura e usam o grafo para o *de novo* (WARD; PONNALA; WEBER, 2012). Os transcritos obtidos na montagem *de novo* são classificados e anotados em bancos de dados públicos como por exemplo o *Gene Ontology* (GO), onde suas funções são classificadas em três grandes categorias: (1) função molecular; (2) processos biológicos; (3) componentes celulares; ou o *Kyoto Encyclopedia of Genes and Genomes* (KEGG) onde são atribuídas funções biológicas e interações em vias bioquímicas (KANEHISA; GOTO, 2000; KANEHISA et al., 2008). Esta caracterização dos transcritos prevê suas prováveis funções e é o ponto de partida para outros estudos na espécie de interesse. Dessa maneira, o RNA-seq tem contribuído de forma significativa na transcriptômica, principalmente em estudos de organismos não-modelo onde não existe um genoma de referência e geralmente existe pouca ou nenhuma informação disponível em banco de dados como, por exemplo, a batata doce (*Ipomoea batatas*) (WANG et al., 2010) e chá (*Camellia simensis*) (SHI et al., 2011).

As plataformas NGS têm provocado um impacto muito grande nos programas de melhoramento, principalmente pela possibilidade de desenvolver milhares de marcadores moleculares (microssatélites e SNPs) localizados em genes. Estes marcadores podem ser usados para saturação de mapas genéticos-moleculares em diversas espécies de interesse econômico, permitindo associar marcadores moleculares ligados a regiões de QTLs para seleção assistida por marcadores (SAM).

Com o advento das tecnologias de sequenciamento NGS, em 2011 foi publicado o primeiro RNA-seq em seringueira, que obteve a descrição geral dos transcriptomas de látex e folha (XIA et al., 2011). Neste mesmo ano, Triwitayakorn et al. (2011) publicaram um transcriptoma do meristema apical, com objetivo de buscar e desenvolver marcadores EST-SSRs para a construção de uma mapa genético e, mais recentemente, Li et al. (2012) descreveram o

transcriptoma de painel e Salgado et al. (2014) decresveram o transcriptoma geral com diversos tecido. A partir destes trabalhos foram desenvolvidos e caracterizados marcadores EST-SSRs (LI et al., 2012) e SNPs (SALGADO et al., 2014).

Mapeamento Genético

Muitas características importantes na agricultura, como produção de grãos, altura da planta e teor de proteína têm natureza quantitativa e são controladas por vários genes. Regiões cromossômicas que controlam caracteres com padrão contínuo de variação fenotípica são chamadas de QTLs (*Quantitative Trait Loci*) (FALCONER; MACKAY, 1996; LIU, 1998; LYNCH; WALSH, 1998). Para identificação destas regiões é de fundamental importância que sejam construídos mapas genéticos.

Um mapa genético é uma sequência linear de marcadores dispostos em grupos de ligação, indicando as regiões do genoma em que estes marcadores se encontram. Atualmente, marcadores moleculares como AFLPs, RFLPs, RAPDs, microssatélites e SNPs são amplamente utilizados em estudos de mapeamento genético, de modo a proporcionar uma ampla cobertura do genoma e, consequentemente, gerar mapas genéticos saturados. A partir de um mapa genético bem construído pode-se realizar o mapeamento de QTLs, permitindo, então, estudos sobre a arquitetura genética de caracteres importantes (FERREIRA; GRATTAPAGLIA, 1998; SIBOV et al., 2003a, 2003b).

O princípio básico para o mapeamento genético foi desenvolvido por Sturtevant (1913), o qual utiliza a porcentagem de recombinantes como indicador quantitativo da distância linear entre dois genes na construção de mapas genéticos. Os mapas desta época mostravam que a posição dos genes correspondia à sua ordem linear nos cromossomos (GARDNER; SNUSTAD, 1986; GRIFFITHS et al., 1998). Dessa forma, marcadores que possuem uma frequência de recombinação maior que 50% podem pertencer a cromossomos diferentes ou estarem ligados ao mesmo cromossomo, mas muito distantes um do outro. De modo contrário, quanto menor a frequência de recombinação entre os marcadores, maior a proximidade entre os mesmos. Baseando-se neste princípio, os marcadores são ordenados e suas frequências de recombinação

são transformadas em distância através do mapeamento, sendo as funções de Haldane (HALDANE, 1919) e a de Kosambi (KOSAMBI, 1944) as mais utilizadas.

A função de Haldane admite que a ocorrência de permuta se dá de modo independente (ausência de interferência), seguindo uma distribuição de Poisson, enquanto a função de Kosambi admite a ocorrência de permutas próximas como eventos não independentes (presença de interferência). Assim, a função de Haldane assume a interferência completa entre regiões arbitrariamente próximas, sendo decrescente para locos mais distantes e iguais a zero para locos independentes (LIU, 1998; WEIR, 1996).

Para a construção de um mapa de ligação torna-se fundamental a escolha de uma população segregante onde os genitores desta população possuam características agronomicamente contrastantes. As populações de mapeamento mais utilizadas são F_2, retrocruzamentos e RILs entretanto, estas populações são originárias de linhagens endogâmicas, as quais não estão disponíveis para todas as espécies. Em espécies arbóreas como a seringueira, há uma grande dificuldade na obtenção de linhagens endogâmicas devido ao longo ciclo de vida e por sofrerem depressão por endogamia. Dessa forma, Grattapaglia & Sederoff (1994) propuseram a utilização de uma estratégia, denominada de *pseudo-testcross*, que permite a construção de mapas de ligação com base na progênie F_1 de genitores heterozigotos. Nessa abordagem, a fase de ligação não é conhecida a priori, mas é inferida a posteriori com base na segregação da progênie. Esta estratégia apresenta a desvantagem de utilizar apenas marcadores com segregação 1:1, ou seja, apenas marcadores que sejam polimórficos em um dos genitores.

A construção de mapas integrados é de extrema importância, pois muitos dos efeitos genéticos interessantes para o melhoramento de plantas (dominâncias e epistasias) só são possíveis de serem estimados com metodologias que construam um mapa único da população. Dessa maneira, Wu et al. (2002) propuseram estimar a fração de recombinação e a fase de ligação em marcadores genéticos moleculares em uma população F1 segregante utilizando o método de máxima verossimilhança. Por isso, marcadores codominantes classificados neste artigo como marcadores do tipo A, B e C, ou seja, marcadores com segregação 1:1:1:1, 1:2:1 e 3:1 respectivamente são considerados mais informativos por apresentarem polimorfismo nos dois genitores e são essenciais para a integração dos mapas.

Em seringueira, Lespinasse et al (2000a) publicaram o primeiro mapa genético para seringueira utilizando a estratégia duplo *pseudo-testcross.* O mapa sintético contém 717 marcas, dentre as quais 301 eram marcadores RFLPs, 388 AFLPs, 18 microssatélites e 10 isoenzimas que foram distribuídos em 18 grupos de ligação.

Mais recentemente, Triwitayakorn et al (2011) e Le guen et al (2011a) publicaram novos mapas genéticos para seringueira. Le Guen et al (2011a) obtiveram um mapa sintético com 110 microssatélites, 1 *sequence target site* (STS) e 2 AFLPs distribuídos em 18 grupos de ligação. Já Triwitayakorn et al (2011) obteve um mapa sintético com 23 grupos de ligação obtidos com 97 marcadores SSRs. O número de grupos de ligação obtidos por Triwitayakorn et al (2011) excede o número haplóide de grupos de ligação esperado (n=18). Como este trabalho utilizou marcadores microssatélites também utilizados por Le Guen et al. (2011a), foram encontrados grupos de homologia, sugerindo que maior saturação deste mapa genético poderá unir os grupos de ligação que estão separados.

Souza et al (2013) publicaram um mapa integrado de *H. brasiliensis*. Este mapa foi construído com 284 marcadores microssatélites, dentre os quais 138 (48,6 %) apresentaram segregação 1:1, 31 (10,9 %) a segregação 1:2:1 e 115 (40,5 %) a segregação 1:1:1:1. Este mapa apresentou um comprimento total de 2668.8 cM distribuídos em 23 grupos de ligação, mesmo número encontrado por Triwitayakorn et al. (2011). O autor sugere que outros tipos de marcadores sejam usados para poder saturar o mapa e poder unir grupos de ligação e chegar em 18 grupos de ligação.

Objetivos

Objetivo Geral

Este projeto teve como objetivo contribuir para o conhecimento genético e genômico da seringueira por meio do desenvolvimento de um mapa genético-molecular com marcadores gênicos e funcionais.

Objetivos Específicos

- Construir bibliotecas enriquecidas e desenvolver marcadores microssatélites
- Caracterizar os locos de microssatélites com relação a otimização das condições de amplificação, tipo de repetição, número e tamanho dos alelos amplificados;
- Montar e caracterizar o transcriptoma de painel de seringueira
- Identificar SNPs e SSRs no transcriptoma
- Desenvolver e validar marcadores SNPs em acessos de *H. brasiliensis*;
- Saturar o mapa genético-molecular da população de mapeamento PB217 x PR255

Capítulo I

Mantello *et al. BMC Research Notes* 2012, **5**:329
http://www.biomedcentral.com/1756-0500/5/329

BMC Research Notes

SHORT REPORT Open Access

Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species

Camila C Mantello[1*], Fernando I Suzuki[1], Livia M Souza[1], Paulo S Gonçalves[2] and Anete P Souza[1,3*]

Abstract

Background: The rubber tree (*Hevea brasiliensis*) is native to the Amazon region and it is the major source of natural rubber in the world. Rubber tree breeding is time-consuming and expensive. However, molecular markers such as microsatellites can reduce the time required for these programs. This study reports new genomic microsatellite markers developed and characterized in *H. brasiliensis* and the evaluation of their transferability to other *Hevea* species.

Findings: We constructed di- and trinucleotide-enriched libraries. From these two libraries, 153 primer pairs were designed and initially evaluated using 9 genotypes of *H. brasiliensis*. A total of 119 primer pairs had a good amplification product, 90 of which were polymorphic. We chose 46 of the polymorphic markers and characterized them in 36 genotypes of *H. brasiliensis*. The expected and observed heterozygosities ranged from 0.1387 to 0.8629 and 0.0909 to 0.9167, respectively. The polymorphism information content (PIC) values ranged from 0.097 to 0.8339, and the mean number of alleles was 6.4 (2–17). These 46 microsatellites were also tested in 6 other *Hevea* species. The percentage of transferability ranged from 82% to 87%. Locus duplication was found in *H. brasiliensis* and also in 5 of other species in which transferability was tested.

Conclusions: This study reports new microsatellite markers for *H. brasiliensis* that can be used for genetic linkage mapping, quantitative trait loci identification and marker- assisted selection. The high percentage of transferability may be useful in the evaluations of genetic variability and to monitor introgression of genetic variability from different *Hevea* species into breeding programs.

Keywords: *Hevea*, *H. brasiliensis*, Microsatellite, Characterization, Transferability

Findings

Hevea brasiliensis (Willd. ex Adr. de Juss.) Muell. -Arg., native to the Amazon rainforest, is a diploid (2n = 36, x = 9), perennial, monoecious and cross-pollinated tree species. It belongs to the genus *Hevea* and the botanical family Euphorbiaceae.

The genus *Hevea* comprises 11 inter-crossable species [1,2] (*H. benthamiana* Muell.- Arg., *H. brasiliensis*, *H. carmagoana* Pires, *H. camporum* Ducke, *H. guianensis* Aubl, *H. microphylla* Ule, *H. nitida* Mart. ex-Muel.-Arg., *H. pauciflora* (Spruce ex-Benth.) Muell.-Arg., *H. rigidifolia* (Spruce ex-Benth.) Muell.-Arg., *H. spruceana* (Benth.) Muell.-Arg. and *H. paludosa* Ule), which have evolved in the Amazon rainforest over 100,000 years [3].

Of all the species in the genus *Hevea*, *H. brasiliensis* is the most economically important, because it is the major source of natural rubber worldwide. Natural rubber is important mainly in the tire industry but also in many other sectors because it is flexible, resistant, impermeable to liquids and abrasion resistant [4]. These singular properties make natural rubber both complementary and competitive to synthetic rubber and furthermore superior to it in varied applications. As is the case for many other perennial trees, rubber tree breeding is time-consuming and expensive. An average duration of 20–25 years of field experiments in large areas is generally required to obtain a new cultivar with reasonably low risks [5].

* Correspondence: camila.mantello@gmail.com; anetepsouza@gmail.com
[1]Centro de Biologia Molecular e Engenharia Genética (CBMEG) - Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6010, CEP 13083-970, Campinas, SP, Brazil
[3]Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6109, CEP 13083-970, Campinas, SP, Brazil
Full list of author information is available at the end of the article

Molecular markers, such as amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPD), simple sequence repeats (SSRs or microsatellites) and single nucleotide polymorphisms (SNPs), have successively become increasingly important in plant breeding. These markers are efficient tools for the assessment of genetic diversity, the identification of quantitative trait loci (QTLs) and/or gene mapping, variety protection and marker-assisted selection (MAS) [6].

Of these markers, microsatellite markers are considered the most suitable for genetic studies, because they combine co-dominance and high polymorphism with abundance, locus specificity and uniform dispersion in plant genomes. Moreover, microsatellite markers can discriminate closely related individuals [7]. In addition, microsatellite analysis is inexpensive, high reproducible and highly transferable across related species.

Following biochemical markers such as isozymes [8], molecular markers have been developed and used since the middle of 1990's for diversity studies [9,10], genetic mapping [11] and the identification of genetic loci implicated in the expression of agronomic traits in *H. brasiliensis* [11–14]. However, most of the markers used have been isozyme, RAPD, RFLP or AFLP markers.

To date, few studies using genomic microsatellites [10,15,16] or microsatellites from expressed sequence tags (EST-SSRs) have been published [6]. Accordingly, we present the development of genomic microsatellites using dinucleotide- and trinucleotide- enriched libraries, the characterization of these microsatellite markers in multiple accessions of *H. brasiliensis* and test their transferability in six other *Hevea* species.

Results

Microsatellite-enriched library analysis

Di- and trinucleotide-enriched libraries were constructed, and the clones from each library were sequenced (576 and 288 clones, respectively).

A total of 291 (50.5%) clones from the dinucleotide-enriched library contained microsatellite sequences. Since dinucleotide probes were used for the library enrichment, these motifs were the most abundant comprising 324 (90.2%) SSRs followed by 20 tetranucleotide (5.6%), 12 trinucleotide (3.3%) and 3 pentanucleotide (0.8%) (Table 1).

In the trinucleotide-enriched library, 62 (21.5%) clones contained microsatellites sequences. Trinucleotides were the most frequent motif with 45 (72.5%) SSRs followed by 13 dinucleotide (20.9%), 9 tetranucleotide (14.5%) and 3 pentanucleotide (4.8%) SSRs (Table 1).

For some sequences, both the dinucleotide- and trinucleotide-enriched libraries had more than one microsatellite, which explains the greater number of SSRs found compared to the number of sequences analyzed.

A total of 130 and 32 primer pairs were designed based on the di- and trinucleotide-enriched libraries, respectively. To remove possible redundancies with published SSRs, each sequence containing SSR was compared against GenBank database using BLASTN. Nine sequences, only from the dinucleotide library, were identical to previously published sequences, which already had primer pairs. These sequences were removed from this study.

Dinucleotide motifs have been reported as the most abundant type of microsatellite in plant genomes [17]. Recent studies relating to expressed sequence tags revealed that trinucleotide motifs are the most abundant motifs in ESTs in many plants, such as sugarcane [18], barley [19], grapes [20], rice [21], wheat [22] and citrus [23], whereas in other plants dinucleotide motifs are the most abundant in ESTs, such as kiwi [24], coffee [25] and apricot and peach [26] In rubber tree, Feng and co-workers analyzed all of the ESTs in NCBI database in search for microsatellites. They found that dinucleotide motifs were the most abundant and were three times more abundant than trinucleotides [6]. The low efficiency of the trinucleotide enrichment in our study could result from the low frequency of trinucleotide motifs in the rubber tree genome.

Polymorphism analysis and cross-species transferability

In total, 153 specific primer pairs were designed. A total of 119 primer pairs produced good amplification products (Additional file 1: Table S1) and 90 of these products were polymorphic among a set of 9 rubber clones indicated in Table 2.

Table 1 Number of sequences and SSRs of each type in each library

	Dinucleotide-enriched library	Trinucleotide-enriched library
Number of sequences	576	288
Number of sequences with SSRs	291	62
Total number of SSRs	359	70
dinucleotide	324	13
trinucleotide	12	45
tetranucleotide	20	9
pentanucleotide	3	3

Mantello *et al. BMC Research Notes* 2012, **5**:329
http://www.biomedcentral.com/1756-0500/5/329

Table 2 Genotypes of *H. brasiliensis* and six species of the genus *Hevea* used for characterization and transferability

Genotypes *H. brasiliensis*			Other species of *Hevea*
RRIM 600	PB 311	IRCA 209	*H. guianensis*
RRIM 606	PB 346	IRCA 230	*H. rigidifolia*
RRIM 701*	PB 260*	IRCA 707	*H. nitida*
RRIM 729	PB 217*	IRCA 1159	*H. pauciflora* -(112 CNSG)
RRIM 728	PC 140	GT 1*	*H. pauciflora* -(116 CNSG)
RRIM 805	RRIC 100	PR255*	*H. benthamiana*
RRIM 809	IAC 306	RO 38*	*H. camargoana*
RRIM 913	IAC 307	Fx 4098	
RRIM 915	IAC 309	CMB 104*	
RRIM 937	IAC 313	CMB 114*	
RRII 105	IAC 318		
PB 233	IAC 500		
PB 235*	IRCA 27		

*: genotypes used to characterize all of the SSRs markers developed.

The observed and expected heterozygosities ranged from 0.1111 to 1 and 0.1111 to 0.9150, respectively, and the PIC values ranged between 0.0994 and 0.8496. The mean number of alleles was 4.46 (2–9 alleles).

Of the 90 polymorphic markers we chose 46 SSRs for characterization among a set of 36 genotypes of *H. brasiliensis* (Table 2). The PIC values of these markers ranged from 0.097 to 0.8339, and the observed and expected heterozygosities ranged from 0.0909 to 0.9167 and 0.1387 to 0.8629, respectively. The mean number of alleles was 6.4 (2–17 alleles) (Table 3).

Six other species from the genus *Hevea* (*H. guianensis*, *H. rigidifolia*, *H. nitida*, *H. pauciflora*, *H. benthamiana and H. camargoana*) being two different genotypes of *H. pauciflora*, were used to evaluate the transferability of the markers (Table 2). All loci were tested under the same PCR conditions used for *H. brasiliensis*.

Of the 46 loci tested, 40 (87%) were amplified for *H. guianensis* and *H. pauciflora*- (112CNSG), 39 (85%) were amplified for *H. camargoana*, *H. nitida* and *H. pauciflora*-(116CNSG) and 38 (82%) were amplified for *H. benthamiana* (Table 4).

For most of the SSR loci, which cross-species amplification were tested, the number of alleles found in the *H. brasiliensis* genotypes in conjunction with the tested species (from 2 to 21, with a mean number of 9.5) was larger when the same loci were analyzed in *H. brasiliensis* alone (Table 3) which means that other species revealed some novel alleles.

Three primers pairs, HB54, HB62 and HB69, did not produce amplification products for the six other species tested, whereas 27 loci were amplified for all species (Table 4).

Saha and co-workers first observed SSR cross-amplification in *H. benthamiana* and *H. spruceana* [27].

Table 3 Characterization of the 46 polymorphic SSR markers

Primer	NA A	size range (bp)	H_e	H_o	PIC	NA B
HB 31	6	182-173	0.753	0.6875	0.7001	12
HB 32	6	256-238	0.6057	0.3243	0.524	12
HB 33	9	188-171	0.6808	0.4	0.6388	10
HB 35	3	159-156	0.3882	0.4054	0.3226	5
HB 36	7	239-217	0.6279	0.6111	0.557	16
HB 37	3	161-154	0.6133	0.3235	0.5325	3
HB 41	6	141-173	0.4855	0.3611	0.459	8
HB 42	7	203-216	0.5313	0.5833	0.4954	10
HB 43	10	219-265	0.7531	0.5405	0.7133	15
HB 45	17	155-238	0.856	0.8333	0.8273	21
HB 47	8	152-201	0.6311	0.4722	0.566	14
HB 50	3	196-210	0.2289	0.1944	0.2124	8
HB 51	2	196-201	0.4909	0.1515	0.3666	7
HB 53	8	204-223	0.7871	0.7568	0.7439	13
HB 54	10	168-203	0.6708	0.5294	0.6201	10
HB 55	9	168-192	0.8	0.4571	0.7645	10
HB 57	9	147-176	0.7169	0.6471	0.6831	12
HB 60	5	148-161	0.3154	0.3514	0.2985	6
HB 61	6	150-172	0.7175	0.9167	0.6662	11
HB 62	5	122-112	0.415	0.1081	0.3888	5
HB 63	8	198-236	0.7553	0.5135	0.7082	15
HB 64	8	146-199	0.6221	0.4444	0.5914	12
HB 66	3	273-279	0.2273	0.25	0.2085	4
HB 68	10	138-167	0.8058	0.8286	0.7636	14
HB 69	10	139-172	0.6962	0.6471	0.6426	10
HB 70	5	155-167	0.6928	0.8	0.6281	6
HB 71	2	186-188	0.1037	0.1081	0.097	2
HB 73	6	206-228	0.7754	0.8056	0.729	8
HB 77	4	135-153	0.1788	0.0811	0.1701	8
HB 78	8	186-215	0.7996	0.8286	0.7605	10
HB 81	9	187-220	0.7469	0.5333	0.6979	13
HB 82	4	175-168	0.5572	0.6757	0.4489	8
HB 83	4	181-165	0.5271	0.7143	0.4631	4
HB 92	7	255-229	0.773	0.7576	0.7246	7
HB 95	6	254-225	0.5742	0.5946	0.4793	12
HB 98	3	190-180	0.2787	0.3143	0.2535	4
HB 100	10	250-214	0.803	0.8649	0.7657	16
HB 101	2	142-132	0.2166	0.2432	0.1908	4
HB 102	4	198-157	0.2714	0.2973	0.2568	10
HB 103	2	164-161	0.1037	0.1081	0.097	2
HB 104	5	173-154	0.6711	0.375	0.6119	14
HB 105	13	258-176	0.8629	0.9091	0.8339	21
HB 106	5	210-230	0.5855	0.0909	0.4975	6

Table 3 Characterization of the 46 polymorphic SSR markers *(Continued)*

HB 109	6	186-210	0.1387	0.1143	0.135	7
HB 110	9	273-256	0.8322	0.6667	0.7981	11
HB 117	8	152-190	0.8052	0.7222	0.7643	12

NA/A, number of alleles in the 36 *H. brasiliensis* clones; NA/B, the total number of alleles in the 36 *H. brasiliensis* clones and in the other *Hevea* species; bp, product size range in base pairs; He, expected heterozygosity; Ho, observed heterozygosity; PIC, polymorphism information content.

Together with the cross-fertility potential, this high SSR transferability supports the consideration of the *Hevea* genus as a species complex with moderate differentiation among the species. These aspects appear to be favorable for genetic introgressions using other *Hevea* species to the rubber breeding population, which is mainly based on *H. brasiliensis*.

Locus duplication

Members of the Euphorbiaceae family have a basic number of chromosomes between 6 and 11; thus, any species with more than this number of chromosomes could be amphidiploid in origin [28,29]. Similar to cassava, plants in the genus *Hevea* have 36 chromosomes and behave as diploid. In these species, it has been assumed that n = 18 and the basic number of chromosomes is x = 9 [30,31].

In different *Hevea* species, chromosomes mainly formed bivalents and tetravalents are rarely produced as a result of pairing between non-homologous chromosomes during prophase I and metaphase I of meiosis [31,32].

Moreover, cytogenetic studies revealed two distinct loci on two different chromosomes, bearing the same18S-5.8 S-25 S rDNA sequence which may have arisen by the hybridization of two unknown diploid species (n = 9), thus suggesting a possible allotetraploid origin, however no potencial diploid ancestor has been described to date [29].

Locus duplication, as revealed by molecular markers, had been reported in *H. brasiliensis* 11] and *H. guianensis*, *H. rigidifolia* and *H. pauciflora* [15].

In *H. brasiliensis*, we observed locus duplication for the markers HB45 and HB109 for clones RRIM 701 and Fx4098, respectively. Cases of locus duplication were observed in all *Hevea* species tested, with the exception of *H. rigidifolia*. We also observed locus duplication of the HB36, HB68, HB100 and HB105 markers in *H. nitida*; HB68 and HB105 markers in *H. benthamiana*; HB101 and HB105 in *H. guianensis*. The marker HB105 exhibited one duplicated locus for *H. pauciflora* (112 CNSG), *H. pauciflora* (116 CNSG) and *H. camargoana*.

Although we have not been mapped the loci we cited as duplicated in the wild species, we observed more than 2 alleles (excluding the stutter bands) in the same base pair range compared with the other genotypes used for characterization (Figure 1). The presence of allele duplication, as determined by polyacrylamide gel electrophoresis suggests that the loci are duplicated for the species involved.

Although locus duplication had been described in *H. brasiliensis*, *H. guianinsis*, *H. rigidifolia* and *H. pauciflora*, this is the first report for *H. benthamiana*, *H. camargoana* and *H. nitida*.

Conclusion

Herein, we report the development of new SSR markers for *H. brasiliensis*, representing a powerful resource for genetic diversity studies and genetic breeding techniques, such as molecular genetic mapping, QTLs identification and MAS.

Due to observed high percentage of cross-amplification and the absence of reproductive barriers between the species within this genus, these markers may be important tools to monitor the genetic variability from other *Hevea* species into the current breeding programs.In addition, these SSR markers can be helpful for the identification of important agronomic characteristics in different *Hevea* species.

Methods

Plant material and DNA extraction

We used 36 genotypes of *H. brasiliensis* to characterize the SSRs (Table 2). These genotypes were kindly provided by the Agronomic Institute of Campinas (IAC) and Michelin Plantation (Brasil). We also used 6 other species of the genus *Hevea* (*H. guianensis*, *H. rigidifolia*, *H. nitida*, *H. pauciflora*, *H. benthamiana* and *H. camargoana*) that were kindly provided by the Brazilian Agricultural Research Corporation (EMBRAPA - Amazônia Ocidental) in Manaus, AM to test the transferability of the SSRs (Table 2). The genomic DNA samples were extracted from lyophilized leaf tissues using a modified CTAB method [33] and their quality and quantity were assessed using 1% agarose gel electrophoresis.

Construction of microsatellite-enriched libraries and sequence analysis

The microsatellite-enriched libraries for *H. brasiliensis* were constructed using the RRIM 600 clone according to the methodology described by Billote and co-workers [34]. The DNA samples were digested with AFAI and enriched using streptavidin magnetic-coated beads (Streptavidin MagneSphere Paramagnetic Particles, Promega, Madison, WI) and $(CT)_8$ and $(GT)_8$ biotinylated microsatellite probes for the dinucleotide-enriched library or $(ATC)_8$ and $(CCT)_8$ for the trinucleotide-enriched library.

Selected DNA fragments were amplified by PCR and then cloned into the pGEM-T vector (Promega, Madison, WI). Competent XL1-blue *Escherichia coli* cells were transformed with the recombinant plasmids and

Mantello *et al. BMC Research Notes* 2012, **5**:329
http://www.biomedcentral.com/1756-0500/5/329

Table 4 Cross-amplification of the 46 polymorphic SSRs markers among the other *Hevea* species

Primer	*H. guianensis*	*H. rigidifolia*	*H. benthamiana*	*H. camargoana*	*H. nitida*	*H. pauciflora-(112 CNSG)*	*H. pauciflora-(116CNSG)*
HB-31	+	+	+	+	+	-	+
HB-32	+	+	+	+	+	+	+
HB-33	+	-	+	+	+	+	+
HB-35	+	+	+	+	+	+	+
HB-36	+	+	+	+	+	+	+
HB-37	+	-	-	+	-	-	-
HB-41	+	+	+	+	+	+	+
HB-42	+	+	+	-	-	+	-
HB-43	+	+	+	+	+	+	+
HB-45	+	+	+	+	+	+	+
HB-47	-	+	+	+	+	+	+
HB-50	-	+	+	+	+	+	+
HB-51	+	+	+	+	+	+	-
HB-53	+	+	+	+	+	+	+
HB-54	-	-	-	-	-	-	-
HB-55	+	+	-	-	-	+	+
HB-57	+	+	+	+	+	+	+
HB-60	+	+	+	+	+	+	-
HB-61	+	+	+	+	+	+	+
HB-62	-	-	-	-	-	-	-
HB-63	+	+	+	+	+	+	+
HB-64	+	+	+	+	+	+	+
HB-66	+	+	+	+	+	+	+
HB-68	-	+	+	+	+	+	+
HB-69	-	-	-	-	-	-	-
HB-70	+	+	+	+	+	+	+
HB-71	+	+	+	+	+	+	+
HB-73	+	+	+	+	+	-	-
HB-77	+	+	+	+	+	+	+
HB-78	+	+	+	+	+	+	+
HB-81	+	-	+	+	+	+	+
HB-82	+	+	+	+	+	+	+
HB-83	+	+	+	+	+	+	+
HB-92	+	-	+	+	+	+	+
HB-95	+	+	+	+	+	+	+
HB-98	+	+	+	+	+	+	+
HB-100	+	+	+	+	+	+	+
HB-101	+	+	+	+	+	+	+
HB-102	+	+	+	+	+	+	+
HB-103	+	+	-	-	+	+	+
HB-104	+	+	+	+	+	+	+
HB-105	+	+	+	+	+	+	+
HB-106	+	-	+	-	-	+	+

Table 4 Cross-amplification of the 46 polymorphic SSRs markers among the other *Hevea* species *(Continued)*

HB-109	+	+	+	+	+	+	+
HB-110	+	+	+	+	+	+	+
HB-117	+	+	+	+	+	+	+

cultivated on agar medium containing ampicillin and 100 μg/ml of X-galactosidase. The clones containing the insert were sequenced using the Big Dye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) and an automated ABI 377 sequencer (Applied Biosystems, Foster City, CA).

All the sequences obtained were aligned, edited and eliminated if redundant using SeqMan (DNASTAR, Madison, WI). The sequences were also evaluated using Microsat software (A. M. Risterucci, CIRAD, personal communication), which removes the adapters and verifies the presence of restriction sites.

The microsatellites were identified using a specific research tool, SSRIT – (Simple Sequence Repeat Identification Tool) [35]. As a criterion for the SSR selection, the sequences that showed at least five dinucleotide repeats; four trinucleotide repeats; and three tetra-, penta- and hexanucleotide repeats were selected. Primers complementary to the sequences flanking the microsatellites were designed using Primer Select Program (DNAStar, Madison, WI) and Primer 3 [36]. To eliminate possible redundancies with the published SSRs, each sequence containing microsatellite was compared against the GenBank database using BLASTN. The redundant sequences which already had primer pairs published were eliminated.

PCR amplification

The microsatellite fragments were PCR amplified in a 15 μl reaction containing 25 ng template DNA, 0.5 μM each primer, 100 μM each dNTP, 1.5 mM $MgCl_2$, 20 mM Tris–HCl, 50 mM KCl and 0.5 U Taq DNA Polymerase. The PCR amplifications were performed with the following four programs: **(1)** initial denaturation at 94°C for 4 min, 30 amplification cycles (1 min at 94°C, 45 s at the specific annealing temperature and 1 min at 72°C), and a final extension at 72°C for 10 min; **(2) TD-1**, initial denaturation for 3 min at 94°C, 10 amplification cycles with a 0.5°C decrease in annealing temperature per cycle starting at 55°C or 60°C (94°C for 1 min, 60°C or 55°C for 30 s, and 72°C for 1 min 15 s); followed by 20 cycles with annealing at 50°C (94°C for 1 min, 50°C for 30 s and 72°C for 30 s) and a final elongation step at 72° C for 10 min; **(3) TD-2**, previously described by Le Guen and co-workers [10]; and **(4) TD-3**, initial denaturation at 94°C for 2 min, 2× 10 cycles with a 1°C decrease in annealing temperature per cycle from 65°C to 55°C (94°C for 1 min, 65°C for 1 min and 72°C for 1 min), followed by 18 cycles at 55°C (94°C for 1 min, 55°C for 1 min and 72°C for 2 min), and a final elongation step at 72°C for 5 min.

The amplification products were resolved by electrophoresis through 3% agarose gels prior to vertical electrophoresis using 6% denaturing polyacrylamide gels and were subsequently silver stained [37]. The product sizes were determined by comparison with a 10 bp DNA ladder (Invitrogen, Carlsbad, CA).

Analysis

The allelic polymorphic information content of each SSR was calculated using the formula $PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$, where n is the number of alleles of the marker among the set of genotypes used for characterizing the SSR polymorphism, and pi and pj are the frequencies of alleles i and j. The observed and expected heterozygosities were calculated using the TFPGA program [38].

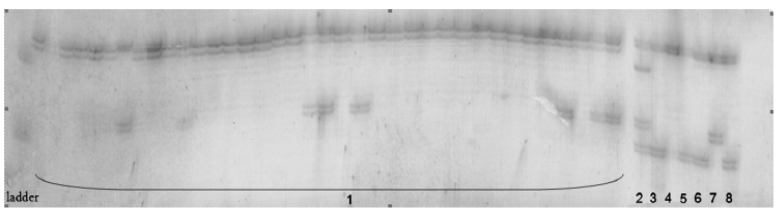


Figure 1 A 6% denaturing silver-stained polyacrilamyde gel for the characterization of the HB101 marker. PCR product for the HB101 marker in the 36 genotypes of *H. brasiliensis* and the 6 other species of *Hevea* which more than 2 alleles can be observed for *H. guianensis*. (**1**) *H. brasiliensis*; (**2**) *H. guianensis*; (**3**) *H. rigidifolia*; (**4**) *H. nitida*; (**5**) *H. pauciflora* - (112 CNSG); (6) *H. pauciflora* - (116 CNSG); (**7**) *H. benthamiana*; and (**8**) *H. camargoana*.

Additional file

Additional file 1: Table S1. Characterization of the developed SSR markers. The table presents the 119 SSR markers developed for *H. brasiliensis*, including the primers sequence, annealing temperature, number of alleles, expected size in base pair, allelic range, observed and expected heterozigosity and polymorphism information content. The nine accession indicated in Table 1 were used for the SSR characterization.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors thank the Brazilian Agricultural Research Corporation (EMBRAPA - Amazônia Ocidental) and the Agronomic Institute for donating the analyzed *Hevea* germplasm. This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, proc. 2007/50562-4). CCM and FIS received undergraduate scholarships (2008/55974 and 2009/51757-9) and, LMS and CCM received graduate scholarships (2007/50392-1 and 2011/50188-0) from FAPESP. APS and PSG are recipient of a research fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Author details

[1]Centro de Biologia Molecular e Engenharia Genética (CBMEG) - Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6010, CEP 13083-970, Campinas, SP, Brazil. [2]Instituto Agronômico de Campinas (IAC), CP 28, Campinas SP, CEP 13012-970, Brazil. [3]Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6109, CEP 13083-970, Campinas, SP, Brazil.

Authors' contributions

CCM and FIS developed the microsatellite-enriched libraries, performed the computational microsatellite identification, designed the flanking primers and performed the microsatellite marker validation. CCM performed the statistical analysis and drafted the manuscript. LMS participated in the design and implementation of the study and the microsatellite identification. APS and PSG conceived the study and participated in its design and coordination. APS helped to draft the manuscript. All of the authors read and approved the final manuscript.

Received: 6 April 2012 Accepted: 15 June 2012
Published: 25 June 2012

References

1. Pires JM, Secco RS, Gomes JI: *Taxonomia e fitogeografia das seringueiras Hevea spp*. Belém: Embrapa Amazônia Oriental; 2002:103.
2. Gonçalves P, Cardoso M, Ortolani AA: **Origem, variabilidade e domesticação da Hevea; uma revisão.** *Pesquisa agropecuária brasileira* 1990, **25**(2):135–156.
3. Clément-Demange A, Priyadarshan PM, Thuy Hoa TT, Venkatachalam P: **Hevea Rubber Breeding and Genetics.** In *Plant Breeding Reviews*. Volume 29th edition. Edited by Janick J. New Jersey: John Wiley & Sons, Inc; 2007:ch 4.
4. Cornish K: **Similarities and differences in rubber biochemistry among plant species.** *Phytochem* 2001, **57**:1123–1134.
5. Webster CC, Baulkwill WJ: *Rubber Longman edn*. Longman Singapore Publishers Ltd; 1989.
6. Feng SP, Li WG, Huang HS, Wang JY, Wu YT: **Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*).** *Mol Breeding* 2009, **23**:85–97.
7. Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK: **Microsatellite markers: an overview of the recent progress in plants.** *Euphytica* 2011, **177**:309–334.
8. Chevallier MH: **Genetic variability of *Hevea brasiliensis* germplasm using isozyme markers.** *J Nat Rubber Res* 1988, **3**:42–53.
9. Besse P, Seguin M, Lebrun P, Chevallier MH, Nicolas D, Lanaud C: **Genetic diversity among wild and cultivated populations of *Hevea brasiliensis* assessed by nuclear RFLP analysis.** *Theor Appl Genet* 1994, **88**:199–207.
10. Le Guen V, Gay C, Xiong TC, Souza LM, Rodier-Goud M, Seguin M: **Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*).** *Plant Breeding* 2011, **130**:294–296.
11. Lespinasse D, Rodier-Goud M, Grivet L, Leconte A, Legnate H, Seguin M: **A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers.** *Theor Appl Genet* 2000, **100**:127–138.
12. Le Guen V, Garcia D, Mattos CRR, Doaré F, Lespinasse D, Seguin M: **Bypassing of a polygenic *Microcyclus ulei* resistance in rubber tree, analyzed by QTL detection.** *New Phytologist* 2007, **173**:335–345.
13. Le Guen V, Lespinasse D, Oliver G, Rodier-Goud M, Pinard F, Seguin M: **Molecular mapping of genes conferring field resistance to South American Leaf Blight (*Microcyclus ulei*) in rubber tree.** *Theor Appl Genet* 2003, **108**:160–167.
14. Lespinasse D, Grivet L, Troispoux V, Rodier-Goud M, Pinard F, Seguin M: **Identification of QTLs involved in the resistance to South American leaf blight (*Microcyclus ulei*) in the rubber tree.** *Theor Appl Genet* 2000, **100**:975–984.
15. Souza LM, Mantello CC, Santos MO, Gonçalves PO, Souza AP: **Microsatellites from rubber tree (*Hevea brasiliensis*) for genetic diversity analysis and cross amplification in six Hevea wild species.** *Cons Genet Res* 2009, **1**:75–79.
16. Yu F, Wang BH, Feng SP, Wang JY, Li WG, Wu YT: **Development, characterization, and cross-species/genera transferability of SSR markers for rubber tree (*Hevea brasiliensis*).** *Plant Cell Rep* 2010, **30**:335–344.
17. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nat Genet* 2002, **30**:194–200.
18. Marconi TG, Costa EA, Miranda HRCAN, Mancini MC, Cardoso-Silva CB, Oliveira KM, Pinto LR, Mollinari M, Garcia AAF, Souza AP: **Functional markers for gene mapping and gnetic diversity studies in sugarcane.** *BMC Research Notes* 2011, **4**:264.
19. Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ: **Analysis of SSRs derived from grape ESTs.** *Theor Appl Genet* 2000, **100**:723–726.
20. Thiel T, Michalek W, Varshney RK, Graner A: **Exploiting EST databases for the development and characterization of genederived SSR-markers in barley (Hordeum vulgare L.).** *Theor Appl Genet* 2003, **106**:411–422.
21. Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T: **Characteristics, development and mapping of Gossypium hirsutum derived EST-SSRs in allotetraploid cotton.** *Theor Appl Genet* 2006, **112**:430–439.
22. La Rota M, Kantety RV, Yu JK, Sorrells ME: **Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley.** *BMC Genomics* 2005, **6**:2.
23. Chen C, Zhou P, Choi YA, Huang S, Gmitter FG Jr: **Mining and characterizing microsatellites from citrus ESTs.** *Theor Appl Genet* 2006, **112**:1248–1257.
24. Fraser LG, Harvey CF, Crowhurst RN, De Silva HN: **EST-derived microsatellites from Actinidia species and their potential for mapping.** *Theor Appl Genet* 2004, **108**:1010–1016.
25. Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L: **Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species.** *Theor Appl Genet* 2007, **114**:359–372.
26. Jung S, Albert A, Jesudurai C, Tomkins J, Main D: **Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs.** *Funct Integr Genomics* 2005, **5**:136–143.
27. Saha T, Roy CB, Nazeer MA: **Microsatellite variability and its use in the characterization of cultivated clones of *Hevea brasiliensis*.** *Plant Breeding* 2005, **124**:86–92.
28. Perry AB: **Chromosome Number and Phylogenetic Relationships in the Euphorbiaceae.** *Am J Bot* 1943, **30**:527–543.
29. Leitch AR, Lim KY, Leitch IJ, O'Neill M, Chye M, Low F: **Molecular cytogenetic studies in rubber, *Hevea brasiliensis* Muell. Arg. (Euphorbiaceae).** *Genome* 1998, **41**:464–467.
30. Majumder SK: **Chromosome studies of some species of *Hevea*.** *J. Rubber Res. Inst. Malays* 1964, **18**:269–273.
31. Ong SN: **Chromosome morphology at pachytene stage in *Hevea brasiliensis*: a preliminary report.** In *Proceedings of the International Rubber Conference*. Edited by Sripathi B. Kuala Lumpur: Rubber Research Institue of Malaysia; 1975:3–12.
32. Bouharmont J: **Recherches taxonomiques et caryologiques chez quelques espèces du genre *Hevea*.** In *Série Scientifique*. Volume 85th edition. Edited by INEAC.; 1960:64.
33. Doyle JJ, Doyle JL: **A rapid DNA isolation procedure for small quantities of fresh leaf tissue.** *Phytochem Bull* 1987, **19**:11–15.

34. Billotte N, Lagoda P, Risterucci AM, Baurens FC: **Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops.** *Fruits* 1999, **54:**277–288.
35. Temnykh S, Declerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S: **2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential.** *Genome Res* 2001, **11:**1441–1452.
36. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM: **Primer3Plus, an enhanced web interface to Primer3.** *Nucleic Acids Res* 2007, **35:**W 71–W74.
37. Creste S, Tulmann-Neto A, Figueira A: **Detection of single repeat polymorphisms in denaturing polyacrilamide sequencing gels by silver staining.** *Plant Mol Biol Rep* 2001, **19:**229–306.
38. Miller MP: *Tools for population genetic analysis (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data.*: Computer Software distributed by the author; 1997.

doi:10.1186/1756-0500-5-329
Cite this article as: Mantello *et al.*: **Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species.** *BMC Research Notes* 2012 **5**:329.

Additional file 1 – Characterization of all of the developed SSR markers.

SSR locus	Gene bank acession n°	Primer sequence (5'-3')	predicted product size (bp)	Repeat motif	TA[b](°C)	NA[c]	size range (bp)	H_e[d]	H_o[e]	PIC[f]
HB 31[a]	JQ655710	CCACTGTTTGTGTCATTTGGA TGGACAAGCAAAACAGCCTA	174	$(GGAA)_3$	60	6	173-182	0,8083	1	0,7234
HB 32[a]	JQ655711	TCCCATTCCATTCCATTCA AAACCGATAGCAAGCACCTC	238	$(CT)_6$ CC $(CT)_6$	58,8	4	236-242	0,7059	0,2222	0,6072
HB 33[a]	JQ655712	GATAATTGCTTCCTGGTCA AATGCCTTCTTTCCCTAAC	188	$(ATG)_3(ATGA)_5$	50	6	181-188	0,7917	0,3750	0,7133
HB 34[a]	JQ655713	TGGGGGATATCACAACACCT CAGCCGATCTCGATTCATAAC	195	$(GAT)_6$	60	1	195	0	0	0
HB 35[a]	JQ655714	ACGCGTGAGTCTTGAGTAAAT AGTGGCCAAGAAAGAATAAAA	157	$(TCC)_6$	60	2	156-159	0,2941	0,3333	0,2392
HB 36[a]	JQ655715	TACTACCCATCCACCAACCTAA TCTCTGCAACGATAATGAACG	230	$(AG)_{18}$	60	4	217-239	0,6732	0,6667	0,5652
HB 37[a]	JQ655716	AGTATTCATCAGCTTCGCACAT TTTATCATCATCACCATCATCC	155	$(GAA)_8$	47,5	3	154-161	0,6333	0,0000	0,5112
HB 39[a]	JQ655625	GAGGCCTTCGAAATGATGG ATGAATGGTTGCTTGGTGAAT	204	$(ATC)_{10}$	60	1	204	0	0	0
HB 41	JQ655599	AACCAGCCACTCTTTTGACCT TCCTTGCATGTTGATTTCTGTT	173	$(AC)_{23}$	56	5	141-173	0,6732	0,5556	0,5995
HB 42	JQ655600	AAATTGGCAGCTTGTTCTTG TCGGTTGGATATGGTGTCA	214	$(TG)_{10}C(GT)_6$	56	5	203-214	0,5556	0,6667	0,4966
HB 43	JQ655601	TTGTCTCCCCTTAATTCTGCTCTT GTGATCTGCCCATAACTACTCCAT	214	$(TC)_{18}$	56	6	221-259	0,7320	0,6667	0,6597
HB 45	JQ655602	GTCAGAAGCAACCCACAAAC ATGCTTACCCAAAAATCAATG	170	$(AG)_{16}$	56	7	156-201	0,7647	0,6667	0,6803
HB 47	JQ655603	ATCCTTGCATTCACTTCTCT ATCCCACTATAATTTGTTTTCTC	170	$(TC)_8$	60	3	152-185	0,5817	0,2222	0,489

HB 48	JQ655604	CGGGAATCGACGACTGAAATGG CGCCGCGTATGCACCTGAAC	223	$(GT)_5$	60	1	215	0	0	0
HB 50	JQ655605	TCCTTGACCCAGAAACATTG TGTTCTTTTCGGCAGGCTAT	210	$(GT)_{10}$	52	3	196-210	0,3072	0,3333	0,2686
HB 51	JQ655606	AACCATCGCTACCTTCTGCTA TTTGGTTAATGTTCGTGCTCA	214	$(TG)_{14}$	52	2	196-201	0,5385	0,1429	0,375
HB 52	JQ655607	ACCCTACTATCCTATCGTCTTG AAAATCGTAGCTTCTTCATCAG	207	$(TC)_{21}$	60	6	176-202	0,8235	1	0,7429
HB 53	JQ655608	CCAGCTTGAGGGAGAGTGTT TAGCAAGGAGGGAGAAAAGAG	216	$(TG)_{11}$	64	6	204-223	0,8170	0,8889	0,7368
HB 54	JQ655609	AAAAAGAAGAAAAGAAAGAAGAGA ATATCCGAGTTTGAAGAGGTAGGT	176	$(AG)_{13}$	60	4	168-183	0,6405	0,4444	0,5501
HB 55	JQ655610	AGATGGGCATGTTTTCGTTGTA TTTTTGCCCCTCCCTTATCA	171	$(GA)_{15}$	60	6	174-192	0,7451	0,5556	0,6611
HB 56	JQ655611	TTGAATTGAGAAGAAGGTGAA CCGACAACTGGAACTAATACT	257	$(AG)_5$	56	1	255	0	0	0
HB 57	JQ655612	AGCCTGCGATCACAACAAT TGCCTCACACCCTCACCT	174	$(AG)_{16}$	60	5	147-174	0,6000	0,7500	0,525
HB 58	JQ655613	GCAGAGATGAGAAGGCAAGTAG AACCCAATAACGTCAACAAATC	189	$(TC)_{12}$	60	5	161-187	0.6928	0.7778	0,6035
HB 60	JQ655614	TGCTTCTTCTCTTCATAACCTCAT AGCAGCCAACAATTCTCAGTG	164	$(CT)_{13}$	52	2	151-161	0,1250	0,1250	0,1103
HB 61	JQ655615	AACACATTTTGACCTGACTTCT AGGTGCTTAGCTTACATTTTTC	171	$(GA)_{14}$	56	4	150-172	0,5948	0,7778	0,5117
HB 62	JQ655616	CTAGGCTGGCTGGGAAATCAAA CCAGGCACAAGGGGAAAAA	130	$(TCT)_6$	60	3	122-112	0,3417	0,1250	0,2944
HB 63	JQ655617	AGGTTACGTCTTGGGATAGTTG TTTTTCACCCAGGTCAGCATAA	253	$(TG)_9(GA)_{11}$	60	5	198-236	0,7059	0,5556	0,6213
HB 64	JQ655618	AATCCACCCAGCCTTACAG AATTGGTGAGCTTCGTTTTT	174	$(AG)_{17}$	60	7	146-199	0,7451	0,8889	0,6789

HB 66	JQ655619	TATGGGAAATGGTAAGACACA AAAGAATATTGCAAAAACAGG	270	$(TTC)_5$	60	3	273-279	0,5033	0,4444	0,4035
HB 68	JQ655620	GTTATTTTCCCCATCCTTTGA CTCCGCATTTGTTTTCTGTG	148	$(CT)_{15}$	60	6	152-167	0,8105	0,5556	0,7297
HB 69	JQ655621	CGTTATTCCCAGGGTAAGATG GCCTAATCCGGTAGCTGTCA	162	$(CT)_{23}$	56	5	155-172	0,6667	0,7500	0,586
HB 70	JQ655622	CAGCAGATGAGTCCGCAGTT TCCAAGAATGATCCCAAGAGC	168	$(CTG)_7$	56	3	161-167	0,5686	0,2222	0,4408
HB 71	JQ655623	ATTTGCTAGAGTTTGGTTGTGA ATGTTATGGGATCTCGTATTC	187	$(CA)_7$	54	2	186-188	0,1111	0,1111	0,0994
HB 73	JQ655624	ATTGCCCATAAAACCCTTCC AACCCCAGATGCACTTCAGAT	226	$(AC)_{10}$	60	3	206-228	0,5686	0,4444	0,4408
HB 77	JQ655626	ACTAGCGGCACAAGAGGTTTTT TTGGGGGTTTATGCAGTGTTCA	150	$(CT)_{18}$	60	3	140-153	0,6209	0,4444	0,5174
HB 78	JQ655627	CCATAGGGAAAGAAAAGAAAAA TCAACCTGTCCAAAAGAGAAG	200	$(TC)_9$	56	7	186-210	0,8497	1,0000	0,7777
HB 79	JQ655628	CTGGTGTGAAAACGGTAGAA AGAATAGAAAAAGCCTCAAAAT	184	$(ATT)_5$	56	1	176	0	0	0
HB 80	JQ655629	AGTTTCCTGGCATCTTCTGA TTTGCATGTCTGTTACTGTTGT	188	$(AT)_5\ (GT)_8$	56	1	177	0	0	0
HB 81	JQ655630	CAAAAACCTTGCCCCAGTAA GTGCTTTGTGTGCATGTGTG	165	$(AG)_{20}$	60	4	187-212	0,7167	0,8750	0,6116
HB 82[a]	JQ655631	ATATGCACGTGCTTGAATGT AAGCGACTAGTGTTATGAAATC	166	$(CCA)_4\ (CCT)_6$	56	4	168-175	0,7000	1,000	0,5945
HB 83[a]	JQ655632	AAGAAACGAAAAAGGAGATAAT TAACCCTAATAATGTGCCAATG	180	$(AAAG)_4$	60	3	165-181	0,5033	0,6667	0,4035
HB 85[a]	JQ655633	GCTGATCTAACTTTGTTACTCG GAATCAACCGTGGACTCAT	195	$(GAG)_7$	56	1	195	0	0	0
HB 86[a]	JQ655634	AGTGGCCAAGAAAGAAAAAG TGTGTAAATGTATGCGTGAGTC	148	$(GGA)_4$	56	1	169	0	0	0

HB 87[a]	JQ655635	TCACTGAGAGGGAGGGGAAGAG GGCCAAACCATCGCAAACA	183	$(AGG)_4$	56	1	181	0	0	0
HB 89[a]	JQ655636	AAAGAAAAGCAATATCCAGTCA TCCTCTTCATCTTTCTTTCATC	133	$(GAT)_8$	56	1	133	0	0	0
HB 91[a]	JQ655637	TGGACATAAAGCAAATAAAAGA GAGAAGCTAGATGAGGAGGAAG	160	$(TGA)_5$	56	1	158	0	0	0
HB 92[a]	JQ655638	TTTTCCTTTTGTCTTTTTCTTT ACCCACATAACCTTCATCACAT	182	$(GTTTT)_3$ ...$(CCT)_7$	TD- 3	4	233-255	0,7516	0,7778	0,6571
HB 94[a]	JQ655639	TGGGGGATATCACAACACC CGATTCATAACGAGATTCAGA	185	$(GAT)_6$	TD- 3	1	185	0	0	0
HB 95[a]	JQ655640	TGGAATTCGCAAAAGAGCAC GAATGATCGAGGGAAAATGGA	219	$(CTC)_6$	60	2	225-245	0,4706	0,2222	0,3457
HB 98[a]	JQ655641	CTTGTGGTCGCCTTTCTT TGATTGTGCTTTGATTCTATTC	184	$(CAT)_6$	TD- 3	3	190-180	0,3072	0,3333	0,2686
HB 99[a]	JQ655642	GTAATGAATGGTTGCTTGGTGA GAGGCTTCGAAATGATGGAG	195	$(GAT)_6$	TD- 3	1	195	0	0	0
HB 100[a]	JQ655643	AGCTGCATTGACTTTTGTTC GCGTGGGTTAATCATCAG	218	$(CTT)_9$	60	6	214-250	0,8170	0,8889	0,7368
HB 101[a]	JQ655644	TCTTCATCCTTCTTCATCAACT TACGTCACAGCTAAAAAGAAAA	144	$(CAT)_9$	60	2	132-142	0,2092	0,2222	0,178
HB 102[a]	JQ655645	TTCCGAATGCATGTTTCTTG AATAAATAAGGCGGCAGCAG	195	$(CTC)_6$	56	2	157-195	0,1111	0,1111	0,0994
HB 103[a]	JQ655646	AATATGGATCAGGCAAAGCAG ATTCTCTTCCTCCGCTCCTC	165	$(GAG)_5$	56	2	161-164	0,2092	0,2222	0,178
HB 104[a]	JQ655647	TGACTTGCTGCCAATATGAA CCATAGCCTCTCCCCTCTCT	158	$(ATA)_5$...$(GGAT)_5$	TD- 3	4	156-173	0,6923	0,5714	0,5849
HB 105[a]	JQ655648	CTACGGTGGGCATTGGTAAT GCCAAGTATTTGCATCTTCG	200	$(GA)_{19}$ $(GAT)_4$	56	9	176-224	0,8954	1,0000	0,8291
HB 106	JQ655649	GGGTGACAGCTTGGTTTCAT CCAATAAGAAGAGAGGGGAACA	235	$(CA)_{16}$ $(CA)_{10}$	56	4	214-230	0,6993	0,4444	0,6111

HB 109	JQ655650	CTAACTAGAGCCTCCGTATGTC TGCCACCCAGTTCAGTAAG	192	$(CATAAA)_3$	60	3	190-198	0,2037	0,2157	0,1939
HB 110	JQ655651	ATGCAGCGATGTAGATAAAAGA TCAAGATGTAAGCACCAGAACT	285	$(GC)_5(CA)_5...(GA)_{14}$	56	7	260-273	0,8562	0,7778	0,782
HB 111	JQ655652	TTCCTTCCAAACCTCAATACT AAAACTAATCAAAGCAAATCTG	149	$(TCTG)_3$	56	1	147	0	0	0
HB 115	JQ655653	ATCTGCATCCCAAGGGACTA AAGCATCAACCCTCTCTTCATC	219	$(GA)_{14}$	56	0	212	0	0	0
HB 117	JQ655654	CATAGCCTCACGCTACCTCA TGGGTGAGTGGATGTTGTTC	197	$(GT)_{22}$	60	6	161-190	0,7647	0,8889	0,6868
HB 118	JQ655655	GCACAAGCAATCACATCC TCTCTTCTTAACTCAACATCAA	177	$(AC)_5(AG)_{16}$	TD -1[g]	5	195-208	0,6732	0,3333	0,5995
HB 121	JQ655656	GCTACTGGGCTGTGACTATGG ATTTTGGTTTTGCAGATGACG	212	$(TG)_{15}$	TD -1[g]	4	218-245	0,7124	1,0000	0,6132
HB 122	JQ655657	TCTCGGAATTTTGCCTCTC AAGCGAAGCGGTTTATTGA	170	$(CA)_5...(TC)_9(TG)_6$	TD -1[g]	3	188-200	0,5917	0,5000	0,4555
HB 123	JQ655658	GGGCCAACATGACAACAG TGAAGCGAGATCTACAAACACT	170	$(TC)_{19}…(AT)_5$	TD -1[g]	7	175-205	0,8170	0,8889	0,744
HB 127	JQ655659	ACCTACGTCAACTACCAAAAAC AAACATCAAATTCCAAACAGA	215	$(CT)_{15}$	TD -1[g]	4	241-260	0,3987	0,3333	0,354
HB 128	JQ655660	CGTGTAAAGTGTTAGCCAAA AGGGTTGCAATATTCTCCTT	216	$(AC)_{12}…(AG)_{17}$	TD -1[g]	2	202-235	0,5033	0,5556	0,3623
HB 132	JQ655661	AATTGGGATCTTAGTGTTTTT TTTTCTGGAGCTAGTTGTGTAA	20	$(CT)_{10}…(CT)_8$	TD -1[g]	2	265-270	0,3660	0,2222	0,2859
HB 133	JQ655662	GCGCTGCAGTTCCTCTTC TCTCATCAATACAAACCCACTC	225	$(TC)_{16}$	TD -1[g]	5	234-243	0,7712	0,6667	0,6844
HB 134	JQ655663	AAAGTTCTACAAAATGGACACA CAGATGACACTCGGGTATGA	231	$(CA)_{22}$	TD -1[g]	1	231	0	0	0
HB 135	JQ655664	CAACTCTCCACCCTTCTTTCTA AATCGCTGTTTCATCTGTCAA	225	$(TC)_{17}(CA)_{19}$	TD -1[h]	6	222-250	0,8627	0,7778	0,7891

HB 137	JQ655665	CTCCCCTTATTTTGGCTTCC TGTGTCATGACCGTATTTCTG	208	$(TC)_{28}…(CT)_{5}…(TC)_{23}$	TD -1[g]	1	208	0	0	0
HB 139	JQ655666	CAAAATGGAGGGATAAATAAAT ACTAGGAGGTACCCATCTTCAG	238	$(GA)_{15}$	TD -1[h]	4	232-260	0,6417	0,3750	0,5471
HB 140	JQ655667	GATGGCAGGGGAGAGTTC CCAAAAGTCGCAAAATAGG	199	$(TAT)_{4}$	TD -1[g]	1	212	0	0	0
HB 141	JQ655668	AGGGCCTTCAAGtACAATCT GCTTCTTTTAGCAGTGTGGT	188	$(GT)_{9}$	TD -1[g]	5	190-205	0,7386	0,3333	0,65
HB 142	JQ655669	CCTGTTGACAATTACTGTCTTG ATCTCTCGCTGGCGATACTA	204	$(AGAC)_{3}$	TD -1[g]	1	204	0	0	0
HB 144	JQ655670	ATTTGGAGCACTCACACACT TGACTCTCTCATTTGGGATT	178	$(TG)_{5,15}$	TD- 2	3	185-200	0,3922	0,4444	0,3402
HB 148	JQ655671	ATAAGGATCATGTCCCAGTG TGATGAAACTGTTGTGGATG	164	$(AC)_{18}$	TD -1[g]	1	160	0	0	0
HB 149	JQ655672	CCCACTTCATTCTCTTCTCA CTTTTGCAGGAACATAGACC	205	$(CT)_{18}$	TD -1[g]	2	205-233	0,4706	0,4444	0,3457
HB 150	JQ655673	AATGGACACACACACATCAG GATCAGATGACACTCGGGTA	222	$(CA)_{22}$	TD -1[g]	4	231-248	0,7417	0,1250	0,6357
HB 151	JQ655674	ACCTTTTGATGAGAGTGTGC CTCTTTGTTGAGCGTGTGT	177	$(TC)_{10}(CA)_{11}(CG)_{6}$	TD -1[g]	7	186-220	0,8497	0,6667	0,7758
HB 152	JQ655675	TATTTTGGAGCTTTGGGTGTTC CTGAGAGCGTTGTATGGGTGTG	197	$(CT)_{17}(CA)_{7}$	TD -1[g]	7	190-224	0,8562	0,6667	0,7858
HB 154	JQ655676	GACCATTTCCACTGCTCAA GGAAAAAGAAAAACCCAAGAAT	245	$(TC)_{5}$	TD -1[g]	1	245	0	0	0
HB 155	JQ655677	AACAAGACATAGCAAAAAGAAT TTAAAGAACCCAAGGAAATG	258	$(TG)_{11}(GA)_{20}$	TD -1[g]	4	248-258	0,7582	0,6667	0,6658
HB 156	JQ655678	TTCTAATACCTTTGGGGACTC AACAAAAGAATCGAAGGAAGA	258	$(CA)_{7}$	TD -1[g]	1	258	0	0	0
HB 157	JQ655679	TGGCATTCGAGGCAAAAA TTAAATGGAAGGAAAGGAAGG	245	$(TG)_{11}(TGTC)_{3}$	TD- 2	1	245	0	0	0

HB 158	JQ655680	CAACCATGGTCAATGACTCG ACTGTGATCCTGCACGTGTT	241	$(AAC)_4$	TD -1[g]	1	241	0	0	0
HB 160	JQ655681	TTGGGTTTAGTGATGCCTATTC GCCACAGCTAAATTGCCTAAA	163	$(TG)_{10}$	TD -1[g]	5	173-188	0,4052	0,4444	0,3677
HB 161	JQ655682	GAGCCATGAACACCAAACAA GTGCTTTCAATCCCCGATAG	221	$(AG)_8...(AG)_7$	TD -1[g]	3	213-234	0,2157	0,2222	0,1939
HB 162	JQ655683	TGACTACTCTCTTCCCGTCTCT ATTTGCCTGCTCTCTTCACC	188	$(TGCA)_3$	TD -1[g]	1	205	0	0	0
HB 163	JQ655684	CTCTATTTGCTGAGAAAGTCCA CCATCAAACGTCAGCAAGA	250	$(TG)_{22}$	TD -1[h]	4	237-265	0,7333	0.8000	0.6102
HB 164	JQ655685	TCTGCATAAGATGGTGAGGGTA TGATCCTGGCTCCATACATTC	178	$(TG)_{20}$	TD- 2	1	178	0	0	0
HB 166	JQ655686	CTGGACCCATTTCGTGCTA GGATATCCTACCCGTTACCC	186	$(GA)_{18}$	TD- 2	6	182-205	0,6275	0,4444	0,5664
HB 167	JQ655687	ATGCAGACTAATTAACAAGAGC TTAAACAATGAACAATGGACAC	225	$(AC)_6...(GA)_{14}$	TD- 2	4	228-240	0,7059	0,4444	0,6072
HB 168	JQ655688	CATTTAACTAGGATTGCTTGTG CATACTGCCCTGATTGACC	177	$(GT)_9$	TD -1[g]	2	180-190	0,1111	0,1111	0,0994
HB 170	JQ655689	GCTTCAGCTATGGACTACCC TTGATGAAACTGTTGTGGATG	141	$(AC)_{18}$	TD -1[g]	7	262-290	0,8627	0,7778	0,79
HB 171	JQ655690	TAGGCCCTGACTTGCTCTGATA AATTGCCCACTCTTGCTTACG	169	$(AC)_{18}$	TD- 2	1	169	0	0	0
HB 172	JQ655691	TTTCAGCTCACTCAACAATAAG ACAGCAGATAGTAGATGAACCT	130	$(AC)_{11}$	TD -1[g]	2	140-150	0,5033	0,3333	0,3623
HB 173	JQ655692	CTCCCTCCCGCATTATCTGT TATCCTACCTTGAAGCACTGAA	190	$(TG)_{20}$	TD -1[g]	3	192-200	0,5033	0,6667	0,4035
HB 174	JQ655692	TTCAGTGCTTCAAGGTAGGAT GAACAGGGAATATAAACGAACA	138	$(TTC)_{11}$	TD -1[g]	7	138-165	0,7386	0,8889	0,6695
HB 176	JQ655693	GAATGGAGGAATTGCGATTTAC TTTTTGCATGCCCATATGTC	229	$(GT)_{15}$	TD -1[g]	5	245-270	0,8333	0,8333	0,726

HB 177	JQ655694	TCTCTTTACTAATTTGCTTTGA ACCACCTCCACTATTCTATCTG	161	$(AC)_{13}$	TD- 2	6	177-190	0,8417	0,5000	0,7588
HB 178	JQ655695	GGTTTCTGCTCCTTGACAC AACTTTTCCTATTGACCCATTA	265	$(CT)_{20}$	TD -1[g]	6	260-285	0,7843	0,7778	0,699
HB 179	JQ655696	TTATTTGCCTTGGGGAACT GTGGATGGGCAGGACAGG	189	$(AC)_{7}$	TD -1[g]	4	210-228	0,7124	0,5556	0,6119
HB 182	JQ655697	TGCAGGGCCAAGAAGAGATG TACAAAGCAAGGGGGATGACC	164	$(TC)_{20}$	TD- 2	5	170-180	0,7333	0,2500	0,6533
HB 183	JQ655698	TCTAATGCCTTTGGGACTCT TTCAAAATAGCTAGGGTGTTC	164	$(CA)_{7}$	TD -1[g]	3	176-183	0,4510	0,5556	0,3709
HB 185	JQ655699	TAGGACCATATCAGAAAAGAAC TTAAACCAGAGGAAACTATCA	173	$(GA)_{17}$	TD- 2	3	180-200	0,5229	0,2222	0,4377
HB 186	JQ655700	CTCTCCCTTTCTCTTGTTTTT TTAGTATTGATCCTCGGCTGAA	264	$(TC)_{19}$	TD- 2	5	250-270	0,7190	0,7778	0,6402
HB 188	JQ655701	ATGACATAACCCTAGAGCAACC TAGGGAGATGAAATGAAGGAAC	249	$(TG)_{7}$	TD- 2	1	249	0	0	0
HB 189	JQ655702	AACTGGATTTGGTCATCTCT TTCAGAAGAACACACAGTGA	225	$(AC)_{9}$	TD- 2	3	220-230	0,3072	0,3333	0,2686
HB 190	JQ655703	TTAGGTATTCTACGGGATCA AGCCCATTAAAACCTTACTC	157	$(CT)_{12}(TC)_{15}$	TD -1[g]	9	178-299	0.9150	0.7778	0,8496
HB 192	JQ655704	CTGATAAATTAGGGTTCCAC ATTGAGATTAGCTGAAGAGC	183	$(TC)_{20}$	TD -1[g]	4	166-190	0,6601	0,4444	0,5567
HB 194	JQ655705	CGTGACTCACAACTTTCTTT AGGAGCCATTTTTGAGAG	222	$(CT)_{11}$	TD -1[g]	4	210-240	0,6863	0,5556	0,5917
HB 195	JQ655706	CACCGTGTGACATAACTTACTC CACTAGCAGGCTTGGGAAA	161	$(AC)_{10}…(CT)_{5}$	TD -1[g]	4	162-167	0,6340	0,6667	0,5392
HB 197	JQ655707	GGAGGGTGAATGTTCCACA TAGGCTCCATTTGCTTCGAT	213	$(AG)_{20}$	TD -1[g]	7	206-238	0,8170	1,000	0,744
HB 198	JQ655708	TCGAAAGTGCTCAGACCAGA CCGAACTTAGGGTGTTACAGAg	158	$(TC)_{20}(CA)_{13}$	TD -1[g]	5	152-174	0,4837	0,2222	0,434

HB 199	JQ655709	ACTTGGAGCACTCACACACC AGAAAGTCCGGCTGAGTTGA	186	$(AT)_5…(GT)_{15}(GA)_{19}$	TD -1[g]	1	186	0	0	0

a-primer pairs developed by trinucleotide enriched library
b-annealing temperature
c-allele number
d-expected heterozygosity
e-observed heterozygosity
f-polimorphism information content
g- 55ºC to 50ºC (-0.5ºC per cycle)

Capítulo II

OPEN ACCESS Freely available online

PLOS ONE

De Novo Assembly and Transcriptome Analysis of the Rubber Tree (*Hevea brasiliensis*) and SNP Markers Development for Rubber Biosynthesis Pathways

Camila Campos Mantello[1]*, Claudio Benicio Cardoso-Silva[1], Carla Cristina da Silva[1], Livia Moura de Souza[1], Erivaldo José Scaloppi Junior[2], Paulo de Souza Gonçalves[3], Renato Vicentini[1], Anete Pereira de Souza[1,4]*

1 Centro de Biologia Molecular e Engenharia Genética (CBMEG) - Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Campinas, São Paulo, Brazil, **2** Agência Paulista de Tecnologia dos Agronegócios, Pólo Regional Noroeste Paulista, Votuporanga, São Paulo, Brazil, **3** Instituto Agronômico de Campinas (IAC), Campinas, São Paulo, Brazil, **4** Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Campinas, São Paulo, Brazil

Abstract

Hevea brasiliensis (Willd. Ex Adr. Juss.) Muell.-Arg. is the primary source of natural rubber that is native to the Amazon rainforest. The singular properties of natural rubber make it superior to and competitive with synthetic rubber for use in several applications. Here, we performed RNA sequencing (RNA-seq) of *H. brasiliensis* bark on the Illumina GAIIx platform, which generated 179,326,804 raw reads on the Illumina GAIIx platform. A total of 50,384 contigs that were over 400 bp in size were obtained and subjected to further analyses. A similarity search against the non-redundant (nr) protein database returned 32,018 (63%) positive BLASTx hits. The transcriptome analysis was annotated using the clusters of orthologous groups (COG), gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Pfam databases. A search for putative molecular marker was performed to identify simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs). In total, 17,927 SSRs and 404,114 SNPs were detected. Finally, we selected sequences that were identified as belonging to the mevalonate (MVA) and 2-C-methyl-D-erythritol 4-phosphate (MEP) pathways, which are involved in rubber biosynthesis, to validate the SNP markers. A total of 78 SNPs were validated in 36 genotypes of *H. brasiliensis*. This new dataset represents a powerful information source for rubber tree bark genes and will be an important tool for the development of microsatellites and SNP markers for use in future genetic analyses such as genetic linkage mapping, quantitative trait loci identification, investigations of linkage disequilibrium and marker-assisted selection.

Citation: Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, et al. (2014) *De Novo* Assembly and Transcriptome Analysis of the Rubber Tree (*Hevea brasiliensis*) and SNP Markers Development for Rubber Biosynthesis Pathways. PLoS ONE 9(7): e102665. doi:10.1371/journal.pone.0102665

Editor: Minami Matsui, RIKEN Biomass Engineering Program, Japan

Received December 20, 2013; **Accepted** June 22, 2014; **Published** July 21, 2014

Funding: This work was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (562979/2010-7 and 478701/2012-8) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (2007/50562-4 and 2012/50491-8). The scholarships were provided by FAPESP to CCM (2008/55974-1, 2011/50188-0), CBCS (12/11109-0), CCS (2009/52975-0) and LMS (12/05473-1). APS, PSG and RV are recipients of a research fellowship from CNPq. 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.

* Email: anete@unicamp.br (APS); camila.mantello@gmail.com (CCM)

Background

Natural rubber is one of the most important polymers that is produced by plants. Rubber is composed of 94% cis-1,4-polyisoprene and 6% proteins and fatty acids [1] and exhibits unique properties including flexibility, impermeability to liquids and abrasion resistance. These singular properties make natural rubber superior to synthetic rubber for use in various applications [2].

Natural rubber is used in more than 40,000 products, including over 400 medical devices, and is of great importance in the tire industry [2]. Approximately 2,500 plant species are known to synthesize natural rubber, but only a few plants, such as *Hevea brasiliensis* (rubber tree), *Parthenium argentatum* (guayule) and *Taraxacum koksaghyz* (Russian dandelion), can produce high-quality natural rubber with molecular weights of greater than 1 million Daltons [3]. Among these species, *H. brasiliensis* (Willd. ex Adr. de Juss.) Muell. –Arg., which is commonly referred to as the rubber tree, is the major source of natural rubber [2] and is planted on a large scale in fields encompassing approximately 11.33 million hectares [4].

H. brasiliensis, which is native to Amazon rainforests, is a diploid (2n = 36, n = 18), perennial, monoecious, cross-pollinated tree species [5], with an estimated haploid genome estimated of 2.15 Gb [6]. The genus *Hevea* belongs to the Euphorbiaceae family, which is comprised of 11 inter-crossable species [7].

Although the Amazon rainforest offers optimal conditions for growth and high rubber yields due to its warm and humid climate, this region also provides optimal conditions for South American leaf blight (SALB) disease, which is caused by the fungus *Microcyclus ulei* (P. Henn.) v. Arx. and was responsible for devastating plantations in northern Brazil in the 1930s. SALB remains a permanent threat to the rubber industry [8]. Because of this disease, rubber tree plantations have expanded throughout the world, in locations such as northeastern India, the highlands and

coastal areas of Vietnam, southern China and the southern plateau of Brazil [9]. These areas are colder and drier than the Amazon rainforest and are not favorable for the growth of this fungus. However, they are associated with other types of stresses, such as low temperatures, strong winds and drought, that are limiting factors for rubber production [5]. Thus, rubber tree breeding programs have focused not only on genotypes that are resistant to SALB disease but also on those that are tolerant to the stress conditions found in these areas and are producers of high quality rubber.

Similar to many perennial trees, rubber tree breeding is time consuming and expensive. An average of 25 to 30 years of field experiments in large areas is generally required to obtain a new cultivar. Thus, molecular biological techniques could optimize field evaluations, thereby reducing the time and area that are required for these experiments.

Over the past two decades, there has been an exponential increase in data acquisition pertaining to the rubber tree, including with regard to genomic microsatellite markers [10,11], expressed sequence tag-simples sequence reapeats (EST-SSRs) [12–14], linkage maps [15,16] and gene expression profiles [17,18]. More recently, a draft genome of the rubber tree was published [19]. High-throughput genomic techniques are associated with innovative bioinformatics tools that can be important to rubber tree breeding and facilitate the development of superior clones that are suited to different agroclimatic conditions [4].

With the reduction in the cost of next generation sequencing (NGS) technologies, RNA sequencing (RNA-seq) has become wide spread because it enables the high-resolution characterization of transcriptomes. This method provides many advantages, including a single-base resolution, enabling the detection of thousands of single nucleotide polymorphisms (SNPs) for further SNP marker development. These markers can be useful for the functional saturation of linkage maps and the identification of markers that are directly related to economic traits for marker assisted selection (MAS). In addition, RNA-seq can be employed to provide information about alternative splicing, to detect rare transcripts and to quantify different levels of expression of individual genes rather than total gene expression, in contrast with microarrays [20].

RNA-seq has become a valuable tool that has been used in the investiagation of many species, such as *Arabidopsis* [21], rice [22] and maize [23]. This technology has also been widely used in non-model species such as the rubber tree [24].

A search for *H. brasiliensis* in the National Center of Biotechnology Information (NCBI) revealed that approximately 40,000 EST sequences had been deposited (as of August 2013). Recently, a transcriptome profile for a mixture of leaves and latex was described [25] in addition to, a bark transcriptome and EST-SSRs markers have been developed [14]. Both of these studies used Illumina HiSeq 2000 technology. RNA-seq employing 454 pyrosequencing technology has also been applied to evaluate the apical meristem transcriptome to facilitate the development of EST-SSR markers and the construction of a genetic linkage map [13].

In the current study, a total of 166,731,798 high-quality reads from bark samples from the GT1 and PR255 clones were obtained through paired-end sequencing using Illumina GAIIx platform to generate a *de novo* assembly. The GT 1 clone, which is male-sterile, and PR 255 are good latex producers in São Paulo State and are parental to two mapping populations. These clones high yielding and cold and wind tolerant, which are important characteristics for rubber tree breeding. The obtained transcripts were submitted for functional annotations, through which it was possible to identify new genes in the *H. brasiliensis* database. The transcripts were also submitted for putative SSR and SNP discovery. A total of 78 SNP markers were validated in the mevalonate (MVA) and 2-C-methyl-D-erythritol 4-phosphate (MEP) pathways, which are two important pathways that are involved in rubber biosynthesis.

Materials and Methods

Ethics statement

We confirm that no specific permits were required for the present study. This work was a collaborative research project that was developed by researchers from the University of Campinas (UNICAMP) and the Agronomic Institute of Campinas (IAC). In addition, we confirm that the field study did not involve endangered or protected species.

Plant materials, and DNA and RNA extractions

Bark samples from the GT1 and PR255 clones were collected at the Agência Paulista de Tecnologia dos Agronegócios/SAA, Votuporanga, São Paulo, Brazil. The selected clones were 18 years old and were tapped once every 4 days. The bark samples were frozen on dry ice and stored at −80°C until use. Total RNA was extracted according to Changet et al. [26]. RNA quality and integrity were evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA).

To validate the SNP markers, genomic DNA from 36 genotypes of *H. brasiliensis* (Table S1) was extracted from lyophilized leaf tissues using the modified CTAB method as described by Doyle JJ and Doyle JL [27], and the quality and quantity of the obtained DNA were measured by electrophoresis using a 1% agarose gel and spectophotometrically using the NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, DE).

cDNA library construction and sequencing

Paired-end Illumina mRNA libraries were generated from 4 µg of total RNA following the manufacturer's instructions for mRNA-Seq Sample Preparation (Illumina Inc., San Diego, CA). Library quality was assessed with the 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Cluster amplification was performed using the TruSeq PE Cluster Kit with the cBot automated system (Illumina Inc., San Diego, CA), and each sample was sequenced in separate GAIIx lanes using the TruSeq SBS 36 Cycle Kit (Illumina, San Diego, CA). Read lengths were 72 bp.

Data filtering and *de novo* assembly

The raw data, which were generated via Illumina sequencing in the BCL format, were converted to qSeq using the Off-Line Basecaller v.1.9.4 (OLB) software. The qSeq files were transformed into FastQ files containing the 72 bp reads using a custom script. The raw reads that were less than 60 bp in length with quality scores of $Q<20$ were trimmed using the CLC Genomics Workbench (v4.9; CLC Bio, Cambridge, MA). For the *de novo* assembly, we employed the CLC Genomics Workbench with the following parameters: the maximum gap and mismatch count were set to 2, the insertion and deletion costs were set to 3, the minimum contig length was set to 200 bp, the length fraction and similarity parameters were set to 0.5 and 0.9, respectively and the word size (k-mer) was set to 29. All of the short reads were deposited in the NCBI Short Read Archive (SRA) under accession number SRX371361.

Characterization through similarity searches and annotations

The contigs were searched against the NCBI non-redundant (nr) and the UniProtKB/Swiss-Prot protein databases using BLASTx with a cut-off e-value of 1e-10. The Blast2GO program [28] was used to obtain gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations. The software WEGO [29] was then employed to perform GO classifications of the annotated contigs to obtain the gene function distributions.

A GO enrichment analysis was conducted to identify the functional categories that were enriched in the bark transcripts. To perform this analysis, we used the Blast2GO program with Fisher's exact test (p-value <0.001).

The contigs were also searched against the STRING database v. 9.05 (http://string-db.org) to predict clusters of orthologous groups (COGs) and classify possible functions at a cut-off e-value of 1e-10. To identify the protein domains, all of the translated sequences were matched against the Pfam database using the InterProScan tool [30].

An *H. brasiliensis* database was constructed using public RNA-seq data [13,14,19,25], the EST database at NCBI (as of August 2013) and data that were provided by Silva et al. (2014) [31] to perform a BLASTn search with a cut-off e-value of 1e-10 for the assessment of the transcriptomic contributions to the publicly available *H. brasiliensis* data and partial and complete open reading frames (ORFs) were predicted using the TransDecoder package (http://transdecoder.sourceforge.net/).

Digital gene expression analysis

Each genotype was mapped separately to the contigs that were obtained in the *de novo* assembly with a minimum number of reads of 10 and a maximum number of mismatches equal to 2. The data were normalized by calculating the reads per kilobase per million mapped reads (RPKM) for each contig. For the statistical analyses, Kal's Z test on proportions was used to determine the significantly differentially expressed genes. Genes showing false discovery rates (FDR) <0.05 and fold changes >2 were considered to be differentially expressed. All of the analyses were performed with the CLC Genomics Workbench.

Variant detection

To identify putative SSRs, the MISA program (http://pgrc.ipk-gatersleben.de/misa/) was used. As a criterion for SSR detection, sequences that showed at least 5 dinucleotide repeats; 4 trinucleotide repeats; and 3 tetra-, penta- and hexanucleotide repeats were considered.

The CLC Genomics Workbench software was first used to map the reads to the transcriptome obtained by *de novo* assembly with length fractions of 0.5 and similarities of 0.9. Then, putative SNP detection was performed using the following criteria: minimum coverage of 10, minimum frequency of 10%, quality value from the central base of Q>30 and quality value from the average base of Q>20.

SNP validation

Primer pairs were designed using the Primer 3 program [32] for at least one putative SNP. PCR amplifications were performed in 20 µl reactions containing 25 ng of genomic DNA, 0.5 µM of each primer, 100 µM of each dNTP, 3 mM $MgCl_2$, 20 mM Tris–HCl, 50 mM KCl and 0.5 U of Pfu Taq DNA Polymerase (recombinant) (Thermo Scientific Inc., San Jose, CA) using the following steps: an initial denaturation at 95°C for 3 min, followed by 35 amplification cycles (30 s at 95°C, 30 s at the specific annealing temperature and 2 min at 72°C), and a final extension at 72°C for 10 min. The PCR products were purified using a solution of 20% (w/v) PEG8000 and 2.5 M NaCl in a 1:1 proportion with the sample volume. The amplification products were resolved via electrophoresis in 1.5% agarose gels prior to the sequencing reaction.

Each amplicon was bidirectionally sequenced (forward and reverse) using the BigDye Terminator v3.1 Kit (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions in an ABI 3500 xL Genetic Analyzer (Applied Biosystems, Foster City, CA). The sequencing chromatograms were visually inspected with the ChromasPro 1.5 software, and SNPs were identified as overlapping nucleotide peaks.

The allelic polymorphic information content of each SNP was calculated using the formula, $PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$ where n is the number of alleles of the marker among the set of genotypes that were used for characterizing the SNP polymorphism, and p_i and p_j are the frequencies of alleles i and j, respectively. The observed and expected heterozygosities were calculated using the TFPGA program [33].

Results and Discussion

Transcriptome sequencing and *de novo* assembly

In total, 179,326,804 raw reads were generated and trimmed to exclude low-quality reads (Table 1). To perform the *de novo* assembly 166,731,798 high-quality reads were used, generating 152,416 contigs. The contigs lengths ranged from 97 to 13,266 bp, with a mean length of 536 bp, an N50 of 720 bp and a GC content of 41.8% (Table 2).

A total of 58,992 contigs longer than 400 bp were selected. Of these, 8,608 shared high identities with non-plant sequences suggesting that 17% of these contigs were contaminant sequences. After removal of these contaminant sequences, a total of 50,384 contigs were used for further analyses (Table S2).

Of the 50,384 contigs, 12,761(25.3%) ranged in size from 1 to 2 kb and 4,515 (8.9%) were longer than 2 kb (Figure 1).

Partial and complete ORFs were predicted from the 50,384 contigs. In total, 23,977 contigs contained ORFs (47.5%), of which 9,247 (18%) were classified as possessing complete ORFs.

Characterization via similarity searches

The 50,384 contigs were searched against the NCBI nr protein and UniProtKB/Swiss-Prot databases using BLASTx employing a cut-off e-value of 1e-10 as the criterion for defining a significant hit.

Of these contigs, 32,018 (63%) showed significant BLASTx matches in the nr database and 23,620 (47%) in the UniProtKB/Swiss-Prot database (Table 3). All of the contigs that were annotated using UniProtKB/Swiss-Prot were also annotated with the nr database.

The proportion of the contigs with BLASTx hits significantly increased for longer contigs (Figure 1). The BLASTx searches yielded hits for 16,383 (49%) contigs that were 400 bp to 1 kb in length, while 4,391 (97%) of the contigs that were longer than 2 kb were annotated in the BLASTx searches. Of the 10 largest contigs, 9 returned BLASTx hits (Table S3).

The top 5 species showing BLASTx hits were *Ricinus communis* (20,522 contigs; 64%), *Populus trichocarpa* (6,310 contigs; 19.7%), *Vitis vinifera* (2,471 contigs; 7.7%), *Glycine max* (535 contigs; 1.7%) and *Hevea brasiliensis* (414 contigs; 1.3%) (Figure 2).

Table 1. Statistical summary of trimmed Illumina sequencing data.

	n° of reads	average length (bp)	total bases
Before trimming			
GT1	85,972,890	68.4	5,880,545,676
PR255	93,353,914	70.2	6,553,444,763
After trimming			
GT1	78,512,628	71.6	5,621,504,165
PR255	88,219,170	71.8	6,334,136,406

doi:10.1371/journal.pone.0102665.t001

To investigate the contributions of novel transcripts to the rubber tree database, a BLASTn search (cut-off e-value of 1e-10) was performed against an *H. brasiliensis* database.

Of the 32,018 contigs showing similarity in the nr database, 1,089 (3.4%) non-redundant contigs presented with no hits against the *H. brasiliensis* database (Figure S1). These results indicate that novel uncataloged genes have been identified for the rubber tree database.

Moreover, the 18,866 contigs with no hit that were subjected to BLASTn revealed significant hits for 10,821 (59%), whereas 7,545 (41%) had no hits. A search for putative ORFs was performed with the contigs with no hits (7,545) in BLASTn. We detected 479 contigs with ORFs, of which 83 were classified as complete ORFs (Figure S1). Future analyses may reveal potential unknown genes in this dataset.

Gene ontology (GO) analysis

The 32,018 contigs showing positive BLAST hits in the nr database were annotated using GO terms. The GO terms allow for the definition and standardization of the properties of gene products in any organism.

Of the 32,018 contigs, 21,725 were annotated with 37,781 GO terms (Table 3). Of the three main subontologies, molecular function was the highly represented, with 19,498 contigs followed by biological process with 13,729 contigs and finally, cellular component with 8,686 contigs (Figure 3).

For molecular function, binding (13,547 contigs) and catalytic activity (12,135 contigs) were the highly represented categories. For biological process, metabolic processes (10,528 contigs) and cellular processes (9,953 contigs) figured prominently. Interestingly, 252 contigs were assigned to the category of biological quality regulation, suggesting that they may be related to processes that modulate qualitative or quantitative traits that are associated with biological qualities such as size, mass or shape, which are important characteristics for bark. In addition, 85 contigs were assigned to the category of cell wall organization and thus play roles in the assembly, arrangement of constituent parts or the disassembly of the cell wall. For the cellular component subontology, cells (8,600 contigs) and organelles (4,196 contigs) were the most highly represented.

A GO enrichment analysis was performed to identify the functional categories that were enriched in the bark-exclusive transcripts.

These suggested bark-exclusive transcripts were identified using a BLASTn search (cut-off e-value of 1e-10) against an *H. brasiliensis* database that did not contain bark transcripts.

A total of 36 GO terms were enriched (Figure 4) among these transcripts, including the following categories: cell wall organization or biogenesis (GO: 0071554) and cell wall organization (GO: 00771555), which are responsible for the assembly, arrangement of constituent parts or disassembly of cell walls, and cytokinin metabolic (GO: 0009690) processes which are related to plant growth.

Categories that are involved in the prevention and/or recovery from an infection that is caused by an attack, such as the defense response (GO: 0006952) and pectinesterase activity (GO: 0030599) were also enriched.

Clusters of orthologous groups (COGs)

The clusters of orthologous groups (COGs) of protein database is used to phylogenetically classify the proteins that are encoded in complete genomes. Each COG includes proteins that are inferred to be orthologs i.e., they are direct evolutionary counterparts [34]. Among the 50,384 contigs, 9,720 were annotated (Table 3) and

Table 2. Statistical summary of the *de novo* assembly for *H. brasiliensis* bark.

Statistics for the *de novo* assembly	
Contig number	152,416
Total read count	166,731,798
Mean read length	71,76
Mean contig length	536
Maximum contig length	13,266
Minimum contig length	97
N50 length	720
GC% content	41,8

doi:10.1371/journal.pone.0102665.t002

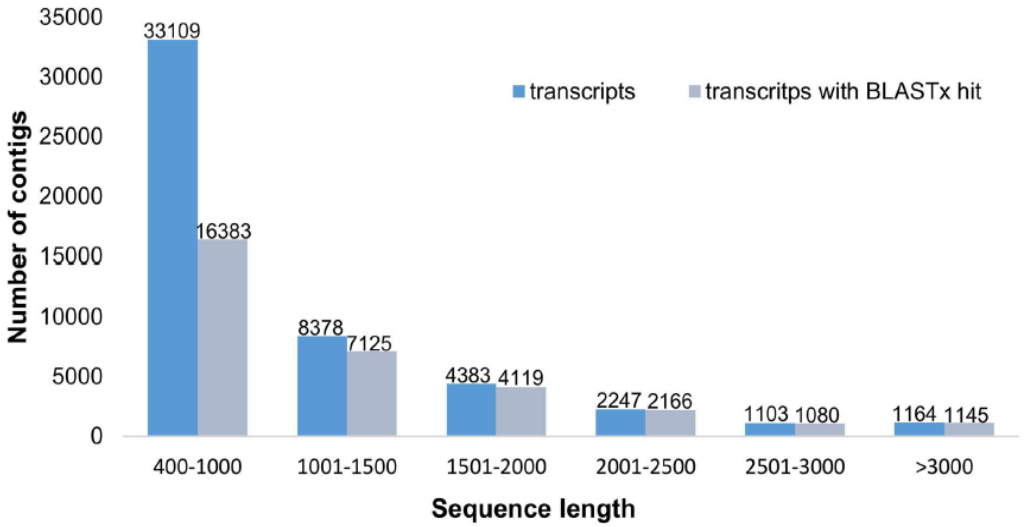


Figure 1. Length distribution of the 50,384 contigs. Histogram of the sequence-length distribution of these transcripts and the transcripts showing BLASTx hits in the nr database with a cut-off e-value of 1e−10.
doi:10.1371/journal.pone.0102665.g001

classified into 23 COG categories (Figure 5). General function prediction was the most highly represented category with 1,732 contigs, followed by replication, recombination and repair with 1,480 contigs and posttranslational modification, protein turnover, and chaperones with 843 contigs.

The smallest groups that were observed in the COG annotation analysis were cell motility, chromatin structure and dynamics and RNA processing and modification (7, 69 and 77 annotated contigs, respectively).

Additionally, the category of secondary metabolite biosynthesis, transport and catabolism was represented by 270 contigs.

Protein domain analysis

A comparison of the 50,384 contigs against the Pfam domain database with a cut-off e-value of 1e-10 resulted in 16,277 contigs matching at least one protein domain model (Table 3). The distribution of the domains ranged from a minimum of one to a maximum of 34 domains per contig.

The 3 most abundant domains that were identified included pentatricopeptide repeat-containing proteins (PPRs) with 3,058 contigs, followed by leucine-rich repeats (LRRs) with 1,479 contigs and WD40 with 967 contigs. The WD40 domain functions as a site of protein-protein interaction, and proteins containing WD40 repeats are known to serve as platforms for the assembly of protein complexes or mediators of transient interplay among other proteins [35] (Figure 6). Furthermore, 112 contigs were associated with WRKY domains which is a DNA-binding transcription factors tnat are found almost exclusively in plants [36] (Figure 6). WRKY containing proteins are thought to play important roles in plant defense responses, plant hormone signaling, secondary metabolism and plant responses to abiotic stress [37].

Moreover, 95 contigs were annotated to the sugar transporter family (Figure 6), 49 to the cellulase synthase family and 11 to cellulase domains (data not shown).

KEGG classification

The KEGG pathways represent collections of manually drawn pathway maps and that are helpful for the understanding if the biological functions and interactions of genes [38,39].

Of the 21,725 contigs that were annotated with GO terms, 8,626 were assigned to 10,355 EC numbers (Table 3). These EC numbers were mapped to the 137 KEGG Pathways (Table S4). Of the 5 main categories, metabolism was the main category represented, with 92% followed by organismal systems, environmental information and genetic information processing with 5%, 2% and 1% respectively.

Table 3. Summary of the annotations of the 50,384 contigs.

Database	Hits	Hits percentage
NCBI non-redundant protein (nr)	32,018	63.54%
UniProtKB/Swiss-Prot	23,620	46.87%
COG	9,720	19.29%
GO	21,725	43.11%
Interpro	16,277	32.30%
KEGG	8,626	17.12%

doi:10.1371/journal.pone.0102665.t003

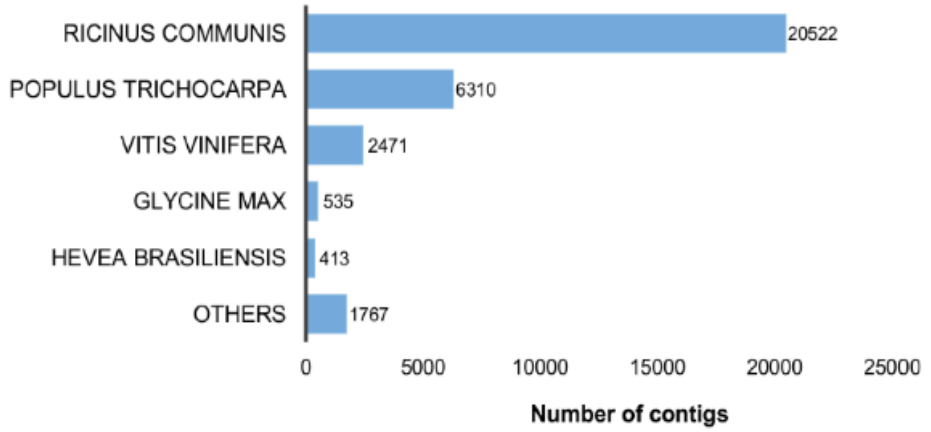


Figure 2. Top-hit species distribution in the BLASTx analysis against the nr database.
doi:10.1371/journal.pone.0102665.g002

In the metabolism category, carbohydrate metabolism (1,988 contigs) and amino acid metabolism (1,262 contigs) were the most prominent classes (Figure 7).

Rubber biosynthesis pathway. Latex is produced in specialized cells known as laticifers or latex vessels, which are located adjacent to the phloem of the rubber tree [4]. The chemical composition of rubber includes high-molecular-weight cis-polyisoprene [1], which is formed through the sequential condensation of isopentenyl diphosphate (IPP) [17]. IPP biosynthesis is related to the mevalonate (MVA) pathway [4], which occurs in the cytoplasm, and the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway which occurs in the plastid [18].

The MVA pathway includes 6 steps, which are catalyzed by the 6 corresponding enzymes, whereas the MEP pathway is catalyzed by 7 enzymes [4,18]. IPP that is synthetized through the MEP pathway was initially thought to be used for carotenoid synthesis in Frey-Wyssling particles [40]. However, the MEP pathway has been shown to serve as an alternative source of IPP for cis-polyisoprene synthesis in mature rubber trees or in clones that do not produce a large amount of carotenoids [18].

Acetyl-CoA is a precursor of the MVA pathways and is produced through the glycolysis/gluconeogenesis pathway. The MEP pathway precursors include glyceraldehyde-3-phosphate, which is produced via the glycolysis/gluconeogenesis pathway, and pyruvate, which is a product of pyruvate metabolism.

For the KEGG annotations, 192 contigs were annotated to 25 enzymes in the glycolysis/gluconeogenesis pathway (Figure S2),

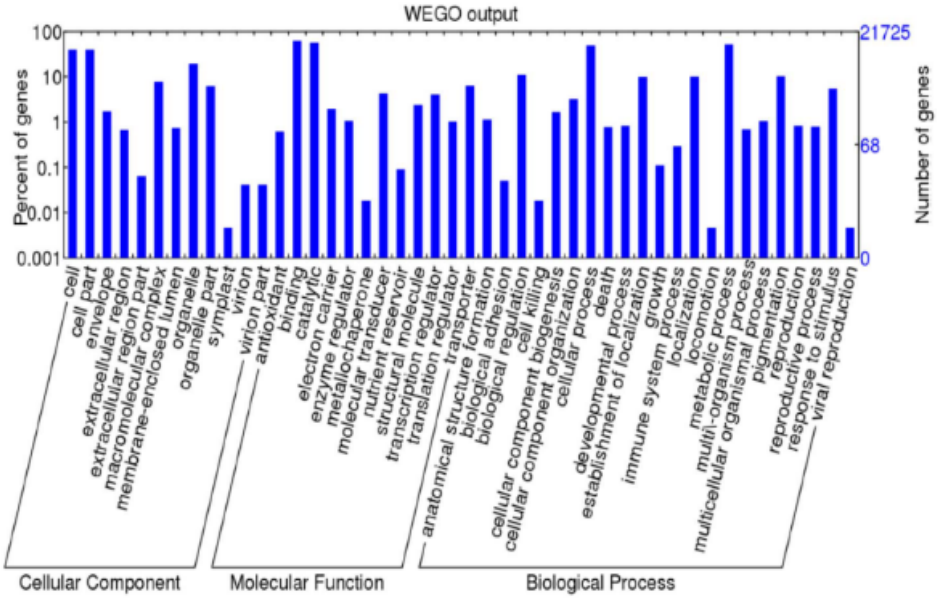


Figure 3. GO classification for the *H. brasiliensis* bark transcriptome.
doi:10.1371/journal.pone.0102665.g003

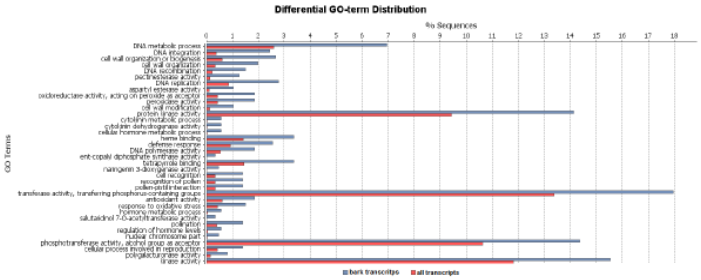


Figure 4. GO enrichment analysis for the bark-exclusive transcripts.
doi:10.1371/journal.pone.0102665.g004

and 116 were annotated to 22 enzymes in pyruvate metabolism (Figure S3).

In addition, we identified all of the key genes that are involved in the MVA and MEP pathways through KEGG annotations (Figure S4). In total, 25 contigs were related to the MVA pathway, and 40 were related to the MEP pathway (Table 4).

Digital gene expression analysis

We conducted a gene expression analysis to evaluate the potential genes that were differentially expressed between the GT1 and PR255 genotypes.

In this analysis, we observed that 716 genes were expressed at higher levels in GT1, and 1,267 were more prominently expressed in PR255 (Figure S5)

The top 20 differentially expressed genes that were found for each genotype are listed in Table S5. Similar to Li et al. (2012) [14], we observed genes that were related to stress/defense responses, such as the chalcone synthase [41], glycine-rich RNA-binding protein [42], ascorbate peroxidase [43] and o-methyl-transferase [44] genes, as these clones were frequently harvested.

Interestingly, the gene enconding carbonic anhydrase was the most highly expressed in PR255. This enzyme is responsible for facilitating the diffusion of carbon dioxide in photosynthesis and is essential for processes such as respiration [45].

The most highly expressed gene in GT1 was phenylalanine ammonia-lyase 2 which is involved in lignin and flavonoid synthesis and is typically highly expressed in response to pathogen attack and tissue wounding [46].

Considering the annotations of all the key genes that are involved in the MVA and MEP pathways according to the KEGG database, we observed that the genes encoding hydroxymethyl-glutaryl-CoA reductase (NADPH) (contig_104848) in the MVA pathway and (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (contig_145940 and contig_146058) and 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (contig_97647) in the MEP pathway, were enhanced in the GT1 genotype, providing a strong evidence that these genes are differentially expressed in these two clones.

Although the experiments did not include replicates, this analysis represents the first step in understanding the unique responses of different genotypes and elucidating possible candidate genes for rubber tree molecular breeding.

Putative SSR marker discovery

The 50,384 contigs were subjected to a search for putative SSR markers. A total of 17,927 SSRs were detected in 13,070 contigs, and 3,433 contigs presented with more than one SSR (Table 5). There were 6,822 di-, 6,098 tri-, 3,033 tetra-, 1,125 penta- and

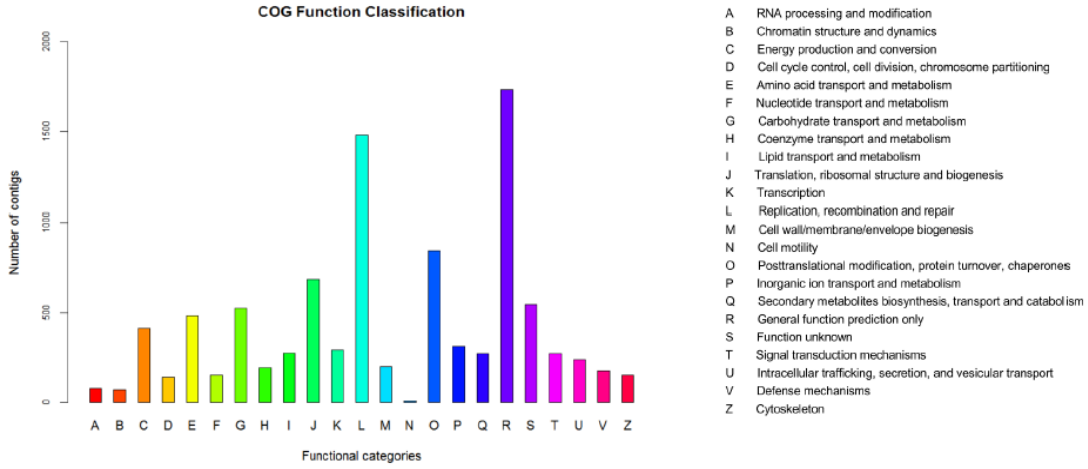


Figure 5. COG functional distribution of the *H. brasiliensis* bark transcriptome.
doi:10.1371/journal.pone.0102665.g005

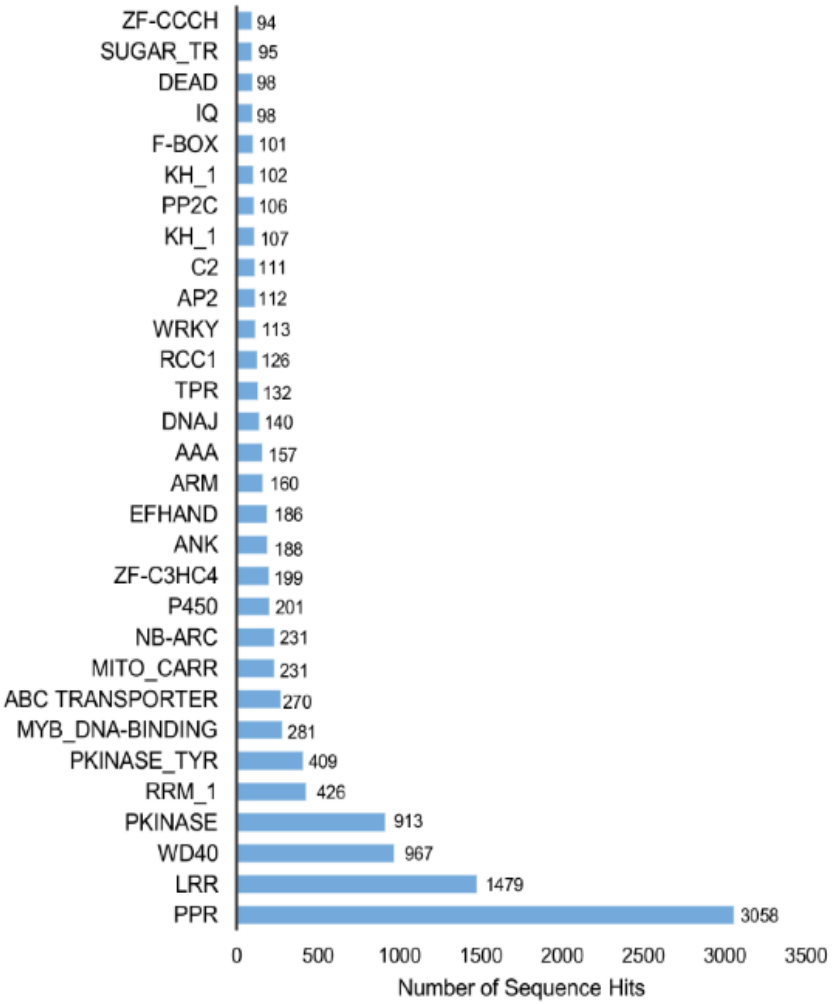


Figure 6. Distribution of the top 30 Pfam domains identified in translated *H. brasiliensis* transcripts.
doi:10.1371/journal.pone.0102665.g006

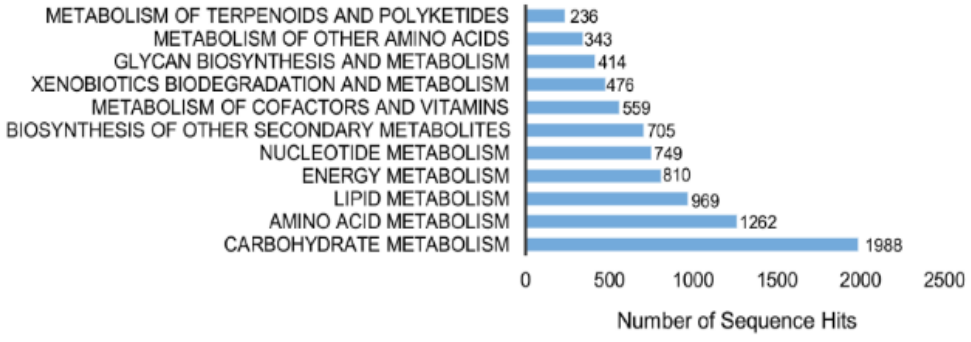


Figure 7. KEGG metabolism pathway distribution for the *H. brasiliensis* contigs.
doi:10.1371/journal.pone.0102665.g007

Table 4. Number of contigs annotated in the MVA and MEP pathways.

MVA pathway	
Enzymes	**number of contigs**
acetyl-CoA C-acetyltransferase (AACT)	4
hydroxymethylglutaryl-CoA synthase (HMGS)	4
hydroxymethylglutaryl-CoA reductase (NADPH)	8
mevalonate kinase (MVK)	3
phosphomevalonate kinase (PMK)	2
diphosphomevalonate decarboxylase (MVD)	4
MEP pathway	
Enzymes	**number of contigs**
1-deoxy-D-xylulose-5-phosphate synthase (DXS)	18
1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR)	3
2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MCT)	2
4-(cytidine 5′-diphospho)-2-C-methyl-D-erythritol kinase (CMK)	3
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS)	2
(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (HDS)	5
4-hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR)	4

doi:10.1371/journal.pone.0102665.t004

849 hexanucleotide potential SSRs (Table 6). A total of 50,608,451 bp were analyzed, and a frequency of one SSR per 2.8 kb was observed, similar to previously described by Feng et al. (one SSR per 2.25 kb) [12] and by Li et al. (one SSR per 2.42 kb) [14].

To investigate the contributions of the novel sequences containing SSRs for future rubber tree studies, we performed a BLASTn search using our identified sequences with SSRs against the *H. brasiliensis* database. We identified 1,709 sequences that showed no similarity with the *H. brasiliensis* database, suggesting that they possess novel SSRs for the rubber tree, and thus 203 sequences were annotated with the nr database.

Dinucleotide SSRs have been reported to be the most abundant SSR type in plant genomes [47]. In contrast with plants such as the sugarcane [48], wheat [49], sweet potato [50] and citrus [51], in which SSRs containing trinucleotide motifs are the most abundant in transcribed regions, it has been reported that dinucleotide motifs figure prominently in *H. brasiliensis* transcripts [12]. Dinucleotide motifs are also abundant in other plants such as sesame [52], kiwi [53] and coffee [54]. In this work, dinucleotide motifs were found to be predominant, corroborating with previous studies in which 38% of the total putative SSRs were shown to possess these motifs (Figure S6).

The most abundant motif in the dinucleotide class was AG/TC (4.674, 68.5%), followed by AT/TA (1298, 19%), AC/TG (829, 12.1%) and GC/CG (21, 0.3%) (Figure S6). The rarity of the CG dinucleotide microsatellites cannot be explained by the low C/G contents. CpG dinucleotides that are not situated in CpG islands can undergo cytosine methylation, and methylated cytosines tend to mutate to thymine, which may explain the underrepresentation of the CpG dinucleotides and, consequently, the low coverage of microsatellites CG motifs [55]. The most frequent trinucleotide motif was AAG/TTC (1876, 30.7%), and the least represented motif was CCG/CGG (90, 1.4%) (Table 6). Previous studies on *Arabidopsis* and soybean also suggested that the trinucleotide AAG motif may figure prominently in dicots [50]. Interestingly, we found only 90 CCG/CGG trinucleotides, which have been reported to predominant in monocots [47,56], such as maize, barley and sorghum [50]. Our results are in accord with previous studies if rubber tree and with the observed rarity of CCG/CGG repeat units that have been reported in a large number of dicotyledonous plants such as *Citrus*, *Coffea* and *Glycine* [57]. Long CCG/CGG sequences could compete with the components of the splicing machinery, resulting in inadequate splicing. Moreover, CCG/CGG repeats, may potentially form higher structures, such as hairpins and quadruplexes, affecting the efficiency and accuracy of splicing and influencing the formation of mature mRNA [56,58].

Our findings correlate with previous studies of the rubber tree, in which AG/TC and AAG/TTC were found to be the most

Table 5. Summary of putative SSRs identified using MISA software.

Number of contigs	50,384
Total bases	50,608,451
Number of sequences with SSRs	13,070
Total number of SSRs	17,927
SSR frequency	1 per 2.8 kb

doi:10.1371/journal.pone.0102665.t005

Table 6. Summary of the distribution of putative SSR motifs.

SSR repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	>15	Total
AC/GT	-	-	364	113	72	56	54	35	34	27	19	13	12	30	829
AG/CT	-	-	1,702	677	362	272	277	307	341	174	109	107	102	244	4,674
AT/AT	-	-	549	240	149	111	66	61	38	46	13	11	7	7	1,298
CG/CG	-	-	9	7	3	2	-	-	-	-	-	-	-	-	21
dinucleotide	-	-	2624	1037	586	441	397	403	413	247	141	131	121	281	6,822
AAC/GTT	-	242	54	19	9	6	5	1	1	-	-	-	-	3	340
AAG/CTT	-	978	369	202	104	99	33	41	29	9	5	5	2	-	1,876
AAT/ATT	-	390	174	104	80	64	22	19	17	10	10	6	4	5	905
ACC/GGT	-	363	108	54	25	7	8	3	2	-	-	-	-	-	570
ACG/CGT	-	61	14	5	2	1	2	-	1	-	-	-	-	-	86
ACT/AGT	-	44	15	2	1	1	-	1	-	-	-	-	-	1	65
AGC/CTG	-	445	131	68	23	13	2	3	-	-	-	-	-	-	685
AGG/CCT	-	360	108	53	27	21	3	2	2	-	-	-	-	-	576
ATC/ATG	-	607	175	57	34	17	7	7	1	-	-	-	-	-	905
CCG/CGG	-	56	22	10	2	-	-	-	-	-	-	-	-	-	90
trinucleotide	-	3,546	1,170	574	307	229	82	77	53	19	15	11	6	9	6,098
tetranucleotide	2456	385	142	34	13	2	1	-	-	-	-	-	-	-	3,033
pentanucleotide	860	205	44	6	6	3	1	-	-	-	-	-	-	-	1,125
hexanucleotide	625	157	49	16	1	1	-	-	-	-	-	-	-	-	849

doi:10.1371/journal.pone.0102665.t006

abundant motifs in the dinucleotide and trinucleotide categories, respectively.

Putative SNP marker discovery

For the putative SNP detection, the 50,384 contigs were first mapped with trimmed short sequence reads using the CLC Genomics Workbench. In total, 404,114 putative SNPs were detected, and an average of one SNP per 125 bp was observed (Table 7), which was similar to the SNP frequencies that were previously reported for *Eucalyptus grandis* (1 SNP per 192 bp) [59], apple (1 SNP per 149 bp) [60] and grapevine (1SNP per 117 bp) [61], in addition to a recent study for rubber tree (1 SNP per 178 bp) [31]. However, the density of putative SNPs was higher than that which was described by Pootakham et al. (2011) [62] and Salgado et al [63] for the rubber tree, who detected one SNP per 1.5 kb and one SNP per 5.2 kb, respectively, using 454 pyrosequencing technology, which has a lower sequencing depth than Illumina sequencing technology.

Transition SNPs were predominant, of which 242,732 (60%) were detected, while 161,382 (40%) transversion SNPs were identified (Table 7). Among the transversion variations, A ↔ T was the most highly represented with 49,283 SNPs detected, and G ↔ C was the least common with 31,376 SNPs identified (Figure S7).

As expected, the transition SNPs were generally observed at higher frequencies than the transversion SNPs. During natural selection, transitions mutations are better tolerated than transversions because they generate synonymous mutations in protein-coding sequences [64].

Because contigs corresponding with genes that are involved in the MVA and MEP pathways were identified in the KEGG annotations, we also searched for SNPs in these sequences. Only 4 contigs that are involved in the MVA pathway did not contain putative SNPs, which were annotated as hydroxymethylglutaryl-CoA synthase (AACT) (1 contig), hydroxymethylglutaryl-CoA reductase (NADPH) (2 contigs) and phosphomevalonate kinase (PMK) (1 contig), while 1 contig from the MEP pathway that did not contain a putative SNP was found, which was annotated as 1-deoxy-D-xylulose-5-phosphate synthase (DXS).

SNP marker validation

Primer pairs were designed for the sequences that were related to the MVA and MEP pathways with putative SNPs to validate the SNPs via Sanger sequencing.

We designed primer pairs for 21 and 31 transcripts from the MVA and MEP pathways, respectively. For some of the sequences, we designed more than one primer pair to validate a greater number of SNPs.

A total of 64 primer pairs were designed and 35 yielded good amplification products for sequencing. However, 9 loci yielded good amplification products in only a few genotypes, and 26 loci were therefore analyzed for SNP marker validations. Some of the loci showed deviations from the expected and observed product sizes because the primers pairs were designed based on transcript regions (exons), whereas the amplification reactions were performed with genomic DNA which contains both exons and intron regions (Table S6).

A total of 78 SNPs were validated in 25 contigs (Table S7). Of these 25 contigs, 9 were annotated to the MVA pathway. Among the 6 enzymes in the MVA pathway, we amplified transcripts that were annotated as the enzymes acetyl-CoA C-acetyltransferase (AACT) (1 contig; 2 SNPs), hydroxymethylglutaryl-CoA synthase (HMGS) (2 contigs; 2 SNPs), hydroxymethylglutaryl-CoA reductase (NADPH) (5 contigs; 12 SNPs) and diphosphomevalonate decarboxylase (MVD) (1 contig; 1 SNP). For the MEP pathway, we evaluated 14 contigs that were annotated as the enzymes 1-deoxy-D-xylulose-5-phosphate synthase (DXS) (10 contigs; 53 SNPs), 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR) (1 contig; 1 SNP), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS) (2 contigs; 6 SNPs) and (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (HDS) (1 contig; 1 SNP). The observed and expected heterozygosities ranged from 0.0294 to 0.9167 and from 0.0000 to 0.5394 respectively, and the PIC values ranged from 0.0286 and 0.4402.

Interestingly, the locus HB_SNP_26 which was annotated as diphosphomevalonate decarboxylase (in the MVA pathway), contained a deletion or insertion (INDEL) polymorphism from positions 161 to 168 bp (Figure 8).

The observed and expected heterozygosities and PIC values were not calculated to the INDEL polymorphisms.

This study provides the first identification and validation of putative SNPs in 2 important pathways for rubber biosynthesis.

RNA-seq for *H. brasilensis* breeding

Crop domestication began more than 10,000 years ago, but the domestication and breeding of forest trees, such as rubber tree, only started approximately 100 years ago. Similar to other forest tree species with long generation times, rubber tree still in the early

Table 7. Summary of putative SNPs identified using CLC Genomics Workbench.

Number of contigs	50,384
Total bases	50,608,451
Number of SNPs	404,114
SNP frequency	1 per 125 bp
Transition	242,732
A ↔ G	120,866
C ↔ T	121,866
Transversion	161,382
A ↔ C	40,681
A ↔ T	49,289
C ↔ G	31,376
G ↔ T	40,036

doi:10.1371/journal.pone.0102665.t007

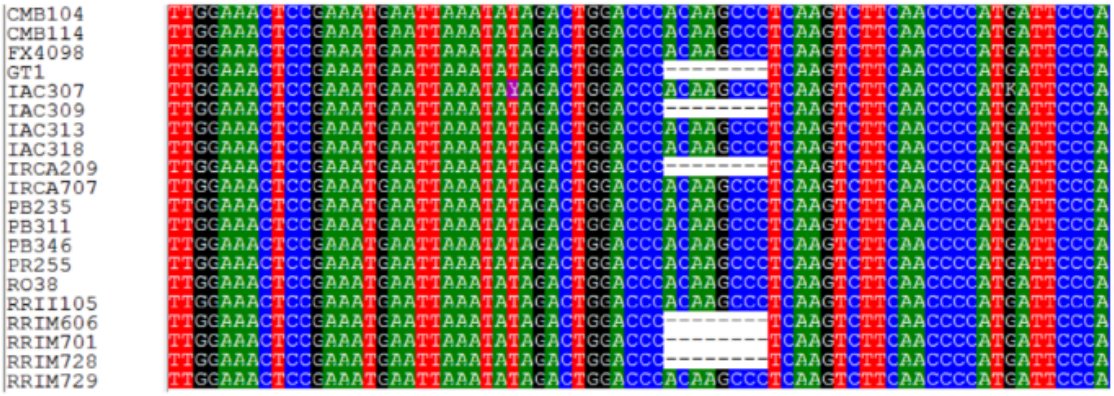


Figure 8. INDEL polymorphism at the HB_SNP_26 locus.
doi:10.1371/journal.pone.0102665.g008

stages of domestication, with most breeding programs producing only two or three generations from the wild-type genotypes [5], whereas the same amount of progress can be accomplished in a single year for many agricultural crops [65].

With the advent of next-generation sequencing technologies, such as RNA-seq, rapid advances have been made in improving the levels of transcriptome coverage for forest trees. These transcripts can be characterized using public databases, and an enormous amount of genetic diversity has been identified in these species.

Since 2011, the publically available RNA-seq data [13,14,19,25] have included an abundance of new information provided on *H. brasiliensis* transcripts and, consequently on rubber tree genetics [25]. These data allowed us to compare and identify novel transcripts (Figure S1) and new sequences with SSRs for the *H. brasiliensis* database to improve this database.

The high genetic variability that is present in *H. brasiliensis* have been demonstrated by the high frequency of polymorphisms that are found in its SSR [11,66,67] and EST-SSR [12,31] markers. SNP markers constitute the most abundant type of DNA polymorphism in genomic sequences and are thought to play major roles in the induction of phenotypic variations [68]. RNA-seq, together with SNP discovery, can be applied to develop new markers in candidate genes for genetic breeding and to investigate the variability of these genes in rubber tree, which has been performed in other tree species. The integration of modern genetics and novel sequencing technologies with conventional breeding can provide additional information and should expedite *H. brasiliensis* domestication.

Conclusions

The use of RNA-seq technology has allowed for a more comprehensive understanding of transcriptional patterns occurring in the bark of *H. brasiliensis*. Furthermore, our data has revealed 1,089 new rubber tree genes and 7,545 potentially novel genes. The RNA-seq data has led to the identification of 1,709 new EST-SSRs for the *H. brasiliensis* database. In addition, SNP analysis elucidated a total 404,114 SNPs that may be associated with potentially important genes. This information may constitute a valuable resource for rubber tree breeding programs and genetic diversity studies. This is the first study in which putative SNPs were identified and validated in genes that are involved in the MVA and MEP pathways.

Supporting Information

Figure S1 Overview of the workflow for investigating the contribution of novel transcripts in the *H. brasiliensis* database.
(TIFF)

Figure S2 Glycolysis/gluconeogenesis KEGG pathway. The annotated contigs are indicated in yellow.
(TIFF)

Figure S3 Pyruvate metabolism KEGG pathway. The annotated contigs are indicated in yellow.
(TIFF)

Figure S4 MVA and MEP KEGG pathways. The annotated contigs are indicated in yellow.
(TIFF)

Figure S5 Digital gene expression analysis. Volcano plot of differentially expressed genes between the GT1 and PR255 genotypes.
(TIFF)

Figure S6 Distribution of putative microsatellite types.
(TIF)

Figure S7 Distribution of putative SNPs that were identified.
(TIF)

Table S1 Genotypes of *H. brasiliensis* that were used for SNP validations and characterizations.
(XLSX)

Table S2 The 50,384 contigs that were longer than 400 bp from the *de novo* assembly.
(XLSX)

Table S3 The 10 longest contigs from the *de novo* assembly.
(XLSX)

Table S4 The 137 pathways that were annotated in the KEGG database.
(XLSX)

Table S5 The 20 most highly expressed genes in the GT1 and PR255 genotypes.
(XLSX)

Table S6 Characterization of all the developed SNP markers. The table presents the SNP markers that were developed for *H. brasiliensis*, including the corresponding primer sequences, annealing temperatures, and expected and observed products sizes in 1.5% agarose gel electrophoresis.
(XLSX)

Table S7 Validation of the SNP markers. The table presents the allelic variants, observed and expected heterozygosities and polymorphism information contents.
(XLSX)

Author Contributions

Conceived and designed the experiments: APS. Performed the experiments: CCM. Analyzed the data: CCM CBCS CCS RV. Contributed reagents/materials/analysis tools: CCS LMS ESJ PSG RV APS. Wrote the paper: CCM.

References

1. Sakdapipanich JT (2007) Structural characterization of natural rubber based on recent evidence from selective enzymatic treatments. J Biosci Bioeng 103: 287–292. doi:10.1263/jbb.103.287.
2. Cornish K (2001) Similarities and differences in rubber biochemistry among plant species. Phytochemistry 57: 1123–1134.
3. Gronover CS, Wahler D, Prüfer D (2009) Natural Rubber Biosynthesis and Physic- Chemical Studies on Plant Derived Latex. 2005.
4. Saha T, Priyadarshan PM (2012) Genomics of Tree Crops. Schnell RJ, Priyadarshan PM, editors New York, NY: Springer New York. doi:10.1007/978-1-4614-0920-5.
5. Priyadarshan PM, Goncalves PDS (2003) Hevea gene pool for breeding: 101–114.
6. Leitch AR, Lim KY, Leitch IJ, Neill MO, et al. (1998) Molecular cytogenetic studies in rubber, Hevea: 464–467.
7. Pires JM, Secco R., Gomes JI (2002). Taxonomia e filogeografia das seringueiras (Hevea spp) Belem: Embrapa Amazonia Oriental p.103.
8. Pushparajah E (2001) Natural rubber. In: Last, F.T. (ed) Tree Crop Ecosystems. Amsterdam, The Netherlands: Elsevier Science.
9. Raj S, Das G, Pothen J, Dey SK (2005) Relationship between latex yield of Hevea brasiliensis and antecedent environmental parameters. Int J Biometeorol 49: 189–196. doi:10.1007/s00484-004-0222-6.
10. Le Guen V, Garcia D, Doaré F, Mattos CRR, Condina V, et al. (2011) A rubber tree's durable resistance to Microcyclus ulei is conferred by a qualitative gene and a major quantitative resistance factor. Tree Genet Genomes 7: 877–889. doi:10.1007/s11295-011-0381-7.
11. Mantello CC, Suzuki FI, Souza LM, Gonçalves PS, Souza AP (2012) Microsatellite marker development for the rubber tree (Hevea brasiliensis): characterization and cross-amplification in wild Hevea species. BMC Res Notes 5: 329. doi:10.1186/1756-0500-5-329.
12. Feng SP, Li WG, Huang HS, Wang JY, Wu YT (2008) Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (Hevea brasiliensis). Mol Breed 23: 85–97. doi:10.1007/s11032-008-9216-0.
13. Triwitayakorn K, Chatkulkawin P, Kanjanawattanawong S, Sraphet S, Yoocha T, et al. (2011) Transcriptome sequencing of Hevea brasiliensis for development of microsatellite markers and construction of a genetic linkage map. DNA Res 18: 471–482. doi:10.1093/dnares/dsr034.
14. Li D, Deng Z, Qin B, Liu X, Men Z (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (Hevea brasiliensis Muell. Arg.). BMC Genomics 13: 192. doi:10.1186/1471-2164-13-192.
15. Lespinasse D., Rodier-Goud M., Grivet L., Leconte A., Legnate H., & Seguin M (2000) A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. Theor Appl Genet 100: 127–138.
16. Souza LM, Gazaffi R, Mantello CC, Silva CC, Garcia D, et al. (2013) QTL mapping of growth-related traits in a full-sib family of rubber tree (Hevea brasiliensis) evaluated in a sub-tropical climate. PLoS One 8: e61238. doi:10.1371/journal.pone.0061238.
17. Chow K-S, Wan K-L, Isa MNM, Bahari A, Tan S-H, et al. (2007) Insights into rubber biosynthesis from transcriptome analysis of Hevea brasiliensis latex. J Exp Bot 58: 2429–2440. doi:10.1093/jxb/erm093.
18. Chow K-S, Mat-Isa M-N, Bahari A, Ghazali A-K, Alias H, et al. (2012) Metabolic routes affecting rubber biosynthesis in Hevea brasiliensis latex. J Exp Bot 63: 1863–1871. doi:10.1093/jxb/err363.
19. Rahman AYA, Usharraj AO, Misra BB, Thottathil GP, Jayasekaran K, et al. (2013) Draft genome sequence of the rubber tree Hevea brasiliensis. BMC Genomics 14: 75. doi:10.1186/1471-2164-14-75.
20. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.
21. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536. doi:10.1016/j.cell.2008.03.029.
22. Lu T, Lu G, Fan D, Zhu C, Li W, et al. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq: 1238–1249. doi:10.1101/gr.106120.110.
23. Hansey CN, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, et al. (2012) Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. PLoS One 7: e33071. doi:10.1371/journal.pone.0033071.
24. Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. Am J Bot 99: 175–185. doi:10.3732/ajb.1200020.
25. Xia Z, Xu H, Zhai J, Li D, Luo H, et al. (2011) RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis. Plant Mol Biol 77: 299–308. doi:10.1007/s11103-011-9811-z.
26. Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Report 11: 113–116. doi:10.1007/BF02670468.
27. Doyle J, Doyle J (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19: 11–15.
28. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676. doi:10.1093/bioinformatics/bti610.
29. Ye J, Fang L, Zheng H, Zhang Y, Chen J, et al. (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34: W293–7. doi:10.1093/nar/gkl031.
30. Zdobnov EM, Apweiler R (2001) signature-recognition methods in InterPro. 17: 847–848.
31. Silva CC, Mantello CC, Campos T, Souza LM, Gonçalves PS, et al. (2014) Leaf-, panel- and latex-expressed sequenced tags from the rubber tree (Hevea brasiliensis) under cold-stressed and suboptimal growing conditions: the development of gene-targeted functional markers for stress response. Mol Breed. doi:10.1007/s11032-014-0095-2.
32. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, et al. (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35: W71–4. doi:10.1093/nar/gkm306.
33. Miller M (1997) 27. Miller MP: Tools for population genetic analysis (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Computer Software distributed by the author 1997.
34. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41. doi:10.1186/1471-2105-4-41.
35. Stirnimann CU, Petsalaki E, Russell RB, Müller CW (2010) WD40 proteins propel cellular networks. Trends Biochem Sci 35: 565–574. doi:10.1016/j.tibs.2010.04.003.
36. Rushton PJ, Somssich IE, Ringler P, Shen QJ (2010) WRKY transcription factors. Trends Plant Sci 15: 247–258. doi:10.1016/j.tplants.2010.02.006.
37. Cheng Y, Zhou Y, Yang Y, Chi Y-J, Zhou J, et al. (2012) Structural and functional analysis of VQ motif-containing proteins in Arabidopsis as interacting proteins of WRKY transcription factors. Plant Physiol 159: 810–825. doi:10.1104/pp.112.196816.
38. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
39. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480–4. doi:10.1093/nar/gkm882.
40. Sando T, Takeno S, Watanabe N, Okumoto H, Kuzuyama T, et al. (2008) Cloning and Characterization of the 2-C-Methyl-D-erythritol 4-Phosphate (MEP) Pathway Genes of a Natural-Rubber Producing Plant, Hevea brasiliensis. Biosci Biotechnol Biochem 72: 2903–2917. doi:10.1271/bbb.80387.
41. Dao TTH, Linthorst HJM, Verpoorte R (2011) Chalcone synthase and its functions in plant resistance. Phytochem Rev 10: 397–412. doi:10.1007/s11101-011-9211-7.
42. Kim JY, Kim WY, Kwak KJ, Oh SH, Han YS, et al. (2010) Glycine-rich RNA-binding proteins are functionally conserved in Arabidopsis thaliana and Oryza sativa during cold adaptation process. J Exp Bot 61: 2317–2325. doi:10.1093/jxb/erq058.
43. Caverzan A, Passaia G, Rosa SB, Ribeiro CW, Lazzarotto F, et al. (2012) Plant responses to stresses: Role of ascorbate peroxidase in the antioxidant protection. Genet Mol Biol 35: 1011–1019.
44. Lam KC, Ibrahim RK, Behdad B, Dayanandan S (2007) Structure, function, and evolution of plant O -methyltransferases. 1013: 1001–1013. doi:10.1139/G07-077.

45. Badger M (2003) The roles of carbonic anhydrases in photosynthetic CO(2) concentrating mechanisms. Photosynth Res 77: 83–94. doi:10.1023/A:1025821717773.
46. Tanaka Y, Matsuoka M, Yamanoto N, Ohashi Y, Kano-Murakami Y, et al. (1989) Structure and characterization of a cDNA clone for phenylalanine ammonia-lyase from cut-injured roots of sweet potato. Plant Physiol 90: 1403–1407.
47. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30: 194–200. doi:10.1038/ng822.
48. Marconi TG, Costa E a, Miranda HR, Mancini MC, Cardoso-Silva CB, et al. (2011) Functional markers for gene mapping and genetic diversity studies in sugarcane. BMC Res Notes 4: 264. doi:10.1186/1756-0500-4-264.
49. La Rota M, Kantety R V, Yu J-K, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6: 23. doi:10.1186/1471-2164-6-23.
50. Wang Z, Fang B, Chen J, Zhang X, Luo Z, et al. (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (Ipomoea batatas). BMC Genomics 11: 726. doi:10.1186/1471-2164-11-726.
51. Chen C, Zhou P, Choi Y a, Huang S, Gmitter FG (2006) Mining and characterizing microsatellites from citrus ESTs. Theor Appl Genet 112: 1248–1257. doi:10.1007/s00122-006-0226-1.
52. Wei W, Qi X, Wang L, Zhang Y, Hua W, et al. (2011) Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics 12: 451. doi:10.1186/1471-2164-12-451.
53. Fraser LG, Harvey CF, Crowhurst RN, De Silva HN (2004) EST-derived microsatellites from Actinidia species and their potential for mapping. Theor Appl Genet 108: 1010–1016. doi:10.1007/s00122-003-1517-4.
54. Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, et al. (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor Appl Genet 114: 359–372. doi:10.1007/s00122-006-0440-x.
55. Meglécz E, Nève G, Biffin E, Gardner MG (2012) Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model eukaryote species. PLoS One 7: e40861. doi:10.1371/journal.pone.0040861.
56. Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967–981.
57. Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. 998: 985–998. doi:10.1139/G05-060.
58. Splicing PRNA, Coleman TP, Roesser JR, Commonwealth V, Uni V, et al. (1998) RNA Secondary Structure: An Important cis -Element in Rat Calcitonin/CGRP: 15941–15950.
59. Novaes E, Drost DR, Farmerie WG, Jr GJP, Grattapaglia D, et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. 14: 1–14. doi:10.1186/1471-2164-9-312.
60. Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, et al. (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS One 7: e31745. doi:10.1371/journal.pone.0031745.
61. Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V, Martínez-Zapater JM (2007) High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics 8: 424. doi:10.1186/1471-2164-8-424.
62. Pootakham W, Chanprasert J, Jomchai N, Sangsrakru D, Yoocha T, et al. (2011) Single nucleotide polymorphism marker development in the rubber tree, Hevea brasiliensis (Euphorbiaceae). Am J Bot 98: e337–8. doi:10.3732/ajb.1100228.
63. Salgado LR, Koop DM, Pinheiro DG, Rivallan R, Le Guen V, et al. (2014) De novo transcriptome analysis of Hevea brasiliensis tissues by RNA-seq and screening for molecular markers. BMC Genomics 15: 236. doi:10.1186/1471-2164-15-236.
64. Allegre M, Argout X, Boccara M, Fouet O, Roguet Y, et al. (2012) Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of Theobroma cacao L. DNA Res 19: 23–35. doi:10.1093/dnares/dsr039.
65. Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12: 111–122. doi:10.1038/nrg2931.
66. Souza LM, Mantello CC, Santos MO, Goncalves PS, Souza A (2009) Microsatellites from rubber tree (Hevea brasiliensis) for genetic diversity analysis and cross-amplification in six Hevea wild species. Conserv Genet Resour 1: 75–79.
67. Le Guen, V Gay, C Xiong, T C., Souza, L M., Rodier-Goud M. and Seguin M (2011) Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (Hevea brasiliensis). Plant Breed 130: 294–296.
68. Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, Aoki K, et al. (2013) Genome-wide SNP genotyping to infer the effects on gene functions in tomato. DNA Res 20: 221–233. doi:10.1093/dnares/dst005.

Supporting Information

Table S1. Genotypes of *H. brasiliensis* used for SNP validation and characterization.

Clone number	Clone	Genealogy [(1)]
1	RRIM 600	Tjir 1 x PB 86
2	RRIM 606	Tjir 1 x PB 49
3	RRIM 701	44/553 x RRIM 501 (Pil A 44 x Lun N)
4	RRIM 728	GT 1 x RRIM 623 (PB 49 x Pil B 84)
5	RRIM 729	RRIM 623 (PB 49 x Pil B 84) x Fx 25 (F 351 x AVROS 49)
6	RRIM 805	RRIM 628 (selfed)
7	RRIM 809	RRIM 600 (Tjir 1 x PB 86) x RRIM 623 (PB 49 x Pil B 84)
8	RRIM 913	PB 5/51 (PB 56 x PB 24) x RRIM 623 (PB 49 x Pil B 84)
9	RRIM 915	RRIM 605 (Tjir 1 x PB 49) x PB 5/51 (PB 56 x PB 24)
10	RRIM 937	PB 5/51 (PB 56 x PB 24) x RRIM 703 (RRIM 600 x RRIM 500)
11	RRII 105	Tjir 1 x Gl 1
12	PB 217	PB 5/51 (PB 56 x PB 24) x PB 69
13	PB 233	PB S/15 x PB 5/63 (PB 56 x PB 24)
14	PB 235	PB 5/51 (PB 56 x PB 24) x PB S/78
15	PB 260	PB 5/51 (PB 56 x PB 24) x PB 49
16	PB 311	RRIM 600 (Tjir 1 x PB 86) x PB 235 [PB 5/51 (PB 56 x PB 24) x PB S/78)]
17	PB 346	PB 235 (PB 5/51 (PB 56 x PB 24) x PB S/78) x PB 218
18	PC 140	PB 5/51 (PB 56 x PB 24) x RRIM 703 (RRIM 600 x RRIM 500)
19	RRIC 100	RRIC 52 x PB 86
20	IAC 306	AVROS 49 x RRIM 509 (Pil A 44 x Lun N)
21	IAC 307	AVROS 1328 (AVROS 214 x AVROS 317) x PR 107
22	IAC 309	RRIM 626 (Tjir 1 x RRIM 600) x Fx 25 (F 351 x AVROS 49)
23	IAC 313	RRIM 511(Pil A 44 x Pil B 16) x AVROS 1518 (AVROS 214 x AVROS
24	IAC 318	RRIM 600 (Tjir 1 x PB 86) x Fx 3899 (F 4542 x AVROS 363)
25	IAC 500	RRIM 600 (Tjir 1 x PB 86) ill.
26	IRCA 27	PB 5/51 (PB 56 x PB 24) x RRIM 623 (PB 49 x Pil B 84)
27	IRCA 209	GT 1 x RRIM 605 (Tjir 1 x PB 49)
28	IRCA 230	GT 1 x PB 5/51 (PB 56 x PB 24)
29	IRCA 707	PB 235 [PB 5/51 (PB 56 x PB 24) x PB S/78] x RO 38
30	IRCA 1159	PB 260 (PB 5/51 x PB 49) x RO 38
31	GT 1*	Primary clone
32	PR 255*	Tjir 1 x PR 107
33	RO 38	Primary clone
34	Fx 4098	PB 86 x FB 110

35	PMB 104	Primary clone
36	PMB 114	Primary clone

Table S3. The 10 longest contigs obtained via *de novo* assembly.

Contigs	Best blast hit	acession number	Contig length
Contig_Hb_93134	Auxin transport protein (BIG) isoform 2	EOY03819.1	13,266
Contig_Hb_90294	conserved hypothetical protein	XP_002520949.1	12,166
Contig_Hb_90326	vacuolar protein sorting-associated protein, putative	XP_002519289.1	11,429
Contig_Hb_90470	conserved hypothetical protein	XP_002516857.1	11,019
Contig_Hb_20993	vacuolar protein sorting-associated protein, putative	XP_002517515.1	9,405
Contig_Hb_20366	heat shock protein binding protein, putative	XP_002515568.1	8,455
Contig_Hb_90348	nucleotide binding protein, putative	XP_002523320.1	8,439
Contig_Hb_95512	no hit	-	8,4
Contig_Hb_106925	protein binding protein, putative	XP_002527141.1	7,996
Contig_Hb_95005	hypothetical protein RCOM_1407450	XP_002523571	7,992

Table S4. The 137 pathways annotated in KEGG database.

1. Metabolism

1.1 Carbohydrate Metabolism	
Glycolysis / Gluconeogenesis	192
Citrate cycle (TCA cycle)	93
Pentose phosphate pathway	124
Pentose and glucuronate interconversions	132
Fructose and mannose metabolism	93
Galactose metabolism	143
Ascorbate and aldarate metabolism	76
Starch and sucrose metabolism	451
Amino sugar and nucleotide sugar metabolism	154
Pyruvate metabolism	166
Glyoxylate and dicarboxylate metabolism	127
Propanoate metabolism	71
Butanoate metabolism	59
C5-Branched dibasic acid metabolism	11
Inositol phosphate metabolism	96
1.2 Energy Metabolism	
Photosynthesis	7
Carbon fixation in photosynthetic organisms	133
Carbon fixation pathways in prokaryotes	88
Methane metabolism	233
Nitrogen metabolism	163
Sulfur metabolism	64
Oxidative phosphorylation	122
1.3 Lipid Metabolism	
Fatty acid biosynthesis	58
Fatty acid elongation	19
Fatty acid metabolism	107
Synthesis and degradation of ketone bodies	11
Cutin, suberine and wax biosynthesis	5
Steroid biosynthesis	35
Primary bile acid biosynthesis	3
Steroid hormone biosynthesis	66
Glycerolipid metabolism	146
Glycerophospholipid metabolism	142
Ether lipid metabolism	40
Sphingolipid metabolism	65
Arachidonic acid metabolism	70

Linoleic acid metabolism	78
alpha-Linolenic acid metabolism	71
Biosynthesis of unsaturated fatty acids	53
1.4 Nucleotide Metabolism	
Purine metabolism	501
Pyrimidine metabolism	238
Caprolactam degradation	10
1.5 Amino Acid Metabolism	
Alanine, aspartate and glutamate metabolism	100
Glycine, serine and threonine metabolism	117
Cysteine and methionine metabolism	147
Valine, leucine and isoleucine degradation	92
Valine, leucine and isoleucine biosynthesis	39
Lysine biosynthesis	20
Lysine degradation	87
Arginine and proline metabolism	139
Histidine metabolism	30
Tyrosine metabolism	84
Phenylalanine metabolism	191
Tryptophan metabolism	107
Phenylalanine, tyrosine and tryptophan biosynthesis	109
1.6 Metabolism of Other Amino Acids	
beta-Alanine metabolism	63
Taurine and hypotaurine metabolism	26
Phosphonate and phosphinate metabolism	4
Selenocompound metabolism	48
Cyanoamino acid metabolism	91
D-Glutamine and D-glutamate metabolism	3
D-Arginine and D-ornithine metabolism	3
D-Alanine metabolism	5
Glutathione metabolism	100
1.7 Glycan Biosynthesis and Metabolism	
N-Glycan biosynthesis	69
Various types of N-glycan biosynthesis	53
Mucin type O-Glycan biosynthesis	1
Other types of O-glycan biosynthesis	2
Glycosaminoglycan biosynthesis - chondroitin sulfate	48
Glycosaminoglycan biosynthesis - heparan sulfate	60

Glycosaminoglycan biosynthesis - keratan sulfate	1
Glycosaminoglycan degradation	33
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	6
Glycosphingolipid biosynthesis - lacto and neolacto series	6
Glycosphingolipid biosynthesis - globo series	21
Glycosphingolipid biosynthesis - ganglio series	33
Lipopolysaccharide biosynthesis	9
Peptidoglycan biosynthesis	8
Other glycan degradation	64

1.8 Metabolism of Cofactors and Vitamins

Thiamine metabolism	156
Riboflavin metabolism	59
Vitamin B6 metabolism	29
Nicotinate and nicotinamide metabolism	51
Pantothenate and CoA biosynthesis	47
Biotin metabolism	6
Lipoic acid metabolism	5
Folate biosynthesis	16
One carbon pool by folate	41
Retinol metabolism	55
Porphyrin and chlorophyll metabolism	66
Ubiquinone and other terpenoid-quinone biosynthesis	28

1.9 Metabolism of Terpenoids and Polyketides

Terpenoid backbone biosynthesis	83
Monoterpenoid biosynthesis	21
Sesquiterpenoid and triterpenoid biosynthesis	4
Diterpenoid biosynthesis	31
Carotenoid biosynthesis	16
Zeatin biosynthesis	12
Limonene and pinene degradation	20
Geraniol degradation	17
Biosynthesis of ansamycins	4
Tetracycline biosynthesis	15
Polyketide sugar unit biosynthesis	8
Biosynthesis of siderophore group nonribosomal peptides	2
Biosynthesis of vancomycin group antibiotics	3

1.10 Biosynthesis of Other Secondary Metabolites

Stilbenoid, diarylheptanoid and gingerol biosynthesis	10
Flavonoid biosynthesis	144
Flavone and flavonol biosynthesis	31
Anthocyanin biosynthesis	51
Isoflavonoid biosynthesis	24
Indole alkaloid biosynthesis	33
Isoquinoline alkaloid biosynthesis	77
Tropane, piperidine and pyridine alkaloid biosynthesis	61
Caffeine metabolism	41
Betalain biosynthesis	6
Glucosinolate biosynthesis	13
Penicillin and cephalosporin biosynthesis	3
Streptomycin biosynthesis	34
Butirosin and neomycin biosynthesis	9
Novobiocin biosynthesis	23
Phenylpropanoid biosynthesis	145
1.11 Xenobiotics Biodegradation and Metabolism	
Benzoate degradation	21
Aminobenzoate degradation	112
Fluorobenzoate degradation	8
Chloroalkane and chloroalkene degradation	18
Chlorocyclohexane and chlorobenzene degradation	10
Toluene degradation	11
Ethylbenzene degradation	4
Styrene degradation	18
Atrazine degradation	3
Naphthalene degradation	6
Metabolism of xenobiotics by cytochrome P450	67
Drug metabolism - cytochrome P450	83
Drug metabolism - other enzymes	106
Steroid degradation	9
2. Genetic Information Processing	
2.2 Translation	
Aminoacyl-tRNA biosynthesis	122
3. Environmental Information Processing	
3.2 Signal Transduction	
Phosphatidylinositol signaling system	139

mTOR signaling pathway	27

5. Organismal Systems

5.1 Immune System

T cell receptor signaling pathway	459

Table S5. The 20 most expressed genes in the GT1 and PR255 genotypes.

Contig ID	Best Blast Hit	Acession number	E-value	RPKM
GT1				
contig_128988	phenylalanine ammonia-lyase 2 [Manihot esculenta]	AF383152_1	0.0	11,384.20
contig_145874	histone H1 [Populus trichocarpa]	XP_002306824.1	6E-19	8,934.83
contig_141892	short chain alcohol dehydrogenase, putative [Ricinus communis]	XP_002511516.1	5.14E-149	4,696.00
contig_107930	*Populus trichocarpa**	AC214596.1	2E-66	3,718.00
contig_55145	o-methyltransferase, putative [Ricinus communis]	XP_002518739.1	7.47E-68	3,254.80
contig_137860	unknown [Lotus japonicus]	AFK46472.1	1.86E-80	2,366.00
contig_137257	adenosylmethionine synthase	Q6GV10.1	6.73E-77	1,917.60
contig_129071	acireductone dioxygenase, putative [Ricinus communis]	XP_002517071.1	8.34E-122	1,598.20
contig_126321	Stem-specific protein TSJT1, putative [Ricinus communis]	XP_002517179.1	5.65E-133	1,352.00
contig_29474	cytochrome P450, putative [Ricinus communis]	XP_002531094.1	8.93E-171	1,305.40
contig_129012	unnamed protein product [Vitis vinifera]	CBI34163.3	5.56E-74	1,164.70
contig_129021	1-aminocyclopropane-1-carboxylate oxidase [Hevea brasiliensis]	AAP41850.1	4.71E-141	1,095.90
contig_109083	conserved hypothetical protein [Ricinus communis]	XP_002534082.1	7.39E-98	1,059.90
contig_129010	cyclophilin [Hevea brasiliensis]	ADV04050.1	7.81E-111	973.90
contig_129603	eukaryotic translation initiation factor 5A isoform I [Hevea brasiliensis]	AAQ08191.1	3.58E-91	947.30
contig_113920	unknown [Glycine max]	ACU20582.1	9.98E-55	866.00
contig_98355	NADPH cytochrome P450 reductase [Citrus maxima]	ACP43317.1	8.17E-101	830.60
contig_125764	rubber peroxidase 1 [Hevea brasiliensis]	ABG46370.1	0.0	826.50
contig_116514	eukaryotic translation initiation factor 1Aa [Hevea brasiliensis]	ADV04051.1	2.65E-57	824.70
contig_146607	5-methyltetrahydropteroyltriglutamate, putative [Ricinus communis]	XP_002525709.1	2.78E-89	779.50
PR255				
contig_137296	carbonic anhydrase, putative [Ricinus communis]	XP_002529418.1	9.01E-57	2,823.80
contig_129198	chalcone synthase [Camellia grijsii]	AAO43487.1	5.36E-89	1,862.01
contig_146189	glycine-rich RNA-binding protein [Ricinus communis]	CAC80549.1	9.28E-49	1,270.50

contig_129318	beta-1,3-glucanase [Manihot esculenta]	ABK58141.1	2.82E-111	1,158.10
contig_129134	aldo/keto reductase AKR [Manihot esculenta]	AAX84672.1	7.53E-146	963.50
contig_118893	AF457210_1ascorbate peroxidase [Hevea brasiliensis]	AF457210_1	3.57E-107	958.50
contig_129036	chalcone synthase 2 [Camellia chekiangoleosa]	ADW11243.1	8.35E-116	897.20
contig_129017	unknown [Medicago truncatula]	AFK40462.1	2.76E-18	756.70
contig_145914	predicted protein [Populus trichocarpa]	XP_002331829.1	4.73E-111	733.90
contig_57137	*Hevea brasiliensis* *	AY247789	0.0	733.50
contig_129099	predicted protein [Populus trichocarpa]	XP_002337967.1	1.45E-125	712.20
contig_129415	predicted protein [Populus trichocarpa]	XP_002322207.1	1.23E-38	711.10
contig_129138	phosphoglycerate kinase, putative [Ricinus communis]	XP_002513352.1	1.16E-166	664.60
contig_55146	hypoia-responsive family protein 1 [Hevea brasiliensis]	ADR30790.1	3.43E-47	661.10
contig_129095	putative polyol transported protein 2 [Hevea brasiliensis]	CAP58707.1	0.0	651.90
contig_137875	serine-threonine protein kinase, plant-type, putative [Ricinus communis]	XP_002527489.1	1.64E-91	628.00
contig_109047	PREDICTED: isoflavone reductase homolog [Vitis vinifera]	XP_002283921.1	9.74E-77	616.00
contig_143735	beta-1,3-glucanase [Manihot esculenta]	ABK58141.1	7.59E-69	609.90
contig_129074	chalcone synthase [Dictamnus albus]	CAH61575.1	0.0	592.80
contig_129209	homocysteine methyltransferase, putative [Ricinus communis]	XP_002525709.1	9.35E-140	559.70

*Information provided from BLASTn NCBI EST database

Table S6. Characterization of all developed SNP marker.

	Sequence name	Primer sequence (5'-3')	best blast hit	predicted product size (bp)	observed product size (bp)[b]	TA(°C)[c]
HB_SNP_01	Contig_Hb_31012	TAATAGTTTGCTCTGGAACCTG	acetyl-CoA C-acetyltransferase	150	850	65,6
		CAATGTTATTCAGTCCCTCTCC				
HB_SNP_02[a]	Contig_Hb_31012	CACTCACCAGAACCAAAGCA	acetyl-CoA C-acetyltransferase	230	1650	65,6
		TATTGCTGCCCAAGACAGTG				
HB_SNP_06[a]	Contig_Hb_95577	AAATGGTAAGTATGGGGTTGC	acetyl-CoA C-acetyltransferase	235	1200	58
		CGTACAATCTCATGCAAATACG				
HB_SNP_09	Contig_Hb_55576	AGCATTCGCATCCCTACTTC	hydroxymethylglutaryl-CoA synthase	238	850-1000	63
		TCGATGCTCCATTAGCTTCA				
HB_SNP_10[a]	Contig_Hb_107134	TTTACTGTGAAGGAGGGATGC	hydroxymethylglutaryl-CoA synthase	247	850	60,4
		GGTGTTCAATGACTTTGTGAGG				
HB_SNP_11	Contig_Hb_110355	CAGAGTCAACGCCTTCAATG	hydroxymethylglutaryl-CoA synthase	229	900	65
		GGTCTTGGACAGGATTGCAT				
HB_SNP_12[a]	Contig_Hb_27098	ATTTGCCACGTCCTTCAATC	hydroxymethylglutaryl-CoA reductase (NADPH)	581	1000	63
		CTTGCCAGTAGAAGGGTTCG				
HB_SNP_13	Contig_Hb_27703	GCAGGGGAGCTGTCTCTTAT	hydroxymethylglutaryl-CoA reductase (NADPH)	300	200-300	62,7
		CAGTCTCACATGGCTCTGGA				
HB_SNP_15	Contig_Hb_34469	TGAAAACAACGGCAATGGTA	hydroxymethylglutaryl-CoA reductase (NADPH)	285	300	60
		ATTGTTGCTGGATGGAAAGG				
HB_SNP_17	Contig_Hb_109288	GAGGGAATACGAAGGGGTAG	hydroxymethylglutaryl-CoA	498	500	62,7

			reductase (NADPH)			
		GCTTCACCGTCTTCTTTTGG				
HB_SNP_18	Contig_Hb_115905	TCACAGCTGCAGGCTTCTTA	hydroxymethylglutaryl-CoA reductase (NADPH)	792	850	65
		AAGCAATGGACCCTATCGTG				
HB_SNP_19	Contig_Hb_125027	GAAGGAATGGAGCCAGAGACT	hydroxymethylglutaryl-CoA reductase (NADPH)	226	200-300	70
		CTAGAGCTTCCACCGATTCC				
HB_SNP_23	Contig_Hb_96672	CAGTTACAGAGGATTGCTTCC	diphosphomevalonate decarboxylase	249	250	58
		AAGATGCCCCAAGGTACAAA				
HB_SNP_26	Contig_Hb_126367	GCGAATCCCCTGACTAAGAC	diphosphomevalonate decarboxylase	207	200	62,7
		CGAGGATTCCGAAATTGAAG				
HB_SNP_27[a]	Contig_Hb_127741	CCACAAGCCCTCAAATCTTC	diphosphomevalonate decarboxylase	300	300	65,6
		TCTGCGCAGTCACCATTATC				
HB_SNP_29	Contig_Hb_30967	AACAAAGGAACCCCACTTGA	1-deoxy-D-xylulose-5-phosphate synthase	364	550	65
		TTTGGTCTTGAGGTGATCCA				
HB_SNP_30	Contig_Hb_31077	TGGGAGACATGAGAGCTGAA	1-deoxy-D-xylulose-5-phosphate synthase	243	250	65
		TGGCAAGGGAAGAATCCTAA				
HB_SNP_31	Contig_Hb_31402	GAGAGCAGGAACCAAAGCAC	1-deoxy-D-xylulose-5-phosphate synthase	526	700	65
		ATTGCCTCATATGCCTGTCC				
HB_SNP_32	Contig_Hb_91529	AAATGGGCAACATGAGATCC	1-deoxy-D-xylulose-5-phosphate synthase	250	250	65
		AGGTTGGAAAAGGCAGGATA				
HB_SNP_33	Contig_Hb_96040	GCATTTCGTCCTCCATACAAA	1-deoxy-D-xylulose-5-phosphate	828	1200	63

			synthase			
		GGCCTAAAACCTTTCTGTGC				
HB_SNP_34	Contig_Hb_96040	TGGAACAAATTAAGCCCTGTG	1-deoxy-D-xylulose-5-phosphate synthase	691	900	65
		GTGGGGAAGGACAATCATGT				
HB_SNP_35	Contig_Hb_96638	TTGTCTCTTCTCCCAGTCAGG	1-deoxy-D-xylulose-5-phosphate synthase	497	1200	65
		CCCGTTGGTTCTCTATTTGC				
HB_SNP_36	Contig_Hb_97000	AGAGGGTTGATCCAGAACCA	1-deoxy-D-xylulose-5-phosphate synthase	661	1450	65
		CTGCGGTCCGATGTTATTTT				
HB_SNP_37	Contig_Hb_97616	GCAGTTGCAAGGGACCTAC	1-deoxy-D-xylulose-5-phosphate synthase	600	700	65
		GAAAGTGTAGGGGATTTTGG				
HB_SNP_38	Contig_Hb_97616	TTGAGATTGGGAAAGGGAGA	1-deoxy-D-xylulose-5-phosphate synthase	482	1000	57
		TGGCTTTTGCTATTCCTGAGA				
HB_SNP_39	Contig_Hb_107075	TTTTGGCTTTGTTCTCTGAGG	1-deoxy-D-xylulose-5-phosphate synthase	247	250	65
		GCACTTGATGGACAGCTTGA				
HB_SNP_42	Contig_Hb_111903	GGACCATCAGACCCTACCAA	1-deoxy-D-xylulose-5-phosphate synthase	382	400	60
		TGCAGGAAGGTTCATTTGTG				
HB_SNP_44[a]	Contig_Hb_111947	GTGGCAACCATGTGAAAAAG	1-deoxy-D-xylulose-5-phosphate synthase	298	350	62
		TTGATGTTGGAATAGCGGAAC				
HB_SNP_47	Contig_Hb_129382	TGAAAGTCTTCACCTGATCTGC	1-deoxy-D-xylulose-5-phosphate reductoisomerase	391	850	58
		CCTGCTTTTACTTTGCTGCAT				
HB_SNP_48[a]	Contig_Hb_143021	GGCACGAGACTATGCTGCTA	1-deoxy-D-xylulose-5-phosphate	390	400	62

			reductoisomerase			
		TCGTCTGAGAAGCCTAAGAAAA				
HB_SNP_54	Contig_Hb_28561	TAATCCTCGTGAAAGCGTCA	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	358	400	60
		TCTTTTGAAGACCCATCTCACA				
HB_SNP_55	Contig_Hb_97044	TAATTGCCCTGAAAGCATGT	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	583	700	65
		GGAAACTTGGATGCCACCT				
HB_SNP_62	Contig_Hb_146058	CAGAAACAGAGCAAAAGCTCAA	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	483	500	65
		GGTATGTTGGTGGTGCTCCT				
HB_SNP_63[a]	Contig_Hb_97647	GGTTTGGAATACTGTTGAGA	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	459	1200	58
		ATAAGATCCAACTTCTCCTC				
HB_SNP_64[a]	Contig_Hb_97900	TGTGCTTCTCAACAGTATTCCA	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	698	1400	60
		AGCTCCACACTTTAGCTTTTCT				

a- locus that amplified in few genotypes; b- observed product size in agarose 1,5%; c- annealing temperature.

Table S7. Validation of the SNP markers.

Primer	Sequence name	SNP	Reference position	PIC[a]	He[b]	Ho[c]	Allelic variant
HB_SNP_01	Contig_Hb_31012	HB_SNP_01_58	58	0.3192	0.409	0.45	A/C
		HB_SNP_01_122	122	0.3338	0.4329	0.5217	C/T
HB_SNP_09	Contig_Hb_55576	HB_SNP_09_85	85	0.2392	0.2874	0.2	C/T
HB_SNP_11	Contig_Hb_110355	HB_SNP_11_226	226	0.3589	0.4787	0.3333	A/G
HB_SNP_13	Contig_Hb_27703	HB_SNP_13_720	720	0.2392	0.2825	0.3333	T/A
		HB_SNP_13_740	740	0.3615	0.3892	0.3226	A/G/C/T
		HB_SNP_13_881	881	0.234	0.275	0.3226	T/A
HB_SNP_15	Contig_Hb_34469	HB_SNP_15_137	137	0.2859	0.3556	0.1111	A/G
HB_SNP_17	Contig_Hb_109288	HB_SNP_17_396	396	0.3589	0.4762	0	C/T
		HB_SNP_17_505	505	0.4208	0.5394	0.3226	T/A
		HB_SNP_17_687	687	0.3544	0.4678	0.2188	A/G
		HB_SNP_17_756	756	0.3589	0.4762	0.4375	A/G
HB_SNP_18	Contig_Hb_115905	HB_SNP_18_1559	1559	0.2818	0.3452	0.2333	G/C
		HB_SNP_18_1466	1466	0.2174	0.2522	0.2903	C/A
HB_SNP_19	Contig_Hb_125027	HB_SNP_19_157	157	0.2298	0.2882	0.3429	G/A
		HB_SNP_19_219	219	0.2437	0.2687	0.3429	C/T
HB_SNP_23	Contig_Hb_96672	HB_SNP_23_442	442	0.3528	0.4672	0.375	A/G
HB_SNP_29	Contig_Hb_30967	HB_SNP_29_389	389	0.293	0.363	0.3214	C/T
		HB_SNP_29_440	440	0.2265	0.3429	0.2857	A/T
		HB_SNP_29_446	446	0.2265	0.2655	0.3077	C/T
		HB_SNP_29_452	452	0.375	0.5098	0.6154	C/T
		HB_SNP_29_460	460	0.1948	0.2227	0.25	A/G
		HB_SNP_29_463	463	0.3478	0.4565	0.5357	A/G
		HB_SNP_29_509	509	0.2559	0.3214	0.3929	C/T

		HB_SNP_29_519	519	0.2805	0.3279	0.3571	A/G
		HB_SNP_29_599	599	0.2659	0.3214	0.3929	C/T
		HB_SNP_29_635	635	0.3047	0.3818	0.5	A/C
		HB_SNP_29_636	636	0.2002	0.2299	0.2593	A/C
HB_SNP_30	Contig_Hb_31077	HB_SNP_30_73.	73	0.3748	0.5067	0.9167	A/G
		HB_SNP_30_205	205	0.3748	0.5067	0.9167	C/T
HB_SNP_31	Contig_Hb_31402	HB_SNP_31_263	263	0.3047	0.3846	0.2	A/G
		HB_SNP_31_298	298	0.3633	0.4857	0.6429	C/T
HB_SNP_32	Contig_Hb_91529	HB_SNP_32_411	411	0.1948	0.2222	0.25	C/T
		HB_SNP_32_442	442	0.1948	0.1726	0.1875	A/T
		HB_SNP_32_496	496	0.1948	0.1726	0.1875	C/T
		HB_SNP_32_453	453	0.1948	0.2222	0.25	A/C
		HB_SNP_32_455	455	0.1948	0.2222	0.25	A/C
		HB_SNP_32_457	457	0.1948	0.2222	0.25	T/G
		HB_SNP_32_508	508	0.2125	0.2455	0.2813	C/T
HB_SNP_33	Contig_Hb_96040	HB_SNP_33_483	483	0.0499	0.0526	0.0526	C/T
		HB_SNP_33_785	785	0.0948	0.1024	0.1053	T/G
HB_SNP_34	Contig_Hb_96040	HB_SNP_34_1282	1282	0.0555	0.0579	0.0588	A/C
		HB_SNP_34_1412	1412	0.329	0.4214	0.4118	A/G
		HB_SNP_34_1418	1418	0.0555	0.0579	0.0588	T/G
		HB_SNP_34_1609	1609	0.0555	0.0579	0.0588	C/T
		HB_SNP_34_1674	1674	0.0286	0.0294	0.0294	C/G
		HB_SNP_34_1796	1796	0.0555	0.0579	0.0588	A/G
HB_SNP_35	Contig_Hb_96638	HB_SNP_35_210	210	0.3398	0.444	0.0909	A/G
HB_SNP_36	Contig_Hb_97000	HB_SNP_36_664	664	0.3676	0.4647	0.4074	C/T
HB_SNP_37	Contig_Hb_97616	HB_SNP_37_195	195	0.4402	0.487	0.2903	A/G/C/T
		HB_SNP_37_255	255	0.2118	0.2457	0.28	A/T

		HB_SNP_37_261	261	0.0624	0.0655	0.0667	A/T
		HB_SNP_37_264	264	0.2688	0.3265	0.4	C/T
		HB_SNP_37_384	384	0.1889	0.2155	0.24	A/G
		HB_SNP_37_604	604	0.0688	0.0727	0.0741	A/G
HB_SNP_38	Contig_Hb_97616	HB_SNP_38_1250	1250	0.1948	0.2244	0.15	A/G
		HB_SNP_38_1301	1301	0.1364	0.1502	0.16	A/T
		HB_SNP_38_1303	1303	0.1364	0.1502	0.16	C/T
		HB_SNP_38_1366	1366	0.2453	0.2919	0.3462	C/T
		HB_SNP_38_1393	1393	0.2453	0.2919	0.3462	A/G
		HB_SNP_38_1450	1450	0.2327	0.2743	0.32	C/T
		HB_SNP_38_1453	1453	0.2058	0.2376	0.2692	C/T
		HB_SNP_38_1469	1469	0.1833	0.2081	0.2308	A/C
		HB_SNP_38_1519	1519	0.1638	0.1846	0.2	A/G
		HB_SNP_38_1547	1547	0.1638	0.1846	0.2	A/C
		HB_SNP_38_1612	1612	0.1638	0.1846	0.2	A/G
		HB_SNP_38_1625	1625	0.1638	0.1846	0.2	A/G
HB_SNP_39	Contig_Hb_107075	HB_SNP_39_255	255	0.3411	0.4442	0.4286	G/T
		HB_SNP_39_377	377	0.3638	0.4592	0.4483	A/C/G
HB_SNP_42	Contig_Hb_111903	HB_SNP_42_847	847	0.375	0.5111	0.5652	C/T
HB_SNP_47	Contig_Hb_129382	HB_SNP_47_469	469	0.3478	0.4565	0.5357	C/T
HB_SNP_54	Contig_Hb_28561	HB_SNP_54_148	148	0.3285	0.4217	0.4483	T/G
HB_SNP_55	Contig_Hb_97044	HB_SNP_55_482	482	0.2567	0.3068	0.3143	A/G
		HB_SNP_55_534	534	0.3648	0.487	0.4571	A/T
		HB_SNP_55_354	354	0.3695	0.4965	0.3824	A/T
		HB_SNP_55_472	472	0.3732	0.5035	0.4	C/T
		HB_SNP_55_623	623	0.373	0.5035	0.3824	C/T
HB_SNP_62	Contig_Hb_146058	HB_SNP_62_439	439	0.3476	0.4553	0.4839	A/C

Figure S1. Overview of the workflow for investigating the contribution of novel transcripts in the *H. brasiliensis* database.

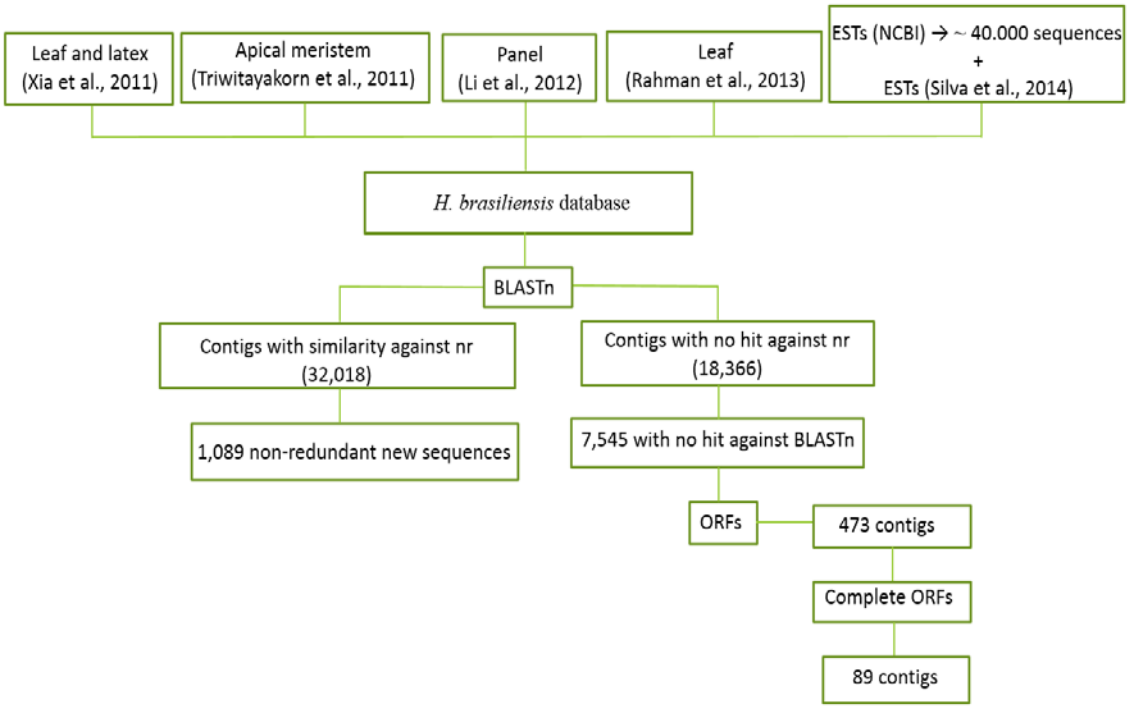


Figure S2. Glycolysis/ gluconeogenesis KEGG pathway. The annotated contigs are indicated in yellow.

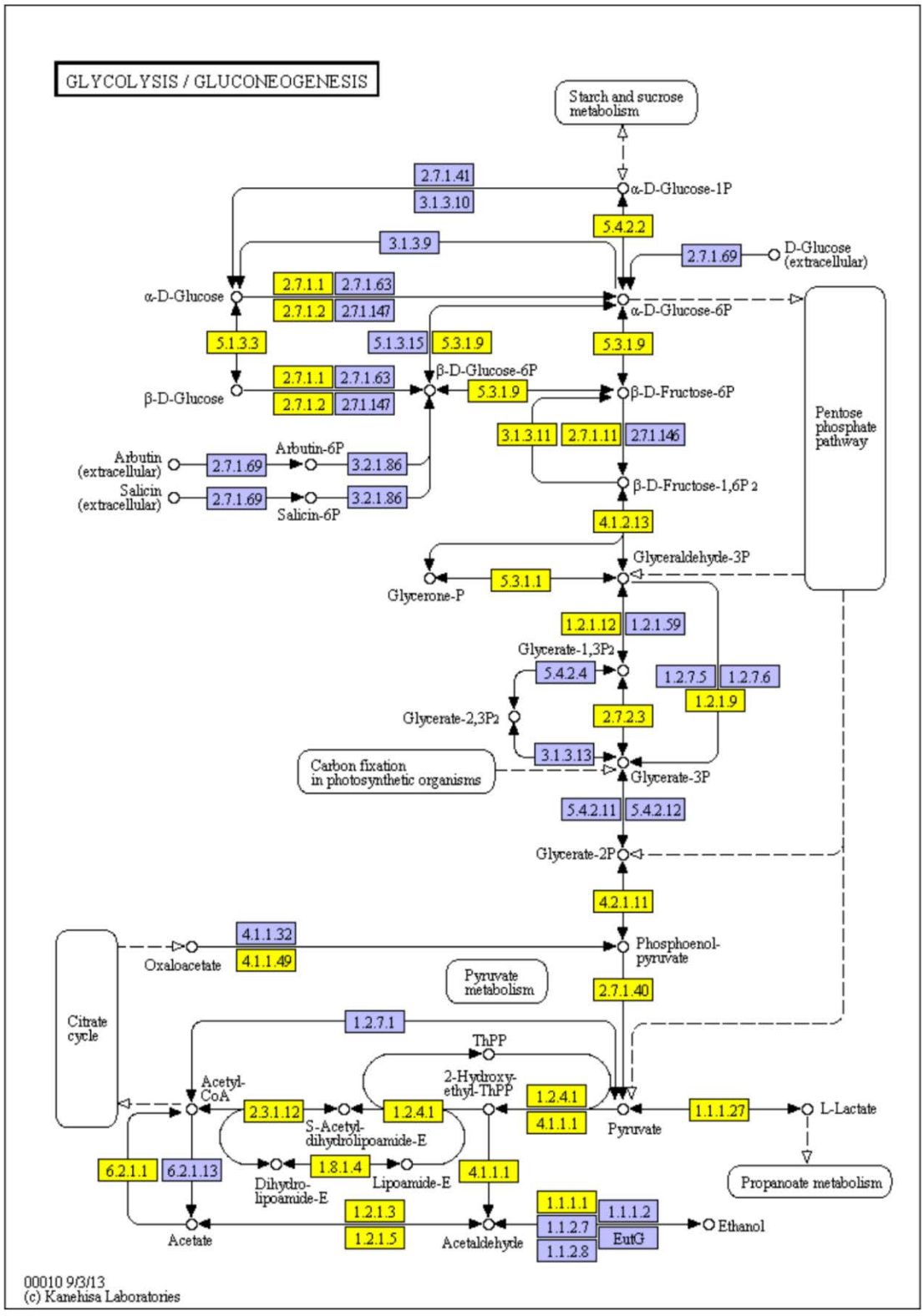


Figure S3. Pyruvate metabolism KEGG pathway.

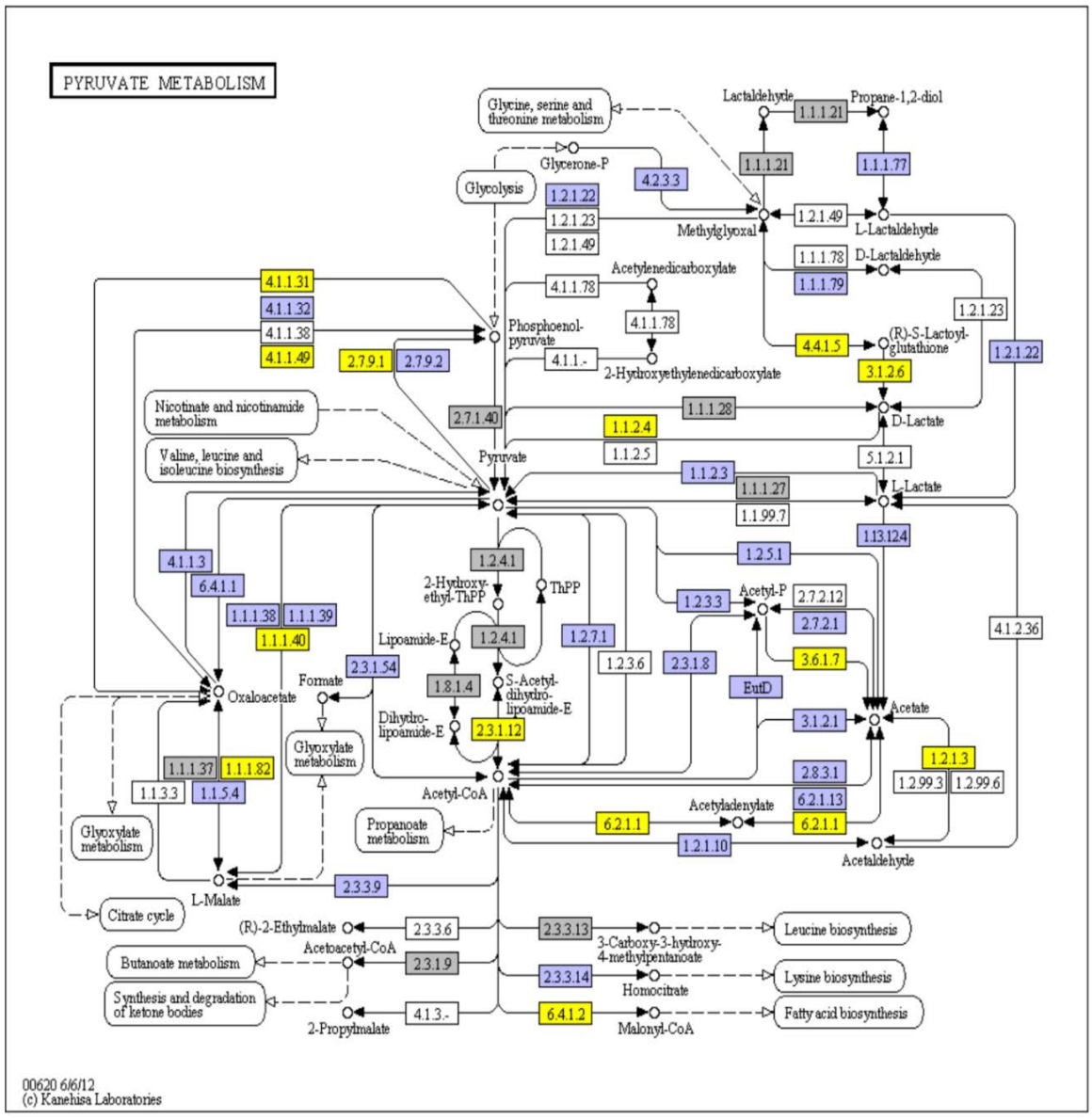


Figure S4. MVA and MEP KEGG pathways.

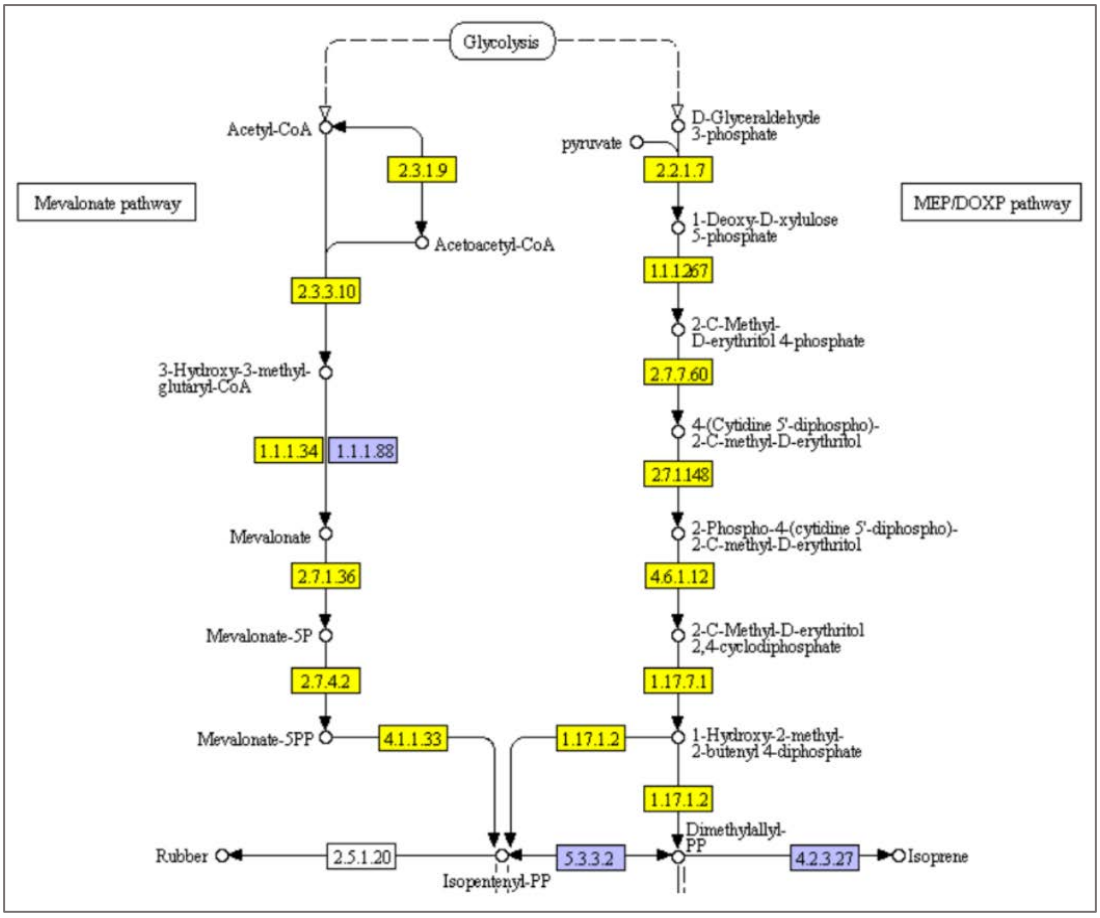


Figure S5. Digital gene expression analysis.

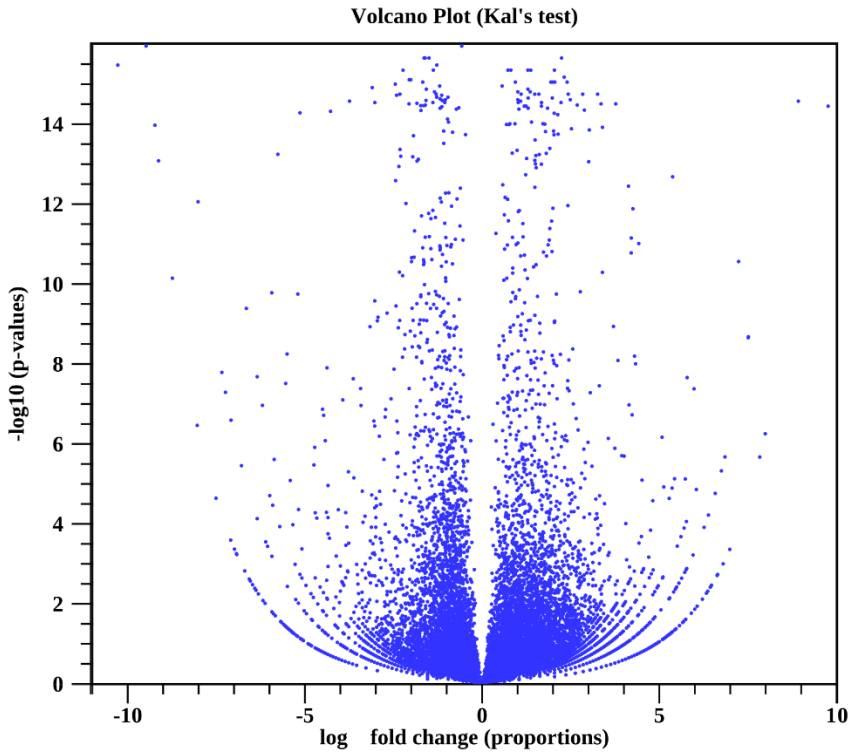


Figure S6. Distribution of putative microsatellite types.

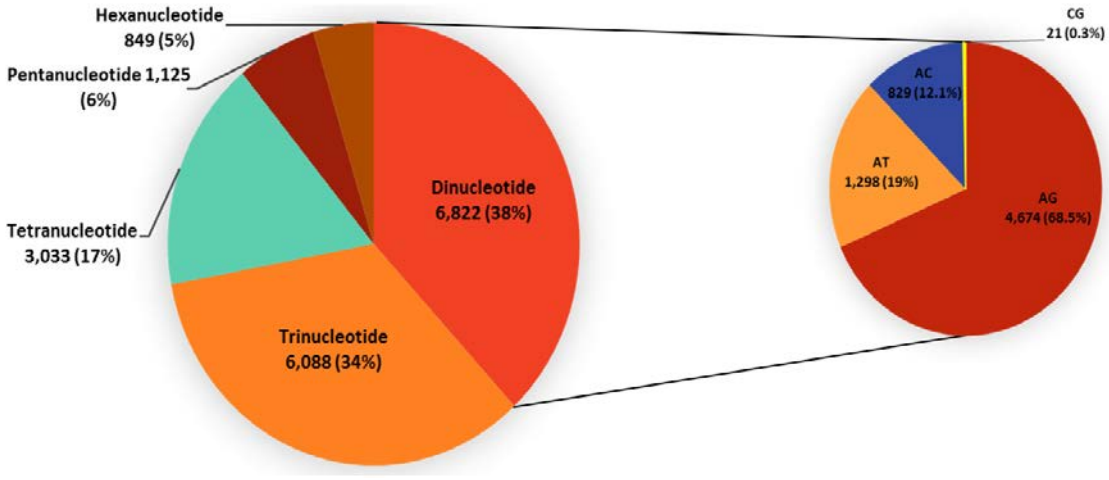
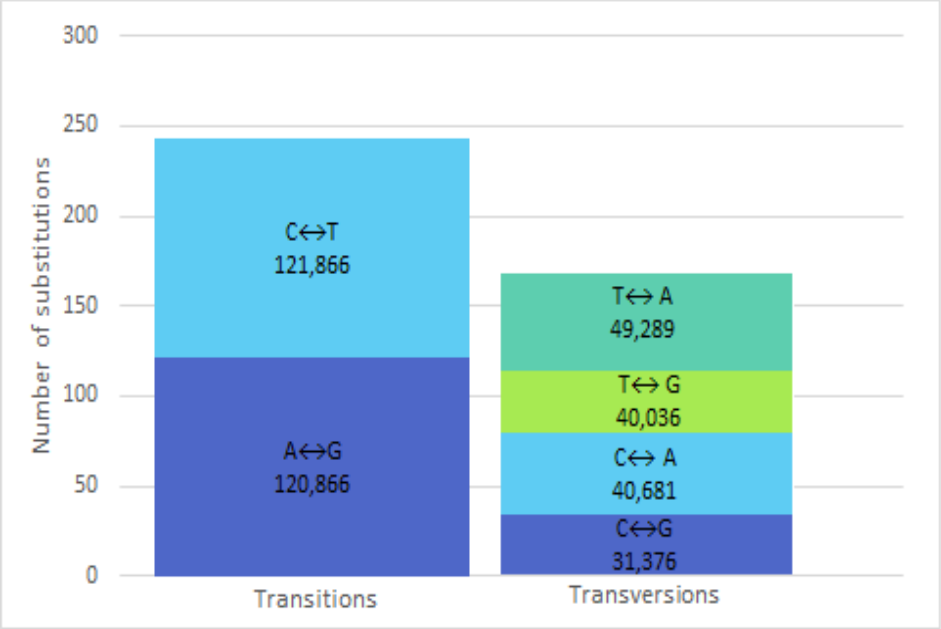


Figure S7. Distribution of putative SNPs that were identified.



Capítulo III

Saturação de um mapa-genético molecular em seringueira (*Hevea brasiliensis*) com marcadores SNPs.

Camila Campos Mantello[1], Carla Cristina da Silva[1], Lívia Moura de Souza[1], Rodrigo Gazaffi[2], Dominique Garcia[3], Vincent Le Guen[3], Saulo Emilio Almeida Cardoso[4], Antônio Augusto Franco Garcia[2], Anete Pereira de Souza[1,5]

[1]Centro de Biologia Molecular e Engenharia Genética (CBMEG) – Universidade Estadual de Campinas (UNICAMP). CEP 13083-970, Campinas, SP, Brasil; [2]Departamento de Genética, ESALQ- Universidade de São Paulo, CEP: 13418-260, Piracicaba, SP, Brasil; [3]CIRAD, UMR, AGAP, Montepellier, Hérault, France; [4]Laboratório de Pesquisa &Desenvolvimento Plantações Michelin da Bahia, Plantações Michelin da Bahia LTDA; [5]Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6109, CEP 13083-970, Campinas, SP, Brasil.

email:

CCM: camila.mantello@gmail.com

CCS: silvacbio@yahoo.com.br

LMS: liviamoura31@gmail.com

RG: rgazaffi@gmail.com

DG: dominique.garcia@cirad.fr

VL: vicent.le_guen@cirad.fr

SEAC: saulo.cardoso@br.michelin.com

AAFG: aafgarci@gmail.com

APS: anete@unicamp.br

Resumo

Hevea brasiliensis é uma espécie nativa da floresta Amazônica e é a principal produtora de borracha natural no mundo. Apesar de a região Amazônica oferecer condições ideais para o crescimento da seringueira, nesta área ocorre o desenvolvimento do fungo *Microcyclus ulei*, causador do mal-das-folhas e é considerada como a principal doença da cultura. Dessa maneira, as plantações de seringueira se expandiram para áreas de escape que oferecem novas condições de estresse como baixa temperatura e umidade. O melhoramento genético vem buscando clones adaptados a estas áreas de escape, porém como o ciclo do melhoramento genético é longo, o desenvolvimento de ferramentas de biologia molecular, como marcadores moleculares e mapas genéticos, podem diminuir o tempo gasto nas seleções do clones.

Este trabalho apresenta um mapa genético integrado saturado com 51 marcadores *single nucleotide polymorphisms* (SNPs). Este mapa contém 383 marcadores distribuídos em 20 grupos de ligação. O comprimento total do mapa é 3.137,6 cM com densidade média de um marcador a cada 8,12 cM. O maior grupo de ligação (LG10) tem 266,2 cM e o menor grupo de ligação (LG19) tem 10,3 cM. Este é o primeiro mapa genético-molecular saturado com marcadores SNPs e representa um avanço para o melhoramento genético de seringueira por ser uma ferramenta que auxilia a identificação de *Quantitative Trait Loci* (QTLs) relacionados a caracteres de importância econômica com maior precisão.

Palavras chaves: seringueira, *Hevea brasiliensis*, mapa genético, SNPs.

Introdução

A seringueira é uma árvore perene, pertencente ao gênero *Hevea* e à família Euphorbiaceae. O gênero é composto por 11 espécies, sendo todas monoicas, alógamas e originadas da floresta Amazônica. [1]. Dentre todas as espécies do gênero *Hevea*, *H. brasiliensis* é a principal fonte produtora de borracha natural do mundo, sendo responsável por 98% da produção mundial [2].

A região Amazônica oferece condições climáticas ótimas para o desenvolvimento e crescimento da seringueira, porém esta é a área de ocorrência do fungo *Microcyclus ulei*, causador do mal-das-folhas que é principal doença que acomete a heveicultura na América. No início do século XXI este fungo, que causa a queda prematura das folhas, dizimou plantações na região norte do Brasil, incluindo as plantações de Henry Ford no estado Pará e até hoje é considerado uma ameaça permanente para indústria da borracha [3]. Dessa maneira, o cultivo da seringueira se expandiu para regiões de escape, como o planalto sul do Brasil. Países asiáticos que cultivam a seringueira também expandiram seu cultivo para regiões de escape como o norte da Índia, Vietnã, sul da China com o objetivo sair das zonas de ocorrência de outras doenças que atacam os seringais nestes países [4].

As regiões de escape proporcionam novas condições de estresse como baixa temperatura e umidade, limitando o crescimento da seringueira e a produção de látex. Dessa forma, o melhoramento genético de seringueira não está buscando apenas genótipos resistentes ao fungo, mas também genótipos que consigam associar boa produtividade e tolerância às novas condições de estresse impostas por estas áreas de escape [5].

O ciclo do melhoramento genético da seringueira é longo, tem duração aproximada de 20 a 30 anos e requer grandes áreas experimentais [6]. Dessa maneira, o desenvolvimento de marcadores moleculares e a construção de mapas genéticos-moleculares são ferramentas de grande importância para os programas de melhoramento e podem ser um ponto de referência para entender a base genética de variações fenotípicas.

Nos últimos anos, o desenvolvimento de marcadores *Single Nucleotide Polymorphisms* (SNPs) tem tido um grande interesse no melhoramento genético [7]. As

plataformas de sequenciamento de nova geração permitiram um avanço na identificação e desenvolvimento de milhares de marcadores SNPs que devido à sua natureza bialélica, têm se mostrado uma ferramenta poderosa para análises genéticas nos programas de melhoramento [8]. Os SNPs desempenham um papel importante nas variações fenotípicas, pois a posição de um SNP pode afetar a expressão de um gene ou até mesmo pode provocar mudança na conformação de uma proteína quando há uma mutação não-sinônima dentro de um gene [9]. Quando comparados a outros marcadores, os SNPs podem ser genotipados em média ou larga escala e têm apresentado alta relação custo-benefício [8].

A construção de mapas genéticos bem saturados permite mapear locos que controlam características quantitativas (QTLs – *Quantitative Trait Loci*) e proporcionam o conhecimento sobre a posição de um gene e sua região adjacente com maior precisão [10]. O desenvolvimento de linhagens homozigotas para construção de mapas de ligação tradicionais (populações F2, RILs) são impraticáveis para seringueira devido ao seu longo ciclo de vida e por sofrer depressão por endogamia. Por isso, o primeiro mapa publicado para seringueira [11] adotou a estratégia de duplo *pseudo-testcross* [12], utilizando uma população segregante F1. Este mapa contém 717 marcas sendo, em sua maioria, RFLPs e AFLPs distribuídos em 18 grupos de ligação [11]. Em 2011, foram publicados dois mapas de ligação para seringueira utilizando a mesma estratégia (duplo *pseudo-testcross*) com marcadores microssatélites [13] e SSRs e AFLPs [14].

Recentemente, Souza et al. (2013) [15] publicaram um mapa genético integrado para seringueira utilizando o software OneMap [16] que utiliza a análise multiponto baseado na cadeia oculta de Markov resultando em um mapa genético mais acurado. Um total de 284 marcadores microssatélites foram mapeados em 23 LGs. O comprimento total do mapa de ligação foi de 2688,8 cM com densidade de um marcador por 9,83 cM. Neste mapa foram detectados 18 QTLs para crescimento durante períodos de verão e inverno. Mais recentemente este mapa foi saturado com 46 marcadores microssatélites de sequências expressas (EST-SSRs) [17] obtendo um mapa com 22 grupos de ligação e comprimento total de 3068,9 cM.

Este estudo teve como objetivo saturar o mapa genético de Souza et al. (2013) [15] e Silva (2014) [17] com marcadores SNPs utilizando a mesma metodologia proposta por

Souza et al. (2013) [15]. Pela primeira vez um mapa genético saturado com marcadores SNPs está sendo apresentado para seringueira. Este mapa representa um avanço importante para o melhoramento e para estudos genéticos na espécie.

Material e Métodos

População de mapeamento e extração de DNA

A população de mapeamento neste estudo é a mesma utilizada por Souza et al (2013) [15]. Esta população foi gerada através do cruzamento entre os clones PB217, considerado um bom produtor de látex e PR255 que apresentou tolerância ao frio. A progênie foi obtida através da polinização controlada dos dois genitores (PB 217 e PR255) e cada indivíduo da progênie F1 foi propagado no campo experimental por enxertia utilizando o clone GT1 como porta-enxerto. A população de mapeamento é composta por 270 indivíduos e foi plantada em 2006 na fazenda Edouard Michelin em Itiquira, Mato Grosso, Brasil (17° 23’ 59.60’’ S e 54° 44’ 53.93’’ O, altitude 519 m), que é caracterizada por um clima sub-tropical com um período frio e seco que dura de maio a outubro. O delineamento experimental foi blocos ao acaso com 4 repetições, utilizando 4 plantas por repetição, totalizando uma área de 7,2 hectares.

A extração de DNA da população de mapeamento e dos genitores foi realizada de acordo com Souza et al. (2013). O DNA foi quantificado com o Nanodrop (Thermo Scientific, Wilmington, DE) e QuantiFluor® dsDNA System (Promega, Madison, WI).

Desenho de *primers* e análise dos marcadores

Foram selecionados 119 sequências com 1 SNP por sequência, dentre os quais 33 foram identificados a partir de bibliotecas de ESTs [18] e 86 identificados no transcriptoma obtido por Mantello et al (2014) [19].

A genotipagem dos SNPs foi realizada através da técnica de espectrometria de massa na plataforma Sequenom MassaARRAY® (Sequenom, San Diego, CA). Para isto, as sequências com os SNPs devidamente identificados foram submetidas no software

MassARRAY® Assay Design. O desenho de *primers* foi feito utilizando os seguintes parâmetros: (1) *High Plex preset* com máximo de 24 SNPs por multiplex, (2) amplicons de no máximo 200 pb e (3) 10 iterações. Foram desenhados 119 pares de *primers* para a reação de caputra e 119 *primers* para a reação de extensão.

As reações de genotipagem dos SNPs foram feitas em três etapas: (1) reação de captura partindo de 10ng/ul, (2) neutralização de dNTPs livres, (3) reação de extensão. O *primer* utilizado na reação de extensão se anela a região adjacente ao SNP e incorpora uma única base. A diferença de massa entre os produtos da reação de extensão permitem detectar a variação alélica.

Os dados obtidos na genotipagem foram analisados no software SuperMASSA [20] que utiliza o método Bayesiano para calcular a probabilidade da dosagem alélica por loco por indivíduo.

Mapa de ligação

O mapa de ligação foi construído com o programa OneMap [16] versão 2.0-1 que utiliza tecnologia multiponto baseada na cadeia de Markov oculta. Este programa permite a análise conjunta de marcadores segregando 1:1, 1:2:1, 3:1 e 1:1:1:1 e a construção de mapas genéticos integrados para populações segregantes F1. Os parâmetros utilizados para este mapa foram os mesmo descritos por Souza et al. (2013) [15], utilizando LOD score de 4,5 e fração de recombinação de 0,4. A ordenação dos marcadores foi obtida através dos algoritmos *compare* (para grupos com até seis marcadores) e *order* (para grupos com mais de seis marcadores). As frequências de recombinação foram convertidas em distância genética (centiMorgans) através da função de Kosambi (1944) [21]. O teste de qui-quadrado com correção de Bonferroni foi feito para testar a hipótese de segregação mendeliana dos marcadores.

Resultados e Discussão

Análise dos marcadores SNPs

Os 119 SNPs foram distribuídos em 3 plex com 24 SNPs cada e um plex com 23 SNPs. Os dados obtidos na genotipagem foram submetidos individualmente na versão online do software SuperMASSA [20].

Do total de 119 SNPs, 53 (44,5%) são polimórficos entre os genitores da população de mapeamento, 29 (24,4%) são monomórficos, 22 (18,4%) falharam na amplificação e 15 (12,6%) apresentaram dosagem alélica maior que dois. Os marcadores polimórficos entre os genitores da população de mapeamento estão listados na tabela S1.

Por saber que a seringueira é um organismo de origem alotetraploide que se comporta como um diploide (2n=36; x=9), a probabilidade de dosagem alélica maior que dois calculada pelo software SuperMASSA é um indicativo de duplicação gênica. Estudos anteriores observaram duplicações de locos para marcadores do tipo RFLPs [11] e SSRs [22,23] em *H. brasiliensis* e esta é a primeira vez que foram observados indícios de duplicação para locos de SNPs em seringueira (Tabela 1).

SNP	Probabilidade[a]	Anotação funcional[b]
Hb_seq_01	0.963	sp|P15252.2|REF_HEVBR - Rubber elongation factor protein
Hb_seq_06	0,999	ref|XP_002262780.1| - membrane steroid-binding protein 2 [Vitis vinifera]
Hb_seq_16	0,999	1-deoxyxylulose-5-phosphate synthase, putative [Ricinus communis]
Hb_seq_45_2	0,999	ABC transporter C family member 3-like [Glycine max]
Hb_seq_70	0,986	hypothetical protein 1 [Hevea brasiliensis]
Hb_seq_85	0,796	chlorophyll A/B binding protein [Jatropha curcas]
Hb_seq_131	1,0	Peroxidase 31 precursor, putative [Ricinus communis]
Hb_seq_153	1,0	Rop guanine nucleotide exchange factor, putative [Ricinus communis]
Hb_seq_160_3	1,0	no hit
Hb_seq_174	1,0	cytochrome P450, putative [Ricinus communis]
Hb_seq_177	0,718	(Di)nucleoside polyphosphate hydrolase, putative [Ricinus communis]
Hb_seq_190_2	0,982	Ran GTPase binding protein, putative [Ricinus communis]
Hb_seq_192	0,979	AMP deaminase, putative [Ricinus communis]
Hb_seq_198	0,970	molybdopterin cofactor synthesis protein A, putative [Ricinus communis]
Hb_seq_204	0,988	Reticuline oxidase precursor, putative [Ricinus communis]

Tabela 1. Locos com marcadores SNPs com elevada probabilidade de duplicação

a-Probabilidade da dosagem alélica ser maior que dois calculada pelo programa SuperMASSA; b- Anotação funcional das sequências no banco de dados *non-redundant* (nr)

Sabe-se que a duplicação gênica tem papel fundamental evolução adaptativa, com a criação de novas funções ou na regulação da expressão gênica [24]. Neste estudo, o SNP Hb_seq_01 anotado para a proteína *rubber elongation fator protein* e o SNP Hb_seq_16 anotado para enzima *1-deoxyxylulose-5-phosphate synthase* estão diretamente ligados à maquinaria de síntese do látex e apresentaram elevada probabilidade da dosagem alélica ser maior que dois, ou seja, há indícios de duplicação nestes locos (Tabela 1). Todos os locos apresentados na Tabela 1 serão investigados para confirmar a hipótese de duplicação. A duplicação dos locos Hb_seq_01 e Hb_seq_16 associada ao polimorfismo encontrado nestes locos podem conferir características únicas que diferenciam a produção de látex entre os cultivares e torna a seringueira a maior produtora de borracha natural.

Mapa de ligação

Inicialmente, foi feito o teste do qui-quadrado nos 53 marcadores SNPs polimórficos e 20 marcadores, todos com segregação 1:2:1, apresentaram distorção na segregação.

Dos 53 marcadores polimórficos entre os genitores, 51 foram mapeados no mapa publicado por Souza et al. (2013) e recentemente saturado com marcadores EST-SSRs por Silva (2014)[17]. Apesar de 20 marcadores terem apresentado distorção no padrão de segregação, estes foram mantidos por não distorcerem os grupos de ligação. Do total de 51 marcadores SNPs, 21 (41%) apresentam segregação 1:2:1 e 30 (59%) a segregação 1:1. Além disso,dois marcadores EST-SSRs desenvolvidos por Silva et al.,(2014) [18] que haviam sido desconsiderados, nesta análise foram incorporados ao mapa.

O mapa genético integrado contém 383 marcadores (332 SSRs e 51 SNPs) distribuídos em 20 grupos de ligação e um comprimento total de 3.137,6 cM. Dentre os grupos de ligação obtidos, LG10 foi o maior grupo com 266,2 cM e o LG19 o menor com 10,3 cM (Tabela 2).

Tabela 2. Distribuição dos marcadores entre os grupos de ligação.

Grupo de Ligacao	N° de marcadores	Comprimento (cM)	Densidade de marcadores
LG1	28	167,9	5,99
LG2a	17	109,4	6,43

LG2b	4	24,3	6,07
LG3	20	178,0	8,9
LG4	15	129,3	8,62
LG5	24	194,7	8,11
LG6	22	160,3	7,28
LG7	25	163,8	6,55
LG8	29	231,0	7.96
LG9	18	144,8	8,04
LG10	38	266,2	7,0
LG11	9	160.8	17,86
LG12	14	202,1	14,43
LG13	20	178,4	8,92
LG14	26	197,0	7,57
LG15	17	155,1	9.12
LG16	19	209.0	11,0
LG17	15	114,6	7.64
LG18	19	140,6	7,4
LG19	4	10,3	2.57

Assim como o mapa de ligação prévio [15] recentemente saturado por Silva [17], os grupos de ligação foram organizados de acordo com o mapa previamente obtido por Lespinasse et al (2000) [11] e com informações de outros mapas ainda não publicados.

O mapa publicado por Souza et al., (2013) [17] obteve 23 grupos de ligação, já o mapa saturado por Silva [18] obteve 22 grupos. O mapa saturado com os marcadores EST-SSRs descritos por Silva [18] aumentou a saturação do grupo de ligação LG11 e permitiu a integração entre os grupos LG11a e LG11b. Neste estudo, a saturação do mapa com marcadores SNPs, permitiu que os grupos de ligação previamente denominados como LG6a e LG6b e LG17a e LG17b se agrupassem e passaram a ser denominados como LG6 e LG17. Apesar deste novo mapa ter sido saturado com novos marcadores, os grupos de ligação LG2a e LG2b ainda se mantêm separados.

Os marcadores utilizados por Souza et al [17] estão destacados em preto, os EST-SSRs mapeados por Silva [18] estão em azul e os novos marcadores SNPs mapeados estão em vermelho (Figura 1). O número e a densidade de marcadores, assim como o tamanho de cada LG estão apresentados na Tabela 2. A densidade média de marcadores foi de um marcador a cada 8,12 cM e está maior quando comparado ao mapa publicado por Souza et

al. (2013) (1/9.8cM) [15] e Silva et al. (2014) (1/9,3) [17], porém esta densidade é menor quando comparada ao mapa publicado por Lespinasse et al. (2000) (1/3cM) [11].

Para seringueira, são esperados 18 grupos de ligação (2n=36, n=18) no mapa genético, entretanto, foram obtidos 20 grupos de ligação. Os grupos LG2b e LG19 se mantiveram com apenas 4 microssatélites cada. A saturação do mapa com novos marcadores SNPs permitiu a união de grupos de ligação que não se agruparam com marcadores SSRs e diminuiu a distância de regiões pouco saturadas (*gaps*) como nos grupos LG 10, LG13 e LG15. Apesar disso, algumas regiões pouco saturadas no mapa de Souza et al. (2013) [15] e Silva (2014) [17] se mantiveram, como observado nos grupos de ligação LG 3, LG11 e LG16, mostrando que ainda há uma cobertura incompleta do genoma.

Mapas previamente publicados [11,14] utilizaram outros tipos de marcadores como AFLPs e RFLPs e demonstraram ser possível chegar em 18 grupos de ligação. Como os marcadores SNPs se apresentam em grande abundância no genoma e desempenham um papel importante na indução de variações fenotípicas [9] estes marcadores podem ser uma alternativa para saturação deste mapa genético e obter os 18 grupos de ligação na busca de novos QTLs. Técnicas de genotipagem de média e larga escala, como as plataformas Sequenom MassARRAY® (Sequenom, San Diego, CA), Golden Gate e Infinuim HD (Illumina, San Diego, CA) e *Genotyping-by-Sequencing* (GBS) [25] mostram uma alternativa rápida e eficiente para obter um mapa de ligação bem saturado.

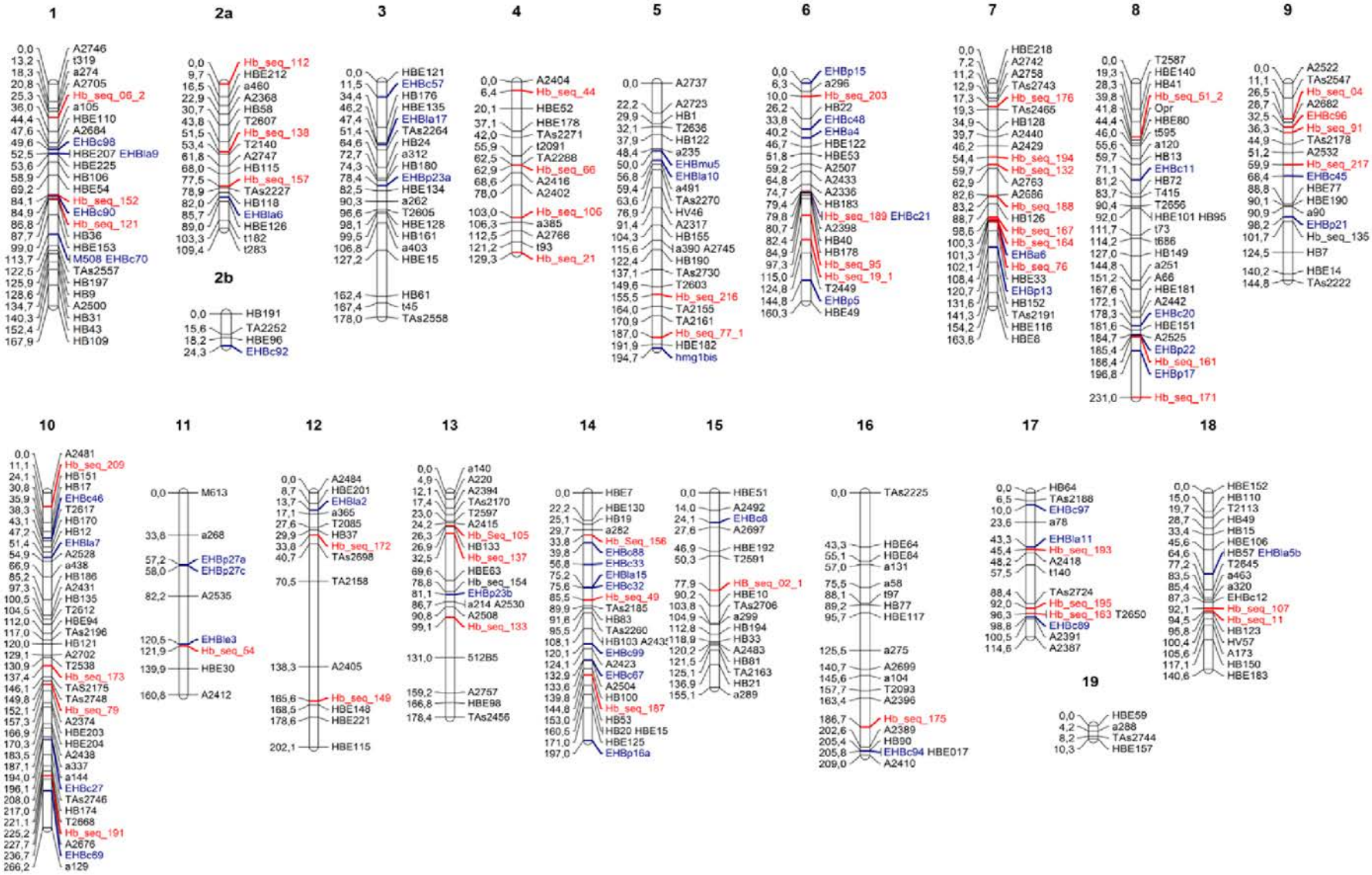


Figura 2. Mapa genético saturado com 383 marcadores.

Contribuição dos autores

CCM genotipou a população de mapeamento. CCM, CCS, LMS e RG foram responsáveis pela análise estatística e construção do mapa genético. DG, VL e SEAC foram responsáveis pela obtenção, cultivo e manutenção da população de mapeamento. AAFG e APS foram responsáveis pelo delineamento e coordenação deste estudo.

Financiamento da Pesquisa

Os autores gostariam de agradecer à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento do projeto e pelas bolsas de estudos de CCM (2008/55974-1 e 2011/50188-0), CCS (2009/52975-0), LMS (2007/50562-4), RG 2009/14068-0). Os autores agradecem ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) pelo auxilio a APS e AAFG.

Referências Bibliográficas

1. Pires JM, Secco R de S, Gomes JI (2002) Taxonomia e Fitogeografia das seringueira: 103.

2. Pootakham W, Chanprasert J, Jomchai N, Sangsrakru D, Yoocha T, et al. (2012) Development of genomic-derived simple sequence repeat markers in Hevea brasiliensis from 454 genome shotgun sequences. Plant Breed 131: 555–562. doi:10.1111/j.1439-0523.2012.01982.x.

3. Pushparajah E (2001) Natural rubber. In: Last, F.T. (ed) Tree Crop Ecosystems. Amsterdam, The Netherlands: Elsevier Science.

4. Raj S, Das G, Pothen J, Dey SK (2005) Relationship between latex yield of Hevea brasiliensis and antecedent environmental parameters. Int J Biometeorol 49: 189–196. doi:10.1007/s00484-004-0222-6.

5. Pushparajah E (1983) Problems and potentials for establishing Hevea under difficult environmental conditions. Planter 59: 242–251.

6. Goncalves, P. de S. and JM (2008) Melhoramento genético da seringueira: passado, presente e futuro. Alvarenga, A. de P. & CAFS do C, editor Viçosa: EPAMIG.

7. Rafalski A (2002) Applications of single nucleotide polymorphism in crop genetics. Curr Opin Plant Biol 5: 94–100.

8. Ganal MW, Altmann T, Röder MS (2009) SNP identification in crop plants. Curr Opin Plant Biol 12: 211–217.

9. Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, Aoki K, et al. (2013) Genome-wide SNP genotyping to infer the effects on gene functions in tomato. DNA Res 20: 221–233. doi:10.1093/dnares/dst005.

10. Shirasawa K, Hirakawa H (2013) DNA marker applications to molecular genetics and genomics in tomato. Breed Sci 63: 21–30. doi:10.1270/jsbbs.63.21.

11. Lespinasse, D., Rodier-Goud, M., Grivet, L., Leconte, A., Legnate, H., & Seguin M (2000) A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. Theor Appl Genet 100: 127–138.

12. Grattapaglia D, Sederoff R, Carolina N (1994) Genetic Linkage Maps. 1137: 1121–1137.

13. Triwitayakorn K, Chatkulkawin P, Kanjanawattanawong S, Sraphet S, Yoocha T, et al. (2011) Transcriptome sequencing of Hevea brasiliensis for development of microsatellite markers and construction of a genetic linkage map. DNA Res 18: 471–482. doi:10.1093/dnares/dsr034.

14. Le Guen V, Garcia D, Doaré F, Mattos CRR, Condina V, et al. (2011) A rubber tree's durable resistance to Microcyclus ulei is conferred by a qualitative gene and a major quantitative resistance factor. Tree Genet Genomes 7: 877–889. doi:10.1007/s11295-011-0381-7.

15. Souza LM, Gazaffi R, Mantello CC, Silva CC, Garcia D, et al. (2013) QTL mapping of growth-related traits in a full-sib family of rubber tree (Hevea brasiliensis) evaluated in a sub-tropical climate. PLoS One 8: e61238. doi:10.1371/journal.pone.0061238.

16. Margarido GR a, Souza a P, Garcia a a F (2007) OneMap: software for genetic mapping in outcrossing species. Hereditas 144: 78–79. doi:10.1111/j.2007.0018-0661.02000.x.

17. Silva CC (2014) Construção de uma para funcional em seringueira (Hevea brasiliensis) Universidade Estadual de Campinas.

18. Silva CC, Mantello CC, Campos T, Souza LM, Gonçalves PS, et al. (2014) Leaf-, panel- and latex-expressed sequenced tags from the rubber tree (Hevea brasiliensis) under cold-stressed and suboptimal growing conditions: the development of gene-targeted functional markers for stress response. Mol Breed.

19. Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, et al. (2014) De Novo Assembly and Transcriptome Analysis of the Rubber Tree (Hevea brasiliensis) and SNP Markers Development for Rubber Biosynthesis Pathways. PLoS One 9: e102665. doi:10.1371/journal.pone.0102665.

20. Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS One 7: e30906.

21. KOSAMBI D (1944) The estimation of map distance from recombination values. Annu Eugenetics 12: 172–175.

22. Souza, LM, Mantello, CC, Santos, MO, Goncalves PS, Souza A (2009) Microsatellites from rubber tree (Hevea brasiliensis) for genetic diversity analysis and cross-amplification in six Hevea wild species. Conserv Genet Resour 1: 75–79.

23. Mantello CC, Suzuki FI, Souza LM, Gonçalves PS, Souza AP (2012) Microsatellite marker development for the rubber tree (Hevea brasiliensis): characterization and cross-amplification in wild Hevea species. BMC Res Notes 5: 329. doi:10.1186/1756-0500-5-329.

24. Fischer I, Dainat J, Ranwez V, Glémin S, Dufayard J-F, et al. (2014) Impact of recurrent gene duplication on adaptation of plant genomes. BMC Plant Biol 14: 151. doi:10.1186/1471-2229-14-151.

25. Elshire RJ, Glaubitz JC, Sun Q, Poland J a, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379. doi:10.1371/journal.pone.0019379.

Material Suplementar

Tabela S1. Lista dos marcadores SNPs polimórficos na população de mapeamento.

SNP	Segregação	variação alelica	Anotação Funcional
HB_seq_02_1	1:2:1	A/G	copper chaperone [Jatropha curcas])
Hb_seq_04	1:2:1	G/T	Pathogenesis-related protein 10.5 [Theobroma cacao])
Hb_seq_06_2	1:2:1	C/T	membrane steroid-binding protein 2 [Vitis vinifera])
Hb_seq_11	1:1	A/G	hydroxymethylglutaryl-CoA reductase (NADPH)
Hb_seq_19_1	1:1	A/G	putative 1-deoxy-D-xylulose 5-phosphate synthase [Hevea brasiliensis]
Hb_seq_21	1:1	G/T	1-deoxyxylulose-5-phosphate synthase, putative [Ricinus communis]
Hb_seq_44	1:1	C/T	(polyubiquitin 14 [Arabidopsis thaliana])
Hb_seq_49	1:1	C/T	(universal stress protein A-like protein [Vitis vinifera])
Hb_seq_51_2	1:1	C/G	(membrane family protein [Populus trichocarpa])
Hb_seq_54	1:1	A/G	plastocyanin, chloroplastic [Solanum lycopersicum]
Hb_seq_66	1:2:1	C/T	metallothionein [Hevea brasiliensis]
Hb_seq_76	1:2:1	A/T	Stress-induced hydrophobic peptide [Theobroma cacao]
Hb_seq_77_1	1:1	A/T	plastocyanin, chloroplastic [Solanum lycopersicum]
Hb_seq_79	1:2:1	A/C	predicted protein [Populus trichocarpa]
Hb_seq_91	1:2:1	G/T	no hit
Hb_seq_95	1:1	G/C	no hit
Hb_Seq_105	1:1	T/C	pentatricopeptide repeat-containing protein, putative [Ricinus communis]
Hb_seq_106	1:2:1	A/C	conserved hypothetical protein [Ricinus communis]
Hb_seq_107	1:1	A/G	conserved hypothetical protein [Ricinus communis]
Hb_seq_112	1:1	C/T	conserved hypothetical protein [Ricinus communis]
Hb_seq_121	1:1	C/T	conserved hypothetical protein [Ricinus communis]
Hb_seq_132	1:2:1	G/C	conserved hypothetical protein [Ricinus communis]
Hb_seq_133	1:1	G/C	conserved hypothetical protein [Ricinus communis]
Hb_seq_135	1:1	A/T	ATP binding protein, putative [Ricinus communis]
Hb_seq_136	1:2:1	A/T	amine oxidase, putative [Ricinus communis]
Hb_seq_137	1:1	A/T	pentatricopeptide repeat-containing protein, putative [Ricinus communis]
Hb_seq_138	1:2:1	C/T	predicted protein [Populus trichocarpa]
Hb_seq_149	1:2:1	C/T	conserved hypothetical protein [Ricinus communis]
Hb_seq_152	1:1	G/C	predicted protein [Populus trichocarpa]
Hb_seq_154	1:1	A/T	hydroxypyruvate reductase, putative [Ricinus communis]
Hb_Seq_156	1:1	A/G	serine/threonine protein kinase, putative [Ricinus communis]

Hb_seq_157	1:1	A/G	predicted protein [Populus trichocarpa]
Hb_seq_161	1:2:1	A/G	GTP binding protein, putative [Ricinus communis]
Hb_seq_163	1:1	A/G	associated receptor kinase 1 precursor, putative [Ricinus communis]
Hb_seq_164	1:2:1	A/T	FACT complex subunit SPT16, putative [Ricinus communis]
Hb_seq_167	1:1	A/G	conserved hypothetical protein [Ricinus communis]
Hb_seq_171	1:2:1	A/G	leucine-rich repeat-containing protein, putative [Ricinus communis]
Hb_seq_172	1:1	C/T	pentatricopeptide repeat-containing protein, putative [Ricinus communis]
Hb_seq_173	1:1	C/T	predicted protein [Populus trichocarpa]
Hb_seq_175	1:1	G/T	hypothetical protein RCOM_1407360 [Ricinus communis]
Hb_seq_176	1:2:1	A/C	transcription factor, putative [Ricinus communis]
Hb_seq_187	1:2:1	A/G	polycomb protein embryonic flower, putative [Ricinus communis]
Hb_seq_188	1:2:1	C/T	heat shock protein binding protein, putative [Ricinus communis]
Hb_seq_189	1:1	G/C	DNA (cytosine-5)-methyltransferase, putative [Ricinus communis]
Hb_seq_191	1:2:1	A/G	Inositol-tetrakisphosphate 1-kinase, putative [Ricinus communis]
Hb_seq_193	1:1	A/C	phospholipase A-2-activating protein, putative [Ricinus communis]
Hb_seq_194	1:1	A/G	pathogenesis-related protein 1 [Vitis hybrid cultivar]
Hb_seq_195	1:1	A/G	UDP-glucosyltransferase, putative [Ricinus communis]
Hb_seq_203	1:1	A/G	no hit
Hb_seq_209	1:2:1	C/T	protein phosphatase 2c, putative [Ricinus communis]
Hb_seq_216	1:2:1	A/T	pentatricopeptide repeat-containing protein, putative [Ricinus communis]
Hb_seq_217	1:1	G/T	mom (plant), putative [Ricinus communis]

*Anotação funcional das sequências contra o banco de dados *non-redundant* (nr).

Considerações Finais

A seringueira é uma espécie nativa do Brasil e desperta um grande interesse econômico por ser a principal fonte produtora de borracha natural do mundo. Desde o início da sua domesticação, iniciada no final do século XIX até os dias atuais, a seringueira passou por dois a três ciclos de melhoramento e, por isso, do ponto de vista do melhoramento genético, a espécie encontra-se em seu estágio selvagem. Dessa forma, o desenvolvimento de ferramentas de biologia molecular tem grande importância quando aliadas ao melhoramento genético clássico para obtenção de indivíduos superiores e diminuição do ciclo de seleção.

Os marcadores moleculares do tipo microssatélite apresentam natureza codominante, possuem ampla distribuição no genoma, alta reprodutibilidade e são ferramentas muito populares em estudos de mapeamento genético e estudos populacionais. No início deste trabalho, o número de marcadores disponíveis para seringueira era restrito, por isso com o objetivo de desenvolver novos marcadores SSRs para a espécie, foram construídas bibliotecas enriquecidas em microssatélites seguindo o protocolo proposto por Billote et al., (1999). As bibliotecas construídas utilizaram a estratégia de enriquecimento em microssatélites dinucleotídeos por saber que este tipo de SSR é abundante no genoma (MORGANTE; HANAFEY; POWELL, 2002) e em trinucleotídeos para tentar selecionar microssatélites em genes.

O capítulo I descreve a construção destas bibliotecas enriquecidas utilizando sondas biotiniladas $(CT)_8$ e $(GT)_8$ para as bibliotecas enriquecidas em dinucleotídeos e $(ATC)_8$ e $(CCT)_8$ para as bibliotecas enriquecidas em trinucleotídeos. Os clones foram sequenciados em suas extremidades com objetivo de formar uma sequência consenso.

No total, foram sequenciados 576 clones das bibliotecas enriquecidas em dinucleotídeos, dos quais 291 continham SSRs, e 288 clones das bibliotecas enriquecidas em trinucleotídeos, dos quais 62 continham SSRs. Quando é feita a comparação do rendimento dos dois tipos de bibliotecas, as bibliotecas enriquecidas em trinucleotídeos apresentaram baixo rendimento (21%) e é um indicativo de que este tipo de SSR encontra-se em menor frequência no genoma.

Com as sequências de microssatélites devidamente identificadas, 121 pares de *primers* foram desenhados a partir das bibliotecas enriquecidas em dinucleotídeos e 32 das bibliotecas em trinucleotídeos, totalizando 153 pares de *primers*. Estes pares de *primers* desenvolvidos foram testados quanto a amplificação e polimorfismo em nove genótipos de *H. brasiliensis* que são genitores de cinco populações de mapeamento em estudo. Do total de 153 pares de *primers*, 119 apresentaram bom produto de amplificação e 90 foram polimórficos. Tendo em vista amostrar a variabilidade genética dos locos de microssatélites, foram escolhidos 46 locos, dentre os 90 que foram polimórficos, para serem caracterizados em 36 genótipos de *H. brasiliensis*. O número de alelos encontrados variou de 2 a 17 mostrando a grande variabilidade genética dentro de *H. brasiliensis*.

Os 46 locos de SSRs que foram caracterizados nos 36 genótipos de *H. brasiliensis* também foram testados quanto a transferibilidade dos marcadores em outras seis espécies do gênero *Hevea* (*H. benthamiana*, *H. nitida*, *H. guianensis*, *H. pauciflora*, *H. camargoana*, *H. rigidifolia*) que foram gentilmente cedidas pela Embrapa Amazônia Ocidental. Os marcadores apresentaram taxa de transferibilidade superior a 80% para todas as espécies corroborando com os resultados encontrados por Souza et al. (2009).

Segundo Barbará et al. (2007) a transferibilidade de marcadores microssatélites polimórficos dentro de gêneros tem taxa de êxito perto de 60% nas dicotiledôneas. Esta taxa de transferibilidade elevada dentro do gênero *Hevea* indica que as regiões flanqueadoras aos microssatélites são bem conservadas entre as espécies de *Hevea* e um indicativo de que o gênero *Hevea* é um complexo de espécies como propõe Pires (2002). Os locos de SSRs também foram avaliados quanto ao número de alelos presentes quando as 6 espécies são incluídas na análise. O número de alelos observados aumentou, variando de 2 a 21 e evidencia alelos exclusivos para estas espécies.

Com a exceção de *H. brasiliensis*, as outras espécies do gênero *Hevea* são encontradas na forma de exemplares nos bancos de germoplasmas, sendo algumas destas espécies não representadas em coleções no Brasil. Estas espécies podem ser utilizadas como fonte de recursos genéticos para características importantes no melhoramento como, por exemplo, resistência a doenças. Dessa forma, faz-se necessário organizar novas

expedições para coletar estas espécies e representá-las nos bancos de germoplasma como fonte de recurso genético nos programas de melhoramento.

Com o objetivo de desenvolver novos marcadores SSRs e SNPs em larga escala, o capítulo II descreve o sequenciamento do transcriptoma de painel na plataforma de sequenciamento de nova geração Illumina GAIIx. Este tecido foi escolhido para ser sequenciado por ter importância fundamental na seringueira, pois contém os vasos laticíferos que sintetizam e estocam o látex. Neste estudo, foram coletadas amostras de painel do clone GT1 por ser bom produtor de látex, tolerante ao frio e genitor de duas populações de mapeamento (GT1 x RRIM 701 e GT1 x PB235) desenvolvidas em parceria com o IAC, e do clone PR255 por ser tolerante ao frio e genitor da população de mapeamento PB 217 x PR255 devenvolvida pela empresa Michelin Ltda em parceria com o CIRAD.

O transcriptoma foi montado através do método *de novo*. Nesta montagem, foram obtidos aproximadamente 152 mil contigs, dos quais um conjunto de sequências com tamanho mínimo de 400 pb foram selecionadas por apresentarem maior confiabilidade na anotação funcional. Aproximadamente 58 mil contigs haviam sido selecionados, porém neste conjunto de sequências foram identificadas sequências contaminantes, ou seja, sequências que apresentaram similaridade com outros organismos como, por exemplo, fungo. Após a remoção destas sequências, um conjunto de 50.384 contigs foram utilizados as análises subsequentes.

Os 50.384 contigs foram anotados em banco de dados como o nr, *Gene Ontology* e KEGG. Além disso, estes contigs foram comparados às sequências disponíveis para seringueira com o objetivo de investigar as contribuições deste estudo para a espécie. Para isto, foi montado um banco de dados com as sequências disponíveis dos trabalhos de RNA-seq (XIA et al., 2011, TRIWITAYAKORN et al., 2011, LI et al., 2012, RAHAM et al., 2013), com banco de ESTs do NCBI e com as sequências ESTs disponibilizadas por SILVA et al. (2014). Através da ferramenta BLASTn foram identificados 1.089 contigs que foram anotados no banco de dados do nr, porém não haviam sido catalogados para *H. brasliensis*.

Após a caracterização do transcriptoma, os 50.384 contigs foram submetidos à busca de microssatélites e SNPs. Na busca de SSRs, foram identificados 17.927 microssatélites em 13.070 sequências, indicando a presença de mais de um microssatélite por sequência. Para identificar novos SSRs em seringueira, foi feito um BLASTn com o conjunto de 13.070 contigs contra o banco de dados de seringueira descrito anteriormente. Dentre os 13.070 contigs, foram identificados 1.709 contigs que não apresentaram similaridade com o banco de dados de seringueira contendo, portanto, novos microssatélites para a espécie. Apesar de os dados terem sido gerados por sequências expressas, os microssatélites dinucleotídeos foram predominantes (6.822; 38%), sendo o motivo AG/TC mais abundante (4.674; 68,5%) dentre os SSRs dinucleotídeos. Estes resultados corroboram com estudos anteriores realizados em *H. brasiliensis*, que haviam reportado que os SSRs dinucleotídeos representavam a maioria dos SSRs em sequências expresas (FENG et al., 2009, TRIWITAYAKORN et al., 2011). Os SSRs dinucleotídeos representam a maioria dos SSRs genômicos de diversos organismos (MORGANTE; HANAFEY; POWELL, 2002) incluindo a seringueira (POOTAKHAM et al., 2012), porém a representatividade deste tipo de SSR em sequências expressas de seringueira pode ser explicar o baixo rendimento das bibliotecas enriquecidas em trinucleotídeos apresentadas no capítulo I.

Em espécies arbóreas, estima-se que a frequência média de SNPs no genoma é de um SNP a cada 100 pb (NEALE; KREMER, 2011). Apesar de ser conhecida a abundância deste tipo de polimorfismo, até o momento, poucos marcadores SNPs foram desenvolvidos para seringueira (POOTKHAM et al., 2011, SILVA et al., 2014, SALGADO et al., 2014). No transcriptoma estudado, foram identificados 404.114 SNPs putativos com frequência de um SNP a cada 125 pb. Quando foram associadas às sequências contendo SNPs com anotação do KEGG, foram identificadas sequências anotadas às seis enzimas da via do mevalonato (MVA) (25 contigs) e às sete enzimas da via C-metil-deritritol 4 fosfato (MEP) (40 contigs) contendo SNPs putativos.

Apesar de as vias MVA e MEP serem muito conhecidas e estudadas por estarem diretamente envolvidas na síntese do látex, nenhum SNP havia sido identificado e validado. Para validar o maior número de SNPs nas sequências anotadas nas vias MVA e MEP, foi escolhido o sequenciamento pelo método de Sanger para sequenciar as regiões de interesse

em 36 genótipos de *H. brasiliensis*. Estes 36 genótipos também foram utilizados para caracterização dos locos de SSRs descritos no capítuo I.

Inicialmente foram feitas reações de amplificação via PCR à partir do DNA genômico e os produtos de amplificação foram verificados quanto à qualidade em gel de agarose 1,5%. Todos os pares de *primers* que haviam sido desenhados apresentaram um bom produto de amplificação na agarose e foram sequenciados. Do total de 64 pares de *primers*, apenas 26 apresentaram boa resolução no sequenciamento e avaliados quanto a validação dos SNPs. Foram validados 78 SNPs, porém também foram observados SNPs e regiões de INDELs que não haviam sido preditos por não estarem representados nos clones PR255 e GT1 e, dessa forma, não foram contabilizados neste trabalho.

O sequenciamento pelo método de Sanger das regiões de interesse apresenta como vantagem avaliar regiões de introns e exons e permite visualizar a grande variabilidade genética entre os genótipos de *H. brasiliensis*, porém esta técnica é custosa, apresentou baixo rendimento e dificuldades em sua execução devido à sensibilidade na detecção das bases do sequenciamento. Plataformas de genotipagem de SNPs como Sequenom MassARRAY® (Sequenom, San Diego, CA), Illumina GoldenGate e Illumina Infinium HD (Illumina, San Diego, CA) avaliam apenas o loco SNP e não permitem avaliar regiões com SNPs muito próximos, ou seja, regiões menores que 50 pares de bases que contenham mais de um SNP, porém apresentam boa relação custo-benefício devido ao baixo custo por SNP genotipado e maior rendimento quando comparado ao sequenciamento de Sanger.

Os resultados obtidos neste trabalho permitiram melhor entendimento e dinâmica do transcriptoma de seringueira e teve contribuição significativa para o banco de dados da espécie. Apesar das regiões gênicas sofrerem maior pressão de seleção, o sequenciamento de algumas destas regiões para a validação de SNPs mostrou que existe grande variabilidade genética em regiões gênicas e indica que a frequência predita de um SNP a cada 125 pb pode ser maior.

O capítulo III encerra este trabalho apresentando pela primeira vez na espécie um mapa genético-molecular saturado com marcadores SNPs. Inicialmente foi publicado um mapa prévio e o mapeamento de 18 QTLs para crescimento na revista *Plos One* intitulado

"*QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (Hevea brasiliensis) Evaluated in a Sub-Tropical Climate*" e apresentado no Anexo I. O mapa publicado contém 284 SSRs tendo este projeto de doutorado contribuído com 52 marcadores SSRs desenvolvidos no capítulo I. Este mapa prévio é composto por 23 grupos de ligação, com comprimento total de 2688,8 cM e densidade de um marcador por 9,83 cM. Mais recentemente, este mapa foi saturado por Silva (2014) com 46 marcadores EST-SSRs, totalizando 330 microssatélites distribuídos em 22 grupos de ligação. A densidade média deste mapa após a saturação com os 46 SSRs é de um marcador por 9,3 cM e comprimento total de 3.068,9 cM. A saturação deste mapa com novos marcadores microssatélites permitiu a integração dos grupos de ligação LG11a e LG11b.

No mapa apresentado no capítulo III, 51 marcadores SNPs desenvolvidos a partir de sequências ESTs (SILVA et al., 2014) e do transcriptoma apresentado no capítulo II e dois EST-SSRs desenvolvidos e genotipados por Silva (2014) foram mapeados. A genotipagem dos SNPs foi feita por espectrometria de massa na plataforma Sequenom MassARRAY® e os dados de genotipagem foram analisados no software SuperMASSA (SERANG et al., 2012) que teve o prof. Dr. Augusto Franco Garcia- ESALQ como um dos colaboradores para seu desenvolvimento.

A metodologia empregada para a construção deste mapa foi a mesma publicada no mapa prévio (SOUZA et al., 2013) apresentado em anexo. O mapa contém 383 marcadores distribuídos em 20 grupos de ligação, dos quais 51 são marcadores SNPs e 332 são microssatélites. Este novo mapa tem comprimento total de 3.137,6 cM e densidade de um marcador a cada 8,12 cM. Comparado ao mapa obtido por Silva (2014), a saturação deste mapa com marcadores SNPs permitiu a integração de outros dois grupos de ligação. Apesar de saturar este mapa com marcadores SNPs, ainda não há uma cobertura homogênea do genoma e, por isso, não foi possível obter os 18 grupos de ligação desejados.

Já é conhecido que os SNPs estão presentes de forma abundante e através do estudo apresentado no capítulo II sabe-se que o SNPs são muito frequentes em regiões expressas na seringueira. O desenvolvimento de novos marcadores SNPs é uma ferramenta poderosa na construção e saturação de mapas genéticos e pode contribuir para um mapa mais homogêneo com 18 grupos de ligação e dete ctar QTLs com maior precisão.

Conclusões

Devido, até então, ao número restrito de marcadores moleculares disponíveis para *H. brasiliensis*, fez-se necessário desenvolver novos marcadores para a construção e saturação de um mapa genético molecular. O desenvolvimento desta ferramenta permitiu o maior conhecimento da genética de espécie e avaliar a elevada variabilidade genética dentro de *H. brasiliensis* evidenciada pelo número de alelos encontrados nos locos de microssatélites e pela frequência de SNPs em regiões gênicas.

Por ser uma espécie perene, o processo de domesticação da seringueira é recente e a variabilidade genética encontrada em poucos acessos de *H. brasiliensis* utilizados neste estudo, revela que há um grande potencial genético a ser explorado nos programas de melhoramento.

A porcentagem de transferibilidade dos marcadores SSRs superior a 80% nas outras espécies do gênero *Hevea* e o aumento no número de alelos quando os locos de microssatélites são avaliados incluindo estas espécies, mostra que os bancos de germoplasma devem colecionar acessos destas espécies como fonte de recursos genéticos.

Perspectivas

O conjunto de marcadores SSRs e SNPs desenvolvidos são ferramentas moleculares importantes para programas de melhoramento e podem ser empregados na construção de novos mapas genéticos.

O transcriptoma obtido neste estudo e sua caracterização abre novos horizontes e novas perspectivas no melhoramento genético, pois auxilia a identificação de genes relacionados à características de interesse para o melhoramento. Além disso, este transcriptoma permite a identificação e desenvolvimento de marcadores SNPs em genes de interesse e se tornou referência para futuros estudos de expressão gênica.

O mapa genético apresentado será saturado com novos marcadores SNPs já desenvolvidos buscando obter uma cobertura mais homogênea do genoma. Dados fenotípicos para produção de látex já foram coletados e futuramente um mapa genético mais saturado será construído e, pela primeira vez na espécie, QTLs para produção serão mapeados.

Literatura Citada

ATAN, S.; LOW, F.; SALEH, N. Construction of a microsatellite enriched library from Hevea brasiliensis. **Journal of Natural Rubber Research**, v. 11, p. 247–255, 1996.

BALDWIN, J. J. . Hevea: a first interpretation. A citogenetic survey of a controversial genus, with a discussion of its implications to taxonomy and to rubber production. **J hered**, v. 38, p. 54–64, 1947.

BARBARÁ, T. et al. Cross-species transfer of nuclear microsatellite markers: potential and limitations. **Molecular ecology**, v. 16, n. 18, p. 3759–67, 2007.

BILLOTE, N. et al. Microsatellite-enriched libraries:applied methodology for the development of SSR markers in tropical crops. **Fruits**, v. 54, p. 277–288, 1999.

BOUYCHOU, J. La biologie de l'Hevea. **Rev. Gen. de Caoutchoucs et des Plastiques**, v. 40, p. 933–1001, 1963.

CARVALHO, M.; SILVA, D. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas; Next generation DNA sequencing and its applications in plant genomics. **Ciênc. rural**, p. 735–744, 2010.

CHAGNÉ, D. et al. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. **PloS one**, v. 7, n. 2, p. e31745, 2012.

CHOW, K.-S. et al. Insights into rubber biosynthesis from transcriptome analysis of Hevea brasiliensis latex. **Journal of experimental botany**, v. 58, n. 10, p. 2429–40, 2007.

CORNISH, K. Similarities and differences in rubber biochemistry among plant species. **Phytochemistry**, v. 57, n. 7, p. 1123–34, 2001.

EGAN, A. N.; SCHLUETER, J.; SPOONER, D. M. Applications of next-generation sequencing in plant biology. **American journal of botany**, v. 99, n. 2, p. 175–85, 2012.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics.** 4th. ed. Harlow, Essex, UK: Longmans Green, 1996. p. 464

FENG, S. P. et al. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (Hevea brasiliensis). **Molecular Breeding**, v. 23, n. 1, p. 85–97, 2008.

FERREIRA, M.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética.** 3rd. ed. [s.l.] Brasilia: Embrapa, 1998. p. 220

GANAL, M. W.; ALTMANN, T.; RÖDER, M. S. SNP identification in crop plants. **Current opinion in plant biology**, v. 12, n. 2, p. 211–7, 2009.

GARDNER, E. J.; SNUSTAD, D. P. **Genética**. 7° ed. ed. Rio de Janeiro: Guanabara Kogan, 1986.

GONÇALVES, P. et al. Desempenho de novos clones de seringueira. III. Seleções promissoras para a região de Votuporanga, Estado de São Paulo. **Bragantia**, v. 34, n. 6, p. 971–980, 1999.

GONÇALVES, P. DE S., AND J. M. **Melhoramento genético da seringueira: passado, presente e futuro**. Viçosa: EPAMIG, 2008. p. 401–407

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature biotechnology**, v. 29, n. 7, p. 644–52, 2011.

GRATTAPAGLIA, D.; SEDEROFF, R.; CAROLINA, N. Genetic Linkage Maps. v. 1137, p. 1121–1137, 1994.

GRIFFITHS, A. et al. **Introdução à genética**. Guanabara ed. Rio de Janeiro: 1998. p. 856

HALDANE, J. The combination of linkage values, and the calculation of distance between the loci of linked factors. **Journal of Genetics**, v. 8, p. 299–309, 1919.

HAMILTON, J. P.; BUELL, C. R. Advances in plant genome sequencing. **The Plant journal : for cell and molecular biology**, v. 70, n. 1, p. 177–90, 2012.

IRSG - Internation Rubber Study Group. Rubber Stastistical Bulletin. Wembley, v 68, n 7-9, 2014

KANEHISA, M. et al. KEGG for linking genomes to life and the environment. **Nucleic acids research**, v. 36, n. Database issue, p. D480–4, 2008.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, v. 28, n. 1, p. 27–30, 2000.

KOSAMBI, D. The estimation of map distance from recombination values. **Annuaire of Eugenetics.**, v. 12, p. 172–175, 1944.

KWOK PY, G. Z. Single nucleotide polymorphism libraries: why and how are we building them? **Molecular Medicine Today**, v. 5, n. 12, p. 538–543, 1999.

LANGMEAD B, TRAPNELL C, POP M, S. S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome biology**, v. 10, p. R25, 2009.

LE GUEN, V. et al. Genetic structure of Amazonian populations of Hevea brasiliensis is shaped by hydrographical network and isolation by distance. **Tree Genetics & Genomes**, v. 5, n. 4, p. 673–683, 2009.

LE GUEN, V. et al. A rubber tree's durable resistance to Microcyclus ulei is conferred by a qualitative gene and a major quantitative resistance factor. **Tree Genetics & Genomes**, v. 7, n. 5, p. 877–889, 2011a.

LE GUEN, V. et al. Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (Hevea brasiliensis). **Plant Breeding**, v. 130, n. 2, p. 294–296, 2011b.

LEITCH, A. R. et al. Molecular cytogenetic studies in rubber, Hevea brasiliensis Muell. Arg. (Euphorbiaceae). **Genome**, v. 41, n. 3, p. 464–467, 1998.

LESPINASSE, D. et al. A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. **Theoretical and Applied Genetics**, v. 100, n. 1, p. 127–138, 2000a.

LESPINASSE, D. et al. Identification of QTLs involved in the resistance to South American leaf blight (Microcyclus ulei) in the rubber tree. **TAG Theoretical and Applied Genetics**, v. 100, n. 6, p. 975–984, 2000b.

LEVINSON, G.; GUTMAN, G. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. **Nucleic acids research**, v. 15, p. 5323–5338, 1987.

LI, D. et al. De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (Hevea brasiliensis Muell. Arg.). **BMC genomics**, v. 13, p. 192, 2012.

LIJAVETZKY, D. et al. High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. **BMC genomics**, v. 8, p. 424, 2007.

LIU, B. H. **Stastistical Genomics: Linkage, Mapping and QTL Analysis.** Boca Raton and New York City: CRC Press, 1998. p. 611

LOW, F. et al. Recent advances in the development of molecular markers for Hevea studies. **Journal of Natural Rubber Research**, v. 11, p. 32–44, 1996.

LYNCH, M.; WALSH, B. **Genetics and Analysis of Quantitative Traits**. Sunderland, Massachusetts: Sinauers Associates, 1998. p. 980

MANTELLO, C. C. et al. Microsatellite marker development for the rubber tree (Hevea brasiliensis): characterization and cross-amplification in wild Hevea species. **BMC research notes**, v. 5, p. 329, 2012.

MARTIN, J. A; WANG, Z. Next-generation transcriptome assembly. **Nature reviews. Genetics**, v. 12, n. 10, p. 671–82, 2011.

MORGANTE, M.; HANAFEY, M.; POWELL, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. **Nature genetics**, v. 30, n. 2, p. 194–200, 2002.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. **Genomics**, v. 92, n. 5, p. 255–64, 2008.

NEALE, D. B.; KREMER, A. Forest tree genomics: growing resources and applications. **Nature reviews. Genetics**, v. 12, n. 2, p. 111, 2011.

NOVAES, E. et al. High-throughput gene and SNP discovery in Eucalyptus grandis , an uncharacterized genome. v. 14, p. 1–14, 2008.

ONG, S. Chromosome morphology at the pachytene stage in Hevea brasiliensis. **IRRBB, Proc Int Rubber Conf**, v. 2, p. 3–12, 1975.

PAIVA, J. **Variabilidade enzimática em populações naturais de seringueira**. [s.l.] ESALQ, 1992.

PIRES, J. M. **Revisão do gênero Hevea: descrição das espécies e distribuição geográfica.IPEAN**. Belém: [s.n.].

PIRES, J. M.; SECCO, R. DE S.; GOMES, J. I. **Taxonomia e Fitogeografia das seringueira**BelémEmbrapa-Amazonia Oriental, , 2002.

POLHAMUS, L. **Rubber: Botany, Production and Utilization**. World Crop ed. London: Leonard Hill Limited, 1962. p. 449

POOTAKHAM, W. et al. Single nucleotide polymorphism marker development in the rubber tree, Hevea brasiliensis (Euphorbiaceae). **American journal of botany**, v. 98, n. 11, p. e337–8, 2011.

POOTAKHAM, W. et al. Development of genomic-derived simple sequence repeat markers in Hevea brasiliensis from 454 genome shotgun sequences. **Plant Breeding**, v. 131, n. 4, p. 555–562, 2012.

PRIYADARSHAN, P. M.; GONÇALVES, P. D. S. Hevea gene pool for breeding. p. 101–114, 2003.

PUSHPARAJAH, E. Problems and potentials for establishing Hevea under difficult environmental conditions. **Planter**, v. 59, p. 242–251, 1983.

PUSHPARAJAH, E. **Natural rubber. In: Last, F.T. (ed) Tree Crop Ecosystems**. Amsterdam, The Netherlands: Elsevier Science, 2001. p. 379–407

RAHMAN, A. Y. A. et al. Draft genome sequence of the rubber tree Hevea brasiliensis. **BMC genomics**, v. 14, n. January, p. 75, jan. 2013.

RAJ, S. et al. Relationship between latex yield of Hevea brasiliensis and antecedent environmental parameters. **International journal of biometeorology**, v. 49, n. 3, p. 189–96, jan. 2005.

ROY, C.; NAZEER, M.; SAHA, T. Identification of simple sequence repeats in rubber (Hevea brasiliensis). **CURRENT SCIENCE-BANGALORE**, v. 87, p. 807–810, 2004.

SAHA, T.; PRIYADARSHAN, P. M. **Genomics of Tree Crops**. New York, NY: Springer New York, 2012.

SAKDAPIPANICH, J. T. Structural characterization of natural rubber based on recent evidence from selective enzymatic treatments. **Journal of bioscience and bioengineering**, v. 103, n. 4, p. 287–92, 2007.

SALGADO, L. R. et al. De novo transcriptome analysis of Hevea brasiliensis tissues by RNA-seq and screening for molecular markers. **BMC genomics**, v. 15, n. 1, p. 236, 2014.

SCHULTES, R. E. The History of Taxonomic Studies In Hevea. **The Botanical review**, v. 36, n. 3, p. 197–274, 1970.

SCHULTES, R. E. Wild Hevea: An untapped source of germplasm. **Journal Rubber Research**, v. 54, p. 227–257, 1977.

SCHULZ, M. H. et al. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics (Oxford, England)**, v. 28, n. 8, p. 1086–92, 2012.

SEIBERT, R. J. A study of Hevea (with its economic aspects) in the Republic of Peru. **Annals of the Missouri Botanical Garden**, p. 261–352, 1947.

SERANG, O.; MOLLINARI, M.; GARCIA, A. A. F. Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. **PloS one**, v. 7, n. 2, p. e30906, 2012.

SHI, C. et al. Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. **BMC Genomics**, v. 12, p. 131, 2011.

SIBOV, S. et al. Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 1. Map construction and localization of loci showing distorted segregation. **Hereditas**, v. 139, p. 96–106, 2003a.

SIBOV, S. et al. Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 2. Quantitative trait loci (QTL) for grain yield, plant heigth, ear height and grain moisture. **Hereditas**, v. 139, p. 107–115, 2003b.

SILVA, C. C. et al. Leaf-, panel- and latex-expressed sequenced tags from the rubber tree (Hevea brasiliensis) under cold-stressed and suboptimal growing conditions: the development of gene-targeted functional markers for stress response. **Molecular Breeding**, 2014.

SILVA, C. C. **Construção de uma para funcional em seringueira (Hevea brasiliensis)**. [s.l.] Universidade Estadual de Campinas, 2014.

SOUZA, L. M. et al. QTL mapping of growth-related traits in a full-sib family of rubber tree (Hevea brasiliensis) evaluated in a sub-tropical climate. **PloS one**, v. 8, n. 4, p. e61238, 2013.

SOUZA, LM, MANTELLO, CC, SANTOS, MO, GONCALVES PS, SOUZA, A. Microsatellites from rubber tree (Hevea brasiliensis) for genetic diversity analysis and cross-amplification in six Hevea wild species. **Conservation Genetics Resources**, v. 1, n. 1, p. 75–79, 2009.

STURTEVANT, A. H. The linear arrangement of six sex-linked factor in Drosophila, as shown by their mode of association. **Journal of Experimental Zoology**, v. 14, p. 43–59, 1913.

TRIWITAYAKORN, K. et al. Transcriptome sequencing of Hevea brasiliensis for development of microsatellite markers and construction of a genetic linkage map. **DNA research : an international journal for rapid publication of reports on genes and genomes**, v. 18, n. 6, p. 471–82, 2011.

VARSHNEY, R. K.; GRANER, A.; SORRELLS, M. E. Genic microsatellite markers in plants: features and applications. **Trends in biotechnology**, v. 23, n. 1, p. 48–55, 2005.

WANG, D. G. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. **Science**, v. 280, n. 5366, p. 1077–1082, 1998.

WANG, Z. et al. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (Ipomoea batatas). **BMC genomics**, v. 11, n. 1, p. 726, 2010.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature reviews. Genetics**, v. 10, n. 1, p. 57–63, 2009.

WARD, J. A; PONNALA, L.; WEBER, C. A. Strategies for transcriptome analysis in nonmodel plants. **American journal of botany**, v. 99, n. 2, p. 267–76, 2012.

WEI, W. et al. Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. **BMC genomics**, v. 12, n. 1, p. 451, 2011.

WEIR, B. **Genetic data analysis**. 2nd. ed. Sunderland: Sinauer Associates, 1996. p. 447

WU, R. et al. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. **Theoretical population biology**, v. 61, n. 3, p. 349–63, 2002.

XIA, Z. et al. RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis. **Plant molecular biology**, v. 77, n. 3, p. 299–308, 2011.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome research**, v. 18, n. 5, p. 821–9, 2008.

Anexo I

OPEN ACCESS Freely available online

PLOS ONE

QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (*Hevea brasiliensis*) Evaluated in a Sub-Tropical Climate

Livia Moura Souza[1☯], Rodrigo Gazaffi[2☯], Camila Campos Mantello[1], Carla Cristina Silva[1], Dominique Garcia[3], Vincent Le Guen[3], Saulo Emilio Almeida Cardoso[4], Antonio Augusto Franco Garcia[2], Anete Pereira Souza[1,5]*

1 Molecular Biology Center and Genetic Engineering, UNICAMP, Campinas, SP, Brazil, **2** Department of Genetics, ESALQ - University of São Paulo, Piracicaba, SP, Brazil, **3** CIRAD, UMR AGAP, Montpellier, Hérault, France, **4** Laboratório de Pesquisa & Desenvolvimento Plantações Michelin da Bahia, Plantações Michelin da Bahia LTDA, Igrapiuna, BA, Brasil, **5** Department of Plant Biology – Biology Institute and Molecular Biology Center and Genetic Engineering, UNICAMP, Campinas, SP, Brazil

Abstract

The rubber tree (*Hevea* spp.), cultivated in equatorial and tropical countries, is the primary plant used in natural rubber production. Due to genetic and physiological constraints, inbred lines of this species are not available. Therefore, alternative approaches are required for the characterization of this species, such as the genetic mapping of full-sib crosses derived from outbred parents. In the present study, an integrated genetic map was obtained for a full-sib cross family with simple sequence repeats (SSRs) and expressed sequence tag (EST-SSR) markers, which can display different segregation patterns. To study the genetic architecture of the traits related to growth in two different conditions (winter and summer), quantitative trait loci (QTL) mapping was also performed using the integrated map. Traits evaluated were height and girth growth, and the statistical model was based in an extension of composite interval mapping. The obtained molecular genetic map has 284 markers distributed among 23 linkage groups with a total length of 2688.8 cM. A total of 18 QTLs for growth traits during the summer and winter seasons were detected. A comparison between the different seasons was also conducted. For height, QTLs detected during the summer season were different from the ones detected during winter season. This type of difference was also observed for girth. Integrated maps are important for genetics studies in outbred species because they represent more accurately the polymorphisms observed in the genitors. QTL mapping revealed several interesting findings, such as a dominance effect and unique segregation patterns that each QTL could exhibit, which were independent of the flanking markers. The QTLs identified in this study, especially those related to phenotypic variation associated with winter could help studies of marker-assisted selection that are particularly important when the objective of a breeding program is to obtain phenotypes that are adapted to sub-optimal regions.

Citation: Souza LM, Gazaffi R, Mantello CC, Silva CC, Garcia D, et al. (2013) QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (*Hevea brasiliensis*) Evaluated in a Sub-Tropical Climate. PLoS ONE 8(4): e61238. doi:10.1371/journal.pone.0061238

Editor: Rongling Wu, Pennsylvania State University, United States of America

Received December 27, 2012; **Accepted** March 6, 2013; **Published** April 19, 2013

Funding: The authors would like to acknowledge funding from the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) for financial support (2007/50562-4) and scholarships to LMS (2009/14068-0), CCM (2008/55974-1), CCS (2009/52975-0) and RG (2009/14068-0). The authors are also thankful to the Technological Development (CNPq) for support to APS and AAFG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Dr. Saulo Emilio Almeida Cardoso, is a researcher from Edouard Michelin Plantation in Brazil. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: anete@unicamp.br

☯ These authors contributed equally to this work.

Introduction

Hevea brasiliensis (Wild. Ex. Adr. de. Juss. Muell. Arg.) is a commercial rubber-producing species that belongs to the *Euphorbiaceae* family. The genus *Hevea* includes 11 species that are native to the Amazon region [1]. Although the growth and rubber yields of *H. brasiliensis* are optimal under hot and humid climate conditions, which prevail in its native region, it is often cultivated in dryer and colder areas throughout the world. In certain native regions, particularly South America, the devastating South American Leaf Blight (SALB) caused by the fungus *Microcyclus ulei* has become an important threat to rubber tree cultivation.

As a result of SALB, rubber production expanded during the late 1970s into worldwide, such as northeast India, the highlands and coastal areas of Vietnam, southern China and the southern plateau of Brazil [2]. All of these areas, however, exhibit various climate constraints. Although these areas satisfy most of the basic growth requirements for rubber production, they also present stressful conditions, such as low temperatures and dry periods [3]. In addition to the damage and the low growth rate caused by the cold temperatures, the production of latex is also halted for approximately 1–3 months of each year in these areas [4,5].

Historically, the breeding of rubber trees has been based on techniques involving statistics and quantitative genetic approaches to determine the best genotypes to be used as new cultivars. The discovery of molecular genetic markers has provided new possibilities for characterizing genotypes for the purpose of

identifying cultivars, analyzing genetic diversity, establishing relationships between agricultural traits and genetic factors (QTLs), and identifying genes of interest [6–16]. Although the genetic basis of disease resistance has been studied, less is known about the genetic basis of cold tolerance and the characteristics related to growth and plant development in the rubber tree.

Genetic maps are highly useful tools in breeding programs for many species. For perennial trees, such as *Hevea* spp., however, it is difficult to obtain homozygous inbred lines due to their extended juvenile period and inbreeding depression. Therefore, the traditional methods used to construct linkage maps for F_2 and backcross populations cannot be used for these species [17].

Traditionally, the construction of genetic maps from a full-sib cross (or F_1 populations) involved the double "pseudo-testcross" strategy, in which only markers that segregate in a 1:1 ratio are considered. In the case of the rubber tree, the first genetic map was generated using a double "pseudo-testcross" strategy [9] and was obtained using several markers. This map was an important step toward obtaining a saturated genetic linkage for the rubber tree.

Considering modern markers technologies available for full sib populations, markers that segregate in 3:1 (dominant), 1:2:1 (codominant), and 1:1:1:1 (codominant) ratios, in addition to 1:1, can be used to integrate individual linkage maps. Wu et al. [18] showed that dominant markers provide less information in linkage analyses than do codominant markers, and loci that segregate in a 1:1:1:1 or 1:2:1 ratio are highly informative. Based on these findings, Wu et al. [18] proposed an alternative strategy that uses maximum likelihood methods that simultaneously estimate linkage and linkage phases. This approach circumvents several of the disadvantages emphasized by Maliepaard et al. [19] when estimating linkage phases in a separated step. This method was successfully used in sugarcane [20,21] and a yellow passion fruit population [22].

OneMap software [23] was initially developed to facilitate linkage analyses in outcrossing species using the methodology proposed by Wu et al. [18]. Later, the software has also been updated to perform multipoint analyses based on the Hidden Markov Models [24]. This methodology allows for the analysis of a mixed set of different marker types that exhibit various segregation patterns, resulting in a genetic map that more accurately represents the polymorphisms of the cross.

Although many statistical methods have been specifically developed to perform QTL mapping in outcrossing species, the general double pseudo-testcross method is the most frequently used approach to studying the QTLs in rubber tree species [9,25,26].

Gazaffi et al. [27] proposed a method in which QTL mapping is performed in an integrated genetic map using Composite Interval Mapping (CIM). This method is based on a combination of different models and incorporate conditional QTL multipoint probabilities, which are more accurate and provide better results. Using this method, it is possible to infer the segregation of QTLs as well as the linkage phase between putative QTL and markers in the integrated genetic map.

In this study, we present an integrated rubber tree map saturated with simple sequence repeats (SSRs) and expressed sequence tags (EST-SSRs). We also conducted QTL mapping using CIM to determine the growth rate during the summer and winter seasons and evaluate growth characteristics (height and girth) of *H. brasiliensis*.

Materials and Methods

Mapping Population

The mapping population has 270 individuals derived from a cross between genotypes PB217 and PR255. The former has a high yield potential, which is expressed throughout the lifetime of its rubber production, and exhibits a low metabolic activity and a high level of sucrose in its latex vessels and the latter shows tolerance to injury and cold. The progeny were obtained through the controlled pollination of the two parental genotypes, and multiple replicates were generated by bud grafting onto rootstock (seedlings of GT1 genotype) for the field experiment.

The mapping population was planted in the Edouard Michelin Plantation (Mato Grosso state, Brazil, 17° 23′ 59.60″ S and 54° 44′ 53.93″ W, altitude 519 m) that is characterized by a sub-tropical climate with a cold and dry period that lasts four months from May to October (Figure S3). Planting was performed in 2006 using a randomized complete block design with four replicates. Each block consisted of 272 elementary plots: 1 plot for each of the 270 progeny individuals, 1 for each of the 2 parents. Each elementary plot consisted of four grafted plants with the same genotype. The statistical analysis considered data from the average of each elementary plot, i.e., average of four plants. Row and column number of each data point were annotated and included in the statistical model of the experiment that used 7.2 hectares, 40 rows and 28 columns. The field management procedures were the same as those procedures used for commercial purposes.

DNA Extraction and Molecular Marker Analysis

Genomic DNA was extracted from 300 mg of lyophilized leaf tissue using the cetyl trimethyl ammonium bromide (CTAB) method, as described by Murray & Thompson [28]. The DNA concentrations were estimated using electrophoresis on ethidium bromide-stained agarose gels with appropriate molecular weight standards.

The polymorphisms from a total of 603 microsatellite markers from different sources were screened for both the PB217 and PR255 parents. From the total number of microsatellite markers, 425 SSR were from an enriched genomic library [13,14,16] and 178 were sequence tag-derived simple sequence repeat (EST-SSR) markers [29].

Two alternative techniques were used to visualize the polymerase chain reaction (PCR) products; both of these techniques required separation by electrophoresis on denaturing acrylamide gels. For visualization by silver staining, PCR was performed as described by Souza et al. [13]. For visualization by fluorescence using a LI-COR 4300 DNA analyzer, oligonucleotides with an M13-19 base extension were added to the forward primers. The PCR amplification was performed according to the methodology described by Le Guen et al. [14].

Linkage Map

Segregation of markers was tested using the common procedure based on the Chi-square test, correcting for multiple tests. Linkage analysis was performed using OneMap software [23] version 2.0-1 using multipoint technology based on Hidden Markov Models [18]. Initially, cosegregation groups were established using an LOD score of 4.5 and a recombination fraction of 0.4. The order of the markers was obtained with algorithms *compare* (for groups with up to six markers) and *order* (for groups with more than six markers) [30,31]. To determine the map, recombination fractions between markers were converted to centiMorgan units using the Kosambi map function [32].

Field Data Analysis

Girth (circumference at 1 m) and height were measured in centimeters (cm) to evaluate the growth of individual rubber trees, and means were calculated based on four plants per plot. These growth characteristics were measured in April and October from 2007 to 2009, when the plants had 18, 24, 30, 36 and 42 months for each plant per plot and the average was calculated for each plot. The traits were defined as the difference in growth between October to April (summer) and April to October (winter) over two years. The values for the summer period were calculated by the subtraction of the measures obtained for 24 minus18 months (first year) and 36 minus 30 months (second year). The growth ratios for first and second year were added to obtain plant development during the summer season. Winter values were obtained by subtracting from 30 to 24 months and 42 to 36 months, and adding these two values was obtained plant development during the winter season. The total development was obtained from the subtraction measures evaluation for 42 and18 months. Six traits were considered: Total Height (TH), the height growth in the summer (SH) and winter (WH), Total Girth (TG), and the girth growth in the Summer (SG) and Winter (WG).

The statistical model used for the analysis was:

$$y_{ijuv} = \mu + g_i + b_j + r_{uj} + c_{vj} + e_{ijuv}$$

where y_{ijuv} is the growth characteristic measured for one season over two years for the i^{th} genotype in the j^{th} block, the u^{th} row and the v^{th} column; μ is the intercept; g_i is the random effect of the i^{th} genotype $\sim N(0, V_g)$; b_j is the fixed effect of the j^{th} block; r_{uj} is the random effect of the u^{th} row within the j^{th} block $\sim N(0, V_{row})$; c_{vj} is the random effect of the v^{th} column within the j^{th} block $\sim N(0, V_{col})$. Here, the usage of rows and columns (indicating the elementary plots) as random effects was an attempt to control local variation, as similarly done by Boer et al. [33] and Thumma et al. [34]; e_{ij} is the error term $\sim N(0, V)$, assuming heteroscedasticity between blocks and no correlation among plots.

Analyses were performed using the Genstat Software [35], and the predictions of genotypic values (BLUPs) were used to perform the QTL mapping. The heritability for each trait was estimated based on the variance components from the phenotypic analysis.

The genetic correlations (r_g) between each pair of traits were obtained using the Pearson's correlations coefficient applied on the individual genotypic values. These correlations were tested assuming global significance level of 0.05. The analyses were done using R software (www.r-project.org).

QTL Mapping

QTL mapping was performed using the integrated genetic map according to the approach described by Gazaffi [27]. The two traits (girth growth and height growth) were analyzed separately for each season. Briefly, the applied methodology extends composite interval mapping (CIM), as presented by Zeng [36], for an outbred scenario. The model has three genetic effects: two for additive effects (one for each parent), and one for dominance. To infer the conditional probabilities of QTL genotypes, given marker genotypes, multipoint probabilities were obtained using Hidden Markov Models.

The mapping strategy is based on two steps. First, the genome scan is performed to detect QTLs (every 1 cM in this study). Second, this mapped region is fully characterized, i.e., the significant effects are identified, along with the linkage phase between markers and QTLs, and finally the QTL segregation pattern was inferred. In diploids, a QTL can exhibit a segregation ratio of 1:1, 1:2:1, 3:1 or 1:1:1:1. Details regarding this methodology are described by Gazaffi [27]. To perform the CIM analysis, cofactors were included in the model to control the QTLs located outside the mapping interval. For this purpose, multiple linear regression analyses using stepwise selection based on the Akaike Information Criterion were considered; a maximum of 10 cofactors was used to avoid over parameterization of the model. The threshold value for considering evidence was obtained using a 0.95 significance level and 1,000 permutations [37] according to the modification proposed by Chen and Storey [38]; i.e., the highest LOD Score peak for each linkage group was obtained, and the distribution of the second highest peak was assumed.

Ethics Statement

We confirm that no specific permits were required for the described field studies. This work was a collaborative research developed by researchers from Edouard Michelin Plantation (Brazil), CIRAD (France), USP (Brazil) and UNICAMP (Brazil). Dr. Saulo Emilio Almeida Cardoso, researcher from Edouard Michelin Plantation in Brazil (where the field experiments were developed) is one of the coauthors of this work, and the Edouard Michelin Plantation is listed in the author addresses. Dr. Cardoso assisted in the field studies. Also, we confirm that this manuscript is a result of a basic research project, developed mainly at the university and mostly funded by public funding agencies, whose aim is to develop new knowledge and, where the generated results should be shared with the scientif and technological community through open university thesis and manuscripts in conventional scientific journals. In this sense, we would like to state that we fully adhere to all the PLOS ONE policies on sharing data and materials.

We confirm that the field studies did not involve endangered or protected species.

Results and Discussion

Marker Polymorphism and Segregation Analyses

Among all of the evaluated microsatellite markers, 51% (308) showed a polymorphism between the parents of the mapping population. This level of polymorphism observed in the studied population was expected because a similar level of polymorphism was previously reported in a study by Lespinasse et al. [9]. Although SSRs derived from ESTs have been shown to be less polymorphic (39%) than SSRs derived from genomic sequences (54.5%), they are present in gene-rich regions of the genome and are frequently detected at a relatively high abundance [39]. A total of 284 polymorphic markers were selected for genotyping of the 270 offspring within the mapping population.

A chi-squared test performed on the genotyped polymorphic loci revealed that 138 (48.6%) loci exhibited a segregation ratio of 1:1, 31 (10.9%) loci exhibited a ratio of 1:2:1, and 115 (40.5%) loci exhibited a ratio of 1:1:1:1. This is a very favorable scenario for building integrated linkage maps. Of the evaluated markers, 3% (9) showed distorted segregation ($P \leq 0.005$, used as a conservative value for the chi-squared test). The linkage analysis revealed that the markers exhibiting distorted segregation were distributed throughout the genome and were not cause distortion in the LGs (linkage groups). The low number of genomic regions showing a skewed segregation in this map is similar to the number reported for several crops, such as melon [40], but differs from the number reported for other crops, such as lettuce [41]. The deviations from the Mendelian segregation ratios detected in this study may be the result of various processes, including gametophyte selection for

sub-lethal genes, i.e., genes controlling the viability of the pollen, zygotes or seedlings that are frequently located within one or more of these linkage groups [42].

Linkage Map

An integrated genetic map was obtained using microsatellite markers. The map contains markers segregating as 1:1:1:1, 1:2:1 and 1:1 fashion. To label the linkage groups, previous maps were used as reference. The estimate map contained 284 markers assigned to 23 linkage groups and spanned a total of 2688.8 cM; the previous map developed for the rubber tree was considerably smaller [9], more informative genetic linkage was determined because it was possible to use markers with segregation ratios of 1:1:1:1, 1:2:1 and 1:1 in order to estimate an integrated map, better results were achieved. Le Guen et al. [26] and Triwitayakorn et al. [43] developed a map for the rubber tree using markers with segregation in both parents, although this map was analyzed as a double pseudo-testcross.

The length of each group ranged from 2.7 cM (LG2b) to 228.7 cM (LG10) (Figure 1). The distribution of the markers in each chromosome is shown in Table S1. In Figure 1, the distorted markers are indicated with an asterisk ($P \leq 0.000172414$), and the origin of each specific marker is indicated by the marker name prefix or is colored. For example, markers with the prefix HBE were identified by Feng et al. [29], markers with the prefix HV were identified by Gouvêa et al. [44], markers with the prefix HB were identified by Souza et al. [13] and Mantello et al. [16] and markers colored red were identified by Le Guen et al. [14].

The linkage groups were organized according to the numbers obtained from the map previously developed by Lespinasse et al. [9], and information from other maps (unpublished) was used to identify syntenic markers. Based on this information, the 23 groups were classified into groups numbered from 1 to 19, and groups 2, 6, 11 and 17 were further divided into subgroups "a" and "b". The density of the markers in our map is lower (one marker for every 10 cM) than the density in the map developed by Lespinasse et al. [9] (one marker per 3 cM), although this previous map was also saturated with many AFLP markers. Additionally, the marker distribution among the linkage groups (LGs) was not uniform, which resulted in several gaps (LG 11a, 12, 13, 15 and 16). The largest gap, 66.3 cM, was observed in LG 12 (Figure 1).

One possible explanation for the observed gaps, especially for those groups that could not be linked together, is that there are fewer SSR and EST-SSR polymorphisms in these regions. For example, the published map indicates that clusters of markers for LG 6, primarily the AFLP markers, are found in certain linkage group regions; however, it is not possible to join LG 6a and 6b. AFLP clustering has been frequently reported in saturated maps for melons [40], lettuce [41] and potatoes [45] and is usually associated with the heterochromatic regions near the centromeres. Although the regions showing AFLP clustering may be indicative of the centromere positions, comprehensive cytogenetic analyses are necessary to demonstrate this association in the rubber tree.

Another possible explanation for the gaps observed in the map is that either the recombination events or mapped loci were not evenly distributed throughout the genome. The low density of markers for certain linkage groups may correspond to highly homozygous regions that have a lower recombination frequency [46].

The expected number of eighteen linkage groups for the linkage map of the rubber tree ($2n = 36$) was exceeded by five linkage groups, all of which showed a low number of markers per linkage group. The small size of some of the linkage groups indicates that the additional observed linkage groups may be the result of an incomplete coverage of the genome by the marker loci. A comparison of our results with the maps published by Lespinasse et al. [9] and Le Guen et al. [26], which used several markers, indicates that the identification of 18 linkage groups is possible. Furthermore, using only SSR and EST-SSR markers, Triwitayakorn et al. [43] published a genetic linkage map in which the number of linkage groups exceeded the haploid number. These findings are consistent with the data in this study, in which the identification of 23 linkage groups suggests that different markers are required to fill the gaps between adjacent markers.

Phenotypic Trait Analysis

The field data analysis indicated that the experiments were well conducted; this reliability is especially evident in the coefficient of variation values that ranged from 13.24 to 27.05% (Table 1). These values are relatively small considering that the experiment was performed using a perennial species and a large experimental area (7.2 ha).

The phenotypic values of the parents (PB217 and PR255) as well as the progeny phenotypic segregation for all of the traits are shown in Figure S1. With the exception of the Summer Height, the phenotypic values for all of the studied traits differed between the parents. These values were higher in PR255 than in PB217 (Table 1). This difference can be attributed to the better adaptation of PR255 to cold weather, such as the winter season with suboptimal conditions for rubber tree growth, used during this study. Transgressive segregation was also observed for all of the traits evaluated. The differences in the phenotypic values analyzed for the progeny was greater than the difference between the phenotypic values of the parents (Table 1). For example, for the Summer Girth, Winter Girth and Total Girth, the progeny differences were 4.3, 3.9 and 3.4 times, respectively, greater than the phenotypic differences of the parents. A high parent heterosis is especially apparent for Height, as indicated by the mapping population distribution that is skewed to the right, showing that the average is higher than the phenotypic values of the parents.

The statistical analysis showed that the heritability varied from 0.13 to 0.32 (Table 1). A comparison between seasons indicated that the estimates for summer (Height, 0.13 and Girth, 0.27) are lower than the estimates for winter (Height of 0.20 and Girth of 0.32). This seasonal variation occurred because the environmental conditions had a greater influence on the phenotypes during the summer season. The genetic variability for Height and Girth was larger during the summer (110.7 and 65.7, respectively) than during the winter (76.9 and 13.19, respectively). This difference occurred because the environmental stress of the winter season reduced the genetic potential for growth in the genotypes; hence, the winter environment does not present an optimal climate for rubber tree growth. A heritability comparison between the traits also indicated that the values obtained for Height ($0.10 < h < 0.20$) were lower than the values obtained for Girth ($0.27 < h < 0.32$). Although similar values were reported by [47], our estimates were generally lower than previously reported values, especially with respect to plant height. Findings from a study on a eucalyptus species support the widely accepted concept that aspects of growth are under polygenic control due to the relatively low heritability of these traits [48].

In general, high correlations (Figure S2) were found when comparing the Summer Girth to the Total Girth and the Summer Height to the Total Height (0.95 and 0.75, respectively). These correlations indicate that a substantial amount of growth occurs during the summer, what was expected in our experimental

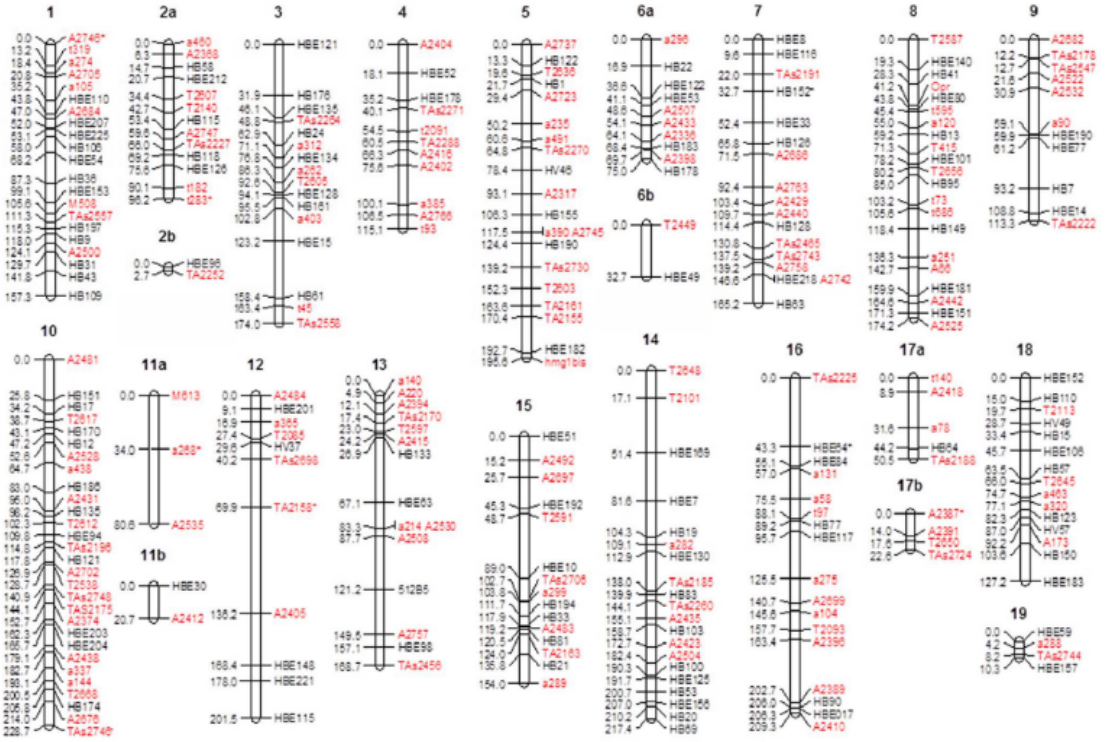


Figure 1. The genetic linkage map of the rubber tree. Note that markers with distorted segregation is indicated with *, the origin of the marker is indicated as follows: the HBE prefix indicates that the marker was identified by Feng et al. [28], the HV prefix indicates that the marker was identified by Gouvea et al. [41], the HB prefix indicates that the marker was identified by Souza et al. [13] and Mantello et al. [16], and the markers in red were identified by Le Guen et al. [14].
doi:10.1371/journal.pone.0061238.g001

conditions. The growth during the winter, however, also contributes to the total Height and Girth, as indicated by the moderate observed correlations (0.45 and 0.69, respectively). The correlations observed between the seasons for Girth (0.43) and Height (−0.22) indicate that the genotypes that performed better during the summer do not necessarily perform as well during the winter. This finding illustrates the difficulties to do rubber trees breeding in the non-traditional regions of Brazil, also referred to as escape areas. The success of a genotype during a period in which there is no environmental stress can differ from the success of that same genotype when it is exposed to a situation in which there is abiotic stress. It should also be noted that moderate weak correlations were observed between Height and Girth during the summer (0.23) and winter (0.18) seasons. These results illustrate that higher correlations were obtained when traits were compared between the seasons and indicate that the environmental influence is an important component to be considered when studying the growth traits of the rubber tree.

Table 1. Mean values for girth and height growth during the summer and winter seasons, measured over a two-year period.

Traits	PB217[(1)]	PR255[(1)]	F_1 (min – max)[(1)]	σ^2g[(2)]	σ^2f[(2)]	$h^{2(2)}$	CV[(2)]
Summer Height (SH)	138.3	139.1	145.48 (124.4–164.2)	110.7 (27.0)	800.2	0.13	17.0
Winter Height (WH)	47.9	55.7	60,9 (41.45–92.57)	76.9 (14.1)	381.6	0.20	27,0
Total Height (TH)	191.7	197.7	205.9 (187.8–227.1)	116.7 (33.2)	1061.9	0.10	13.7
Summer Girth (SG)	75.4	82.4	84.5 (64.31–101.80)	65.7 (9.9)	238.9	0.27	14.1
Winter Girth (WG)	14.3	18.6	19,1 (12.4–31.7)	13.19 (1.8)	41.1	0.32	25.9
Total Girth (TG)	89.6	100.6	103.7 (78.6–127.5)	112.5 (15.2)	344.1	0.32	13.2

(1)Mean values in cm.
(2)Genotypic (σ^2g) and phenotypic (σ^2f) variances, heritability (h^2) and coefficient of variation (CV). Values in parenthesis are the confidence intervals.
doi:10.1371/journal.pone.0061238.t001

QTL Mapping

QTL mapping was performed for Summer Height (SH), Winter Height (WH), Total Height (TH), Summer Girth (SG), Winter Girth (WG) and Total Girth (TG) by applying a composite interval mapping (CIM) model to the integrated genetic map [27]. Using this method, 18 QTLs were detected in 11 linkage groups (Figure 2; Table 2). Of the detected QTLs, 9 were identified for summer (7 for Height and 2 for Girth), 5 for winter (2 for Height and 3 for Girth) and 4 for the total values (2 for Height and 2 for Girth). The part of the phenotypic variation explained by the QTLs ranged from 2.72% to 8.97%, and the QTLs segregated according to the ratios 1:1:1:1, 1:2:1, 3:1 and 1:1 (Table 2). To detect a QTL, a threshold value for the LOD Score was obtained based on 1,000 permutations, and these values were similar for all of the traits (3.74 for SH, 3.71 for WH, 3.85 for TH, 3.81 for SG, 3.83 for WG and 3.73 for TG).

For SH, 7 QTLs were identified (Table 2) in linkage groups 1 (129.73 cM), 5 (21 cM), 6a (64.06 cM), 8 (23 cM and 174 cM), 9 (78 cM), and 16 (10 cM), which explained 30.61% of the phenotype variation (R^2). The QTL located in LG 5 had the highest peak for SH with an LOD Score of 7.33. LG 5 also had the highest peak observed for all Height traits. The phenotypic variation explaining by each QTL ranged from 2.72 (LG 9) to 6.6 (LG 5). QTLs in linkage groups 1, 5 and 8 (174 cM) had a 1:2:1 segregation pattern, the QTLs in LGs 6a, 8 (23 cM) and 16a had a 1:1 type of segregation pattern and QTL located in LG9 segregated in a 1:1:1:1 fashion. Generally, the parents contributed equally to the trait variability. Of the 7 mapped QTLs, 3 had additive effect for PB217 (LG 6a, LG 8 at 174 cM and LG 9), 3 for PR255 (LG 1, LG 8 at 23 cM and LG 16a) and one QTL had an additive effect for both parents (LG5). A dominance effect was also detected in three mapped QTLs (LG 1, LG 8 at 174 cM and LG 9). This type of result can only be verified by performing QTL mapping using an integrated map, showing the advantage of using the procedure hereby implemented.

QTL mapping for Winter Height indicated that two QTLs in linkage groups 2a (34.43 cM) and 10 (82.99 cM) explained 6.42% and 6.48% of the phenotypic variation, respectively. The former has a significant additive effect from both parents with a 1:2:1 segregation pattern. The latter only has a significant additive effect from PR 255 and has a 1:1 segregation pattern.

For Total Height, two additional QTLs were mapped in linkage groups 5 (60.57 cM) and 9 (54 cM), accounting for 10.64% of the phenotypic variation. The first QTL has a LOD Score of 6.98 and an R^2 value of 8.07%. Furthermore, in agreement with its significant genetic effects and statistical similarities, this QTL has a 3:1 segregation pattern. The second QTL has a lower LOD Score and R^2 value (5.05 and 3.19, respectively) and has a 1:1 segregation pattern due to its significant dominance effect.

Summer Girth has mapped QTLs in linkage group 5 (56 cM) that account for 5.25% of the phenotypic variation, and its 1:1 segregation pattern is the result of a significant additive effect from parent PR255. Another QTL was mapped in linkage group 6b (6 cM) and has a LOD Score of 5.41, an R^2 value of 7.66% and a 1:2:1 segregation pattern that is the result of two similar and significant additive effects from the PR255 and PB217 parents.

Three QTLs for Winter Girth were mapped in linkage groups 3 (64 cM), 16a (145.63 cM) and 17b (82.26 cM) and account for 15.80% of the phenotypic variation. The first QTL mapped in linkage group 3 has the highest LOD Score and R^2 value for this trait (6.28 and 7.39, respectively) and has a 3:1 segregation pattern. This QTL segregated in a 3:1 ratio because all of its genetic effects are significant and similar. The second QTL mapped in linkage group 16a has a LOD Score of 4.01 and a R^2 value of 2.84%. This QTL only has a significant additive effect from PR255 and segregated in a 1:1 ratio. The QTL mapped in linkage group 17b has a LOD Score of 4.70 and a R^2 of 6.68. This QTL has an additive effect from PB217 and a significant dominance effect that results in a 1:2:1 segregation pattern.

For Total Girth there were two QTLs in linkage groups 6b (8 cM) and 16a (40 cM). The first QTL in linkage group 6b has a LOD Score of 7.32 and the highest R^2 value (8.97%). It segregates in a 1:2:1 ratio because both of the additive effects were significant. The second QTL has a LOD Score of 5.49 and an R^2 value of 5.14. This QTL has only one significant additive effect (PB217) that results in a 1:1 segregation pattern.

A comparison between the summer and winter seasons showed that the QTLs detected during the summer were different from the QTLs detected during the winter for Height and Girth growth (Figure 2). Differences were also observed when comparing the QTL mapped for Winter Height versus Total Height and Winter Girth versus Total Girth. For linkage group 6b, a QTL was mapped for Summer Girth at 6 cM and another QTL was mapped for Total Girth at 8 cM, indicating that the same QTLs control both traits. Furthermore, the 1:2:1 segregation pattern was the result of the significant additive effects from both parents.

In general, the phenotypic analysis indicated that higher correlations were found for growth between the summer and total period. This type of relationship was also observed in the QTL mapping because two QTLs could be linked for Summer Height and Total Height, and a potentially pleiotropic QTL was identified for Summer Girth and Total Girth. These results are interesting because they emphasize that when QTL mapping is conducted for the total growth period, the observed variability frequently occurs during the summer, during which rubber trees growth was maximized. Furthermore, it is interesting that no QTLs were found to occur during both the summer and a winter season, suggesting differential gene control under environmental stress. Such differential gene control would inhibit the expression of the regions responsible for growth, which causes the observed suboptimal response during the winter period. The study of growth during the winter season, however, did reveal other regions on the genetic map that control growth for this progeny. These QTLs may be important in future studies that attempt to identify specific polymorphisms that provide tolerance against cold and/or water stress.

There were QTLs that were only mapped during the summer period. The linkage group 5 has one QTL for each trait, spaced by 35 cM. The additive effect from PR255 is conserved for both traits, although the additive effect from PB217 is only present for Height. In the PR255 parent, we believe that both QTLs are linked by coupling when they show the same signal. Previous studies in *Pinus radiata* [49] and *E. globulus* [50] have shown that a single co-located QTL affects traits, and there are several independent QTLs for growth. Taken together, these findings suggest that the growth rate and wood property traits could be simultaneously improved through a proper breeding program.

A direct comparison of the regions mapped in different studies on rubber trees is challenging because neither the investigated markers and the populations and the traits nor the methodologies are comparable. For example, Lin et al. [51] proposed a QTL mapping model for outbred species using genetic maps with molecular markers showing different segregation patterns. Our approach was different because we did not model the linkage phase between QTL and markers with parameters from the likelihood, thus reducing substantially the complexity of EM algorithm. In our model the linkage phase between QTL and markers was inferred by the interpretation of the sign of additive effects, and the QTL segregation

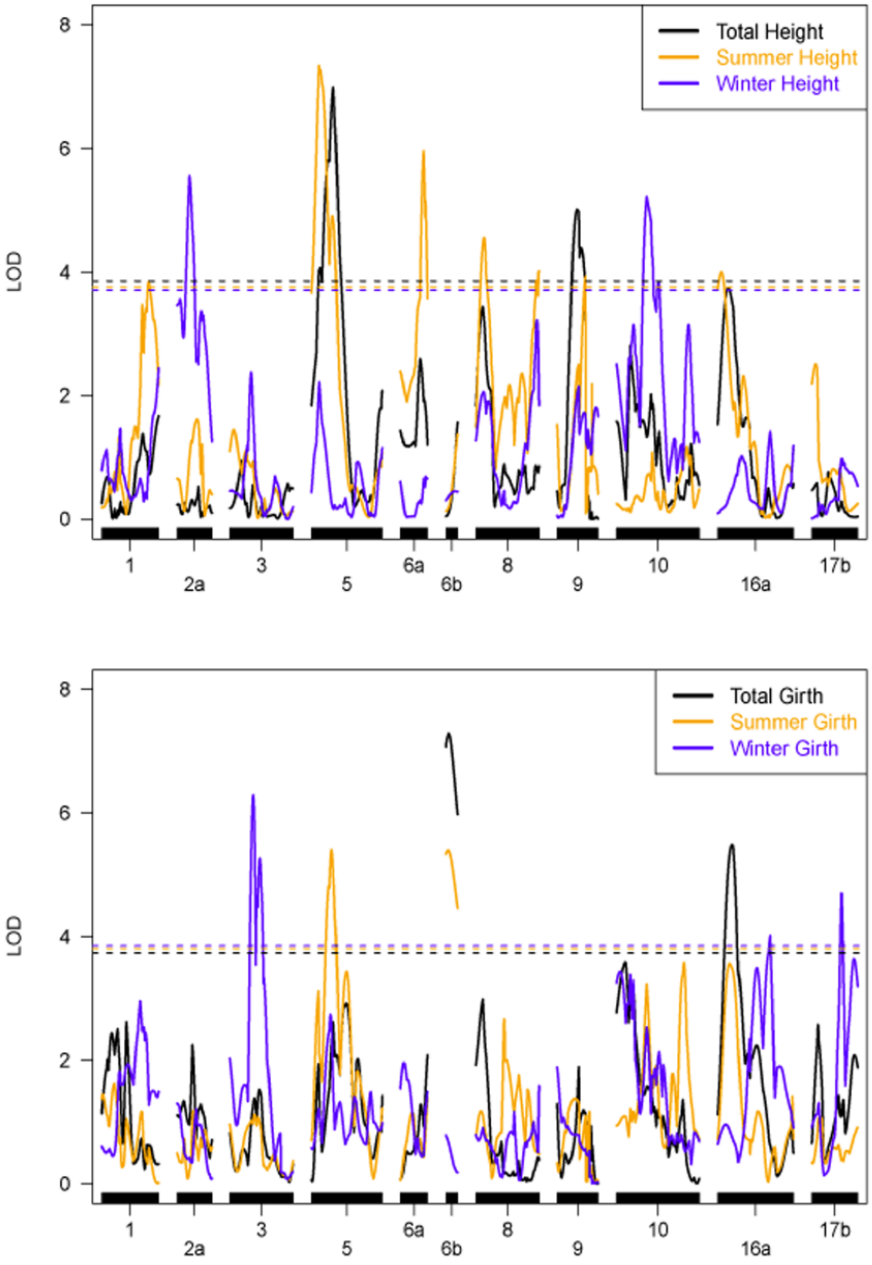


Figure 2. QTL mapping for height (upper panel) and girth (lower panel) for summer (orange), winter (violet) and total development (black). Note that dashed lines represents threshold values obtained with 1000 replicates.
doi:10.1371/journal.pone.0061238.g002

was inferred for mapped QTL. One of the advantages of our approach was the possibility to include cofactors (previously selected by multiple linear regressions) in the model, increasing the statistical power [36,52]. Another difference of our model was the usage of multipoint probabilities based on Hidden Markov Models [23,24], which is useful specially for interval flanked by not fully informative markers.

Some previous QTL mapping studies in rubber trees have been performed to characterize *Microcyclus ulei* resistance [10,25,26]. To our knowledge, this present work is the first QTL mapping study to use an integrated genetic map for the rubber tree. This method allowed for the identification of QTLs with different segregation patterns as well as QTLs with a dominance effect, which was not possible with the approaches used in previous studies. Here, ten

Table 2. Mapped QTLs for height and girth growth during the summer and winter seasons.

QTLs name	Flanquing marquers	LG[(1)]	Position (cM)	Global LOD	$R^{2(1)}$	Additive effect PB217	LOD[(2)]	Additive effect PR255	LOD[(2)]	Dominance effect	LOD[(2)]	Segregation
SH.1	HB31	1	129.73	3.85	5.51	0.6343	0.62	−1.1830	2.12	−1.0541	1.71	1:2:1
SH.2	T2636-HB1	5	21.00	7.33	6.60	1.3617	2.89	1.6181	4.06	0.2631	0.11	1:2:1
SH.3	A2336	6a	64.06	5.95	5.23	1.6770	4.13	0.7106	0.81	0.6478	0.66	1:1
SH.4	HBE140-HB41	8	23.00	4.56	5.58	0.3421	0.14	1.8253	4.30	−0.3767	0.15	1:1
SH.5	HBE151-A2525	8	174.00	4.02	3.82	1.0036	1.51	0.4135	0.26	−1.3086	2.45	1:2:1
SH.6	HBE77-HB7	9	78.00	4.41	2.72	−1.1673	1.89	0.5608	0.48	−4.0434	3.69	1:1:1:1
SH.7	TAs2225-HBE64	16a	10.00	4.01	4.07	1.3488	0.76	1.5550	1.03	0.2046	0.03	1:1
WH.1	T2607	2a	34.43	5.55	6.42	−1.2659	2.53	−1.2586	2.50	−0.0066	0.00	1:2:1
WH.2	HB186	10	82.99	5.21	6.48	−0.4545	0.32	−1.7141	4.47	0.6631	0.63	1:1
TH.1	a491	5	60.57	6.98	8.07	1.7309	4.52	0.8814	1.21	0.8265	1.08	3:1
TH.2	A2532-a90	9	54.00	5.05	3.19	−0.3608	0.13	0.3308	0.15	−3.0877	4.89	1:1
SG.1	a235–a491	5	56.00	5.40	5.25	0.0952	0.01	1.8027	5.13	0.2791	0.10	1:1
SG.2	T2449-HBE49	6b	6.00	5.41	7.66	1.7824	3.24	−1.7277	3.46	0.3765	0.17	1:2:1
WG.1	HB24-a312	3	64.00	6.28	7.39	0.3638	1.01	0.6775	3.76	0.6172	2.76	3:1
WG.2	a104	16a	145.63	4.01	2.84	−0.0832	0.07	0.6368	3.77	−0.1774	0.30	1:1
WG.3	HB123	17b	82.26	4.70	6.68	−0.3362	0.99	−0.1617	0.25	0.6684	4.00	1:2:1
TG.1	T2449-HBE49	6b	8.00	7.32	8.97	2.8462	3.89	−2.9124	4.77	0.6612	0.26	1:2:1
TG.2	TAs2225-HBE64	16a	40.00	5.49	5.14	2.2344	4.12	0.6982	0.42	−1.0117	0.78	1:1

[(1)]LG indicate linkage group and R^2 is a phenotype variation.
[(2)]To detect a QTL, a threshold value for the LOD Score was obtained based on 1,000 permutations, and these values were similar for all of the traits (3.74 for SH, 3.71 for WH, 3.85 for TH, 3.81 for SG, 3.83 for WG and 3.73 for TG).
doi:10.1371/journal.pone.0061238.t002

(55.5%) of the 18 mapped QTLs has a segregation pattern of either 1:1:1:1, 1:2:1 or 3:1, which may have contributed to the R^2 values obtained here.

The parents PR255 and PB217 had contrasting and complementary phenotypes. The significant additive effects detected in this study indicate that both parents contributed equally to the polymorphisms in the investigated traits. Among the significant marginal effects that were identified, 11 were associated with PB217 and 12 were associated with PR255. The results obtained using QTL mapping provides novel insights into the genetic control of growth characteristics in the rubber tree. Among the 18 QTLs that were detected, 7 (38.8%) are the result of dominance effects. Although additive effects are the primary cause of the polymorphisms in the progeny, the dominance effects play an important role in the genetic architecture of these traits.

The results hereby presented indicate that complex traits in rubber trees are controlled by many genes and the individual effect of one of these genes on the phenotype is small. Furthermore, these results indicate that broad genome-wide searches will be required to identify all of the genes that control a complex trait as well as all of the segregating variation in a population [53].

Height and Girth are considering as quantitative traits and are influenced by many variables, finding QTLs explaining these changes is difficult. The low heritability is also difficult to identify QTLs, thus, few regions are detected, so the chance to detect QTLs in commons between summer and winter is also reduced.

The approach used here is informative and provides interesting insights about growth traits in rubber tree. This could be combined in the future with ideas related to functional mapping [54–57]. By considering a series of repeated measures collected over time, these authors suggested a logistic model for growth pattern. This would allow a better understanding of QTL expression over time and will be considered for future studies.

The QTLs identified in this work, particularly those related to phenotypic variation during the winter season, can be used to begin investigating the potential for marker-assisted selection, especially when the objective of a breeding program is to obtain phenotypes adapted to sub-tropical climate areas. Studying the behavior of plants in winter will be very interesting to find polymorphism that allows better adaptation to winter, for example, some resistance to cold and dry, than finding genotypes that develop in the cold with the same intensity as they would in the summer. These studies are possible because the winter climate (April to October) in the location where these experiments were performed is dry and cold compared to the summer climate (Figure S3); thus, the fungus *M. ulei* is not well adapted to this type of climate. Further studies using other crosses are required to identify additional QTLs that affect frost tolerance in *H. brasiliensis*. The long developmental cycle of rubber trees presents a challenge to QTL mapping, and the size of the experiment is a limiting factor for establishing mapping populations in different environments. Therefore, further studies investigating traits associated with the development of these plants are necessary for the selection of genotypes that are adapted to sub-tropical climate areas.

Supporting Information

Figure S1 Distribution of the phenotypic data for the F1 population and the genitor (PB217 and PR255).
(TIF)

Figure S2 Genotypic correlation coefficients and dispersion of the phenotypic data. Genotypic correlation coefficients and

dispersion of the phenotypic data for each trait measured during the summer and winter seasons. (*significant at the 5%; **significant at the 1%; ***significant at 5% of global level - Bonferroni correction for multiple tests).
(TIF)

Figure S3 The average temperatures and precipitation. The average temperatures (maximum and minimum) and precipitation for the years 2006 to 2009 in Itiquira-MS.
(TIF)

Table S1 The distribution and type of SSR and EST-SSR markers in the linkage groups.
(TIF)

Author Contributions

Conceived and designed the experiments: LMS DG VLG APS AAFG. Performed the experiments: LMS CCM CCS SEAC. Analyzed the data: LMS RG. Contributed reagents/materials/analysis tools: APS AAFG. Wrote the paper: LMS RG.

References

1. Pires JM, Secco RS, Gomes JI (2002) Taxonomia e fitogeografia das seringueiras (*Hevea spp.*). Belém: Embrapa Amazônia Oriental. 103 p.
2. Raj S, Das G, Pothen J, Dey S (2005) Relationship between latex yield of *Hevea brasiliensis* and antecedent environmental parameters. International Journal of Biometeorology 49: 189–196.
3. Priyadarshan PM, Goncalves PS (2003) *Hevea* gene pool for breeding. Genetic Resources and Crop Evolution 50: 101–114.
4. Rao P, Saraswathyamma C, Sethuraj M (1998) Studies on the relationship between yield and meteorological parameters of para rubber tree (*Hevea brasiliensis*). Agr Forest Meteorol 90: 235–245.
5. Jacob J, Annmalainathan K, Alam BM, Sathick MB, Thapaliyal AP, et al. (1999) Physiological constraints for cultivation of *Hevea brasiliensis* in certain unfavorable agro climatic regions of India. Indian J Nat Rub Res 12: 1–16.
6. Besse P, Seguin M, Lebrun P (1994) Genetic diversity among wild and cultivated populations of *Hevea brasiliensis* assessed by nuclear RFLP analysis. Theoretical and applied genetics 88: 199–207.
7. Low F, Atan S, Jaafar H, Tan H (1996) Recent advances in the development of molecular markers for *Hevea* studies. Journal of Nartural Rubber Research 11: 32–44.
8. Varghese Y, Knaak C, Sethuraj M, Ecke W (1997) Evaluation of random amplified polymorphic DNA (RAPD) markers in *Hevea brasiliensis*. Plant Breeding 116: 47–52.
9. Lespinasse D, Rodier-Goud M, Grivet L, Leconte A, Legnate H, et al. (2000) A saturated genetic linkage map of rubber tree (*Hevea spp.*) based on RFLP, AFLP, microsatellite, and isozyme markers. Theoretical and Applied Genetics 100: 127–138.
10. Le Guen V, Lespinasse D, Oliver G, Rodier-Goud M, Pinard F, et al. (2003) Molecular mapping of genes conferring field resistance to South American Leaf Blight (*Microcyclus ulei*) in rubber tree. Theoretical and applied genetics 108: 160–167.
11. Roy C, Nazeer M, Saha T (2004) Identification of simple sequence repeats in rubber (*Hevea brasiliensis*). Current Science 86: 807–811.
12. Saha T, Roy CB, Nazeer MA (2005) Microsatellite variability and its use in the characterization of cultivated clones of *Hevea brasiliensis*. Plant Breeding 124: 86–92.
13. Souza LM, Mantello CC, Santos MO, Gonçalves PS, Souza AP (2009) Microsatellites from rubber tree (*Hevea brasiliensis*) for genetic diversity analysis and cross-amplification in six *Hevea* wild species. Conservation Genetics Resource 1: 75–79.
14. Le Guen V, Gay C, Xiong T (2010) Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*). Plant Breeding 130: 294–296.
15. Yu F, Wang B-H, Feng S-P, Wang J-Y, Li W-G, et al. (2011) Development, characterization, and cross-species/genera transferability of SSR markers for rubber tree (*Hevea brasiliensis*). Plant Cell Reports 30: 335–344.
16. Mantello CC, Suzuki FI, Souza LM, Gonçalves PS, Souza AP (2012) Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species. BMC research notes 5: 329.
17. Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137: 1121–1137.
18. Wu R, Ma C, Painter I, Zeng Z (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theoretical Population Biology 61: 349–363.
19. Maliepaard C (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genetical Research 70: 237–250.
20. 20. Garcia AAF, Kido EA, Meza AN, Souza HMB, Pinto LR, et al. (2006) Development of an integrated genetic map of a sugarcane (*Saccharum spp.*) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theoretical and applied genetics 112: 298–314.
21. Oliveira KM, Pinto LR, Marconi TG, Margarido GRA, Pastina MM, et al. (2007) Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum spp.*) commercial cross. Molecular Breeding 20: 189–208.
22. Oliveira EJ, Vieira MLC, Garcia AAF, Munhoz CF, Margarido GRA, et al. (2008) An integrated molecular map of yellow passion fruit based on simultaneous maximum-likehood estimation of linkage and linkage phases. Journal of the American Society for Horticultural Sciences 133: 35–41.
23. Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. Hereditas 144: 78–49.
24. Wu R, Ma C-X, Wu SS, Zeng Z-B (2002) Linkage mapping of sex-specific differences. Genetical Research 79: 85–96.
25. Le Guen V, Garcia D, Mattos CRR, Doaré F, Lespinasse D, et al. (2007) Bypassing of a polygenic *Microcyclus ulei* resistance in rubber tree, analyzed by QTL detection. New Phytologist 173: 335–345.
26. Le Guen V, Garcia D, Doaré F, Mattos CRR, Condina V, et al. (2011) A rubber tree's durable resistance to *Microcyclus ulei* is conferred by a qualitative gene and a major quantitative resistance factor. Tree Genetics & Genomes 7: 877–889.
27. Gazaffi R (2009) Desenvolvimento de modelo genético-estatístico para mapeamento de QTLs em progênie de irmãos completos, com aplicação em cana-de-açúcar. PhD Thesis: Universidade de São Paulo. 103 p.
28. Murray M, Thompson W (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Research 8: 4321–4325.
29. Feng SP, Li WG, Huang HS, Wang JY, Wu YT (2009) Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Molecular Breeding 23: 85–97.
30. Lander E, Green P, Abrahamson J, Barlow A (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174–181.
31. Mollinari M, Margarido G, Vencovsky R, Garcia A (2009) Evaluation of algorithms used to order markers on genetic maps. Heredity 103: 494–502.
32. Kosambi DD (1944) The estimation of map distances from recombinant values. Annals of Eugenics 12: 172–175.
33. Boer MP, Wright D, Feng L, Podlich DW, Luo L, et al. (2007) A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics 177: 1801–1813.
34. Thummam BR, Baltunis BS, Bell JC, Emebiri LC, Moran GF, et al. (2010) Quantitative trait locus (QTL) analysis of growth and vegetative propagation traits in *Eucalyptus nitens* full-sib families. Tree Genetics & Genomes 6: 877–889.
35. Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2010) GenStat for Windows 13th Edition. VSN International, Hemel Hempstead,UK, GenStat.-co.uk.
36. Zeng Z (1994) Precision mapping of quantitative trait loci. Genetics 136: 1457–1468.
37. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.
38. Chen L, Storey J (2006) Relaxed significance criteria for linkage analysis. Genetics 173: 2371–2381.
39. Scott K, Eggler P, Seaton G (2000) Analysis of SSRs derived from grape ESTs. Theoretical and applied genetics 100: 723–726.
40. Diaz A, Fergany M, Formisano G (2011) A consensus linkage map for molecular markers and Quantitative Trait Loci associated with economically important traits in melon (*Cucumis melo* L.). BMC plant biology 11: 111.
41. Truco MJ, Antonise R, Lavelle D, Ochoa O, Kozik A, et al. (2007) A high-density, integrated genetic linkage map of lettuce (*Lactuca spp.*). Theoretical and applied genetics 115: 735–746.
42. Yan Z, Denneboom C, Hattendorf A (2005) Construction of an integrated map of rose with AFLP, SSR, PK, RGA, RFLP, SCAR and morphological markers. Theoretical and Applied Genetics 110: 766–777.
43. Triwitayakorn K, Chatkulkawin P, Kanjanawattanawong S, Sraphet S, Yoocha T, et al. (2011) Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. DNA research 18: 471–482.
44. Gouvêa LRL, Rubiano LB, Chioratto AF, Zucchi MI, De Souza Gonçalves P (2010) Genetic divergence of rubber tree estimated by multivariate techniques and microsatellite markers. Genetics and Molecular Biology 33: 308–318.
45. van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, 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.
46. Castiglioni P, Ajmone-Marsan P, van Wijk R, Motto M (1999) AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. Theoretical and Applied Genetics 99: 425–431.

47. Gonçalves PS, Moraes M (2005) Genetic variation in growth traits and yield of rubber trees (*Hevea brasiliensis*) growing in the Brazilian state of São Paulo. Genetics and Molecular Biology 28: 765–772.
48. Volker P, Potts B, Borralho N (2008) Genetic parameters of intra- and inter-specific hybrids *of Eucalyptus globulus* and *E. nitens*. Tree Genet Genom 4: 445–460.
49. Devey M, Carson S, Nolan M (2004) QTL associations for density and diameter in *Pinus radiata* and the potential for marker-aided selection. Theoretical and applied genetics 108: 516–524.
50. Bundock PC, Potts BM, Vaillancourt RE (2007) Detection and stability of quantitative trait loci (QTL) in *Eucalyptus globulus*. Tree Genetics Genomes 4: 85–95.
51. Lin M, Lou X, Chang M, Wu R (2003) A General Statistical Framework for Mapping Quantitative Trait Loci in Nonmodel Systems: Issue for Characterizing Linkage Phases. Genetics 165: 901–913.
52. Zeng ZB (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America 90: 10972–10976.
53. Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12: 111–122.
54. Ma C-X, Casella G, Wu R (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 161: 1751–1762.
55. Wu R, Ma C-X, Lin M, Casella G (2004) A general framework for analyzing the genetic architecture of developmental characteristics. Genetics 166: 1541–1551.
56. Tong C, Wang Z, Zhang B, Shi J, Wu R (2011) 3FunMap: full-sib family functional mapping of dynamic traits. Bioinformatics 27: 2006–2008.
57. Wu R, Lin M (2006) Functional mapping - how to map and study the genetic architecture of dynamic complex traits. Nature Reviews Genetics 7: 229–237.

Supporting Information

Figure S1. The distribution and type of SSR and EST-SSR markers in the linkage groups.

Linkage group	1:1:1:1	1:2:1	1:1	Total	Length	Markers density
LG1	10	1	10	21	157.30	7.49
LG2a	5	1	7	13	96.24	7.40
LG2b	-	2	-	2	2.72	1.36
LG3	5	1	10	16	174.00	10.88
LG4	7	2	2	11	115.10	10.46
LG5	10	1	9	20	195.60	9.78
LG6a	2	1	7	10	74.95	7.50
LG6b	-	1	1	2	32.65	16.33
LG7	6	2	9	17	165.20	9.72
LG8	7	3	11	21	174.20	8.30
LG9	1	1	9	11	113.30	10.30
LG10	14	6	9	29	228.70	7.89
LG11a	3	-	-	3	80.63	26.88
LG11b	2	-	-	2	20.70	10.35
LG12	2	2	7	11	201.50	18.32
LG13	7	-	8	15	168.70	11.25
LG14	8	-	12	20	217.40	10.87
LG15	6	4	5	15	154.00	10.27
LG16	9	2	6	17	209.30	12.31
LG17a	2	-	3	5	50.46	10.09
LG17b	-	1	3	4	22.64	5.66
LG18	7	-	8	15	172.20	11.48
LG19	2	-	2	4	10.25	2.56
Total	**115**	**31**	**138**	**284**	**2837.74**	**9.99**

Figure S2. Genotypic correlation coefficients and dispersion of the phenotypic data.

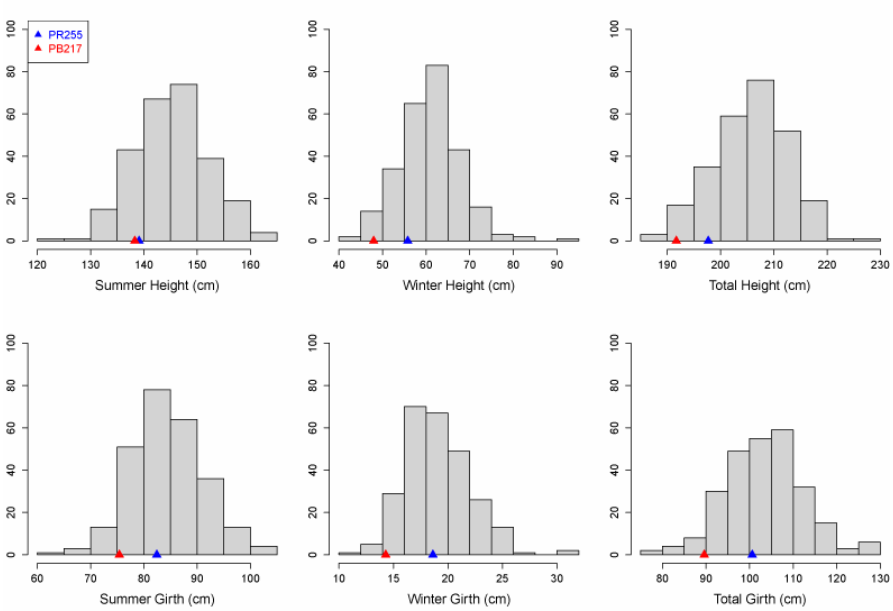
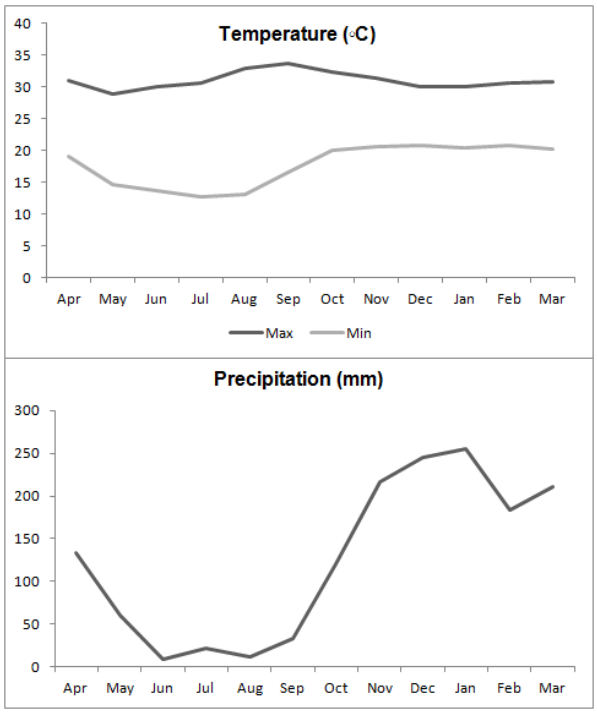


Figura S3. The average temperatures and precipitation.



Anexo II

DECLARAÇÃO

Declaro para os devidos fins que o conteúdo de minha dissertação de Mestrado/tese de Doutorado intitulada "Mapeamento Genético Molecular em *Hevea brasiliensis*":

( ) não se enquadra no § 4º do Artigo 1º da Informação CCPG 002/13, referente a bioética e biossegurança.

Tem autorização da(s) seguinte(s) Comissão(ões):

(X) CIBio – Comissão Interna de Biossegurança , projeto No. 05|2003, Instituição: _____.

( ) CEUA – Comissão de Ética no Uso de Animais , projeto No. _____, Instituição: _____.

( ) CEP - Comissão de Ética em Pesquisa, protocolo No. _____, Instituição: _____

* *Caso a Comissão seja externa ao IB/UNICAMP, anexar o comprovante de autorização dada ao trabalho. Se a autorização não tiver sido dada diretamente ao trabalho de tese ou dissertação, deverá ser anexado também um comprovante do vínculo do trabalho do aluno com o que constar no documento de autorização apresentado.*

Aluno: Camila Campos Mantello

Orientador: Anete Pereira de Souza

Para uso da Comissão ou Comitê pertinente:
(X) Deferido ( ) Indeferido

Carimbo e assinatura

DRA. EDI LUCIA SANTONIC
Lab. Genética Humana
CEMEC - UNICAMP

Para uso da Comissão ou Comitê pertinente:
( ) Deferido ( ) Indeferido

Carimbo e assinatura