Mol Breeding (2016) 36:115
DOI 10.1007/s11032-016-0534-3

Development of single nucleotide polymorphism markers in the large and complex rubber tree genome using next-generation sequence data

Livia Moura de Souza · Guilherme Toledo-Silva · Claudio Benicio Cardoso-Silva · Carla Cristina da Silva · Isabela Aparecida de Araujo Andreotti · Andre Ricardo Oliveira Conson · Camila Campos Mantello · Vincent Le Guen · Anete Pereira de Souza

Received: 8 February 2016 / Accepted: 16 July 2016 / Published online: 30 July 2016

Abstract The development of single nucleotide polymorphism (SNP) markers provides the opportunity to improve many areas of plant breeding and population genetics. Unfortunately, for species such as the rubber tree (*Hevea brasiliensis*), the use of next-generation sequencing for genomic SNP discovery is very difficult because of the large genome size and the abundance of repeated sequences. Access to a set of validated SNP markers is a significant advantage for rubber researchers who wish to apply SNPs in scientific research. Here, we performed genomic sequencing of *H. brasiliensis* and generated 10,993,648 short reads, which were assembled into 10,071 contigs (N50 = 3078) by a *de novo* assembly strategy. A total of 2446 contigs presented no hits in the current *H. brasiliensis* genome assembly and may therefore be considered novel genomic sequences of rubber tree. A total of 143 putative polymorphic positions were selected, gene annotations were available for 58.7 % of the markers, and all of the sequences could be anchored to the released *H. brasiliensis* genome. These SNPs were validated in eight genotypes of *H. brasiliensis* and 15 F1 plants from a mapping population, resulting in 30 (20.9 %) positions correctly classified. The analysis revealed key candidate genes responsible for defence mechanisms and provided markers for further genetic improvement of *Hevea* in breeding programmes.

Keywords Single nucleotide polymorphism · Next-generation sequencing · Molecular marker · *Hevea brasiliensis*

Electronic supplementary material The online version of this article (doi:10.1007/s11032-016-0534-3) contains supplementary material, which is available to authorized users.

L. M. de Souza · G. Toledo-Silva · C. B. Cardoso-Silva · C. C. da Silva · I. A. de Araujo Andreotti · A. R. O. Conson · C. C. Mantello · A. P. de Souza (✉)
Molecular Biology and Genetic Engineering Center (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil
e-mail: anete@unicamp.br

G. Toledo-Silva
Department of Biochemistry, Federal University of Santa Catarina, Florianópolis, SC, Brazil

V. Le Guen
UMR AGAP, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpellier, Hérault, France

A. P. de Souza
Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP), Campinas, SP, Brazil

Introduction

Until recently, single nucleotide polymorphism (SNP) marker development was expensive and time-consuming. With the advent of next-generation sequencing (NGS) and faster genotyping technologies, SNPs have emerged as the marker of choice in crop breeding (Varshney et al. 2009). SNP marker development and genotyping have provided insight into the genetics of model organisms; however, studies of non-model species have lagged behind because of the scarcity of sequences and markers. SNPs and insertions/deletions (InDels) are the most abundant types of DNA sequence polymorphisms and can be theoretically found within every genomic sequence (Rafalski 2002). SNP markers have many applications, such as cultivar identification, construction of genetic maps, assessment of genetic diversity, detection of genotype/phenotype associations, and marker-assisted breeding (Flint-Garcia et al. 2005).

There are few studies involving whole genome sequencing (WGS) efforts that have led to the successful discovery of SNPs; nonetheless, many studies in the rubber tree (*Hevea brasiliensis* (Willd. ex Adr. de Juss.) Muell.-Arg.) have focused on transcriptome analysis (Li et al. 2012; Mantello et al. 2014; Triwitayakorn et al. 2011). Although these RNA-seq data have added an abundance of new information on the rubber tree, the non-coding regions of the genome are also essential for understanding the regulatory elements controlling gene expression, as well as other genomic features, allowing the development of a more comprehensive set of molecular markers.

NGS technologies have also been extended to SNP discovery in large and complex genomes that lack an assembled reference genome (Bachlava et al. 2012; You et al. 2011), and additional findings from SNP research in non-model organisms are especially timely. In the case of the rubber tree, a limited number of SNPs have been described. The first study of this type on the rubber tree developed only ten SNP markers (Pootakham et al. 2011). In 2014, Salgado et al. (2014) developed 172 new SNP markers; Silva et al. (2014) characterized 43 SNP markers in 13 sequences that showed similarities to those encoding proteins involved in stress response, latex biosynthesis and developmental processes; Mantello et al. (2014) 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, and validated 78 SNP markers. However, there are many SNPs in non-coding regulatory regions that can be used as molecular markers, and the exact functions of these SNPs are not yet clear.

As the majority of SNPs occur in non-coding sequences of the genome, the frequency of nucleotide substitutions is almost three times higher in non-coding regions than in coding sequences (Ching et al. 2002), and their influences on phenotype may occur through biological mechanisms such as transcription factor binding and alternative splicing. Therefore, the importance of developing markers in these regions is evident for *H. brasiliensis*, which is the prime source of commercial rubber, and several other economically important species.

Hevea brasiliensis is a deciduous perennial tree of the family Euphorbiaceae. The genus *Hevea* occurs naturally only within the Amazon rainforest and its distribution extends over six million square kilometres, which is more than half of the territory of Brazil, at the boundaries of the Amazon forest. With a wide adaptation range for different ecological environments, the rubber tree exhibits considerable morphological variability, ranging from tall trees to shrubs (Gonçalves and Fontes 2012).

Due to the long generation time and the large size of the crop, new tools need to be developed to manage the germplasm bank variability and assist breeders in their recombination strategies. The development of DNA-based markers is important for the selection and improvement of varieties and hybrids in plant breeding programmes (Gupta et al. 2001; Kota et al. 2003).

We performed genomic shotgun sequencing in *H. brasiliensis* to develop SNP markers in regions of possible regulatory cis-elements of promoters and regulators. These regions are not accessible with RNA-seq; however, they are important in gene regulation and are therefore interesting targets for plant breeding programmes. This study reports the results from the characterization of rubber tree SNP markers in eight genotypes of *H. brasiliensis* and 15 F1 plants from the cross between PR255 and PB217, which were part of a mapping population.

Materials and methods

DNA preparation, library construction and sequencing

Leaves of two *H. brasiliensis* genotypes (GT1 and RRIM701 genotypes) were sampled, and their total DNA was extracted following the protocol of Doyle and Doyle (1987). DNA samples were sonicated (Bioruptor, Diogenode, Liege, Belgium) to obtain fragments ranging from 300 to 400 bp. Illumina libraries were constructed using the Paired End DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA) in 50 µl reactions containing 2 µg of DNA, as recommended by the manufacturer. The quality of the libraries was assessed on a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and samples were clustered using a TruSeq PE Cluster Kit on cBot (Illumina). Resulting libraries were sequenced on an Illumina GAIIx platform with paired-end reads of 72 bases in length using TruSeq SBS 36-Cycle kits (Illumina).

Raw data analysis and de novo assembly

The BLC files generated by Illumina sequencing from GT1 and RRIM701 genotypes were first converted to qSeq format using Off-Line Basecaller v.1.9.4 (OLB) software and then transformed into FASTQ files using a custom script. Short read data were filtered and trimmed using the CLC Genomics Workbench package (CLC Bio, Aarhus, Denmark). Reads with length <60 bases, single reads (lacking a read pair) and bases with low-quality scores ($Q < 20$) were discarded. De novo assembly was performed using CLC Genomics Workbench with the following parameters: word size (k-mer) was defined as 29, maximum gap and mismatch count were set to 2, insertion and deletion penalties were set to 3, minimum contig length was 500 bp and similarity and length fraction values were 0.8 and 0.5, respectively.

The contigs resulting from the *de novo* assembly were cleaned, removing chloroplast- and mitochondria-derived sequences. This step consisted of a similarity search against organelle sequences of *H. brasiliensis*, *Manihot esculenta*, *Ricinus communis* L., *Vitis vinifera*, *Populus trichocarpa* and *Glycine max* via BlastN using the BLAST+ suite (Camacho et al. 2009), considering a cut-off e-value of 1e−06. The resulting dataset was used for all subsequent analysis. Selected contigs were analysed using CENSOR (Kohany et al. 2006), a tool designed to rapidly identify repetitive elements by comparison with known repeats. The microsatellite identification tool (MISA) (Thiel et al. 2003) was used to detect simple sequence repeats (SSRs).

Variant detection

For SNP detection, reads were mapped to the filtered set of contigs to evaluate putative polymorphisms using Burrows–Wheeler Aligner (BWA) (Langmead et al. 2009). First, contigs were indexed through the index function of BWA, and then they were aligned as paired-end reads using the BWA-MEM function, using default parameters for all procedures. For variant calling, we used a strategy in which only sites detected by two different programs were selected. The mpileup pipeline of SAMtools (Li et al. 2009) was used as the first caller, with a minimum mapping quality of 30, minimum base quality of 20, minimum coverage of 10 and maximum coverage of 100. Freebayes (Garrison and Marth 2012) was used as the second caller with the following parameters: minimum read counting for variant calling 2, minimum base quality 20, minimum mapping quality 30, minimum coverage 10 and no InDels or multiple nucleotide polymorphisms (MNPs) called. Variants from sequenced genotypes were compared using the vcf-isec program from VCFtools (Danecek et al. 2011), which generated a list of unique and shared SNPs. Selection of putative polymorphic sites for validation was performed under the criteria of minimum alternate allele observations (AO) of 5, mean mapping quality of observed alternate alleles (MQM) of 50, mean mapping quality of observed reference alleles (MQMR) of 50, proportion of observed alternate alleles that are supported by properly paired read fragments (PAIRED) of 0.50 and proportion of observed reference alleles that are supported by properly paired read fragments (PAIRED) of 0.50.

Characterization of putative SNP-containing contigs

To evaluate whether novel *H. brasiliensis* genomic sequences were obtained with this *de novo* assembly, contigs were aligned to the published draft genome of *H. brasiliensis* (Rahman et al. 2013) using BlastN with a cut-off e-value of 1e−06.

Additionally, we aimed to find contigs containing coding regions. For this purpose we used Illumina short reads of RNA-seq data from a study performed by Mantello et al. (2014). Reads from two *H. brasiliensis* genotypes (GT1 and RRIM701) were combined and then mapped to *de novo* assembled contigs using the splice-aware aligner TopHat2 (Kim et al. 2013), choosing Bowtie2 (Langmead and Salzberg 2012) as the mapper and other parameters as default. Transcripts were reconstructed from the alignment using Cufflinks 2.2.1 (Trapnell et al. 2009) with default parameters and were submitted to a BlastX search against the UniprotKB/Swiss-prot database, using BLAST+ suite (Camacho et al. 2009).

The complete coding sequences (CDS) from cassava (*Manihot esculenta*) were retrieved from http://www.phytozome.net/cassava (Prochnik et al. 2012) and mapped against the *de novo* assembly contigs to detect putative coding regions using sim4 software (Florea et al. 1998). BLAST hits with at least 70 % alignment were considered putative coding regions.

Manual characterization was performed for the sequences bearing the SNP markers developed here. The BLAST tool was used to search against the *H. brasiliensis* genomic sequences deposited in NCBI GenBank (Whole Genome Shotgun—WGS-database) to identify the matching sequences in the rubber tree draft genome. Annotation was performed using the BLAST tool to search against the Uniprot database (www.uniprot.org) and against the genomes of *Populus trichocarpa*, *Ricinus communis* and *Manihot esculenta* in the Phytozome database (Goodstein et al. 2012), using a cut-off e-value of e−10 in both databases.

KASPar assays

Only SNP positions free of other variants within the 50 bp upstream and downstream flanking sequences were selected (Table S1) for developing the KASPar assays. Furthermore, KASPar primer pairs for the targeted SNPs were ordered from KBioscience (Hertfordshire, UK); two allele-specific forward primers and one common reverse primer were designed.

Molecular markers were validated on eight genotypes of *H. brasiliensis* that are parents of studied mapping populations (GT1, PB235, RRIC100, PB217, PR255, PB260, RRIM701 and RRIM600) and 15 F1 plants from the cross between PB217 × PR255 (Souza et al. 2013). DNA samples were extracted from lyophilized leaf tissues using a modified CTAB method (Doyle and Doyle 1987).

A 4.08 µl KASP assay reaction contained 2 µL KASPar 2× Reagent Mix (KBioscience, PN KBS-1004-001), 0.05 µL assay mix (12 µM each allele-specific forward primer and 30 µM reverse primer), 0.03 µL $MgCl_2$ (2.2 mM) and 2 µL genomic DNA (10 ng/µL). The cycling conditions were as follows: 15 min at 94 °C; 10 touchdown cycles of 20 s at 94 °C and 20 s at 65 °C (the annealing temperature for each cycle being reduced by 0.8 °C per cycle); and 40 cycles of 20 s at 94 °C and 60 s at 57 °C. Fluorescence detection of the reactions was analysed using LightCycler 480 1.5.0 SP3 software (Roche, Basel, Switzerland).

Allele-calling data were viewed graphically as a scatter plot for each marker assayed using the LightCycler 480 II (Roche). Alleles of each SNP were scored as present, absent, duplicated or missing (failed to amplify) and converted into a binary matrix to determine minor allele frequencies (MAFs) for each SNP locus.

SNP descriptive analyses

Key descriptive statistics for measuring the informativeness of the SNP markers were calculated. The allelic polymorphism information content (PIC) and expected heterozygosity (He) values were calculated using the PIC calculator (https://www.liverpool.ac.uk/~kempsj/pic.html).

Results and discussion

Sequencing and assembling

The shotgun sequencing was a paired-end run which generated ~75.9 million short reads from *H. brasiliensis* DNA (GT1 and RRIM701 genotypes) using Illumina sequencing technology. The sequencing run was performed in paired-end mode, which generated reads of 72 bp (Table 1). According to Hillier et al. (2008), paired-end reads clearly increase the power to properly interpret problematic areas of the genome, including collapsed or misassembled repeats, and to detect structural variations. As genomes increase in size and complexity, paired ends

Table 1 Summary of data resulting from sequencing and the *de novo* assembly of Illumina paired-end reads from the genome of *H. brasiliensis*

Sequencing	
Total read count	75,820,926
Mean read length	72
Total high-quality reads	71,971,604
% high-quality reads	94.92
De novo assembly	
Contig count	10,071
Number of reads used	10,993,648
% reads used	14.49
Mean contig length	791
Total contig length	7,968,458
"N" occurrences in contigs	247
CpG sites	126,491
GC content in %	37.89
N50 contig length	3078
Long contigs (>10 kb)	38

will also be more efficiently placed than single-end reads. After filtering and trimming, approximately 94.92 % of data were considered high-quality (HQ) reads.

The *de novo* assembly (Table 1) used 10,993,648 short reads resulting in 10,071 contigs (Additional file 1) with a total length of 7,968,458 bp (92 % Q20 bases and 37.89 % GC content), with a mean length of 791 bp and a N50 contig length of 3078 bp. The N50 metric of *H. brasiliensis* contigs was higher than that of the sequence assembly of the barley genome (5.1 Gb), as the WGS assembly based on Illumina short reads resulted in relatively small contigs (N50 = 1425 bp) (The International Barley Genome Sequencing Consortium 2012). Direct comparisons of assembly metrics are challenging because the methods used for contig definition and/or minimum contig settings have not been standardized. Nevertheless, these metrics showed that the current assembly was successful in obtaining useful genomic contigs of *H. brasiliensis* for genomic SNP discovery.

Contig dataset filtering consisted of the removal of non-genomic-derived sequences through a homology search via BlastX against organelle sequences (chloroplast and mitochondria) of closely related species, resulting in less than 2 % (168) of contigs containing organelle sequences that were removed from the dataset. The remaining 9903 contigs were used as a reference for subsequent analysis. First, we aimed to check whether the *de novo* assembly generated new genomic information about the species. The BlastN homology search of the assembled sequences against the *H. brasiliensis* draft genome (Rahman et al. (2013) showed that 2446 (22.7 %) of contigs were novel, while 7457 contigs exhibited high homology to existing rubber tree genomic sequences. Thus, *de novo* assembly is a useful method by which to both assemble new information and re-analyse existing public data.

Repetitive DNA was searched in the genome using CENSOR (Jurka et al. 1996), and 20.8 % of the contigs comprised repetitive sequences (Table 2). The majority of the repetitive elements were long interspersed nuclear elements and long-terminal repeat elements (LTRs, 6.4 %). Rahman et al. (2013) estimated that repeat sequences represent ~78 % of the genome, similar to the percentage in barley (84 %) (Schnable et al. 2009), and concluded that for the rubber tree less than 2 % of the total repeat elements are DNA transposons. Many repeat elements (50.24 %) could not be associated with any known family. This difference can be explained because a large percentage of reads of the repetitive regions were abandoned due to being difficult to assemble, and only 0.3 % of the estimated *Hevea* genome was sequenced in this work, which is likely not sufficient to represent the complexity of the genome.

MISA analysis identified 4684 putative microsatellites (Table 2). The most abundant repeat motifs were mononucleotides (1674), followed by dinucleotides (1157) and tetranucleotides (777). NGS clearly offers a rapid means of acquiring the sequences needed to detect SSRs.

SNP calling and molecular marker validation

To detect putative variants in the dataset, we mapped all reads from the GT1 and RRIM701 sequenced genotypes onto the *de novo* assembled contigs (10,071 contigs) using Burrows–Wheeler Aligner (BWA). Approximately 8.2 million reads were successfully mapped, resulting in ~35× depth (Table 3). Using the mapping file as a reference, two different strategies (SAMtools mpileup and Freebayes) were applied to uncover putative SNP variants (Table 4).

A total of 57,820 and 6221 putative SNPs were called by the Freebayes and SAMtools methodologies,

Table 2 Main classes of repeat elements identified in the *H. brasiliensis* genome assembly

Repeat class	Fragments	Total length (bp)
Integrated virus	2145	16,498
DNA transposon	1578	209,909
LTR retrotransposon	2314	509,096
Non-LTR retrotransposon	385	59,409
Transposable element	4297	782,094
Repetitive element		
Simple sequence repeat (SSR)	4684	83,504
Total	15,403	1,660,510

Table 3 Summary of BWA mapping of high-quality reads onto the *de novo* assembled *H. brasiliensis* contigs

Metrics	Results
Total reads	71,971,604
Mapped reads	8,223,980
% of mapped reads	11.42
Both pairs mapped	4,474,291
Singletons	3,749,689
Average depth	35.40X

Table 4 Summary of SNP calling in de novo assembled *H. brasiliensis* contigs

Method/metric	No. of SNPs
Freebayes (1)	57,820
SAMtools (2)	6221
Intersection of (1) and (2)	3779
Transitions (Ts)	2546
Transversions (Tv)	1233
Ts/Tv	2.06
SNP frequency	1 SNP per 2.1 kb

respectively. In search of a more conservative approach, we used the intersection between the two strategies, resulting in a final SNP dataset of 3779 variants, corresponding to an average occurrence of one SNP for every 2.1 kb. Pootakham et al. (2011) and Salgado et al. (2014) detected one SNP per 1.5 kb and one SNP per 5.2 kb for the rubber tree, respectively, using Roche 454 pyrosequencing technology, whereas Mantello et al. (2014) used Illumina sequencing technology and detected an average of one SNP per 125 bp. It is difficult to make a comparison because the methodologies are different with regard to sequencing technique, sequenced material (genomic DNA vs. mRNA) and detection of SNPs.

A total of 2546 transitions and 1233 transversions were detected, with A↔G being the most common variation (1304; 34.48 %) and A↔C the least common variation (263; 6.9 %) observed (Table 4). The ratio of the number of transitions (Ts) to the number of transversions (Tv) is particularly helpful for assessing the quality of SNP calls (DePristo et al. 2011). The Ts/Tv ratio in our study was 2.06, indicating that the conservative approach resulted in higher accuracy regarding variant detection.

Considering the total of 3779 SNPs that were filtered, only those free of other variants within the 50 bp upstream and downstream regions flanking the central base position were selected. Based on this strategy, a total of 143 putative polymorphic positions were selected for further evaluation. Putative SNPs were validated on eight genotypes of *H. brasiliensis* which are parents of currently studied mapping populations (GT1, PB235, RRIC100, PB217, PR255, PB260, RRIM701 and RRIM600) and 15 F1 plants from the cross between PB217 × PR255 (Souza et al. 2013).Thirty (20.9 %) positions presented profiles that could be correctly classified and worked well as potential molecular markers (Additional file 2 Table S1). An example of a correctly classified SNP is the segregation of the locus sHbUNI497 (Fig. 1a), where PB217 was classified as heterozygous (XY), PR255 was classified as homozygous (XX), and the F1 genotypes segregation presented two classes (XX and XY) as expected.

Of the positions, 31 were monomorphic across the samples and 48 were no-call SNPs (failed to amplify). These monomorphic markers may have resulted from errors in sequencing, which then led to the misidentification of the SNP, and some genotyped SNPs may

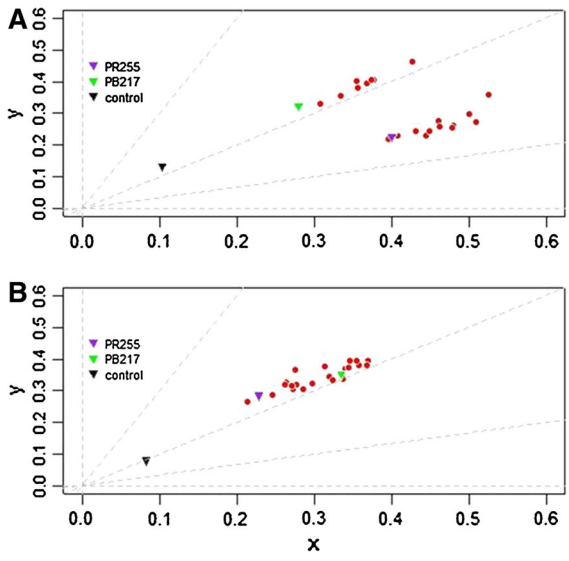

Fig. 1 SNP assays using the Fluidigm platform in parentals of mapping populations and 15 hybrids (*red colour*). **a** Expected segregation of sHbUNI497 and **b** likely duplication segregation of sHbUNI294

have failed due to the presence of occult variants (Bentley et al. 2008) or due to poor PCR amplification and low signal intensities resulting in missing data.

The mean PIC was 0.283, ranging from 0.111 to 0.368. The mean expected heterozygosity was 0.353, ranging from 0.118 to 0.486 (Additional file 2 Table S1). These values are lower than those found by Mantello et al. (2014). However, it should be noted that the present set of markers was analysed on a limited number of genotypes (parentals of mapping populations), while Mantello et al. (2014) screened 36 genotypes from a germplasm bank of *Hevea*.

Approximately 23.77 % (34) of the 143 markers analysed showed a segregation profile in which all analysed genotypes were heterozygotes, and were considered a duplication (Additional file 2 Table S1). All genotypes were heterozygotes (including the F1 genotypes from PB217 × PR255 crossing) at the locus sHbUNI294, which can be observed in Fig. 1b. The same results were obtained by Salgado et al. (2014); ten SNPs displaying the same heterozygous combination for all the genotypes were found. This scenario was not expected, considering the parental origin and because some hybrids of the mapping populations were analysed. Our hypothesis is that this segregation profile was caused by the presence of duplicated regions throughout the *H. brasiliensis* genome. These duplicated regions may have gone through independent mutations over the course of evolution, although they maintain homozygosity in each duplicated locus.

The corresponding reads of these regions were mapped as if they were from the same locus, showing a SNP-type variation in the in silico analysis. In the KASPar technique, both loci would have been amplified; thus, each locus would be responsible for 50 % of the amplicons, which created a heterozygous profile of all genotypes evaluated.

Locus duplication, as revealed by microsatellite molecular markers, has been reported in *H. brasiliensis* and other *Hevea* sp. (Le Guen et al. 2011; Mantello et al. 2012; Silva et al. 2014; Souza et al. 2009); however, such evidence highlights the importance of focusing efforts on elucidating these possible duplications.

Functional annotation and SNP marker characterization

To improve the quality of annotations, we aligned ~94 million paired-end reads from RNA-seq data (Mantello et al. 2014) to the 10,071 contigs using TopHat2 (Kim et al. 2013), which allows gaps in the read-to-reference alignment at putative splice sites. Approximately 3.8 % of the total reads were successfully mapped on the *de novo* assembled genomic regions. Cufflinks was used to aid the refinement of gene structures by creating transcript "fragments" with sharply defined exon boundaries (Trapnell et al. 2009). Cufflinks reconstruction yielded 4374 different transcripts belonging to 3934 assembled genomic contigs.

The total set of contigs was aligned against the cassava (*Manihot esculenta*) complete CDS using sim4 software (Florea et al. 1998). We identified 1447 contigs with predicted coding regions, with at least 70 % of the correspondence in the alignment (Additional file 2 Table S2).

A large number of molecular markers isolated from genomic DNA during the last few decades have been located in the intergenic or genic regions of the genome without any information available on their functions (Varshney et al. 2007). To evaluate the utility of the 143 SNP markers developed, the contig sequences of these markers were manually annotated and compared using the BLAST tool to search against the Uniprot database (www.uniprot.org) and against

the genomes of *Populus trichocarpa*, *Ricinus communis* and *Manihot esculenta* in the Phytozome database (Additional file 2 Table S3).

All contigs containing SNP markers showed significant sequence homology to the draft of the *Hevea* genome (Rahman et al. 2013) (Additional file 2 Table S4). From the 143 SNP markers developed, 86 SNP markers (60.1 %) are inserted in sequences that presented similarities to those encoding proteins with known functions. The majority of these sequences are involved in metabolic processes (27 sequences—e.g.: sHbUNI409), followed by developmental processes (19—e.g.: sHbUNI289, sHbUNI399) and stress response (19—e.g.: sHbUNI371, sHbUNI419). Other identified functions were transport (6—e.g.: sHbUNI380), cell cycle (4—e.g.: sHbUNI495), gene silencing (3—e.g.: sHbUNI377) and respiration (2—sHbUNI403).

sHbUNI402 was identified in contig_4931, which was annotated in Swiss-prot as a probable disease-resistance protein (*Arabidopsis thaliana* At4g27220) (Additional file 2 Table S3). In contig_5088, the SNP marker sHbUNI419 was associated with a disease-resistance protein (*TMV* resistance protein in *Nicotiana glutinosa*).

For the sHbUNI439 marker sequence, there was no matching protein in either the Uniprot or Phytozome databases; nevertheless, the sequence was highly similar to genomic sequences from two *M. esculenta* scaffolds (scaffold05875 and scaffold12525, e-values = 0 and 3.4e−101, respectively) and one scaffold from the *R. communis* genome (29706, e-value = 8e−46). In all three scaffolds the matching sequences were close to transcripts that are similar to telomeric repeat binding protein (TRBP) 1. In *R. communis*, this protein is encoded by a single copy gene, while in *M. esculenta* there are three copies (Phytozome database, as of December 2015). Moreover, *M. esculenta* scaffold12525 presents a structure similar to *R. communis* scaffold29706 around the matching sequence; however, scaffold05875 of *M. esculenta* differs from the previous scaffolds. Because it is a genomic region that is conserved among different species from the same family and is close to transcripts that are similar to the same protein, it is possible that the region in which marker sHbUNI439 is inserted is a cis-element that regulates the expression of a similar protein in *H. brasiliensis*.

Of the SNP markers developed, 34 exhibited a profile which suggested that the loci analysed were duplicated. The sequences being analysed may be members of conserved gene families, or the regions may in fact have been duplicated. For sHbUNI320, the sequence used for the assay design of this marker presented similarity to an Argonaute (AGO) protein. AGO proteins are essential for small RNA silencing pathways, as they bind to different siRNAs and miRNAs and mediate the repression of specific RNAs by degradation or translation inhibition (Höck and Meister 2008). The AGO protein family has several members in different plant species, including *Arabidopsis thaliana* (10), *Populus trichocarpa* (15), *Ricinus communis* (9) and *Manihot esculenta* (13) (Phytozome database, as of December 2015). A possible explanation for the duplicated profile that marker sHbUNI320 presented is that two members of the AGO family were amplified by the same primers.

Likewise, marker sHbUNI324 presented a duplicated profile; however, the sequence in which this SNP was detected presented similarity to that encoding chromosome transmission fidelity 7 (Ctf7) protein, which is a single gene in the majority of plant species (Phytozome database, as of December 2015). Ctf7 is required for the establishment of sister chromatid cohesion, regulation of chromosome segregation and DNA repair. Although Ctf7 is a single-copy gene in most plant genomes, the presence of another copy of this sequence in the rubber tree genome is possible. Another possibility is that the designed primers for the sHbUNI324 marker amplified another HSF1 TFBS region in the *H. brasiliensis* genome.

This study demonstrates that high-throughput DNA sequencing is a powerful approach for the identification of novel sequences and the rapid development of SNP markers in non-model organisms. In addition, new SNP markers possibly involved with different plant molecular mechanisms, such as plant responses to abiotic and biotic stresses and plant developmental processes, were developed. These SNPs are an extremely useful source of markers in rubber tree breeding for marker-assisted selection and gene-based cloning.

The new SNP markers identified in this work provide an excellent tool for the enrichment of genetic linkage maps and the identification of candidate genes for traits of interest (QTL) in progeny from crosses

developed by breeding programmes, as well as for studying genetic diversity in the rubber tree.

Acknowledgments The authors gratefully acknowledge the Fundação de Amparo a Pesquisa do Estado de São Paulo, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Computational Biology Program and Agropolis Program) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico for financial support and scholarships and a research fellowship.

Authors' contributions LMS performed the molecular genetic studies, helped to perform the biocomputational analysis and drafted the manuscript. GTS, CBCS and CCS performed a biocomputational analysis and drafted the manuscript. GTS, ARC, CCM and IAAA assisted in the molecular genetics studies. VLG participated in the evaluations of the molecular data and helped to draft the manuscript. APS conceived the study, participated in its design and coordination and helped to draft the manuscript. All of the authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bachlava E, Taylor CA, Tang S, Bowers JE, Mandel JR, Burke JM, Knapp SJ (2012) SNP discovery and development of a high-density genotyping array for sunflower. PLoS One 7:e29814. doi:10.1371/journal.pone.0029814

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. doi:10.1038/nature07517

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:421. doi:10.1186/1471-2105-10-421

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19–32

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi:10.1093/bioinformatics/btr330

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. doi:10.1038/ng.806

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064. doi:10.1111/j.1365-313X.2005.02591.x

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8:967–974

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. Preprint at arXiv:1207.3907[q-bio.GN]

Gonçalves PS, Fontes JRA (2012) Domestication and breeding of the rubber tree. In: Borém A, Lopes MTG, Clement CR, Noda H (eds) Domestication and breeding: Amazon species. UFV, Viçosa, pp 393–420

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186. doi:10.1093/nar/gkr944

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods 5:183–188. doi:10.1038/nmeth.1179

Höck J, Meister G (2008) The Argonaute protein family. Genome Biol 9:210. doi:10.1186/gb-2008-9-2-210

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–121. doi:10.1016/S0097-8485(96)80013-1

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. doi:10.1186/gb-2013-14-4-r36

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: repbaseSubmitter and censor. BMC Bioinform 7:474. doi:10.1186/1471-2105-7-474

Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). Mol Genet Genomics 270:24–33. doi:10.1007/s00438-003-0891-6

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923

Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) Searching for SNPs with cloud computing. Genome Biol 10:R134. doi:10.1186/gb-2009-10-11-r134

Le Guen V, Gay C, Xiong TC, Souza LM, Rodier-Goud M, Seguin M (2011) Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*). Plant Breed 130:294–296. doi:10.1111/j.1439-0523.2010.01774.x

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

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

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

Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, de Souza Gonçalves P, Vicentini R, de Souza AP (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

Pootakham W, Chanprasert J, Jomchai N, Sangsrakru D, Yoocha T, Therawattanasuk K, Tangphatsornruang S (2011) Single nucleotide polymorphism marker development in the rubber tree, *hevea brasiliensis* (Euphorbiaceae). Am J Bot 98:e337–e338. doi:10.3732/ajb.1100228

Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S (2012) The cassava genome: current progress, future directions. Trop Plant Biol 5:88–94

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100. doi:10.1016/S1369-5266(02)00240-6

Rahman AY, Usharraj AO, Misra BB, Thottathil GP, Jayasekaran K, Feng Y, Hou S, Ong SY, Ng FL, Lee LS, Tan HS, Sakaff MK, Teh BS, Khoo BF, Badai SS, Aziz NA, Yuryev A, Knudsen B, Dionne-Laporte A, Mchunu NP (2013) Draft genome sequence of the rubber tree *hevea brasiliensis*. BMC Genomics 14:75. doi:10.1186/1471-2164-14-75

Salgado LR, Koop DM, Pinheiro DG, Rivallan R, Le Guen V, Nicolás MF, de Almeida LG, Rocha VR, Magalhães M, Gerber AL, Figueira A, Cascardo JC, de Vasconcelos AR, Silva WA, Coutinho LL, Garcia D (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

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534

Silva CC, Mantello CC, Campos T, Souza LM, Gonçalves PS, Souza AP (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 34:1035–1053. doi:10.1007/s11032-014-0095-2

Souza LM, Mantello CC, Santos MO, de Souza Gonçalves P, Souza AP (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. doi:10.1007/s12686-009-9018-7

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

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422. doi:10.1007/s00122-002-1031-0

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. doi:10.1093/bioinformatics/btp120

Triwitayakorn K, Chatkulkawin P, Kanjanawattanawong S, Sraphet S, Yoocha T, Sangsrakru D, Chanprasert J, Ngamphiw C, Jomchai N, Therawattanasuk K, Tangphatsornruang S (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

Varshney RK, Beier U, Khlestkina EK, Kota R, Korzun V, Graner A, Börner A (2007) Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency, and applications for genome mapping and diversity studies. Theor Appl Genet 114:1105–1116. doi:10.1007/s00122-007-0504-6

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530. doi:10.1016/j.tibtech.2009.05.006

You FM, Huo N, Deal KR, Gu YQ, Luo M-C, McGuire PE, Dvorak J, Anderson OD (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genomics 12:59. doi:10.1186/1471-2164-12-59