Harnessing the future natural rubber supply – Hevea brasiliensis Genome Sequence as a foundation for targeted crop improvement

TGAC The Genome Analysis Centre™ BBSRC Greater Norwich Development Partnership

Ewan M. B. Mollison[1,5], Bernardo J. Clavijo[2], Gonzalo Garcia Accinelli[2], Meena Anissi[2], Siti Arija M. Arif[3], Safiah Bt. Atan[3], Azlina Bahari[3], David Baker[2], Tom Barker[2], Gawain Bennett[2], Shabhonam Caim[2], Keng-See Chow[3], Fiorella Cugliandolo[2], Alessandra Di Cola[1], Nizar Drou[2], Darren Heavens[2], Ishtiaq Khaliq[1], Kirsten McLay[2], Rachel Piddock[2], Ricardo Ramirez[2], Cristina Ribeiro[1], Vania Serkiova[1], David Swarbreck[2], Chloë E. Thompson[1], Darren Waite[2], Diane Warner[1], Chris Watkins[2], Mario Caccamo[2] and Maria A. Kolesnikova-Allen[1†]

[1]Tun Abdul Razak Research Centre, Brickendonbury, Hertford SG13 8NL, UK; [2]The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, UK; [3]Rubber Research Institute of Malaysia (RRIM), 47000 Sungai Buloh, Selangor, Malaysia; [5]The James Hutton Institute, Errol Road, Invergowrie, Dundee DD2 5DA, UK

† Corresponding author: mkallen@tarrc.co.uk

Introduction

Hevea brasiliensis, the Pará rubber tree, is the principal source of the world's natural rubber (NR) and a valuable source of hardwood timber. It is hugely important to the economies of several south-east Asian countries, global leaders in the supply of NR, including Malaysia, Indonesia, Thailand, and Vietnam.

In Malaysia over 200 clonal lines of *H. brasiliensis* have been generated through traditional breeding strategies, with only a small pool of elite clones being suitable for large-scale planting. Different clones are selected and grown for their specific properties including latex yield, timber quality and strong rootstock. In this study the clone RRIM928, a dual purpose clone with good latex and timber properties, was selected to be the basis of a reference genome for *H. brasiliensis*.

An assembled genome for *H. brasiliensis* will be of value as a resource for research into this crop's genetics and breeding. The objective is to assemble a high-quality draft of the *H. brasiliensis* gene-space that can be used for genome annotation, study of target gene families associated with key traits, and as a resource for marker-assisted selection.

Assembly Pipeline

The *H.brasiliensis* RRIM928 clone genome sequence assembly was generated from approximately 100x Illumina paired-end data, a 4Kbp Illumina Nextera Mate Pair library (~25x), and a 16.5Kbp Long Mate Pair (LMP) library generated by Eurofins (sequenced to ~25x on Illumina).

The data generated from the Illumina Nextera Mate pair library was processed using **NextClip [1]** to categorise reads into those that contain the junction adapter one (B or C), both (A), or neither read (D). Only categories A,B, and C were taken forward into the assembly. The assembly pipeline is summarised in **Fig 1**.

Figure 1. Pipeline for assembling *H.brasiliensis* genome

Annotation Pipeline

Work is currently underway to annotate the *H. brasiliensis* genome assembly using an evidence guided gene prediction approach. Bark, latex, and leaf transcriptome data for clone RRIM928 along with root transcriptome data for another clone and publicly available datasets for other *H.brasiliensis* clones, Cassava (*Manihot esculenta*), Castor Bean (*Ricinus communis*), Flax (*Linum usitatissimum*), and Poplar (*Populus trichocarpa*) will be used in the process. **AUGUSTUS [4]**, a program that predicts genes in eukaryotic genomic sequence, will be trained using assembled *H. brasiliensis* EST, cDNA and RNA-Seq data. **Figs 5 & 6** summarise the Augustus training and prediction processes.

A strict filtering approach will be applied to identify a subset of genes with bona fide gene structures to use as a training set. Cross species protein alignments will be generated with exonerate protein2genome and provided to AUGUSTUS as external hints in conjunction with exon and intron features derived from the aligned RNA-Seq data. Genes predicted by AUGUSTUS will be further improved by running **PASA** to add additional splice variants and UTR extensions.

Figure 5. Augustus training process

Figure 6. Gene prediction process

K-mer Analysis and Assembly Statistics

Figure 2. Assembly duplication histogram including singletons

Figure 3. Assembly duplication histogram excluding singletons

The spectra in **Figs 2 & 3** show a *k*-mer content consistent with a heterozygous diploid. Singletons are included in the final release version of assembly. For more information on *k*-mer spectra analysis see the **KAT [2]** walkthrough. **Table 1** shows assembly contiguity statistics (scaffolds/sequences <200bp in length are filtered out).

Assembly	No of sequences	No of N50 sequences	N80	N50	N20	Max sequence length	Total length of sequences
Including Singletons	954,373	26,813	978	8,883	34,469	341,514	1,237,000,000
Excluding Singletons	395,351	15,598	2,120	13,299	42,111	341,514	963,100,000

Table 1. Assembly contiguity statistics

Estimation of gene-space coverage using conserved genes

The **CEGMA [3]** core gene set of 458 conserved genes was searched against the assembled *H.brasiliensis* scaffolds and singletons using **TBLASTN**; 444 genes (96.94%) were identified in the assembly with confidence of 1e-10 or better.

Fosmid Library

As a part of the *H.brasiliensis* genome project a fosmid library for clone RRIM928 was generated. The average insert size is 40Kbp with around 2X genome coverage. Sanger sequenced fosmid ends will be used for gap closure, and refinement of the assembled genome. So far over 5,000 fosmid clones have been end-sequenced at TARRC. **Fig 4** shows distance between a set of fosmid-ends after mapping to the assembly.

Figure 4. Fosmid-end mapping distance on assembly

Conclusions and Further Work

The 1.24Gbp *H.brasiliensis* assembly (approximately 56% total genome size) will allow the genome information to accelerate rubber tree improvement through molecular breeding, and efficient exploitation of genetic resources. Data generated by this and other projects supported by Malaysian Rubber Board (MRB), along with publicly available data, will provide a foundation for further in-depth study of *H.brasiliensis* at genomic and proteomic levels. Subsequent functional annotation will allow discovery of molecular markers linked to genes playing key roles in the performance of commercially important traits prioritised by industry.

Acknowledgements

This work is supported by MRB and carried out at Tun Abdul Razak Research Centre (TARRC) in collaboration with The Genome Analysis Centre (TGAC). The authors thank **Mr. Zarawi Abdul Ghani** (RRIM, MRB) for provision of plant materials from MRB Germplasm collection, and **Dr. Christopher Middleton** and **Mr. Shaun Clare** (TARRC) for editorial aid.

References

[1] Leggett, R., Clavijo, B., Clark, M., and Caccamo, M. (2013). *NextClip: an analysis and read preparation tool for Nextera long mate pair libraries.* Bioinformatics. doi: 10.1093/bioinformatics/btt702

[2] Clavijo, B., Mapleson, D., Ayling, S., and Caccamo, M. (2013). KAT – K-mer Analysis Tool. Available online at: http://www.tgac.ac.uk/kat

[3] Parra, G., Bradnam, K., Korf, I. (2007). *CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.* Bioinformatics 23 (9): 1061 – 1067

[4] Stanke, M., Waack, S. (2003). Gene prediction with a Hidden Markov model and a new intron submodel. Bioinformatics, 19 (Suppl 2): ii215 – ii225

Working for the future of the industry - *Hevea brasiliensis* clonal identification using SNP marker genotyping

Rachel Greenhill[1], Chloë Thompson[1], Ewan Mollison[1,4], Shaun Clare[1], Bernardo Clavijo[3], Safiah Atan[2], Diane Warner[1], Nor Azira Abu Bakar[2], Vania Serkiova[1], Cristina Ribeiro[1] and Maria Kolesnikova-Allen[1†]

[1] Tun Abdul Razak Research Centre (TARRC), Brickendonbury, Hertford, Herts, SG13 8NL, United Kingdom
[2] Rubber Research Institute of Malaysia (RRIM), Malaysian Rubber Board, RRIM Experimental station, 47000 Sungai Buloh, Selangor D.E., Malaysia
[3] The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, NR4 7UH, United Kingdom
[4] The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, United Kingdom
† *Corresponding author:* **mkallen@tarrc.co.uk**

Introduction

Pará rubber tree (*Hevea brasiliensis*) is the only commercial source of natural rubber (NR), the raw material for the rubber industry, which is used to produce thousands of different products. In the last decade fast developing economies have expanded the market and caused a surge in demand for NR. However, demand still exceeds supply of NR.

Clonal misidentification is found in many tree crops, like the cocoa (*Theobroma cacao*)[1][2][3] or African palm oil (*Elaeis guineensis*)[personal communication], and, therefore, could potentially be present within rubber. This could have significant effects on the productivity of the plantation if incorrect clones are used in unsuitable environments. That is why it is important to maintain stringent clonal identification, certification and tracking of planting materials.

Single nucleotide polymorphisms (SNPs) were selected as the most appropriate molecular markers to develop a genotyping panel. SNPs are the most frequent type of variation within the genome (from 1 SNP per 21bp in potato to 1 SNP per 7000bp in tomato[4]), highly conserved, their detection is easy to automate for highthroughput screening.

Materials and Methods

Two different approaches were used for the discovery of SNP candidates and marker assay design (**Fig 1**).

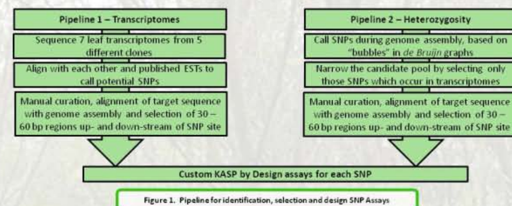

Figure 1. Pipeline for identification, selection and design SNP Assays

Over 700,000 potential SNP candidates were discovered in genome sequence of RRIM928 clone of *Hevea brasiliensis*. More than 900 SNP assays have been designed and 344 of them used so far to screen 909 individual samples representing 210 clones from Malaysian, UK and Indonesian germplasm collections. The KASP™ genotyping technology was used for screening (LGC, Hoddesdon, UK). Allelic discrimination (AD) plots (**Fig 2**) were used to automatically assign genotypes, which then were confirmed via manual curation. A database of genotypes for the entire germplasm collection was assembled, enabling their comparative analysis.

Allelic Discrimination Plot

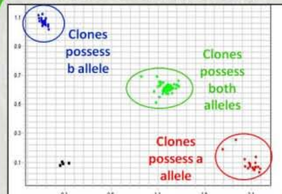

Figure 2. Allelic discrimination (AD) plot showing tight clusters. This SNP has all three genotypes present in the selection of clones. Those producing FAM fluorescence cluster along the X axis, while those producing HEX fluorescence cluster along the Y axis. Heterozygous samples cluster in the centre of the plot.

SNP marker Panel

SNP #	SNP Name	PIC value	EST Accession #	Alleles FAM/HEX
37	A898_127	0.49	EC606469	C/G
112	W569_244a	0.50	EC602560	A/G
169	LatexY07F06_416	0.48	EC604135	C/T
171	LatexY07F06_542	0.50	EC604135	G/T
175	LatexY11G07_76	0.50	EC600710	C/G
189	LatexY13E02_379	0.49	EC600603	A/G
201	Y06G11_582	0.49	EC609038	C/T
241	Y36E06_275	0.49	EC601842	A/G
268	LatexY13C07_504	0.47	EC609315	A/C
280	Y36B06_187	0.45	EC607393	A/G
302	Y48H05_16	0.40	EC607764	C/T
304	Y48H12_235	0.41	EC607211	C/T

Table 1. Final panel of 12 SNP markers used for genotyping, including their PIC values, corresponding EST accession (NCBI GenBank) numbers and allele base information.

SNP Assay Panel Selection

SNP markers for which clones segregated into tight clusters showing clear genotypes were selected for the next round of validation. The following validation process included:

- **Analysis of parental and F1 genotypes from clonal pedigrees** for correct allelic segregation (210 clones),
- **Polymorphism information content** (PIC) value limit of 0.4 for biallelic markers to assess the allelic frequency in the representative population (210 clones),
- **Confirmation of allelic segregation** in CP mapping population (PB5/51 x IAN873) in accordance with parental genotypes to ensure marker is genuine (110 progenies)

The current panel consists of 12 SNPs (Table 1) and has the potential to differentiate between more than 530,000 clones. Further panel refinement step includes:

- **Linkage mapping** to ensure SNPs are evenly distributed throughout the genome and inherited independently.

Clonal Discrimination

Clonal identities of each clone were confirmed with a larger sets of SNPs on multiple Malaysian, UK and Indonesian leaf tissue samples to pool into a consensus reference genotype.

Sample Name	Labelled Identity	Genotypic Identity	37	41	47	112	169	171	175	189	201	204	205	212	221	223	224	225	226	240	241	268	280	302	303	304	305
A1	IRCA 41	IRCA 41	b	a	h	h	a	a	h	b	a	b	b	a	b	b	a	a	a	a	h	h	b	h	b	h	a
A2	IRCA 41	IRCA 41	b	a	h	h	a	a	h	b	a	b	b	a	b	b	a	a	a	a	h	h	b	h	b	h	a
A3	IRCA 41	IRCA 41	b	a	h	h	a	a	h	b	a	b	b	a	b	b	a	a	a	a	h	h	b	h	b	h	a
A4	IRCA 41	IRCA 41	b	a	h	h	a	a	h	b	a	b	b	a	b	b	a	a	a	a	h	h	b	h	b	h	a
A5	IRCA 19	IRCA 41	b	a	h	h	a	a	h	b	a	b	b	a	b	b	a	a	a	a	h	h	b	h	b	h	a
A6	IRCA 19	IRCA 41	b	a	h	h	a	a	h	b	a	b	b	a	b	b	a	a	a	a	h	h	b	h	b	h	a
T1	IRCA 19	IRCA 19	h	h	h	b	a	b	h	a	b	b	h	h	h	h	h	h	h	h	h	a	b	h	b	b	h
T2	IRCA 19	IRCA 19	h	h	h	b	a	b	h	a	b	b	h	h	h	h	h	h	h	h	h	a	b	h	b	b	h

Sample Name	Labelled Identity	Genotypic Identity	37	112	169	171	175	189	201	241	268	280	302	304
A1	IRCA 41	IRCA 41	b	h	a	a	h	b	a	h	h	b	h	h
A2	IRCA 41	IRCA 41	b	h	a	a	h	b	a	h	h	b	h	h
A3	IRCA 41	IRCA 41	b	h	a	a	h	b	a	h	h	b	h	h
A4	IRCA 41	IRCA 41	b	h	a	a	h	b	a	h	h	b	h	h
A5	IRCA 19	IRCA 41	b	h	a	a	h	b	a	h	h	b	h	h
A6	IRCA 19	IRCA 41	b	h	a	a	h	b	a	h	h	b	h	h
T1	IRCA 19	IRCA 19	h	b	a	b	h	a	b	h	a	b	h	b
T2	IRCA 19	IRCA 19	h	b	a	b	h	a	b	h	a	b	h	b

Table 2. Selective downsizing of clonal identification panel

Above: Genotypic data for two suspected mislabelled samples on the 25 SNP panel. Samples in white are correctly labelled for their corresponding clone while samples in yellow are suspected to be mislabelled. Genotype coding includes: heterozygotes (h) and homozygotes (a or b).

Left: The same samples as above using the 12 SNP panel.

After confirming allelic segregation within the mapping population, the number of SNPs in the genotyping panel was reduced from initial 25 markers to 12. The final panel retained the differentiation power, this is shown with samples A5 and A6 still suspected of being mislabelled as IRCA 19 from their corresponding clone of IRCA 41 (**Table 2**).

Linkage Mapping

JoinMap® 4.1 (Kyazma® B.V.) was used for genetic linkage mapping of 140 SNP markers with a mapping population of 110 progenies from parental clones **PB5/51 (female)** and **IAN873**. Maps were generated using the maximum likelihood mapping function at LOD 5.0. Currently, 18 linkage groups (LG) incorporating 98 markers correspond to *Hevea brasiliensis* 18 chromosomes with a total length of 834.7cM (**Fig 3**).

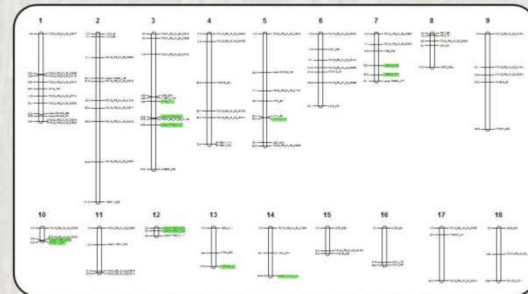

Figure 3. Current *Hevea brasiliensis* SNP genetic map with panel markers distribution. Linkage maps incorporating SNP markers were created using JoinMap® 4.1 with maximum likelihood mapping function at LOD 5.0. Highlighted SNP markers are included in the 12 SNP panel.

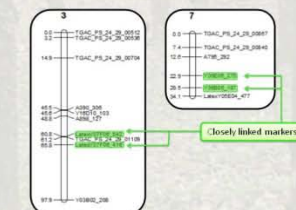

Figure 4. Examples of closely linked markers on linkage groups 3 & 7, currently included in 12 marker genotyping panel.

Closely linked markers within the current panel will be removed . Markers are considered linked if within 40cM of one another, LG3 and LG7 have markers that are 5cM and 6.6cM apart respectively (**Fig 4**).

One marker from each uncovered LG that meets the selection criteria will be added to increase the panel to at least 18 markers. This will improve marker distribution throughout the genome to increase the reliability and differentiation power of the panel.

Conclusions and Further Work

The current genotyping panel shows how effectively genuine clones can be determined. The following refinement work will aim to increase the panel to at least 18 markers across 18 saturated linkage groups, each corresponding to a *Hevea* chromosome. Future research will focus on generating ***reference profiles*** for key progenitor and commercial clones, with samples for each clone sourced from multiple locations worldwide.

Acknowledgments

This work is supported by Malaysian Rubber Board (MRB) and carried out at Tun Abdul Razak Research Centre (TARRC). We would like to thank **Mr. Donald J. Murray**, **Mr. Alistair Griffiths** and **Mr. Christopher Bisson** for provision of *Hevea* samples from Eden Project (Bodelva, Cornwall, PL24 2SG, UK), **Ms. Kate Pritchard** and **Ms. Lisa Campbell** for provision of *Hevea* samples from University of Oxford Botanic Garden (Rose Lane, Oxford, OX1 4AZ, UK), **Dr. Siti Arija Mad Arif** and **Mr. Zarawi Abdul Ghani** for provision of plant materials from MRB Germplasm collection. We are also very grateful to **Dr. Ishtiaq Khaliq** (TARRC) for editorial assistance.

References

[1] Takrama, J., *et al.* (2012) **Applying SNP marker Technology in the Cacao breeding programme in Ghana**. African Crop Science Journal, 20 (1):67-75.
[2] Livingstone, D.S., *et al.* (2012) **Optimization of a SNP assay for genotyping *Theobroma cacao* under field conditions**. Molecular Breeding, 30 (1): 33-52.
[3] Livingstone, D.S., *et al.* (2010) **Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones**. Molecular Breeding, 27 (1): 93-106
[4] Henry, RJ (ed.) (2008) **Plant genotyping II: SNP technology**. CABI Publishing, Wallingford, UK.