Discovery of novel genic-ssR markers from transcriptome dataset of an important non-human primate, *Macaca fascicularis*

Wendy Chang¹, J. EE-ULI², W. L. NG³, Jeffrine J. Rovie-Ryan⁴, S. G. Tan² & Christina S. Y. Yong ¹

*Macaca fascicularis*, also known as the cynomolgus macaque, is an important non-human primate animal model used in biomedical research. It is an Old-World primate widely distributed in Southeast Asia and is one of the most abundant macaque species in Malaysia. However, the genetic structure of wild cynomolgus macaque populations in Malaysia has not been thoroughly elucidated. In this study, we developed genic-simple sequence repeat (genic-SSR) markers from an in-house transcriptome dataset generated from the Malaysian cynomolgus macaque via RNA sequencing, and applied these markers on 26 cynomolgus macaque individuals. A collection of 14,751 genic-SSRs were identified, where 13,709 were perfect SSRs. Dinucleotide repeats were the most common repeat motifs with a frequency of 65.05%, followed by trinucleotide repeats (20.55%). Subsequently, we designed 300 pairs of primers based on perfect di- and trinucleotide SSRs, in which 105 SSRs were associated with functional genes. A subset of 30 SSR markers were randomly selected and validated, yielding 19 polymorphic markers with an average polymorphism information content value of 0.431. The development of genic-SSR markers in this study is indeed timely to provide useful markers for functional and population genetic studies of the cynomolgus macaque and other related non-human primate species.

*Macaca fascicularis* (Raffles 1821), also known as the ‘long-tailed macaque’ or the ‘cynomolgus macaque’, is a macaque species that is native to Southeast Asia and widely distributed in Malaysia, Thailand, Myanmar, Laos, Cambodia, Vietnam, Indonesia, Timor Leste, and the Philippines¹. Despite being one of the predominant macaque species in Malaysia, the genetic structure of their wild populations remains unclear. In Malaysia, most studies were conducted to examine the distribution², behaviour³, human-macaque conflict⁴,⁵ and their association with zoonotic diseases⁶,⁷. Only a few genetic studies involving phylogeography and population genetics of cynomolgus macaques were conducted thus far. Most genetic studies conducted were based on the maternally inherited mtDNA marker⁸–¹⁰, and a few reports were based on the Y-chromosome¹¹ and genomic SSR markers¹²,¹³.

Simple sequence repeats (SSRs) are repetitive DNA sequences, generally with motifs of 2–6bp long, and present abundantly in eukaryotic genome. Its codominant and multi-allelic properties are highly valued by geneticist and evolutionary biologist, and are commonly used as DNA markers in genetic studies. Despite the recent thriving of single nucleotide polymorphism (SNP) markers, SSR markers are still relevant in many applications¹⁴–¹⁸. SSRs can be broadly categorized into genomic SSR and genic-SSR, depending on their locations in the genome. SSRs sited in the transcribed region are generally known as genic-SSRs. As more SSRs associated with protein coding genes are found, it is more evident now that the previously presumed junk-DNA possibly play a crucial role in adaptive evolution¹⁹. While genic-SSR is not as abundant and as polymorphic as genomic SSR, it offers several advantages over genomic SSR marker – higher probability of finding association with functional gene,
higher degree of transferability across related species, and lower occurrence of null alleles\textsuperscript{20,21}. Despite its lower polymorphism level compared to genomic SSR, genic-SSR has been used successfully in population genetic and evolutionary studies in many species\textsuperscript{20–23}.

In recent years, advancement in sequencing technologies has made whole-genome or transcriptome sequencing of both model and non-model organisms feasible. The massive amount of transcriptome data obtained via RNA sequencing can be used in various applications, from gene identification to comparative functional analysis and differential gene expression. It also serves as an excellent sequence resource for marker development. Transcriptome sequencing coupled with established bioinformatic pipeline have been used effectively for high throughput identification of genic-SSR markers from various organisms\textsuperscript{24–26}. Some of the tools used for SSR mining include MicroSatellite identification tool (MISA)\textsuperscript{27}, FullSSR\textsuperscript{28} and Genome-wide Microsatellite analysing tool package (GMATA)\textsuperscript{29}.

There are fewer reported studies on the development of SSR markers for the cynomolgus macaque compared to \textit{Macaca mulatta}, another non-human primate model. Hitherto, development of genic-SSR markers from whole transcriptome sequencing data of cynomolgus macaque has yet to be attempted. Therefore, the present study aimed to develop genic-SSR markers from an in-house transcriptome dataset of the Malaysian cynomolgus macaque generated from our previous studies\textsuperscript{30,31}. This study is the first comprehensive report on the development of genic-SSR markers from the transcriptome of cynomolgus macaque. Here, we mined sequences containing SSRs from the transcriptome dataset, designed primers flanking pure di- and trinucleotide SSRs, and identified their associations with functional genes. Some randomly selected markers were further validated. The genic-SSR markers reported in this study are useful for population, functional genomic and comparative mapping studies of cynomolgus macaque and other related species.

Results and Discussion

\textit{De novo} assembly and functional annotation. \textit{De novo} assembly of the transcriptome data generated a total of 597,457 contigs with an average contig length of 400 bp; minimum and maximum contig lengths of 178 bp and 21,411 bp, respectively. Of the total contigs generated, 356,560 (~60\%) of the contigs had an average coverage of more than 10 reads, and annotation of these contigs revealed 73,880 (~21\%) contigs associated with functional genes. Out of the 73,880 annotated contigs, 67,399 contigs matched to \textit{M. fascicularis} (GCF 000364345.1) RNA sequences. Subsequent protein sequence similarity searched against \textit{M. mulatta} (GCF 000772875.2), \textit{Homo sapiens} (GRCH38) and SwissProt databases, further annotated 1,461, 742 and 4,278 contigs respectively.

Identification and classification of genic-SSRs. We identified a total of 14,751 genic-SSRs in this study, reflecting the effectiveness of SSR mining from the transcriptome dataset. Of the total identified genic-SSRs, 13,709 (92.94\%) were perfect repeats; while complex and compound repeats constituted the remaining 7.07\% (Fig. 1). Among the perfect SSRs, dinucleotide repeats were the most abundant (8,918; 65.05\%), followed by tri- (2,817; 20.55\%), tetra- (1,062; 7.75\%), penta- (767; 5.59\%) and hexa- (145; 1.06\%) nucleotide repeats. Di- and trinucleotide repeats constituted the largest groups of repeat motifs in our dataset, concurring with the results reported in other animal species such as human\textsuperscript{32}, chicken\textsuperscript{33} and fish\textsuperscript{34}.
Among the dinucleotide repeats, AC/GT (64.03%) accounted for the highest proportion, while CG/CN repeats were the lowest in proportion (0.29%). Amongst the ten types of trinucleotide repeats identified, AAC/GTT repeats were the most abundant (23.86%), and ACG/CGT repeats were the least common (~0.1%). The distributions of di- and trinucleotide SSRs according to motif are shown in Fig. 2. As for tetra-, penta- and hexanucleotide repeats, the most common motifs were AAAC/GTTT (8.3%), AAAAC/GTTTT (15.5%) and AAAAAC/GTTTTT (8.3%) respectively. Analysis of SSR densities in the human genome revealed that dinucleotide (AC/GT and AT/AT) and trinucleotide (AAC/GTT, AAT/ATT, AAG/CTT and AGG/CCT) repeats were the most common in humans. AC/GT repeats were also reported to be the most common dinucleotide repeat in other organisms, including fish and sheep. CG-rich SSR motifs are very rare in the transcriptome of M. fascicularis, occurring less than 1%, which corroborated the results reported in the genomes of humans and other primate species. CG/CN dinucleotide repeats are significantly low in vertebrates due to the methylation of cytosine, which favours the deamination of cytosine to thymidine.

**Table 1. Twenty validated genic-SSR markers for M. fascicularis.**

| SSR locus | Amplicon size range (bp) | Motif and repeat number based on transcriptome data | Motif observed based on sequenced PCR product | Polymorphic |
|-----------|--------------------------|---------------------------------------------------|---------------------------------------------|-------------|
| MF016     | 350–375; 23–35           | (GT)$_{16}$                                      | Yes                                         | Ye         |
| MF017     | 410–430; 17–27           | (TC)$_{12}$                                      | No                                          | Ye         |
| MF021     | 435–450; 26–33           | (TG)$_{12}$                                      | Yes                                         | Ye         |
| MF029     | 350–385; 18–35           | (GT)$_{13}$                                      | Yes                                         | Ye         |
| MF069     | 400–425; 9–21            | (CA)$_{12}$                                      | No                                          | Ye         |
| MF080     | 285–315; 9–24            | (TG)$_{12}$                                      | Yes                                         | Ye         |
| MF102     | 465–525; 6–36            | (CA)$_{12}$                                      | Yes                                         | Ye         |
| MF113     | 400–450; 10–31           | (CA)$_{12}$                                      | Yes                                         | Ye         |
| MF121     | 475–525; 15–38           | (TG)$_{13}$                                      | Yes                                         | Ye         |
| MF130     | 440–450; 7–12            | (GT)$_{16}$                                      | Yes                                         | Ye         |
| MF147     | 350–415; 7–39            | (AC)$_{12}$                                      | Yes                                         | Ye         |
| MF188     | 445–460; 5–15            | (AC)$_{12}$                                      | Yes                                         | Ye         |
| MF197     | 315–375; 6–35            | (AT)$_{10}$                                      | Yes                                         | Ye         |
| MF225     | 325–375; 5–30            | (AC)$_{10}$                                      | Yes                                         | Ye         |
| MF242     | 500–525; 18–30           | (TG)$_{14}$                                      | Yes                                         | Ye         |
| MF255     | 335–400; 4–37            | (GT)$_{12}$                                      | Yes                                         | Ye         |
| MF259     | 315–335; 8–18            | (CA)$_{10}$                                      | Yes                                         | Ye         |
| MF261     | 385–425; 14–24           | (CT)$_{14}$                                      | Yes                                         | Ye         |
| MF272     | 365–385; 5–14            | (TG)$_{10}$                                      | Yes                                         | Ye         |
| MF273     | 435–450; 8–15            | (TG)$_{10}$                                      | Yes                                         | Ye         |

**Table 1.** Twenty validated genic-SSR markers for M. fascicularis.

**Figure 2.** Frequency of di- and trinucleotide repeat motifs in the transcriptome of M. fascicularis.
polymorphic, demonstrating that more than 60% (19 out of 30) of the markers screened in this study were polymorphic. Alignment of the sequences obtained from the PCR amplicons with the contig sequences used to design the primers also verified successful amplification of the targeted DNA regions.

Data analysis. Genetic diversity assessment was performed based on 19 polymorphic markers (Table 1) amplified across 26 *M. fascicularis* individuals, which were divided into the West Coast and East Coast populations. Heterozygosity assessment was performed on individual population (Supplementary Table S2), with West Coast and East Coast populations showed similar mean *H*<sub>e</sub> values of 0.481 and 0.484, respectively. The mean *N*<sub>st</sub> values for the West Coast population was 3.316 and East Coast population was 2.684. For overall genetic diversity assessment in all 26 individuals, *N*<sub>st</sub>, *H*<sub>e</sub>, *H*<sub>o</sub>, and PIC ranged from 2 to 6, 0 to 1, 0.125 to 0.713, and 0.110 to 0.653, respectively (Supplementary Table S3). The overall mean *N*<sub>st</sub>, *H*<sub>e</sub>, *H*<sub>o</sub>, and PIC were 3.630, 0.269, 0.495, and 0.431, respectively. *F*-statistics calculated from the 19 polymorphic loci revealed a mean *F*<sub>ST</sub> of 0.059. Out of the 19 loci, three loci (MF121, MF259, and MF272) were the most polymorphic with six alleles each. Seven of the 19 polymorphic SSR loci had PIC values of >0.5, and thus, they were considered as highly informative. Compared to previous population studies, where genomic SSR markers were used, the genomic-SSR markers used in this study generated lower *N*<sub>st</sub>, *H*<sub>o</sub>, *H*<sub>e</sub>, and PIC values for the same species. As the SSR markers developed in our study were generated from transcriptome, it was anticipated that the genetic diversity of these markers would be lower than those of SSR markers derived from genomic DNA regions. The lower values could also be contributed by the lower number of samples (n = 26) and sampling sites in the current work compared to those studies. Nonetheless, the identification of 19 polymorphic SSRs out of 30 markers screened based on 26 individual samples is promising. We are confident that higher *N*<sub>st</sub>, *H*<sub>o</sub>, *H*<sub>e</sub>, and PIC values would be obtained with more samples.

The average PIC value of 0.431 for the 19 polymorphic loci validated in this study was comparable to those genomic-SSRs developed for the Korean quail (mean PIC value = 0.494) and crab (mean PIC value = 0.49) in *C. asiatica*. Although not all the 19 genomic-SSR markers showed high polymorphism and PIC values, all showed the reproducibility and specificity highly desired in genotyping by PCR.

There were very few reported studies on the development of SSR markers for cynomolgus macaque. The first study conducted to develop SSR markers for the cynomolgus macaque was reported in 2007 by Kikuchi et al. In their work, they crossed-amplified 148 SSR markers selected from human genome database, and discovered 66 (44%) polymorphic SSR markers in the cynomolgus macaque. Later, Higashino et al. identified an additional 499 polymorphic SSR markers from the BAC library of *M. fascicularis*. They analysed the genetic polymorphisms of cynomolgus macaques originated from Indonesia, the Philippines and Malaysia using these SSR markers. In both studies, the SSR markers employed were derived from genomic DNA regions. The polymorphic genomic-SSR markers identified in this study is a good addition to complement existing SSR markers to provide more markers for the investigation of the genetic structure of wild macaque populations.

Materials and Methods

Ethical clearance. The usage of *M. fascicularis* samples in this investigation complied with the animal care regulations and all relevant national laws of Malaysia. Sampling protocols were approved by the Institutional Animal Care and Use Committee (IACUC), University of California, Davis, USA as adopted by the PREDICT Programme in Malaysia, under which the Department of Wildlife and National Park (DWNP) Malaysia is working collaboratively with the EcoHealth Alliance, the Ministry of Health Malaysia, and the Veterinary Services Department, Malaysia.

De novo assembly and functional annotation. A transcriptome dataset was generated from a previous RNA sequencing project of the *M. fascicularis* on liver, kidney, lymph node, spleen and thymus. We subjected the raw sequencing reads to quality assessment using FASTQC v0.11.2. Illumina co-sequencing positive control (PhIX) sequences were filtered and cleaned sequence reads were subjected to base quality checking (Q ≥ 30). De novo sequence assembly was performed using CLC Genomics Workbench version 8.5.1 (CLC Bio-Qiagen, Aarhus, Denmark). We subjected the assembled contigs (average coverage ≥ 10 reads) to annotation by sequence similarity searches with BLAST+ version 2.2.31 using Blast against database built with *M. fascicularis* (GCF_000364345.1) RNA sequences from the NCBI RefSeq database. Contigs with no match to *M. fascicularis* RNA sequences were further searched against database built with *Macaca mulatta* (GCF_000772875.2) protein

| No. | State | GPS |
|-----|-------|-----|
| 1   | Perlis | 247716 734307, 251413 736659, 243353 722458 |
| 2   | Kedah | 209354 699281, 208481 698262, 210009 704813 |
| 3   | Pulau Pinang | QU263935 WMR615425, QU258571 WMR589429, QU258571 WMR589429 |
| 4   | Perak | 281927 548666, 335311 503907, 285403 468407 |
| 5   | Selangor | 398663 351330, 395891 344432, 389528 356202 |
| 6   | Negeri Sembilan | 485958 311210, 476472 305602, 480356 318551 |
| 7   | Kelantan | QU435536 WMR605751, QR437054 WMR552482, QR438147 WMR539413 |
| 8   | Pahang | 489719 398997 |
| 9   | Terengganu | 586162 555216, 564609 591740 |

Table 2. GPS locations of the *M. fascicularis* samples used.
sequences using Blastp program. Sequences with no match to *M. mulatta* protein sequences were then searched against database built with *Homo sapiens* (GRCh38) protein sequences. Contigs with no match were further examined using protein similarity search against SwissProt database.

| Primer ID | Primer sequence 5′→3′ | (Repeat motif)n | Product size (bp) | T<sub>a</sub> (°C) |
|-----------|------------------------|----------------|------------------|------------------|
| MF016     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF017     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF021     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF029     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF069     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF075     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF080     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF095     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF102     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF113     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF121     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF130     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF147     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF149     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF176     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF188     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF197     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF209     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF225     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF248     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF255     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF259     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF261     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF267     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF268     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF272     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF273     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |
| MF279     | F:CTTTATGAGAGACAAAAAGG  |
|           | R:GATGAGAGACAAAAAGG     |

Table 3. The 30 genic-SSR primers used in preliminary screenings on 26 macaque DNA samples. F, Forward primer; R, Reverse primer; NA, no amplification; T<sub>a</sub> annealing temperature.
Identification and classification of genic-SSRs. Genic-SSR identification and classification were performed on the filtered contigs (average coverage ≥ 10 reads) using MicroSAtellite identification tool (MISA)44. The minimum number of repeats for di-, tri-, tetra-, penta-, and hexanucleotides were set at six, five, four, and three, respectively. Categorization of perfect, compound and complex SSRS were as follows. Perfect: consisting of a single repeat of n units; compound perfect: consisting of two or more alternate tandem repeats of n units each; complex: consisted of repeat motifs that varied in profiles by a single unit/consisted of alternate repeat motifs interspersed within a single region/consisted of two simple perfect motifs separated by nonrepeating sequences of variable length.

Primer design. Contig sequences containing SSRs identified from the transcriptome dataset were employed for primer design using Primer3 software45. We focused on candidate SSR sequences of perfect di- and trinucleotides with repeat numbers ≥ 10 and with only one SSR present in each contig for primer design. All contig sequences used for primer design were checked against genomic sequences to predict the location of introns. Three-hundred SSR primer pairs were designed. All the contigs used for SSR primers design were checked for functional annotation where a cut-off value of $E < 10^{-15}$ was used.

Sampling, DNA extraction, PCR amplification, and electrophoresis. Thirty genic-SSR primer pairs selected randomly from the 300 pairs designed were used for initial screening on the DNA samples of 26 M. fascicularis individuals. Primers were selected randomly among those that have self- and cross- primer complementarity values of less than 3, low tendency to form secondary structures and 0.9−1.0 complementary value of less than 3.

To test the robustness of the markers, samples were obtained from nine states in Peninsular Malaysia (Table 2) with the permission and collaboration of DWNP. Three samples from each state were obtained except Terengganu (2 samples). Genomic DNA samples of 24 M. fascicularis individuals were provided in the form of extracted DNA. Two DNA samples were isolated from liver tissue samples provided by DWNP using QIAamp DNA mini kit (Qiagen, Germany) according to the manufacturer’s protocol.

PCR was performed in 10 µl reaction volumes containing 10ng of genomic DNA, 2.0 µM of each primer, 1× PCR buffer, 2.5 mM MgCl₂, 0.2 mM dNTPs, and 1 U Taq polymerase (Promega, USA), in a thermal cycler T100 (Bio-Rad, USA). Gradient PCR protocol under the following conditions was employed: a single cycle of initial denaturation at 95°C for 5 minutes, followed by 35 cycles of denaturation at 95°C for 1 minute, annealing at X°C for 30 seconds, extension at 72°C for 5 minutes, and ended with a single cycle of final extension at 72°C for 5 minutes. X denotes the different annealing temperatures (°C) used for different primers (Table 3). Primers that were not successfully amplified or produced multiple bands were further tested using touchdown PCR with 1°C decrements starting from 60°C. PCR products were separated on a 2.0% agarose gel and 8.0% non-denaturing polyacrylamide gel stained with EtBr (0.5µg/ml). To further confirm the presence of targeted SSRs in the amplified products, PCR products with the expected fragment sizes were sequenced on an ABI 3730 through services provided by First Base Sdn. Bhd. (Seri Kembangan, Malaysia). Sequences obtained were compared with the contig sequences that the primers were designed from and the targeted SSR repeats were also identified.

Genetic diversity analysis. The 26 individual macaque samples from the nine states in Peninsular Malaysia were arbitrarily divided into two populations, the East Coast and West Coast populations, taking into consideration the Titiwangsa Range as a potential geographical barrier for gene flow between populations on both coasts. The East Coast population comprised of eight individuals from three states (Kelantan, Terengganu and Pahang), while the West Coast population consisted of 18 individuals from six states (Perlis, Kedah, Pulau Pinang, Perak, Selangor and Negeri Sembilan). SSR banding patterns were analyzed with PopGene version 1.3246 and Cervus version 3.0.747 to calculate the number of alleles ($N_a$), observed heterozygosity ($H_e$), expected heterozygosity ($H_e$), fixation index ($F_{ST}$), and polymorphic information content (PIC).

Data Availability
The transcriptome data have been deposited in the NCBI Short Read Archive database under accession SRP096937 and SRX2499144-SRX2499147.

References
1. Fooden, J. & Field Museum of Natural, H. Systematic review of Southeast Asian long-tailed macaques, Macaca fascicularis (Raffles, [1821]). Vol. n.s. no. 81 (1995) (Field Museum of Natural History 1995).
2. Saaban S, K. K. Population Status of Long-Tailed Macaque (Macaca fascicularis). PeerJ PrePrints 3, e1253v1, https://doi.org/10.7287/peerj.preprints.1253v1 (2015).
3. Karimullah & Anuar, S. Condition and population size of long-tailed Macaque (Macaca fascicularis) in Kuala Selangor nature park. Tropical Natural History 12, 189–205 (2012).
4. Md-Zain, B. M., Ruslin, F. & Idris, W. M. R. Human-macaque conflict at the main campus of Universiti Kebangsaan Malaysia. Pertanika Journal of Tropical Agricultural Science 37, 73–85 (2014).
5. Akter, R. et al. Simian malaria in wild macaques: first report from Hulu Selangor district, Selangor, Malaysia. Malar J 14, 386, https://doi.org/10.1186/s12936-015-0856-3 (2015).
6. Abdul-Latif, M. A. et al. Continental monophony and molecular divergence of Peninsular Malaysian Macaca fascicularis fascicularis. Biomed Res Int 2014, 897682, https://doi.org/10.1155/2014/897682 (2014).
7. Khajaeian, P. et al. Inter simple sequence repeats (ISSRs): Neglected DNA markers for molecular dissection of Plasmidum species in long-tailed Macaque (Macaca fascicularis). PeerJ PrePrints 3, e1253v1251, https://doi.org/10.7287/peerj.preprints.1253v1 (2015).
8. PeerJ PrePrints 3, e1253v1251, https://doi.org/10.7287/peerj.preprints.1253v1 (2015).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool.

47. Kalinowski, S. T., Taper, M. L. & Marshall, T. C. Revising how the computer program CERVUS accommodates genotyping error

46. Yeh, F. C.

44. Thiel, T., Michalek, W., Varshney, R. K. & Graner, A. Exploiting EST databases for the development and characterization of gene-

40. Ellis, J. R. & Burke, J. M. EST-SSRs as a resource for population genetic analyses.

45. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers.

38. Schorderet, D. F. & Gartler, S. M. Analysis of CpG suppression in methylated and nonmethylated species.

37. Loire, E., Higuet, D., Netter, P. & Achaz, G. Evolution of coding microsatellites in primate genomes.

22. Zhou, T.

21. Perez, F.

17. Maia, T. A., Vilaca, S. T., Silva, L. R. D., Santos, F. R. & Dantas, G. P. M. DNA sampling from eggshells and microsatellite genotyping

18. Vieira, M. L., Santini, L., Diniz, A. L. & Munhoz Cde, F. Microsatellite markers: what they mean and why they are so useful.

19. Liu, F.

23. Arya, L., Verma, M., Gupta, V. K. & Seetharam, A. Use of genomic and genic SSR markers for assessing genetic diversity and

24. Ma, H.

25. Deng, T. et al. De Novo Transcriptome Assembly of the Chinese Swamp Buffalo by RNA Sequencing and SSR Marker Discovery. PLoS One 11, e0147132, https://doi.org/10.1371/journal.pone.0147132 (2016).

26. Bai, J. Y., Pang, Y. Z., Qi, Y. X., Zhang, X. H. & Yun, X. Y. Development and Application of EST-Ssr Markers in Genus of Asian Swamp Buffalo (Bubalus bubalis). Molecules 21, 66, https://doi.org/10.3390/mons21030166 (2016).

27. Beier, S., Münch, T., Scholz, U., Mascher, M. & Thiel, T. MISA-web: a web server for microsatellite prediction. Bioinformatics 33, 2583–2585, https://doi.org/10.1093/bioinformatics/btx198 (2017).

28. Wang, X. & Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front Plant Sci 7, 1350, https://doi.org/10.3389/fpls.2016.01350 (2016).

30. Ee Uli, J. et al. RNA sequencing (RNA-Seq) of lymph node, spleen, and thymus transcriptome from wild Peninsular Malaysian cylomolgus macaque (Macaca fascicularis). PeerJ 5, e3566, https://doi.org/10.7717/peerj.3566 (2017).

31. Ee Uli, J. et al. RNA sequencing of kidney and liver transcriptome obtained from wild cylomolgus macaque (Macaca fascicularis) originating from Peninsular Malaysia. BMC Genet 11, 923, https://doi.org/10.1186/s12863-014-0104-1 (2016).

32. Subramanian, S. et al. SSRD: simple sequence repeats database of the human genome. Comp Funct Genomics 4, 342–345, https://doi.org/10.1002/cfg.289 (2003).

33. Bakhtiarizadeh, M. R., Arefnejad, B., Ebrahimie, E. & Ebrahim, M. Application of functional genomic information to develop efficient EST-SSRs for the chicken (Gallus gallus). Genet Mol Res 11, 1558–1574, https://doi.org/10.4238/2012-May.21.12 (2012).

34. Liu, H. G., Yang, Z., Tang, H. Y., Gong, Y. & Wan, L. Microsatellite characterization of Saurucomys gabriellae Bleeker, 1871 in a Yangtze river-connected lake, China. J Genet 96, e1–e4, https://doi.org/10.1007/s12041-016-0753-x (2017).

35. Zhang, J. et al. Characterization and development of EST-ssr markers derived from transcriptome of yam catfish. Molecules 19, 16402–16415, https://doi.org/10.3390/molecules191016402 (2014).

36. Zhang, W. et al. Using Bioinformatic Methods to Develop EST-SSR Markers from Sheep’s ESTs. Journal of Animal and Veterinary Advances 9, 2759–2762, https://doi.org/10.3923/javaa.2016.2759.2762 (2010).

37. Loire, E., Huguet, D., Netter, P. & Achaz, G. Evolution of coding microsatellites in primate genomes. Genome Biol Evol 5, 283–295, https://doi.org/10.1093/gbe/evt003 (2013).

38. Schorderet, D. F. & Gartler, S. M. Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA 89, 957–961 (1992).

39. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32, 314–331 (1980).

40. Ellis, J. R. & Burke, J. M. EST-SSRs as a resource for population genetic analyses. Heredity (Edinb) 99, 125–132, https://doi.org/10.1038/sj.hdy.6801001 (2007).

41. Kikuchi, T., Hara, M. & Terao, K. Development of a microsatellite marker set applicable to genome-wide screening of cynomolgus monkeys (Macaca fascicularis). Primates 48, 140–146, https://doi.org/10.1007/s10329-006-0008-z (2007).

42. Tighe, T., Michalek, W., Varshney, R. K. & Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 106, 411–422, https://doi.org/10.1007/s00122-003-0371-0 (2003).

43. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132, 365–386 (2000).

44. Yeh, F. C. et al. PopGene, the user-friendly shareware for population genetic analysis, molecular biology and biotechnology center. POGGENE (1997).

45. Kalinowski, S. T., Taper, M. L. & Marshall, T. C. Revising the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16, 1099–1106, https://doi.org/10.1111/j.1365-294X.2007.03089.x (2007).
Acknowledgements
This project was funded by Universiti Putra Malaysia, through the Putra Group Initiative grant, project no. GP-IPB/2013/9413602. The funder had no role in study design, conducting the study, decision to publish and preparation of the manuscript.

Author Contributions
W.C. conceived the study, performed the experiments, and wrote the paper. J.E.-U. performed the experiments, data analyses and edited the paper. W.L.N. helped with data analysis and edited the paper. S.G. Tan edited the paper. J.J.R.-R. helped with sample collection and edited the paper. C.S.Y.Y. oversaw the experiments and data analysis, and co-wrote the paper.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-44870-4.

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019