plant biology

Plant Biology ISSN 1435-8603

RESEARCH PAPER

Molecular evolution and functional characterisation of haplotypes of an important rubber biosynthesis gene in *Hevea brasiliensis*

T. K. Uthup, A. Rajamani, M. Ravindran & T. Saha

Genome Analysis Laboratory, Rubber Research Institute of India, Kottayam, Kerala, India

Keywords
Haplotype structure; *Hevea brasiliensis*; HMGCoA synthase; isoprenoid biosynthesis; SNPs.

Correspondence
Thomas K. Uthup, Genome Analysis Laboratory, Rubber Research Institute of India, Rubber Board PO, Kottayam 686009, Kerala, India.
E-mail: thomasku79@gmail.com

Editor
U. Wittstock

Received: 26 August 2015; Accepted: 12 January 2016

doi:10.1111/plb.12433

ABSTRACT

Hydroxy-methylglutaryl coenzyme-A synthase (HMGS) is a rate-limiting enzyme in the cytoplasmic isoprenoid biosynthesis pathway leading to natural rubber production in *Hevea brasiliensis* (rubber). Analysis of the structural variants of this gene is imperative to understand their functional significance in rubber biosynthesis so that they can be properly utilised for ongoing crop improvement programmes in *Hevea*. We report here allele richness and diversity of the HMGS gene in selected popular rubber clones. Haplotypes consisting of single nucleotide polymorphisms (SNPs) from the coding and non-coding regions with a high degree of heterozygosity were identified. Segregation and linkage disequilibrium analysis confirmed that recombination is the major contributor to the generation of allelic diversity, rather than point mutations. The evolutionarily conserved nature of some SNPs was identified by comparative DNA sequence analysis of HMGS orthologues from diverse taxa, demonstrating the molecular evolution of rubber biosynthesis genes in general. *In silico* three-dimensional structural studies highlighting the structural positioning of non-synonymous SNPs from different HMGS haplotypes revealed that the ligand-binding site on the enzyme remains impervious to the reported sequence variations. In contrast, gene expression results indicated the possibility of association between specific haplotypes and HMGS expression in *Hevea* clones, which may have a downstream impact up to the level of rubber production. Moreover, haplotype diversity of the HMGS gene and its putative association with gene expression can be the basis for further genetic association studies in rubber. Furthermore, the data also show the role of SNPs in the evolution of candidate genes coding for functional traits in plants.

INTRODUCTION

Hevea brasiliensis (Euphorbiaceae) is the chief source of natural rubber, a critically strategic industrial raw material that is indispensable for the production of a wide range of products having diverse applications, ranging from healthcare to space science. Natural rubber is practically pure poly-cis-1,4 isoprene. It is synthesised in the latex vessels of rubber trees by the mevalonate or 2-C-methyl-D-erythritol 4-phosphate pathway (Chow *et al.* 2012). The limitations in conventional breeding practices to increase rubber productivity prompted investigations into the underlying molecular mechanisms of natural rubber biosynthesis, leading to identification of several important genes and enzymes involved (Tang *et al.* 2013). 3-Hydroxy-3-methylglutaryl coenzyme-A synthase (HMGS) is an important gene in the mevalonate pathway, which catalyses the condensation of acetyl-CoA and acetoacetyl-CoA to form HMG-CoA. This HMG-CoA acts as the substrate for HMG-CoA reductase (HMGR) to yield mevalonate, which is further converted to different isoprenoid compounds, e.g. growth regulators, chlorophyll, phytoalexins, natural rubber, etc. (Suwanmaneea *et al.* 2002). The involvement of HMGS and HMGR in the early steps of rubber biosynthesis and the positive correlation between their activity and the dry rubber content of the latex from rubber trees is well established (Wititsuwannakul 1986; Suvachittanont & Wititsuwannakul 1995). Even though clear-cut evidence at the gene level is not available so far to link expression with the quality and quantity of latex produced, the rate-limiting role of these two enzymes in the mevalonate pathway and results from the aforementioned studies prompted us to carry out further investigations.

Assessment of genetic diversity using molecular markers is required, not only for crop improvement programmes but also for suitable management and conservation of plant genetic resources in gene banks (Varshney *et al.* 2007). Among the different types of molecular marker used, single nucleotide polymorphisms (SNPs) have gained popularity recently because of their abundance and ease of automation. Furthermore, an SNP is of great importance if it affects gene function, thereby directly influencing the phenotype in terms of yield or stress response. Such information can be used in breeding programmes to develop better varieties using approaches such as marker-assisted recurrent selection (MARS) or gene pyramiding. In *Hevea*, the advent of next generation sequencing platforms significantly broadened the scope of SNP markers through identification of thousands of SNPs from RNA

sequencing and whole genome sequencing projects, which may aid in genome-wide association studies (Pootakham *et al.* 2015; Shearman *et al.* 2015). A more focused approach to examine potential SNP markers directly linked to a specific phenotype is the candidate gene-based method, where in depth sequence structure analysis of individual genes and their alleles is performed. An example of this approach in rubber is the study on structural variants of the farnesyl diphosphate synthase gene in *Hevea* reported recently (Uthup *et al.* 2013). In order to comprehend the functional significance of all rubber biosynthesis genes and to use them for crop improvement programmes in *H. brasiliensis*, extensive SNP information from all major genes in this pathway is essential. The present study is an attempt to understand the sequence structure and function of HMGS in *H. brasiliensis* clones.

Prior studies in *H. brasiliensis* revealed that there are two isoforms of HMGS, namely HMGS1 and HMGS2, having high similarity to each other, which might have evolved recently through gene duplication (Sirinupong *et al.* 2005; Sando *et al.* 2008). Similar mRNA expression profiles of HMGS1 and HMGS2 obtained from semi-quantitative PCR by this group further confirmed this assumption. In the present study, the allelic diversity of the candidate gene HMGS involved in rubber biosynthesis was assessed in selected popular *Hevea* clones, followed by an in-depth structural analysis of the various haplotypes. SNP markers linked to this gene were also developed with the ultimate aim of using them for crop improvement in *Hevea*. We detected high genetic variation within the HMGS gene and established the role of intragenic recombination in the induction of these variations using segregation analysis of specific loci. The downstream impact of the non-synonymous SNPs at the protein level was also estimated with computational methods.

MATERIAL AND METHODS

Phylogenetic analysis

A phylogenetic tree was constructed using HMGS protein sequences, downloaded from the NCBI database, to understand their sequence diversity and evolutionary relationships. One or two representative sequences, each from organisms from major domains of life other than plants, were taken for the analysis. In the case of plants, HMGS sequences from related as well as distant genera/family/species were considered for a relationship study with *Hevea* sequences. A phylogenetic tree was constructed using the online software Phylogeny.fr (http://www.phylogeny.fr/index.cgi) with the John Taylor Thornton substitution model and bootstrap value of 500. Details of the sequences used for the analysis are provided as Fig. S1.

Initial sampling, PCR amplification and re-sequencing of full-length HMGS gene

The initial sample material consisted of five genetically diverse popular *Hevea* clones having divergent yield patterns: RRII-105, RRII-118, RRIM-600, RRIC-52 and GT1. These clones, which are cultivated extensively in the Asia-Pacific region, were selected for this study with the objective of identifying SNPs from genes involved in the rubber biosynthesis pathway. Leaf genomic DNA was isolated following the CTAB protocol (Doyle & Doyle 1990). Three sets of overlapping primers spanning the entire gene were designed initially based on the complete cDNA sequence information of the HMGS gene of around 1.8 kb (Genbank Acc No: AB294688.1). Gap filling was done using additional primers (list of primers used is given in Table S1). PCR amplification was performed as per standard protocol using AdvantageTaq (Clontech, Mountain View, CA, USA). The amplified products were checked on 1% agarose and gel bands of interest were eluted using the Illustra GFX gel band purification kit (GE Healthcare, Little Chalfont, UK). Purified products were TA-cloned in pGEMT easy vector (Promega, Madison, WI, USA). Direct sequencing of the PCR products, as well as sequencing of the cloned products (multiple colonies), was performed using the ABI 3500xl sequencing system.

Additional sampling and sequencing of the putative variable region

Since the five genotypes could easily be differentiated on the basis of allele status of six highly heterozygous SNPs (locus 3059–3817) from a 758-bp region, variability in 14 additional clones of different parent hybrid combinations was analysed to check the possibility of developing this sequence as a bar-coding region to differentiate popular clones. In addition to information on distribution and inheritance patterns of these SNPs in Wickham clones, the results could also estimate whether the SNPs occurring in this region were induced by reciprocal recombination (crossing over) or through gene conversion (non-reciprocal, which does not follow Mendelian inheritance). Details of all clones studied are provided in Table S2. The region of interest was amplified from all the clones using sequence-specific flanking primers HbHMGS-Var-F-5′-CTACCTCA TGGCTCTTGATTCC-3′and HbHMGS-Var-R-5′-AGGACTA AGCCCTTATGTTGCAT-3′. The products were gel-purified and subjected to direct sequencing as well as sequencing by cloning.

Primary sequence data analysis

The DNA sequences were aligned using the multiple sequence alignment module of DNASIS MAX (Hitachi Solutions, San Bruno, CA, USA). Gaps as well as SNPs were identified from the aligned sequences and later confirmed by checking the chromatograms. Heterozygous SNPs were identified as double peaks on the chromatogram and through cloning and sequencing of plasmids that harbour the respective alleles.

Haplotype reconstruction was carried out using DnaSP 4.0 software using the SNP information generated (Librado & Rozas 2009). The minimum number of recombination events (RM) was calculated using the four gamete test. Statistical analysis of SNP polymorphism and haplotype reconstruction was also carried out using Haploview (Barrett *et al.* 2005) and SNiPlay (http://sniplay.southgreen.fr/cgi-bin/home.cgi; Dereeper *et al.* 2011). In order to identify the conserved nature of non-synonymous SNPs during HMGS gene evolution in plants, a comparative study with respect to *Hevea* haplotypes was undertaken using the DNA and protein sequence of various plant species included in the phylogenetic analysis. The haplotypes Hap_1, Hap_2, Hap_4, Hap_6 and Hap_8 were selected based on their unique combination of six non-synonymous

SNPs. The nucleotide as well as amino acid sequences derived from these haplotypes were aligned using the software 'muscle 3.8' to identify amino acid changes resulting from the six non-synonymous SNPs. Changes other than those present in the *Hevea* haplotypes were not counted.

Sequence analysis of the putative variable region

The 19 clones were broadly classified based on their geographic region of origin as Indian, Sri-Lankan and Southeast Asian (Malaysia/Indonesia). DnaSP 4.0 and SNiPlay software was used to reconstruct the haplotypes. The software DARwin version 5 (Perrier *et al.* 2003) was used to build a phylogenetic tree applying the unweighted neighbour-joining method to visualise genetic relationships among the detected haplotyes. The significance of each node was evaluated by bootstrapping data for 1000 replications of the original matrix.

Segregation analysis of an SNP locus from the variable region

Genotyping of full-sib progeny (F1 progeny) of 46 individuals derived from a controlled cross between the cultivars RRII-105 and RRII-118 was performed using the SNP HbHMGS3059AG to assess mode of inheritance of the haploblock in which it resides and to identify the role of genetic recombination in induction of these SNPs. The female parent of the cross, RRII-105, was heterozygous (A/G) while the male parent, RRII-118, was homozygous (G/G). Genotyping was performed using the high-resolution melt (HRM) analysis technique (Lochlainn *et al.* 2011). The HRM analysis was performed using the Type-it HRM™ PCR kit (Qiagen, Crawley, UK) following the manufacturer's instructions. Both parents were included in triplicate so as to be considered as melting standard. The LightScanner Data Analysis software (version 2.0; Idaho Technology, Boise, ID, USA) was used to analyse the data and to produce normalised disassociation curves and difference plots.

Relative gene expression studies

Gene expression patterns were studied in five popular rubber clones to identify putative associations between haplotypes and HMGS expression. For RNA isolation, fresh latex was collected in the early morning from 25- to 30-year-old mature tapping trees of the five clones having comparable girth. The collected latex was mixed with an equal volume of RNA extraction buffer (Chang *et al.* 1993) and immediately transferred to the lab in ice. Total RNA was isolated as described in the pine tree method (Chang *et al.* 1993). After DNAse treatment, the integrity and concentration of the RNA was checked on 0.8% agarose gels followed by quantification using NanoDrop (Thermo Scientific, Waltham, MA, USA). An aliquot of 500 ng DNAse-treated RNA was used for cDNA synthesis using Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA, USA) following the manufacturer's instructions. Relative gene expression analysis was carried out using the primer pairs QHMGS1-F-5′-TCTATGCCCAGAAGGCTGTTG-3′ and QHMGS1-R-5′-TCCTGGCATGCTACATGACTTCC-3′ on a Light Cycler 480 II Real Time PCR system (Roche, Basel, Switzerland). qPCR was performed in a 15-µl reaction mixture using 30 ng DNA and 133 nM primers, followed by melt curve analysis. Each PCR reaction was performed in triplicate. Actin was used as endogenous control for the qPCR analysis. The relative quantification (RQ) values were analysed using Light Cycler 480 software, and the expression rate of genes was represented as fold change in transcript level normalised to the actin gene, relative to that in RRII 105 plants.

Comparative protein structural conformation analysis

The major aim of the homology modelling study was to verify whether the non-synonymous SNPs identified in the HMGS gene from *Hevea* genotypes lead to structural modifications in the active site of HMGS. Since no crystal structure was available for HMGS from *Hevea*, the structure was constructed using the SWISS-MODEL server (Schwede *et al.* 2003). Six non-synonymous SNPs were identified at positions 172 (N → D), 176 (V → A), 276 (A → V), 308 (P → L), 316 (P → S) and 383 (T → A). Protein with these SNP variations at all combinations observed in the different haplotypes was subjected to molecular modelling using the crystal structure (PDB id: 2FA3) of HMGS from *Brassica juncea* as template. Since the structure complexes with acetyl-CoA, the impact of SNPs on this binding site was examined in detail using Deep View-spdbv 3.7 (Guex & Peitsch 1996) and chimera-1.10.1 (Pettersen *et al.* 2004).

RESULTS

Phylogenetic tree

The hierarchical linkage of the HMGS gene across major kingdoms and divisions was apparent from the phylogenetic tree constructed based on the amino acid sequences (Fig. 1). The organisms from different strata of life were clearly differentiated based on their amino acid sequence identity. The *H. brasiliensis* HMGS sequence was clustered along with its orthologues in species such as *Ricinus communis* and *Populus trichocarpa*, which formed a subgroup within the major cluster of the plant kingdom. Cereals formed another subgroup within this cluster. Bacterial species including Archae, Proteo- and Acinetobacter formed a separate group. Fish, birds and animals, including humans, formed another major cluster.

Identification of SNPs and haplotype structuring of the full-length HMGS gene

Twenty bi-allelic SNPs were detected from the 5437 bp sequenced genomic region of HMGS from five popular rubber clones, resulting in an average of one SNP every 272 bp. Characteristics of the 20 SNPs across the studied popular clones are listed in Table S3. Among the 20 SNPs, seven were located in the coding region. The SNP HbHMGS2167AG resulted in a change in the encoded amino acid (aspartic acid to asparagine), SNP HbHMGS2180CT (alanine to valine), SNP HbHMG S3513CT (alanine to valine), SNP HbHMGS4052CT (proline to leucine), SNP HbHMGS4075CT (proline to serine) and SNP HbHMGS4427AG (alanine to threonine). SNP HbHMG S4345CT was synonymous in nature. The remaining 13 SNPs were from intronic regions. Furthermore, 14 SNPs led to

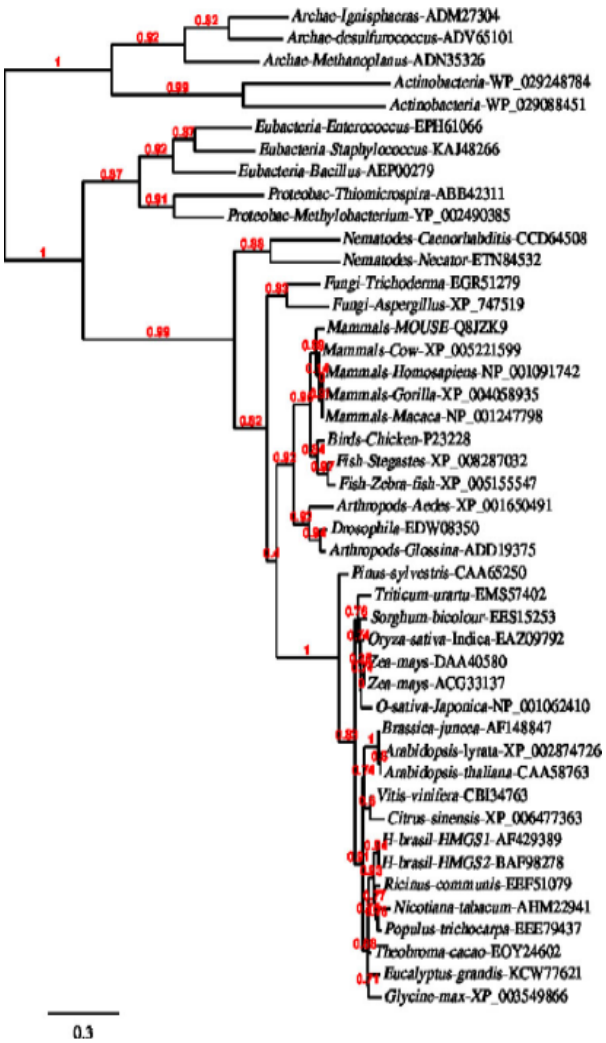

Fig. 1. HMGS phylogenetic tree. Phylogenetic tree constructed using the amino acid sequences of HMGS from selected plant species and representative organisms belonging to major domains of life. The tree depicts the evolutionary relationship of the HMGS gene across different organisms.

transitions and six SNPs to transversions. Nucleotide diversity (Pi) was 0.00119. Out of the 20 SNP loci, 11 were single variable sites, whereas the remaining nine were parsimony informative sites. The distribution of 20 SNPs in the five clones is depicted as a Venn diagram (Fig. S2). Haplotype re-construction using DnaSP yielded eight haplotypes. Haplotypes with their allele frequency in each clone are shown in Table 1. Three recombinations were identified between the sites: (2010, 2704) (2704, 3059) and (3766, 3817). Full-length HMGS genomic sequences from the five clones were submitted to Genbank, with the accession numbers KM272629 and KT447224 to KT447227. The SNPs reported above were submitted to the dbSNP database, with the submitted SNP numbers (SS No.) 1839574848–1839574867.

Comparative analysis of HMGS protein sequences in diverse plant species with that of *Hevea* haplotypes revealed that the region consisting of non-synonymous SNPs, HbHMGS2167AG and HbHMGS2180CT, was highly conserved across different phyla and classes. Conversely, the remaining four SNPs were within partially conserved regions. DNA and protein multiple sequence alignments are shown in Figs S3 and S4. The amino acid changes resulting from the six non-synonymous SNPs in *Hevea* haplotypes and selected plant species are shown in Fig. S5.

Table 1. Haplotypes with their respective frequency in five *Hevea* genotypes.

haplotype name	haplotype	frequency in clones
HAP_1	CTGCTCTATACTTGGCCCGA	[RRII-105-1]
HAP_2	GTATCCTATGCTGATCCCGA	[RRII-105-2, RRII-118-1]
HAP_3	GTGCTCTGTGCTGAGCCCGA	[RRII-118-2]
HAP_4	CTGCTATATACCTGGCTTGA	[RRIM-600-1]
HAP_5	GTGCTCTATACTGAGCCCGA	[RRIM-600-2, GT1-2]
HAP_6	GTGCTATATATTGGTTCCGA	[RRIC-52-1]
HAP_7	GAGCTATATGCTGAGCCCGA	[RRIC-52-2]
HAP_8	GTATCCCAAGCTGATCCCAT	[GT1-1]

Sequence data analysis of the variable region

The 754 bp variable region comprised four small exons. The six SNPs falling within this region are HbHMGS 3059AG, HbHMGS3513CT, HbHMGS3598CT, HbHMGS 3734GT, HbHMGS3766AG and HbHMGS3817GT. Apart from the non-synonymous SNP HbHMGS3513CT, all others were of intronic origin. No new SNPs were identified from the additional 14 clones sequenced. The allelic status of the six SNPs in 19 popular clones with parental details is shown in Table 2. A chromatogram showing the allele status of the loci HbHMGS3059AG in one parent hybrid combination is shown in Fig. S6. The estimated haplotype (gene) diversity and nucleotide diversity using DnaSP was Hd 0.677 and Pi 0.00236, respectively. Haplotype reconstruction using DnaSP resulted in seven types, which are shown in Table 3, together with the genotypes representing them. In the haplotype-based phylogenetic tree constructed, Hap_3_Var formed a separate branch (Branch 3) evolutionarily distant from the other two branches (Fig. S7). Branch 1 had the haplotypes Hap_1_Var, Hap_5_Var and Hap_7_Var; Branch 2 had the major haplotypes Hap_2_Var, Hap_4_Var and Hap_6_Var. The geographic origin distribution of the six SNPs in the 19 clones is shown in Fig. 2.

Segregation analysis with the HRM genotyping technique

The segregation analysis of the locus HbHMGS3059AG clearly showed the presence of two genotypes in a 1:1 ratio as expected. Based on the melting curve analysis, clear differentiation was obtained between the AG and GG genotypes (Fig. S8). A total of 24 progeny were genotyped as AG and the remaining 22 as GG. The expected and observed frequencies for the alleles and the genotypes, together with chi-square values, are shown in Table 4.

Relative gene expression studies

The RRII 105 plants showed a five-fold or more increase in HMGS expression than that in the other four clones. The

Table 2. Allelic status of six SNPs within the putative variable region of 19 popular *Hevea* clones.

genotype no	parents	3059	3513	3598	3734	3766	3817
RRII-105	Tjir-1x Gl-1	AG	CC	TT	GT	AG	GT
Tjir-1		AG	CC	TT	GT	AG	TT
Gl-1		AG	CC	TT	GT	AG	GT
RRII-203	PB86 x MIL3/2	AG	CC	TT	GT	AG	GT
PB-86		AG	CC	CT	GG	AA	GT
MIL3/2		AG	CC	TT	GT	AG	GT
RRIM-600	Tjir-1x PB86	AA	CC	CT	GT	AG	GG
RRII-414	RRII-105 X RRIC-100	AG	CC	TT	GT	AG	GT
RRII-422	RRII-105 X RRIC-100	AG	CC	TT	GT	AG	GT
RRII-429	RRII-105 X RRIC-100	GG	CC	TT	GG	AA	TT
RRII-430	RRII-105 X RRIC-100	AG	CC	TT	GT	AG	GT
RRIC-100		GG	CC	TT	GT	AA	TT
RRIC-104	Tjir-1x RRIC-52	AG	CT	TT	GG	AG	GT
PB-314	RRIM-600 X PB-235	AG	CC	TT	GG	AA	GT
PB-235		GG	CC	TT	GG	AA	TT
RRIC-52		AG	CT	TT	GG	AG	TG
RRII-118		GG	CC	TT	GG	AA	TG
PR-107		AG	CC	TT	GT	AG	GT
GT-1		AG	CT	TT	GG	AA	TG

Allelic status of six SNPs in the selected parents confirms the identity of hybrids. The numbers on top denote the respective position of each SNP loci in the gene.

expression levels in clones RRII 118, RRIM 600 and RRIC 52 were comparable, whereas expression in GT1 plants was much lower (Fig. 3). The qPCR raw data file is included as Table S4.

Comparative protein structural conformation analysis

The HMGS protein structure from *B. juncea* in complex with acetyl-CoA (2FA3) was identified as the best available template for the modelling study. Interestingly, this structure shared over 85% sequence identity with HMGS from *Hevea*. Binding site analysis of 2FA3 revealed that the amino acids Ser 31, Lys 32, Pro 155, Tyr 151, His 247, Lys 256 and Arg 296 are part of the binding cavity. Analysis of each SNP with respect to the location of the affected amino acid residue in HMGS revealed that they do not have any direct interaction with the acetyl-CoA binding site, and none of them was found inside or near the protein–small molecule interaction interface of HMGS. The three-dimensional protein structure showing the estimated location of the amino acids affected by six non-synonymous SNPs is depicted in Fig. 4.

DISCUSSION

Comparative analysis of HMGS protein and nucleic acid sequences

Due to the highly dynamic nature of plant nuclear genomes, they vary considerably in structure and size, leading to the evolution of new alleles and species at micro and macro levels, respectively. The phylogenetic tree constructed using HMGS protein sequences can be considered as an indirect representation of evolutionary changes at gene level that

Table 3. Haplotypes re-constructed using the six SNPs from the variable region in *Hevea* with their respective frequencies in 19 clones.

HAP-1_Var ACTTGG	HAP_2_Var GCTGAT	HAP-3_Var ACCGAG	HAP_4_Var GCTTAT	HAP_5_Var ATTGGG	HAP_6_Var GCTGAG	HAP_7_Var ATTGAG
RRII-105-a	RRII-105-b	PB-86-b	RRIC-100-b	RRIC-104-b	RRII-118-b	GT-1-b
RRII-203-a	RRII-203-b	RRIM-600-b		RRIC-52-b		
Tjir-1-a	Tjir-1-b	PB-314-b				
GL-1-a	GL-1-b					
Mil-a	Mil-b					
RRII-414-a	RRII-414-b					
RRII-422-a	RRII-422-b					
RRII-430-a	RRII-430-b					
RRIM-600-a	RRII-429-a					
PR-107-a	RRII-429-b					
	PR-107-b					
	RRIC-100-a					
	RRIC-104-a					
	PB-86-a					
	PB-314-a					
	PB-235-a					
	PB-235-b					
	RRIC-52-a					
	RRII-118-a					
	GT-1-a					

a and b represent the two chromosomes of the diplod *Hevea brasiliensis*.

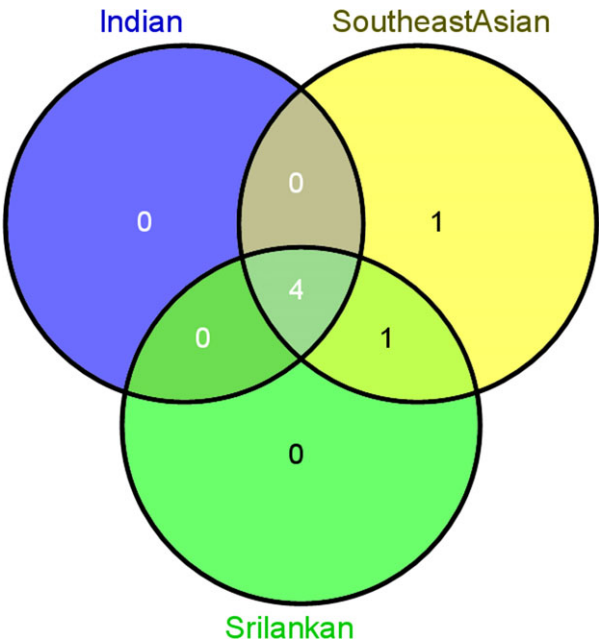

Fig. 2. Distribution of six variable region SNPs. The 19 popular *Hevea* clones were grouped into three based on their geographic origin as Indian, Southeast Asian (Malaysian or Indonesian) and Sri Lankan. The Venn diagram shows the unique as well as common SNPs among these regions.

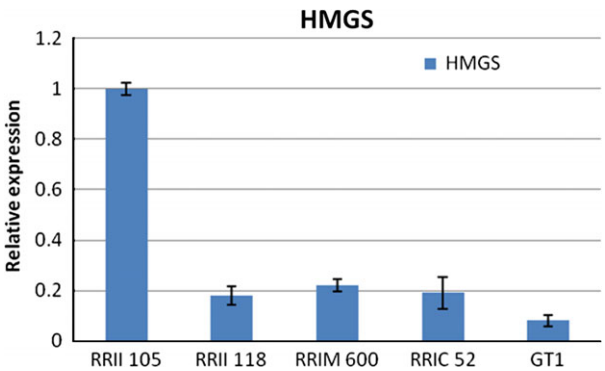

Fig. 3. Relative gene expression levels. Graph shows the fold change in transcript level in five clones normalised to the actin gene, relative to that in RRII 105 plants. RRII 105 showed around five-fold increase in HMGS expression than the other four clones.

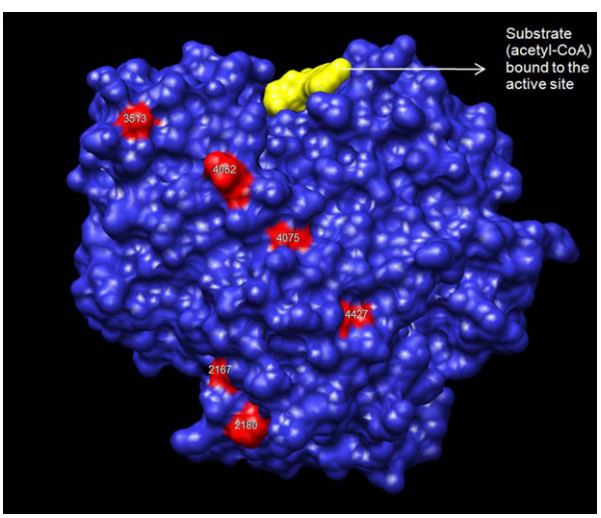

Fig. 4. Predicted three-dimensional structure of *Hevea* HMG-CoA synthase. The three-dimensional structure of *Hevea* HMG-CoA synthase bound to its substrate (acetyl-CoA). Amino acid positions that are affected by non-synonymous substitutions are highlighted. Numbers refer to position of substitution in the corresponding nucleotide sequence.

Table 4. Segregation ratios of two alleles and genotypes in a *Hevea* progeny population.

	parents		
	RRII-105 (A/G) × RRII-118 (G/G)		
	allele frequency		
alleles	expected frequency	observed frequency	contribution to chi- square
A	23 (25%)	24 (26.0%)	0.04
G	69 (75%	68 (73.9%)	0.01
Total			0.05 (not significant)
	genotype frequency		
genotypes	expected frequency	observed frequency	contribution to chi-square
AG	23 (50%)	24 (52.1%)	0.04
GG	23 (50%)	22 (47.8%)	0.05
Total			0.09 (not significant)

Allele and genotype frequencies of SNP HbSNP3059 (A/G) in a full sib family progeny of size 46. The alleles 'A' and 'G' and the genotypes 'AG' and 'GG' showed no statistically significant deviation from Mendelian inheritance.

occurred in the organisms over millions of years, and enabled this enzyme to participate in diverse pathways (Boucher & Doolittle 2000; Lange *et al.* 2000). The clustering and branching of the tree also represents the evolutionary ladder of life, from lower simple organisms to higher more complex forms. Correct assessment of this diversity at genome level is possible only through comparative genetic mapping and synteny analysis, which could reveal conservation of gene content, order and function among closely related taxa, as well as distantly related organisms (Bennetzen 2000). Recently, microsynteny has been investigated across several plant species using whole-genome sequences and selected discrete sequences to infer shared ancestry (Dohm *et al.* 2009). Although not as extensive as the above studies, comparative DNA or protein sequence analysis using orthologous sequences from different species is an efficient way to understand the functional significance, evolution and distribution of nucleotide variations at the gene level. For example, Jung *et al.* (2012) reported the existence of conserved as well as phylum- or group-specific SNPs in gene sequences among ciliates. Although not specific to any phylum, the wide distribution of the coding SNPs noted in *Hevea* HMGS also provides valuable information on the evolution of this gene in plants. For example, the presence of SNP HbHMGS2167AG across plant species belonging to different phyla and classes is a good indication of their separate lines

of differentiation from a common ancestor. With regard to HbHMGS4427AG (alanine to threonine), the threonine-coding version seems to be the primitive allele that is still retained at low frequency in *Hevea*, as seen in Hap_8 and in *Pinus*. Similarly, valine in the case of HbHMGS3513CT was noted only in *Nicotiana*. In contrast, HbHMGS4052CT (proline to leucine) and SNP HbHMGS4075CT (proline to serine) were assumed to have evolved after the speciation of *Hevea*, since these variants were unique to *Hevea*. It is these SNPs that may aid in differentiating *Hevea* clones, whereas the remainder hold evolutionary information of the HMGS gene in plants.

A reduction in genetic diversity is generally expected in cultivated plants rather than their wild counterparts due to breeding practices followed during domestication (Li *et al.* 2011). However, significant genetic diversity in popular rubber clones has been reported in molecular genetic studies (Le Guen *et al.* 2011; Roy *et al.* 2012). Based on recent findings from Shearman *et al.* (2015), the frequency of single nucleotide substitutions found in the rubber tree transcriptome was approximately one in 270 nucleotides. Even though not at the genome level, SNP density of one for every 272 bp in the HMGS gene reported here is congruent with this report. In the case of non-synonymous SNPs, rare alleles like the 'T' at position 3513 in HMGS had low frequency in a population and were considered non-functional because of the heterozygous condition, apparently having less downstream impact on gene expression. This phenomenon is supported by earlier reports from Koehn & Eanes (1976) on a strong correlation between rarer alleles and their heterozygous status, as well as in later studies of Chen & Sun (2013) and Toka *et al.* (2013) on the effects of rare mutations on low allele frequencies in a population.

Sequence diversity and haplotype analysis of the HMGS gene in *Hevea*

A major contributing factor for genetic diversity in domesticated plants is genetic recombination resulting from selective hybridisation during crop improvement (Watt 1972). This is well proved in studies in R genes of *Arabidopsis* and the self-incompatibility locus of *Petunia inflata* (McDowell *et al.* 1998; Wang *et al.* 2001). Also in the case of *Hevea* HMGS, the three intragenic recombination sites predicted within the gene might have substantially contributed towards the high variability. Although massive vegetative propagation through bud grafting may induce somatic mutations (Fournier-Level *et al.* 2010), this alternative can be partially ruled out here since the origin of SNPs within a single haploblock could be traced in the parents based on study of parental combinations discussed elsewhere. Haplotype analysis in HMGS revealed that the incidence of any one haplotype in more than one clone was only apparent in the case of Hap_2 and Hap_5, where they were distributed among clones belonging to the same geographic origin. Interestingly, despite the significant differences in parentage of RRII 105 and RRII 118 they shared one haplotype, indicating a single collection point by Sir Henry Wickham in 1876 in Amazon forests. The Venn diagram depicting distribution of 20 SNPs in five genotypes shows that the Southeast Asian clones possessed the maximum number of unique SNPs compared to the Indian and Sri Lankan clones. Alternatively, the uniqueness of RRIC 52 may be attributed to its evolution through ortet selection from seedlings not shared by Sri Lanka to other countries. The existence of genetic diversity within the existing Wickham population is further highlighted by the presence of just three common SNPs out of 20.

Sequence diversity of the putative hypervariable region within the HMGS gene

The number of SNP haplotypes identified in a population may vary based on the length of the sequence analysed and the size of the population. Also in the case of variable region, when population size increased, the frequency of haplotypes like Hap_1_Var and Hap_2_Var increased significantly, thereby reducing the total number of haplotypes (Table 3). Therefore, the possibility of developing this region as a bar-coding sequence for the differentiation of popular rubber clones may not be feasible. However, the haplotype information of this small region provided valuable inputs on the variability as well as inheritance pattern of haplotypes prevailing among popular rubber clones in India. For example, except for RRII 118, all Indian clones shared Hap_1_Var and Hap_2_Var, indicating their low variability compared to clones from the other two groups. Similarly, the most popular Malaysian clone, RRIM 600, appears to be unique since it did not have Hap_2_Var whereas all the other clones possessed it. Similarly, the complete homozygosity in all six loci observed only in PB235 may be due to the genetic relatedness of its Malaysian parents (PB-5/51 9 PB S/78).

From the haplotype-based phylogram generated (Fig. S7), it appears that high frequency haplotypes like Hap_1_Var and Hap_2_Var evolved separately, forming two distinct clusters, and the selective breeding of popular rubber clones inadvertently resulted in the mixing of these two distinct haplotypes as they are seen together in the majority of the popular clones. Similarly, the presence of Hap_3_Var exclusively in the three Southeast Asian clones may be attributed to the direct genetic relationship of Malaysian clones RRIM 600, PB 314 and PB 86. This is also evident from the unique distant branch formation of Hap_3_Var in the phylogram. Likewise, the relation between RRIC52 and its offspring RRIC 100 and RRIC 104 may be responsible for the uniqueness of Hap_4_Var and Hap_5_Var observed only among Sri Lankan clones. Despite the high allelic frequencies of intronic SNPs in this region, the low incidence of the minor 'T' allele of the non-synonymous HbSNP3513CT in 19 popular clones shows that coding regions in *Hevea* are also more conserved and less prone to mutations than non-coding regions such as the SNP pattern within conserved structures in the human genome (Bejerano *et al.* 2004).

Intragenic recombination is known to contribute to the generation of allelic diversity and serves as the primary source of genetic diversity (Van der Hoorn *et al.* 2001; Wang *et al.* 2001). One predicted recombinant event between positions 3766–3817 indicates that the same process may be responsible for the sequence diversity and high heterozygosity rate present in the neighbouring loci. Furthermore, segregation analysis of the locus HbHMGS3059AG showed no statistically significant deviation from Mendel's law of inheritance for alleles and genotypes ($P > 0.10$, 0.05, respective;y). The results further suggest the role of intragenic recombination in triggering allele induction rather than somatic mutations, which does not fol-

low Mendelian inheritance. Since the neighbouring SNPs, except for HbSNP3817, fall under a single haploblock, the same explanation is also applicable to these loci. Together, the sequence data exclusively from this region in 19 clones suggest that despite their relationship to each other, there is still rich genetic diversity.

Downstream impact of non-synonymous SNPs

Association between haplotypes and gene expression is well established in plants and animals (Kelly *et al.* 2013). For example, He *et al.* (2006) detected haplotype-specific expression variation in the gene cluster associated with a quantitative trait locus for improved yield in rice. In the present study, we presume that if any correlation between HMGS gene expression and the identified haplotypes exists, the best candidate will be a combination of Hap_1 and Hap_2 due to the significantly higher expression levels noted in RRII 105. The effect of Hap_2 alone on expression was ruled out on the grounds that its presence in RRII 118 had no significant impact, whereas the possibility of Hap_1 alone could not be neglected, as it is unique to RRII 105. Our findings are only preliminary in nature in indicating potential haplotype gene expression associations, and further experimental proof from a large divergent population is imperative to establish it. Beyond gene expression, the downstream effects of non-synonymous SNPs may also ultimately lead to structural and functional changes at the protein level. Protein structural changes associated with non-synonymous SNPs are known to alter agronomic traits in plants (Weng *et al.* 2013). In this context, *in-silico* prediction of the impact of six non-synonymous SNPs on protein conformation is relevant. The putative structure of *Hevea* HMGS was predicted first by Sirinupong *et al.* (2005) using the crystal structure of ACP synthase III from *Mycobacterium tuberculosis* as template. Unfortunately, the model generated was of inferior quality due to the low sequence identity with this template. Since HMGS from *B. juncea* shared over 85% sequence identity with that of *Hevea*, the model generated in our study has higher reliability. As none of the non-synonymous SNPs affected the composition of the acetyl-CoA binding region, we assume that the non-synonymous SNPs identified in the study might not influence the catalytic function of the protein apart from minor structural variations in the N and C terminals. As the current work presents the first structural modelling carried out on HMGS with plant protein as template, it can be considered a reliable theoretical model for any furture HMGS structural analysis studies in *Hevea*.

In conclusion, the current study is a novel attempt of its kind in *Hevea*, providing comprehensive sequence structure information for an important gene like HMGS. Gene-linked SNP markers and information on the structural variability of various HMGS haplotypes prevailing among popular rubber clones that might have direct or indirect influence on rubber production were obtained. Haplotype structural analysis revealed the high rate of polymorphism and allele richness within this gene. Segregation analysis and prediction of recombinant sites confirmed that these intragenic sequence variations were derived from genetic recombinations rather than somatic mutations. Gene expression studies on different genotypes indicated the possibility of association between selected haplotype/haplotype combinations with HMGS gene expression. Comparative protein structure conformational studies verified the functional role of the coding SNP alleles. From a broader perspective, it is anticipated that information from this study could be used for selection of better rubber clones for crop improvement. Moreover the data suggest that Wickham clones still hold rich genetic diversity that can be exploited, and further crossings with the Indonesian and Sri Lankan high-yielding clones may increase the genetic base of Indian clones. Concurrently, the structural information provides insights into the evolution of this important gene in rubber, together with a window for crop improvement through increasing the variability. Future studies should include verifying the influence of Hap_ and Hap_1/Hap_2 combinations in more clones with varying yield characteristics to establish their association with gene expression and subsequently latex yield.

ACKNOWLEDGEMENTS

We thank Dr James Jacob, Director of Research, Rubber Research Institute of India, for his constant encouragement. Plant material for DNA isolation was contributed by Dr. Kavitha K. Mydin, Dr. C. Narayanan and Dr. Jayashree Madhavan. TKU and TS conceived and designed the research; TKU and MR conducted the experiments; TKU and AR analysed the data; TKU and AR wrote the manuscript. This study was supported by research funds from the Rubber Board, Ministry of Commerce and Industry, Government of India.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Protein sequences used for the phylogenetic analysis.

Figure S2. Venn diagram showing distribution of the 20 SNPs in the five clones.

Figure S3. HMGS protein multiple sequence alignment for the selected plants.

Figure S4. HMGS DNA multiple sequence alignment for the selected plants.

Figure S5. Comparative amino acid distribution study.

Figure S6. Sequence chromatogram depicting the allele status of parents and hybrids.

Figure S7. Phylogenetic tree of *Hevea* haplotypes based on the variable region.

Figure S8. Melting curve variation in parents and a progeny population.

Table S1. Primers used for the amplification of the full-length HMGS gene.

Table S2. Origin and parentage details of popular *Hevea* clones studied.

Table S3. Characteristics of SNPs obtained from the full-length gene sequence of HMGS.

Table S4. Raw qPCR data file of relative HMGS gene expression.

REFERENCES

Barrett J.C., Fry B., Maller J., Daly M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Bennetzen J. (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell*, **12**, 1021–1029.

Boucher Y., Doolittle W.F. (2000) The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Molecular Microbiology*, **37**, 703–716.

Chang S., Puryear J., Cairney J. (1993) A simple and efficient method for isolating RNA from Pine trees. *Plant Molecular Biology Reporter*, **11**, 113–116.

Chen Q., Sun F. (2013) A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms. *BMC Genomics*, 14 (Suppl. 1), S1. doi:10.1186/1471-2164-14-S1-S1.

Chow K.S., Mat-Isa M.N., Bahari A., Ghazali A.K., Alias H., Mohd-Zainuddin Z., Hoh C.-C., Wan K.-L. (2012) Metabolic routes affecting rubber biosynthesis in *Hevea brasiliensis* latex. *Journal of Experimental Botany*, **63**, 1863–1871.

Dereeper A., Nicolas S., Lecunff L., Bacilieri R., Doligez A., Peros J.P., Ruiz M., This P. (2011) SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics*, **12**, 134.

Dohm J.C., Lange C., Reinhardt R., Himmelbauer H. (2009) Haplotype divergence in *Beta vulgaris* and microsynteny with sequenced plant genomes. *The Plant Journal*, **57**, 14–26.

Doyle J.J., Doyle J.L. (1990) Isolation of plant DNA from fresh tissue. *Focus*, **12**, 13–15.

Fournier-Level A., Lacombe T., Le Cunff L., Boursiquot J.M., This P. (2010) Evolution of the *VvMybA* gene family, the major determinant of berry colour in cultivated grapevine (*Vitis vinifera* L.). *Heredity*, **104**, 351–362.

Guex N., Peitsch M.C. (1996) Swiss-PdbViewer: a fast and easy-to-use PDB viewer for Macintosh and PC. *Protein Data Bank Quaterly Newsletter*, **77**, 7.

He G., Luo X., Tian F., Li K., Zhu Z., Su W., Qian X., Fu Y., Wang X., Sun C., Yang J. (2006) Haplotype variation in structure and expression of a gene cluster associated with a quantitative trait locus for improved yield in rice. *Genome Research*, **16**, 618–626.

Jung J.H., Kim S., Ryu S., Kim M.S., Baek Y.S., Kim S.J., Choi J.K., Park J.K., Min G.S. (2012) Development of single-nucleotide polymorphism-based phylum-specific PCR amplification technique: application to the community analysis using ciliates as a reference organism. *Molecules and Cells*, **34**, 383–391.

Kelly R.D., Rodda A.E., Dickinson A., Mahmud A., Nefzger C.M., Lee W., Forsythe J.S., Polo J.M., Trounce I.A., McKenzie M., Nisbet D.R., St John J.C. (2013) Mitochondrial DNA haplotypes define gene expression patterns in pluripotent and differentiating embryonic stem cells. *Stem Cells*, **31**, 703–716.

Koehn R.K., Eanes W.F. (1976) An analysis of allelic diversity in natural populations of Drosophila: the correlation of rare alleles with heterozygosity. In: Karlin S., Nevo E. (Eds), *Population genetics and ecology*. Academic Press, New York, pp 377–390.

Lange B.M., Rujan T., Martin W., Croteau R. (2000) Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 13172–13177.

Le Guen V., Gay C., Xiong T.C., Souza L.M., Rodier-Goud M., Seguin M. (2011) Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*). *Plant Breeding*, **130**, 294–296.

Li Z., Zheng X., Ge S. (2011) Genetic diversity and domestication history of African rice (*Oryza glaberrima*) as inferred from multiple gene sequences. *Theoretical and Applied Genetics*, **123**, 21–31.

Librado P., Rozas J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **11**, 1451–1452.

Lochlainn S., Amoah S., Graham N.S., Alamer K., Rios J.J., Kurup S., Stoute A., Hammond J.P., Ostergaard L., King G.J., White P.J., Broadley M.R. (2011) High Resolution Melt (HRM) analysis is an efficient tool to genotype EMS mutants in complex crop genomes. *Plant Methods*, **7**, 43.

McDowell J.M., Dhandaydham M., Long T.A., Aarts M.G., Goff S., Holub E.B., Dangl J.L. (1998) Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of *Arabidopsis*. *The Plant Cell*, **10**, 1861–1874.

Perrier X., Flori A., Bonnot F. (2003) Data analysis methods. In: Hamon P., Seguin M., Perrier X., Glaszmann J.C. (Eds), *Genetic diversity of cultivated tropical plants*. Enfield Science Publishers, Montpellier, pp 43–76.

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. (2004) UCSF Chimera – a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**, 1605–1612.

Pootakham W., Ruang-Areerate P., Jomchai N., Sonthirod C., Sangsrakru D., Yoocha T., Theerawattanasuk K., Nirapathpongporn K., Romruensukharom P., Tragoonrung S., Tangphatsornruang S. (2015) Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Frontiers in Plant Science*, **6**, 367.

Roy C.B., Ravindran M., Saha T. (2012) Efficient screening of AFLP primer combinations for evaluating genetic diversity among cultivated rubber (*Hevea brasiliensis*) clones. *Rubber Science*, **25**, 21–30.

Sando T., Takaoka C., Mukai Y., Yamashita A., Hattori M., Ogasawara N., Fukusaki E., Kobayashi A. (2008) Cloning and characterization of mevalonate pathway genes in a natural rubber producing plant, *H. brasiliensis*. *Bioscience, Biotechnology and Biochemistry*, **72**, 2049–2060.

Schwede T., Kopp J., Guex N., Peitsch M.C. (2003) Swiss-model: an automated protein homology modeling server. *Nucleic Acids Research*, **31**, 3381–3385.

Shearman J.R., Sangsrakru D., Jomchai N., Ruang-areerate P., Sonthirod C., Naktang C., Theerawattanasuk K., Tragoonrung S., Tangphatsornruang S. (2015) SNP identification from RNA sequencing and linkage map construction of rubber tree for anchoring the draft genome. *PLoS One*, **10**, e0121961.

Sirinupong N., Suwanmanee P., Doolittle R.F., Suvachitanont W. (2005) Molecular cloning of a new cDNA and expression of 3-hydroxy-3-methylglutaryl-CoA synthase gene from *Hevea brasiliensis*. *Planta*, **221**, 502–512.

Suvachittanont W., Wititsuwannakul R. (1995) 3-Hydroxy 3-methylglutaryl-CoA synthase in *Hevea brasiliensis*. *Phytochemistry*, **40**, 757–761.

Suwanmaneea P., Suvachittanonta W., Fincherb G.B. (2002) Molecular cloning and sequencing of a cDNA encoding 3-Hydroxy -3-Methylglutaryl coenzyme A synthase from *Hevea brasiliensis* (HBK) mull arg. *Science Asia*, **28**, 29–36.

Tang C., Xiao X., Li H., Fan Y., Yang J., Qi J., Li H. (2013) Comparative analysis of latex transcriptome reveals putative molecular mechanisms underlying super productivity of *Hevea brasiliensis*. *PLoS One*, **8**, e75307.

Toka H.R., Genovese G., Mount D.B., Pollak M.R., Curhan G.C. (2013) Frequency of rare allelic variation in candidate genes among individuals with low and high urinary calcium excretion. *PLoS One*, **8**, e71885.

Uthup T.K., Saha T., Ravindran M., Bini K. (2013) Impact of an intragenic retrotransposon on the structural integrity and evolution of a major isoprenoid biosynthesis pathway gene in *Hevea brasiliensis*. *Plant Physiology and Biochemistry*, **73**, 176–188.

Van der Hoorn R.A., Kruijt M., Roth R., Brandwagt B.F., Joosten M.H., De Wit P.J. (2001) Intragenic recombination generated two distinct Cf genes that mediate AVR9 recognition in the natural population of *Lycopersicon pimpinellifolium*. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10493–10498.

Varshney R.K., Chabane K., Hendre P.S., Aggarwal R.K., Graner A. (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Science*, **173**, 638–649.

Wang X., Hughes A.L., Tsukamoto T., Ando T., Kao T. (2001) Evidence that intragenic recombination contributes to allelic diversity of the S-RNase gene at the self-incompatibility (*S*) Locus in *Petunia inflata*. *Plant Physiology*, **125**, 1012–1022.

Watt W.B. (1972) Intragenic recombination as a source of population genetic variability. *The American Naturalist*, **106**, 737–753.

Weng J., Li B., Liu C., Yang X., Wang H., Hao Z., Li M., Zhang D., Ci X., Li X., Zhang S. (2013) A non-synonymous SNP within the isopentenyl transferase 2 locus is associated with kernel weight in Chinese maize inbreds (*Zea mays* L.). *BMC Plant Biology*, **13**, 98.

Wititsuwannakul R. (1986) Diurnal variation of HMG-CoA reductase in latex of *Hevea brasiliensis*. *Experientia*, **42**, 45–46.