RESEARCH ARTICLE

De novo transcriptome analysis and microsatellite marker development for population genetic study of a serious insect pest, *Rhopalosiphum padi* (L.) (Hemiptera: Aphididae)

Xinle Duan[1,2], **Kang Wang**[1], **Sha Su**[1], **Ruizheng Tian**[1], **Yuting Li**[1], **Maohua Chen**[1]*

1 State Key Laboratory of Crop Stress Biology for Arid Areas, Key Laboratory of Crop Pest Integrated Pest Management on the Loess Plateau of Ministry of Agriculture College of Plant Protection, Northwest A&F University, Yangling, China, **2** College of Bee Science, Fujian Agriculture and Forestry University, Fuzhou, China

* maohua.chen@nwsuaf.edu.cn

Citation: Duan X, Wang K, Su S, Tian R, Li Y, Chen M (2017) *De novo* transcriptome analysis and microsatellite marker development for population genetic study of a serious insect pest, *Rhopalosiphum padi* (L.) (Hemiptera: Aphididae). PLoS ONE 12(2): e0172513. doi:10.1371/journal.pone.0172513

Editor: J. Joe Hull, USDA Agricultural Research Service, UNITED STATES

Received: November 8, 2016

Accepted: February 5, 2017

Published: February 17, 2017

Data Availability Statement: All relevant data are available from the NCBI Short Read Archive (SRA) with accession numbers of SPR095727, SPR5133458 and SPR5133459.

Funding: This work was supported by grants of the National Natural Science Foundation of China (No. 31272036 and No. 31471766). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The bird cherry-oat aphid, *Rhopalosiphum padi* (L.), is one of the most abundant aphid pests of cereals and has a global distribution. Next-generation sequencing (NGS) is a rapid and efficient method for developing molecular markers. However, transcriptomic and genomic resources of *R. padi* have not been investigated. In this study, we used transcriptome information obtained by RNA-Seq to develop polymorphic microsatellites for investigating population genetics in this species. The transcriptome of *R. padi* was sequenced on an Illumina HiSeq 2000 platform. A total of 114.4 million raw reads with a GC content of 40.03% was generated. The raw reads were cleaned and assembled into 29,467 unigenes with an N50 length of 1,580 bp. Using several public databases, 82.47% of these unigenes were annotated. Of the annotated unigenes, 8,022 were assigned to COG pathways, 9,895 were assigned to GO pathways, and 14,586 were mapped to 257 KEGG pathways. A total of 7,936 potential microsatellites were identified in 5,564 unigenes, 60 of which were selected randomly and amplified using specific primer pairs. Fourteen loci were found to be polymorphic in the four *R. padi* populations. The transcriptomic data presented herein will facilitate gene discovery, gene analyses, and development of molecular markers for future studies of *R. padi* and other closely related aphid species.

Introduction

The bird cherry-oat aphid, *Rhopalosiphum padi* (L.), is a notorious insect pest that devastates wheat crops globally [1–4]. This species can reduce both the yield and quality of wheat by sucking sap and transmitting barley yellow dwarf virus (BYDV), which leads to serious economic damage to wheat production [1]. In the past few years, due to global climate change, farming systems, wheat varieties, anthropogenic effects, and other factors, the damage to wheat caused

Competing interests: The authors have declared that no competing interests exist.

by *R. padi* has increased and its distribution in China has expanded [5–8]. Due to its potential to cause serious damage to crops, the life cycle, host specificity, potential harmfulness, and methods of management of *R. padi* have been the subject of intensive studies [9–12]. Furthermore, the dual sexual and asexual reproductive mode of *R. padi* make this species an important model for the evolution of life-history traits and sympatric speciation [13–15].

Population genetic studies can provide insights into the evolution of reproduction modes, adaptive strategies of aphid species in agroecosystems, and the influence of environmental and anthropogenic factors on the genetic diversity and genetic structure of, as well as gene flows between, aphid populations [8]. These studies also facilitate the design and optimization of sustainable pest management strategies [8,13,16]. Molecular markers—such as restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), mitochondrial DNA (mtDNA), and simple sequence repeats (SSRs)—are useful for studies of insect taxonomy, phylogeography, genetic diversity, and population structure at various taxonomic levels [16–18]. Microsatellite markers are special repetitive DNA sequences that have a high information content, co-dominance, locus specificity, and are readily amplified [19,20]. Therefore, microsatellite markers are used widely in meta-population genetics studies at various scales [3, 21–24]. Simon et al. (2001) isolated eight polymorphic microsatellite loci in *R. padi* and confirmed their utility for population genetics studies of parthenogenetic organisms [25]. Subsequently, these microsatellite loci were used in population genetics studies of French [25], Australian [17], and Chinese [8] *R. padi* populations. However, only five microsatellite loci were highly polymorphic in Chinese *R. padi* populations, which are insufficient for *R. padi* population genetics research [26].

Despite the abundance of microsatellite markers, traditional microsatellite development methods require expertise and substantial investment of time and other resources, particularly for non-model species lacking reference genomes and transcriptome data [27–30]. Next-generation sequencing (NGS) technologies are considered to be powerful, rapid, cost-effective, and reliable tools for transcriptional analysis, novel gene discovery, and molecular marker development [31–33]. Unlike traditional methods, transcriptome sequences generated by NGS facilitate rapid mining of microsatellite markers in non-model organisms [30,34]. Transcriptome sequences are coding DNA, and so a high proportion contains functional information [35,36]; therefore, transcriptome-derived microsatellite markers are situated close to or within functional genes [37–39] and increase cross-species transferability [40–42]. Transcriptomic sequencing for mining of microsatellite markers has been applied to several insect species, such as *Rhagoletis pomonella* [43], *Aphis glycines* [44], *Dolerus aeneus* [45], *Maruca vitrata* [46] and *Spodoptera exigua* [47].

In this study, the complete transcriptome of adult *R. padi* was characterized using the Illumina HiSeq 2000 platform, and all transcriptome sequences were assembled, BLAST searched, and annotated using public databases. Based on the transcriptome databases, several thousand *R. padi* microsatellite loci were mined, some of which were used to design primers to validate and estimate the intraspecific genetic diversity in four *R. padi* populations. To our knowledge, this is the first transcriptome analysis of this non-model species. The transcriptome data and polymorphic microsatellite markers will facilitate further studies of the population genetics and molecular biology of *R. padi*.

Materials and methods

Ethics statement

No specific permissions were required for the described field studies for this wide spread agriculture pest. We confirm that the locations were not privately owned or protected in any way. The field studies did not involve endangered or protected species.

Insect materials, RNA extraction, cDNA library construction, and Illumina sequencing

A apterous parthenogenetic female of *R. padi* was collected from a wheat field at Northwest A&F University, Yangling, Shaanxi, China in July 2013 (108˚05'E, 34˚17'N) to set a clone (parthenogenetic line) on seedlings of wheat (*Triticum aestivum*) cultivar "Xiaoyan 22" at 24 ± 1˚C, 40% RH, and a 16:8 h (L:D) photoperiod. Ten aphids were randomly selected from the clone for RNA extraction. The total RNA of *R. padi* was isolated using TRIzol reagent (Tiangen Biotech, Beijing, China) with minor modifications at the recovery step, in which RNase-free filter columns (Sangon Biotech, Shanghai, China) were used. RNA quantity and quality were assessed by gel electrophoresis and spectrophotometry, respectively. Ribosomal RNA (rRNA) was depleted from RNA samples using the Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) (Epicentre, Madison, USA) following the manufacturer's instructions. Around 20 μg purified RNA samples were sent to Beijing Genomics Institute (BGI) (Shenzhen, China) for cDNA library construction. Briefly, poly-T oligo-attached magnetic beads (Illumina, San Diego, CA, USA) were used to isolate poly (A) RNA from total RNA. Then a SuperScript Double-Stranded cDNA Synthesis kit (Invitrogen, Camarillo, CA) was employed for double-stranded cDNA synthesis with random hexamer (N6) primers (Illumina, San Diego, CA, USA). To normalize cDNA, the frequency of abundant cDNA species was reduced using Trimmer-2 cDNA Normalization Kit (Evrogen, Moscow, Russia). The T4 DNA polymerase, Klenow DNA polymerase and T4 polynucleotide kinase were used for end-repair and phosphorylation of synthesized cDNA. These repaired cDNA fragments were 3' adenylated using Klenow Exo- (Illumina, San Diego, CA, USA). Illumina paired-end adapters were ligated to the ends of these 3'-adenylated cDNA fragments. The products of this ligation reaction were electrophoresed on a 2% (w/v) Tris-acetate-EDTA-agarose gel for downstream enrichment with different sizes. cDNA fragments of 200 (± 25 bp) were excised from the gel. Fifteen cycles of PCR amplification were performed with PCR primers (PE 1.0 and PE 2.0) and Phusion DNA Polymerase to enrich the quantity of purified cDNA template. The majority of the amplified fragments in the Illumina library were about 200 bp in size. Four samples were sequenced per lane on an Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA). Paired-end sequencing was used to sequence both ends and the cDNA library was deep-sequenced generating four gigabytes of data and a total of 114,428,314 raw reads. After cleaning and trimming, a total 108,340,100 clean reads were used for assembly and analysis.

De novo assembly and analysis of Illumina reads

To ensure the quality requirement for *de novo* transcriptome sequencing, a stringent filtering process was carried out. Initially, Illumina reads that passed the Failed-Chastity filter (Illumina) based on a setting of "failed-chastity less than or equal to 1" with a chastity threshold of 0.6, were reserved on the first 25 cycles. Then, all reads with adaptor contamination and low-quality reads with ambiguous sequences "N" were discarded. Finally, we ruled out reads with more than 10% $Q < 20$ bases. The cleaned reads were assembled de novo using SOAPdenovo2 and contigs with length less than 200 bp were discarded due to a low annotation rate [48]. The paired-end Illumina reads were first combined to produce longer fragments (*i.e.*, contigs) and then mapped back to the contigs. The paired-end reads and contigs were assembled to form longer sequences that originated from the same transcript, with N indicating unknown bases (*i.e.*, scaffolds). The paired-end reads were used for gap filling of the scaffolds to obtain unigenes with the least Ns that could not be extended at either end. All unigenes assembled were compared with the non-redundant protein database (nr) of the National Center for Biotechnology Information (NCBI), non-redundant nucleotide sequence (nt) database (NCBI) (http://www.ncbi.nlm.nih.gov), UniProt/Swiss-Prot (http://www.expasy.ch/sprot), and the Clusters of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG/) using BLASTx [49] with E-values of less than 1e^{-5} and 1e^{-10}, respectively. The COG functional classification was used to analyze the gene composition, as well as predict and classify possible functions of transcriptome sequences. With the nr annotation, Blast2GO was used for gene ontology (GO) annotation of the unigenes according to molecular function, biological process and cellular component ontologies (http://www.geneontology.org). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and the online KEGG Automatic Annotation Server (http://www.genome.jp/tools/kaas/) were used to map unigenes to pathways applying an E-value threshold of 1e^{-5}. During the determination of the sequence direction of unigenes, the priority order of nr, Swiss-Prot, KEGG and COG was followed when the alignment results of four databases conflicted with each other.

SSR mining and primer design

The MIcroSAtellite (MISA, http://pgrc.ipk-gatersleben.de.sci-hub.org/misa/) Perl script was employed to identify microsatellites in the unigenes. In this study, cDNA-based SSRs (cSSRs) were considered to contain motifs with two to six nucleotides and a minimum of four contiguous repeat units. The criterion of no fewer than six repeat units for di-, five for tri- to tetra-, and four for penta- to hexa-nucleotide repeats was adopted. Based on the MISA results, repeat motifs were set randomly and searched in sequences longer than 200 bp. The frequency of cSSRs refers to kilobase pairs of cDNA sequences containing one SSR. The output was then parsed by the Primer3-2.3.4 program (http://sourceforge.net/projects/primer3/files/primer3/2.3.4/primer3-2.3.4.tar.gz/download) for design of PCR primers in the flanking regions of SSRs. Primers were designed based on the following criteria: (1) length of 18–23 bp; (2) annealing temperature of 55–65°C with a maximum discrepancy of <2°C among primers; and (3) a PCR product size of 80–300 bp. Finally, 60 pairs of primers were designed and screened for their PCR-amplification efficacy.

PCR amplification and validation of selected SSRs

To evaluate the amplification efficacy, specificity, and polymorphisms of all selected SSRs, we collected *R. padi* samples from Lanzhou in the Gansu Province (the sample was coded as GSL; the coordinates are 103°41′ E and 36°05′ N), Xianyang in the Shaanxi Province (SAX; 108°05′ E, 34°17′ N); Wuhan in the Hubei Province (HUW; 112°47′ E, 32°08′ N), and Baicheng in the Jilin Province (JLB; 122°52′ E, 45°39′ N). Apterous parthenogenetic individuals from the four regions were brought to the laboratory in individual tubes. Twelve individuals from each region were used to set up 12 respective clones (parthenogenetic lines) on seedlings of wheat cultivar ‘Xiaoyan 22’ at 24±1°C, 40% RH, and a 16:8 h (L: D) photoperiod. A total of 48 clones were set up for the four regions. One individual was randomly taken from the first generation of each clone for DNA extraction. Genomic DNA was extracted from each individual using the EasyPure Genomic DNA Kit (TransGen Biotech, BeijingChina) according to the bench protocol for animal tissues. DNA was eluted in 30 μL of ultrapure water and stored at –20°C. Three primers-a forward primer with an M13 (-21) at the 5’ end, a reverse primer, and an FAM fluorscent dye-labeled M13 (-21) primer [50] were used for PCR amplification of the microsatellite loci. PCR was performed in a total volume of 25 μL, comprising 12.5 μL of 2× Taq Master Mix (containing 0.05 U/μL Taq DNA Polymerase, 2× Taq PCR Buffer, 3 mM $MgCl_2$, and 400 μM dNTP mix) (CoWin Biotech, Beijing, China), 0.5 μL of each forward primer (10 μM), 2 μL of each reverse primer (10 μM), 2 μL of M13 primer (10 μM), and 1.5 μL of genomic DNA (10–30 ng/μL). PCR amplification involved denaturation at 95°C for 2 min, followed by 30 amplification cycles consisting of 95˚C for 20 s, 20 s at the primer-specific annealing temperature (Table 1), and 72˚C for 20 s. This was followed by eight cycles of 95˚C for 30 s, 53˚C for 45 s, and 72˚C for 45 s, and a final step at 72˚C for 10 min. To examine the length of the amplified PCR products, an ABI3730XL automated DNA sequencer (Applied Biosystems, Foster City, CA, USA) was used, and all genotypes were called with GeneMapper v4.0 software (Applied Biosystems, Foster City, CA, USA).

Statistical analysis

Population genetic parameters for polymorphic loci, such as the observed heterozygosity (Ho), the expected heterozygosity (He), and number of alleles (Na) were calculated using FSTAT v2.9.3.2 software [51]. The Excel Microsatellite Toolkit v3.1 program (MS Tools) [52] was utilized to calculate the polymorphism information content (PIC) of each SSR; the inbreeding index (Fis) was assessed using Popgene v1.3.1 software [53]; Hardy-Weinberg equilibrium (HWE) and *P* values of HWE were evaluated using GENEPOP v4.0 software [54]. The statistical power of microsatellite markers for detecting genetic differentiation at various Fst (fixation index) levels was evaluated using POWSIM v4.1 software [55]. Based on an effective population size of Ne = 1000, simulations were carried out with Fst values of 0.0001, 0.0025, 0.005, 0.0075, 0.01, 0.0125, 0.015, 0.0175, and 0.02, and 1000 replicates [55]. Chi-squared tests and Fisher's exact tests were used to analyze the allele number and frequency at randomly selected microsatellite loci, and the power of the analysis was indicated by the proportion of tests that were significant at $P < 0.05$ [8].

Table 1. Characteristics of 24 microsatellite loci developed for *R. padi*.

Locus	Primer sequence (5'-3')	Repeat motif	Ta (˚C)	Size range (bp)
RP02	F: TACACACTTTGCCGTAAATCAGA; R: CGTTGTATTGGCATGATATAGG	$(AC)_8$	56	218–234
RP05	F: GCGGTCTTTCGTTTCTCTCTATC; R: GGGAAATTGAAATAACATCTCGG	$(TC)_{11}$	60	145–167
RP06	F: ATAAAGGTACCTACGCGAAATCC; R: CCTCGTGACTCGACATGATAGTA	$(GT)_{10}$	60	148–168
RP08	F: TCATTTGCGTATAAGACATGGAT; R: CATCACTGCATATCAGTCTGACG	$(AC)_{10}$	54	133–153
RP11	F: GGGTCGGGTATAGTCAGAGTCTT; R: CTCGACGACAATTCTACGTCTTT	$(GT)_{10}$	56	138–158
RP13	F: AGTTGTATTGTTTTGAACGGTCG; R: TCGTGGATTATCGTTACAATACTGA	$(AT)_{11}$	58	146–166
RP14	F: TTGTGAAATCGGTTTTACGTTTT; R: CTCTACACTCAAGCCCCAATTTA	$(GT)_{10}$	56	146–166
RP17	F: AGTAAGTCCGTCCCGTCGTCT; R: AATATCGCGTATCGTACCAGTGT	$(CAG)_7$	58	140–170
RP22	F: TTCGATCTGTTTCTTGAGCG; R: CGCGGTATAAGGTCACCG	$(GCC)_8$	56	147–171
RP23	F: CTGAGCTGCAGTATTTTCCGAGT; R: CGCGTGCATAAATGTATAACGTA	$(ATAC)_5$	54	150–170
RP24	F: AGTTGGGGAATTTATAGTGAGCG; R: TGGTGTCTCAGAGTAAAGAAAAGAA	$(TATG)_5$	58	146–166
RP25	F: ATTTTGAGTTTTCACCATCGTTT; R: TTTTTAGTACGCCAAATTGTTGA	$(GGTA)_5$	58	145–165
RP29	F: TTAAATAAAAAGGCCAAACCCAT; R: ATTGGCGCTAAATTAAATGACTC	$(ATCAGA)_4$	56	146–170
RP30	F: AACGACCCAGATTATGTAGTCCA; R: ACCAACACCAACACCAACATC	$(TGGTGT)_4$	56	144–168
RP31	F: AGCAGCTTCTTGAGCTGGAC; R: AACAACAACACCAAAGGGTCTAC	$(GCC)_8$	56	143–167
RP36	F: AATCATACAAACGAGATTCTTCCC; R: GCAAGTATCGAAACGCTAACCTA	$(TA)_6$	56	224–236
RP37	F: AGTGTAAATGATATCGCGGCTTA; R: TCCGGTTCACAGTACAAAAATATC	$(AT)_8$	56	210–226
RP38	F: ATACATAGGACACAGGCGAAAAC; R: CACCACCTAAATTGTATTAAGGAAA	$(AT)_7$	56	203–217
RP42	F: GATCACAATAATATTTACCGGCG; R: ATAACGTACGCGTGTGGATAAAA	$(TA)_{10}$	54	129–149
RP43	F: GTATGCGGTCAACCTATTTTACG; R: TATTGTTGGAATTAGAAAGGCCA	$(AT)_{10}$	56	120–140
RP45	F: ACCATGATCAACTTAAGAGGTGC; R: ATGTTTACGTTTAGAGCGCATGT	$(TA)_{10}$	60	125–145
RP47	F: TATTTACGACGGGAATGTACAGC; R: TCAATATATCTCTCACCCCCTCA	$(TATC)_5$	60	143–163
RP48	F: GTTCCTCCGGCCGTACAT; R: GATATCTCGTTCTGGCGAGTCT	$(CAC)_8$	58	141–165
RP60	F: ATTTAGATTTCATCTCATCGTCGG; R: GCTGAGGACGAAAACTATTTTGA	$(CGT)_7$	56	145–172

doi:10.1371/journal.pone.0172513.t001

Results

Illumina sequencing and de novo assembly of short reads

Illumina sequencing resulted in 114,428,314 raw reads of which 108,340,100 were of acceptable quality. Total GC content was 40.03%, with a Q20 value of 98.37% and a final length of 9,750,609,000 bp (S1 Table). The transcriptomic data are available via NCBI Short Read Archive (SRA) with accession numbers of SPR095727, SPR5133458 and SPR5133459. The average raw read length was 90 bp, which is consistent with the Illumina sequencing capacity. The raw reads were assembled into contigs using the SOAPdenovo2 software, and empty reads and low-quality sequences were filtered out. These contigs were further assembled into scaffolds by paired-end joining and gap-filling. Finally, the scaffold sequences were assembled into clusters, which yielded 29,467 unigenes with an average length of 990 bp (N50 = 1,580) and main lengths between 200-3000bp. Of these unigenes, 5,699 (19.34%) were 200–300 bp, 6,832 (23.19%) were 301–500 bp, 6986 (23.4%) were 501–1,000 bp, 6448 (21.89%) were 1,001–2,000 bp, 2273 (7.71%) were 2,001–3,000 bp, and 1319 (4.48%) were >3,000 bp in length (Fig 1).

Functional annotation of the unigenes

All *R. padi* unigenes were aligned to the nr, nt, UniProt, KEGG, COG, and GO databases to predict and classify possible functions. A total of 24,302 (82.47%) unigenes were annotated by BLAST searches of the databases. Among these unigenes, 21,139 (71.74%), 22,939 (77.85%), 16,024 (54.38%), 14,586 (49.5%), 8,022 (27.22%), and 9,895 (33.58%) had homologous sequences in the nr, nt, Swiss-Prot, KEGG, COG, and GO databases, respectively. In addition, 5,165 (17.53%) unigenes showed no homology to known sequences.

In the COG functional classification, 8,022 generated 91,152 functional annotations across 25 COG categories (Fig 2). Among these COG categories, general function prediction

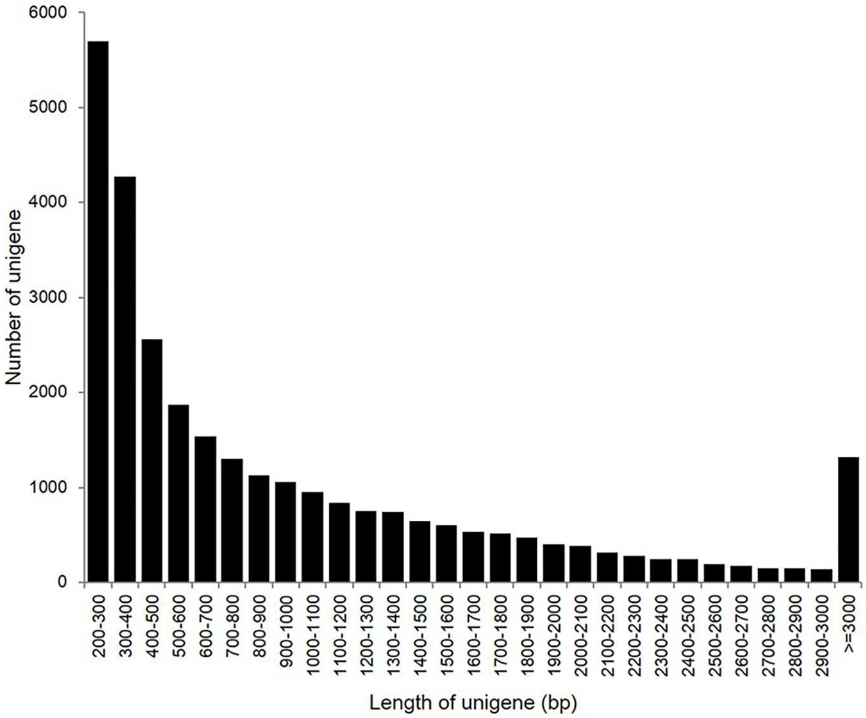

Fig 1. Frequencies length distribution of Illumina read sequences.

doi:10.1371/journal.pone.0172513.g001

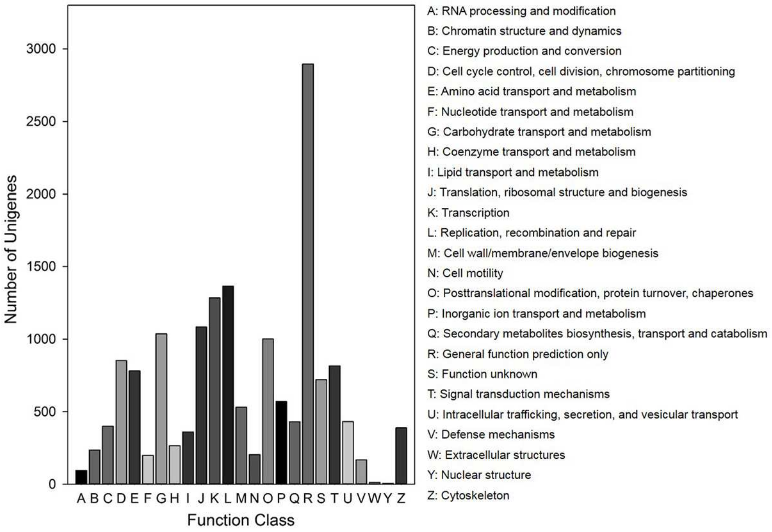

Fig 2. Histogram presentation of clusters of orthologous groups (COG) classification.

doi:10.1371/journal.pone.0172513.g002

contained the largest number of unigenes (2,896, 36.10%), followed by replication, recombination, and repair (1,365, 17.02%), transcription (1,285, 16.02%), translation, ribosomal structure, and biogenesis (1,083, 13.50%), and other groups. Interestingly, only six and 14 unigenes were related to nuclear structures and extracellular structures, respectively.

A total of 9,895 unigenes with BLAST matches to known proteins were annotated with 76,184 GO functions based on sequence similarity, with an average of 7.7 GO annotations per unigene (Fig 3). The three main GO annotation categories were biological process (42,447, 55.72%), cellular component (21,425, 28.12%), and molecular function (12,312, 16.16%). The annotations could be further subdivided into 59 subcategories (Fig 3). For sequences that initially sorted to the biological process classification, cellular process, single-organism process, metabolic process, and biological regulation were among the most represented matches. The major subcategories for the cellular component classification were cell and cell part, whereas in the molecular function classification the major subcategories were binding and catalytic activity.

In addition, 14,586 unigenes aligned with 257 KEGG pathways (S2 Table). These pathways contained 1–2,105 unigenes. The pathways with the greatest number of unigenes were metabolic pathways (2,105 unigenes, 14.43%), RNA transport (505, 3.46%), regulation of actin cytoskeleton (474, 3.25%), focal adhesion (466, 3.19%), and pathways in cancer (461, 3.16%). Only one unigene was involved in each of the D-glutamine and D-glutamate metabolism, allograft rejection, and graft-*versus*-host disease pathways.

Development and characterization of microsatellite markers

To develop novel molecular markers, all of the 29,476 unigenes were used to mine potential microsatellite loci consisting of di- to hexa-nucleotide SSRs with at least four repetitions. Using the MISA Perl script, a total of 7,936 potential SSRs were identified in 5,564 unigenes, 3,960 contained only one SSR while 1,604 contained more than one SSR. Of these SSRs, 714 were present in compound form (with adjacent tandem simple repeats of a different sequence), and 7,222 SSRs were in perfect form (without interruptions in the runs of repeats) [56]. In addition, the frequency, type, and distribution of the potential 7,936 SSRs were analyzed. On average, one SSR was present every 0.99 kb in unigenes, and the frequency of cSSRs was 26.93%.

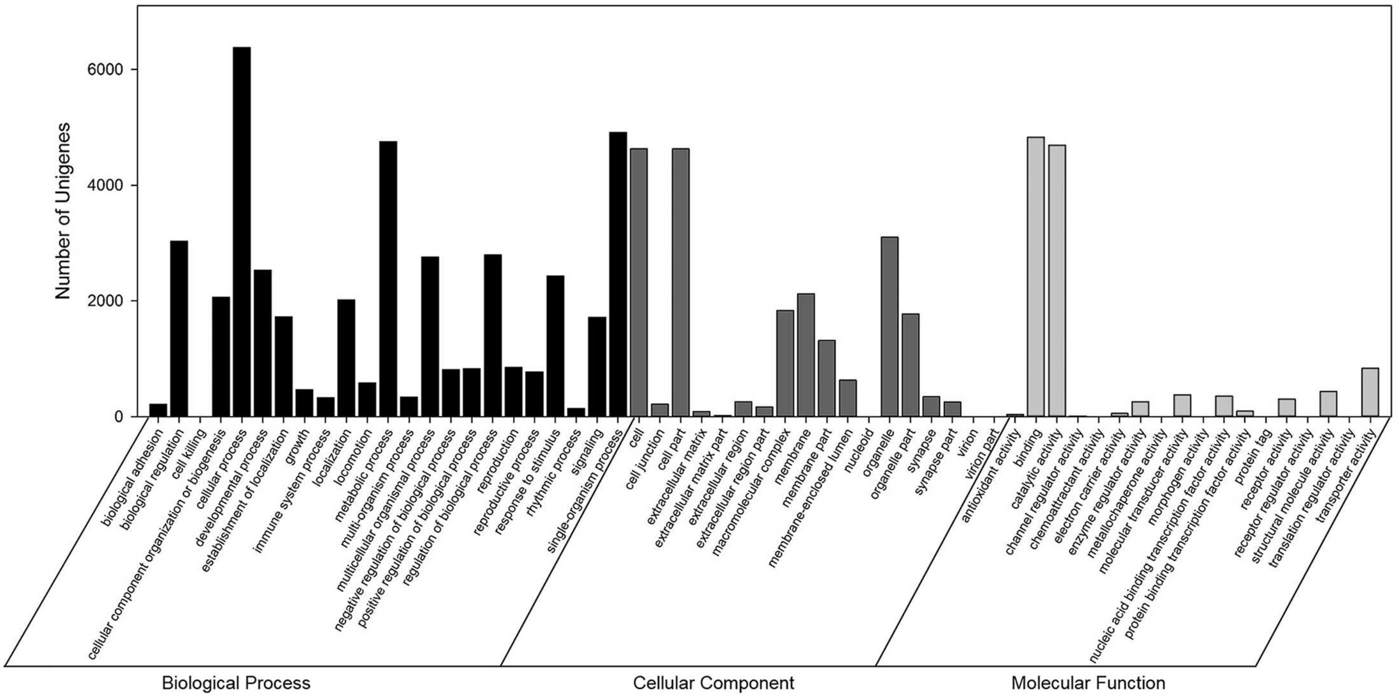

Fig 3. Gene Ontology classification of assembled unigenes.

doi:10.1371/journal.pone.0172513.g003

Among the 7,936 SSRs, the tri- and mono-nucleotide repeat motifs were most abundant (3,631, 45.75% and 2290, 28.86%, respectively), followed by di- (1,857, 23.40%), tetra- (70, 0.88%), penta- (49, 0.62%), and hexa-nucleotide (39, 0.49%) repeat motifs (Fig 4). Di- to hexa-nucleotide motifs were further analyzed in terms of the number of repeat units (or SSR length, Table 2). The most frequent number category of repeat units was 4–9, accounting for 68.18% of total SSRs, followed by the 10–15 repeat unit category (1793 SSRs, 22.59%). There were 732 SSRs with more than 15 repeat units. Among the detected SSRs, 163 motif types were identified, that comprised 4, 12, 59, 32, 28, and 28 types of mono- di-, tri-, tetra-, penta-, and hexa-nucleotide repeats, respectively. The A/T mono-nucleotide repeat was the most abundant motif type (2,267, 28.57%), followed by AAT/ATT (1,494, 18.83%), AT/TA (1,103, 13.9%), AC/GT (612, 7.71%), ACG/CGT (447, 6.01%), AAC/GTT (417, 5.25%), CCG/CGG (332, 4.18%), AGG/CTG (281, 3.54%), and ACC/GGT (240, 3.02%). The remaining 154 motif types accounted for 8.98% of the total (Fig 4).

Validation of microsatellite markers

A total of 60 microsatellite loci were selected at random for verification of their utility in various *R. padi* populations. Twenty-four of these loci were amplified successfully and yielded bands of the correct sizes; these loci were used to analyze the genetic diversity of *R. padi* (Table 1). Fourteen loci (RP06, RP08, RP13, RP14, RP22, RP23, RP24, RP30, RP31, RP42, RP43, RP45, RP48 and RP60) were polymorphic in four different geographical populations of *R. padi* (Table 3). The number of alleles (Na) per locus ranged from 2 to 9, with an average of 3.98. The observed heterozygosity (Ho) value was 0.417–1.000, whereas the expected heterozygosity (He) value was 0.409–0.823. The the polymorphism information content (PIC) of each locus ranged from 0.221 to 0.765. The inbreeding index (Fis) values ranged from −1.000 to 0.367, most of which were negative values indicating the heterozygous excess in the four *R.*

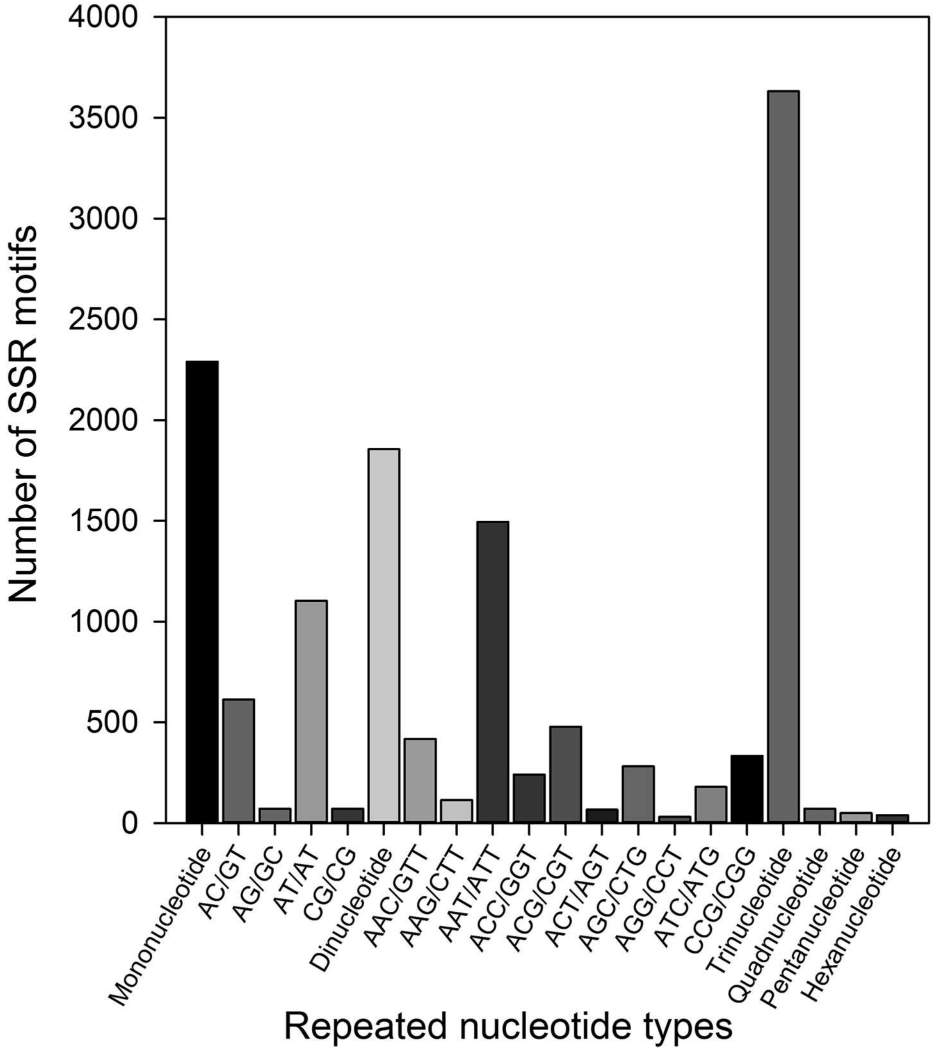

Fig 4. Total numbers of SSRs based on motif types in *R. padi*.

doi:10.1371/journal.pone.0172513.g004

Table 2. Frequency of SSRs based on repeat types in *R. padi* transcriptome.

Repeat type	No. of repeat														
	4	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**>15**	**Total**	**Percentage**
Mononucleotide	0	0	0	0	0	0	0	0	696	413	262	187	732	2290	28.86%
Dinucleotide	0	0	715	441	263	203	121	105	9	0	0	0	0	1857	23.40%
Trinucleotide	0	1876	1039	608	108	0	0	0	0	0	0	0	0	3631	45.75%
Quadnucleotide	0	62	8	0	0	0	0	0	0	0	0	0	0	70	0.88%
Pentanucleotide	43	6	0	0	0	0	0	0	0	0	0	0	0	49	0.62%
Hexanucleotide	39	0	0	0	0	0	0	0	0	0	0	0	0	39	0.49%
Total	82	1944	1762	1049	371	203	121	105	705	413	262	187	732	7936	
Percentage	1.03%	24.50%	22.20%	13.22%	4.67%	2.56%	1.52%	1.32%	8.88%	5.20%	3.30%	2.36%	9.22%		

doi:10.1371/journal.pone.0172513.t002

Table 3. Genetic diversity revealed by 14 microsatellites in four populations of *R. padi* from China.

Locus	P 1					P 2					P 3					P 4				
	Na	Ho	He	PIC	Fis	Na	Ho	He	PIC	Fis	Na	Ho	He	PIC	Fis	Na	Ho	He	PIC	Fis
RP06 (KU325537)	7	0.750	0.725	0.665	-0.037	6	0.583	0.801	0.730	0.280***	5	0.583	0.565	0.488	-0.034	5	1.000	0.670	0.575	-0.526**
RP08 (KU325538)	3	0.833	0.554	0.428	-0.538	2	0.917	0.518	0.373	-0.833*	2	0.917	0.518	0.373	-0.833*	3	0.667	0.554	0.428	-0.214
RP13 (KU325539)	4	1.000	0.652	0.559	-0.571*	5	0.833	0.634	0.571	-0.333*	3	1.000	0.594	0.477	-0.737**	4	0.917	0.656	0.579	-0.424
RP14 (KU325540)	4	0.667	0.540	0.482	-0.248	3	0.250	0.359	0.307	0.313*	3	0.750	0.562	0.480	-0.356	2	0.583	0.431	0.328	-0.375
RP22 (KU325541)	4	0.833	0.612	0.535	-0.384	5	0.583	0.819	0.750	0.297**	4	0.636	0.680	0.594	0.067	5	0.455	0.706	0.620	0.367*
RP23 (KU325542)	5	0.833	0.710	0.643	-0.183	6	0.917	0.725	0.644	-0.280***	4	0.833	0.717	0.638	-0.170*	5	0.667	0.580	0.513	-0.158
RP24 (KU325543)	2	0.750	0.489	0.359	-0.571	3	0.750	0.518	0.408	-0.478	2	1.000	0.522	0.375	-1.000**	2	0.417	0.344	0.275	-0.222
RP30 (KU325544)	3	0.750	0.554	0.468	-0.375	2	0.667	0.464	0.346	-0.467	3	0.500	0.409	0.341	-0.234	3	0.667	0.522	0.449	-0.294
RP31 (KU325545)	5	1.000	0.699	0.627	-0.459	4	0.917	0.714	0.633	-0.301*	4	0.917	0.685	0.595	-0.360	5	0.583	0.623	0.553	0.067
RP42 (KU325546)	7	0.909	0.823	0.755	-0.111	9	1.000	0.823	0.765	-0.228	6	0.833	0.710	0.643	-0.183	4	0.833	0.641	0.557	-0.317
RP43 (KU325547)	3	0.500	0.409	0.341	-0.234	4	0.583	0.486	0.432	-0.213	3	0.500	0.409	0.341	-0.234	4	0.250	0.239	0.221	-0.048
RP45 (KU325548)	5	1.000	0.732	0.659	-0.389	6	1.000	0.696	0.612	-0.467**	3	0.917	0.583	0.470	-0.613*	6	0.546	0.537	0.491	-0.017
RP48 (KU325549)	3	0.833	0.540	0.420	-0.583	3	0.417	0.467	0.397	0.113	3	0.583	0.475	0.410	-0.242	4	0.667	0.533	0.469	-0.266
RP60 (KU325550)	2	1.000	0.524	0.375	-1.000**	4	0.667	0.511	0.431	-0.323	4	0.833	0.605	0.523	-0.401	3	0.750	0.518	0.408	-0.478

Na, numbers of alleles; Ho, observed heterozygosity; He, expected heterozygosity; PIC, polymorphism information content; Fis, the inbreeding index; Deviation from Hardy-Weinberg equilibrium (HWE) are indicated by asterisks (*, $P < 0.05$, **,$P < 0.01$, ***, $P < 0.001$). GenBank accession number for each locus was showed in the parenthesis. The population P 1 was from Gansu Province, P 2 was from Shaanxi Province, P 3 was from Hubei Province, and P 4 was from Jiilin Province. Twelve individuals from each populations were used in the test. Detailed information of the sampling locations was described in the text.

doi:10.1371/journal.pone.0172513.t003

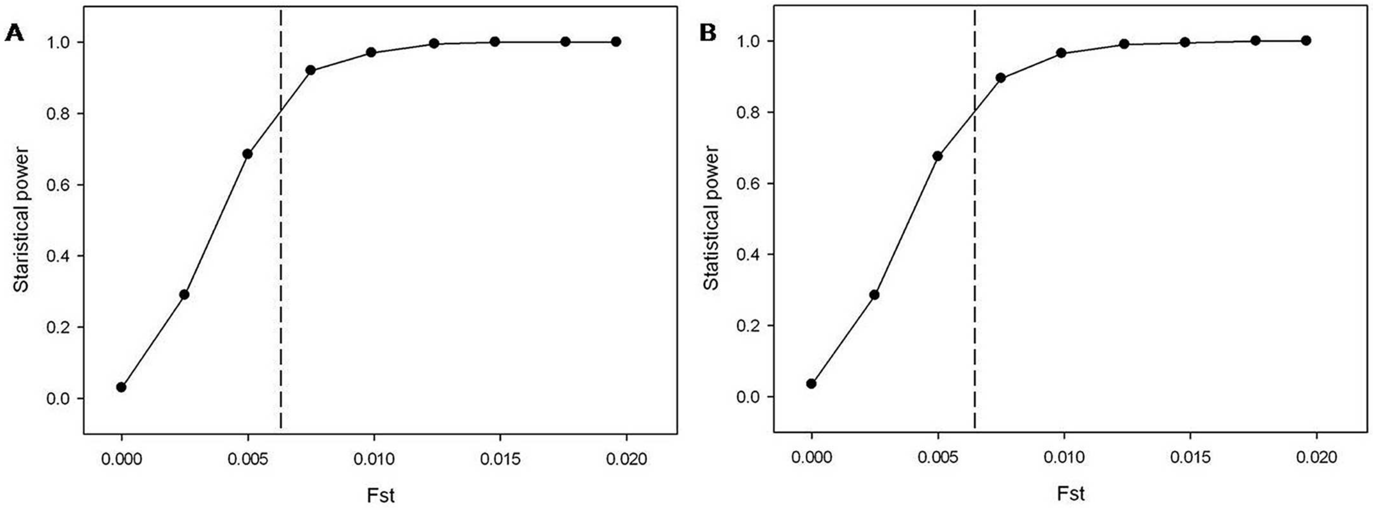

Fig 5. POWSIM analyses for testing power to detect genetic differentiation at different Fst values for 14 microsatellite loci as showed by Chi-square test (A) and Fisher's exact test (B).

doi:10.1371/journal.pone.0172513.g005

padi populations. Among the 14 analyzed microsatellite loci in the four *R. padi* populations, 10 showed deviation from Hardy-Weinberg equilibrium (HWE) ($P < 0.005$) (Table 3). Power calculations using a chi-squared test (Fig 5A) and Fisher's exact test (Fig 5B) showed that Fst (fixation index) values as low as 0.0062 could be detected with more than 80% probability. Therefore, the 14 microsatellite loci provided sufficient statistical power to detect population differentiation.

Discussion

Transcriptome analysis

In this study, the complete *R. padi* transcriptome was sequenced on the Illumina HiSeq 2000 platform, yielding a total of 9,750,609,000 bp with 114.43 million clean reads. These sequences also produced longer unigenes (mean = 990 bp) than those assembled in *Ipomoea batatas* (765 bp) [57], *Eucalyptus grandis* (197 bp) [58], *Acropora millepora* (440 bp) [59], *Bemisia tabaci* (266 bp) [60], and *Dialeurodes citri* (539 bp) [61]. Furthermore, the N50 length of the unigenes was 1,580 bp, longer than that of *Ipomoea batatas* (765 bp) [57], *D. citri* (632 bp) [61], and *Grapevine phylloxera* (936 bp) [62]. This result indicated that the *R. padi* transcriptome sequences were of high quality, which was likely due to the improved transcript construction and scaffolding and low heterozygosity of the new paired-end sequencing technology.

For functional annotation, we utilized several complementary approaches to annotate the assembled sequences using several public databases. About 82.47% of the unigenes had orthologs or homologs in these databases and were assigned at least one functional annotation, which is higher than previous reports of other species using the same sequencing platform [39,60,61,63]. The high level of annotations is due to the availability of complete functional information in all of the public databases, the high mean length of unigenes [39], and the availability of an aphid genome database (http://www.aphidbase.com/aphidbase/). A total of 17.53% unigenes were unmapped in any of the databases, possibly due to the short sequence reads generated [61], the presence of non-coding transcripts among the unigenes [31], and/or the incompleteness of the public sequence databases [30]. In COG and GO functional classification, a large proportion of unigenes (27.22 and 33.58%) were assigned to a wide range of

COG and GO classifications (Figs 2 and 3), indicating that the transcriptome data included a wide diversity of transcripts. In the KEGG pathway analysis, a high proportion of unigenes were mapped to metabolic pathways, the RNA transport pathway, and regulation of the actin cytoskeleton pathway (S2 Table). In addition, several pathways related to pesticide resistance —such as ABC transporters, drug metabolism-cytochrome P450, and metabolism of xenobiotics by cytochrome P450—were identified. These annotated unigenes will facilitate more in-depth investigations of population genetics and functional genomics of *R. padi* and other closely related aphid species.

Microsatellite loci characterization

The major traditional methods for microsatellite loci development are the hybrid capture method [64], loci selection from available genetic/genomic information [65], and loci transferable from closely related species [28,66]. Compared with traditional methods, *de novo* transcriptome sequencing technology is a rapid, cost-effective, and reliable tool that enables microsatellite markers to be developed directly from transcriptome sequences, particularly for non-model species [33,42,67]. Among the 29,467 assembled unigenes, 5,564 (18.89%) possessed 7,936 potential microsatellite loci. This is higher than the values for other insect pests, such as *Bombyx mori* [68], *Tomicus yunnanensi* [69], *Bactrocera dorsalis* [70], and *Phenacoccus solenopsis* [71]. Six types of microsatellite loci repeat type were identified among the unigenes; the most common were trinucleotide (45.75%) and mononucleotide (28.86%) repeats, in agreement with the results for *D. aeneus* [45], *Timema cristinae* [72], and *B. dorsalis* [70]. Wang et al. (2012) found that trinucleotide microsatellite loci were abundant in the transcriptome data of *Tetrao tetrix*, and predicted that tri-nucleotides can remain in coding regions without causing reading frame shifts [73]. Therefore, the abundance and frequency of the various microsatellite loci repeat types were related to the size of the transcriptome database, the microsatellite loci search software used, and the parameter criteria [74]. A/T, AAT/ATT, and AT/TA were the most abundant SSR motif types in the *R. padi* transcriptome database. A/T homopolymers are also more abundant than C/G homopolymers in *Schistosoma mansoni* [75], *Tenebrio molitor* [76], and *P. solenopsis* [71]. Tóth et al. (2000) examined the abundance of microsatellites with repeated unit lengths of 1–6 base pairs in several eukaryotic taxonomic groups [77]. They found poly (A/T) tracts are more abundant in each taxon than poly(C/G) sequences and the plausible explanation for the higher proportion of A/T-rich SSRs is the poly-A tails of retroposed sequences and processed pseudogenes. In 154 non-model eukaryote species, the previous reports found that the GC/CG motif was rare, and that the GC/CG was absent in several eukaryote species [78–83]. Indeed, the CG/CG motif was detected at a low frequency (0.89%) in this study. This phenomenon cannot be explained only by the low C/G content of the genome and thus may represent a genuine pattern [78,84,85].

Microsatellite loci development and validation

Among 60 randomly selected potential microsatellite markers, 24 loci (40%) were amplified successfully, and 14 loci exhibited polymorphisms in four *R. padi* populations. The low amplification rate may be caused by the special structure at the primer (s) location or between the primers, for example, the presence of large intron between primers, or unrecognized splice sites disrupting primer positions. The chimeric primers and assembly errors also could result in failed amplification. However, the other microsatellites obtained from our transcriptome data can provide a larger pool for mining more polymorphic loci. SSR polymorphisms are positively correlated with the number of motif repeats [70]. In the transcriptome database, the number of motif repeats of most SSRs (68.18%) ranged from 4 to 9, and only 9.23% of SSRs had more than 15 repeats. Hence, SSRs in the transcriptome were less polymorphic than genomic SSRs, but possessed potential polymorphisms [30, 86]. Among the four *R. padi* populations, 14 loci had fewer alleles (Na) than genomic SSRs, but similar Ho, He, and Fis values [8]. The PIC is an important parameter for microsatellite polymorphisms [87]. The average PIC values of 14 loci in four *R. padi* populations were 0.334–0.680, suggesting that 14 were moderately or highly polymorphic. A chi-squared test and Fisher's exact test confirmed that 14 loci had strong statistical power to detect low Fst levels, while amplification at various annealing temperatures using the remaining 36 primer pairs failed.

Conclusions

To our knowledge, this is the first report of the assembly and characterization of the transcriptome of *R. padi* using the Illumina HiSeq 2000 platform. A total of 2,9467 unigenes were generated and 7,936 EST-SSRs were identified, which will facilitate development of molecular markers for *R. padi*. Sixty of these loci were selected randomly, and 24 were amplified successfully and validated experimentally in four *R. padi* populations. Our results will enable development of microsatellite markers and population genetic studies of *R. padi*.

Supporting information

S1 Table. Summary of transcriptome data for *R. padi* and bioinformatics annotation.
(PDF)

S2 Table. KEGG biochemical mappings for *R. padi*.
(PDF)

Acknowledgments

This work was funded by the National Natural Science Foundation of China (grant nos. 31471766 and 31272036). We would like to thank Hui-min Shen, Jin Huang, Geng Tao, Li-jun Chen and Shu-xia Tao for their help with the collection of *Rhopalosiphum padi* samples, and to thank two anonymous reviewers for useful comments to the manuscript.

Author Contributions

Conceptualization: XD MC.

Data curation: XD MC.

Formal analysis: XD MC.

Funding acquisition: MC.

Investigation: XD MC.

Methodology: XD.

Project administration: MC.

Resources: XD KW SS YL.

Software: XD.

Supervision: MC.

Validation: MC.

Visualization: XD KW MC.

Writing – original draft: XD MC.

Writing – review & editing: XD RT KW MC.

References

1. Zhang XC, Zhou GH, Shi M, Fang JZ, Zhao ZP, Li SH, et al. The rules of long distance migration and virus transmission of wheat aphids. J Plant Protect. 1985; 12: 9–16.
2. Simon JC, Blackman R, Le Gallic JF. Local variability in the life cycle of the bird cherry-oat aphid, *Rhopalosiphum padi* (Homoptera: Aphididae) in western France. Bull Entomol Res.1991; 81: 315–322.
3. Simon JC, Carrel E, Hebert PDN, Dedryver CA, Bonhomme J, Le Gallic JF. Genetic diversity and mode of reproduction in French populations of the aphid *Rhopalosiphum padi* L. Heredity. 1996; 76: 305–313.
4. Martinez-Torres D, Moya A, Hebert P, Simon JC. Geographic distribution and seasonal variation of mitochondrial DNA haplotypes in the aphid *Rhopalosiphum padi* (Hemiptera: Aphididae). Bull Entomol Res. 1997; 87: 161–167.
5. Chen MH, Han ZJ, Qiao XF, Qu MJ. Resistance mechanisms and associated mutations in acetylcholinesterase genes in *Sitobion avenae* (Fabricius). Pestic Biochem Phys. 2007a; 87: 189–195.
6. Chen MH, Han ZJ, Qiao XF, Qu MJ. Mutations in acetylcholinesterase genes of *Rhopalpsiphum padi* resistant to organophosphate and carbamate insecticides. Genome. 2007b; 50: 172–179.
7. Lu Y, Gao X. Multiple mechanisms responsible for differential susceptibilities of *Sitobion avenae* (Fabricius) and *Rhopalosiphum padi* (Linnaeus) to pirimicarb. Bull Entomol Res. 2009; 99: 611–617. doi: 10.1017/S0007485309006725 PMID: 19413913
8. Duan XL, Peng X, Qiao XF, Chen MH. Life cycle and population genetics of bird cherry-oat aphids *Rhopalosiphum padi* (L.) (Hemiptera: Aphididae) in China: an important pest on wheat crops. J Pest Sci. 2017; 90: 103–116.
9. Leather SR, Dixon AFG. Growth, survival and reproduction of the bird-cherry aphid, *Rhopalosiphum padi*, on its primary host. Ann Appl Biol. 1981; 99: 115–118.
10. Pike KS, Schaffner RL. Development of autumn populations of cereal aphids, *Rhopalosiphum padi* (L.) and *Schizaphis graminum* (Rondani) (Homoptera: Aphididae) and their effects on winter wheat in Washington State. J Econ Entomol. 1985; 78: 676–680.
11. Tatchell GM, Plumb RT, Carter N. Migration of alate morphs of the bird cherry aphid (*Rhopalosiphum padi*) and implications for the epidemiology of barley yellow dwarf virus. Ann Appl Biol. 1988; 112: 1–11.
12. Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann Entomol Soc Am. 1994; 87: 651–701.
13. Delmotte F, Leterme N, Gauthier JP, Rispe C, Simon JC. Genetic architecture of sexual and asexual populations of the aphid *Rhopalosiphum padi* based on allozyme and microsatellite markers. Mol Ecol. 2002; 11: 711–723. PMID: 11972759
14. Carter MJ, Simon JC, Nespolo RF. The effects of reproductive specialization on energy costs and fitness genetic variances in cyclical and obligate parthenogenetic aphids. Ecol Evol. 2012; 2: 1414–1425. doi: 10.1002/ece3.247 PMID: 22957150
15. Ollivier M, Gabaldón T, Poulain J, Gavory F, Leterme N, Gauthier JP, et al. Comparison of gene repertoires and patterns of evolutionary rates in eight aphid species that differ by reproductive mode. Genome Biol Evol. 2012; 4: 155–167. doi: 10.1093/gbe/evr140 PMID: 22215803
16. Karsten M, van Vuuren BJ, Barnaud A, Terblanche JS. Population Genetics of *Ceratitis capitata* in South Africa: Implications for Dispersal and Pest Management. PLoS ONE. 2013; 8: e54281. doi: 10.1371/journal.pone.0054281 PMID: 23342117
17. Valenzuela I, Ridland PM, Weeks AR, Hoffmann AA. Patterns of genetic variation and host adaptation in an invasive population of *Rhopalosiphum padi* (Hemiptera: Aphididae). Ann Entomol Soc Am. 2010; 103: 886–897.
18. Karsten M, Addison P, van Vuuren BJ, Terblanche JS. Investigating population differentiation in a major African agricultural pest: evidence from geometric morphometrics and connectivity suggests high invasion potential. Mol. Ecol. 2016; 25: 3029–3032.
19. Tautz D, Renz M. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 1984; 12: 4127–4138. PMID: 6328411

20. Zhou Q, Luo D, Ma LC, Xie WG, Wang Y, Wang YR, et al. Development and cross-species transferability of EST-SSR markers in *Siberian wildrye* (Elym sibiricus L.) using Illumina sequencing. Sci Rep. 2016; 6: 20549. doi: 10.1038/srep20549 PMID: 26853106

21. Sunnucks P, England PR, Taylor AC, Hales DF. Microsatellite and chromosome evolution of parthenogenetic *Sitobion* aphids in Australia. Genetics. 1996; 144: 747–756. PMID: 8889535

22. Wilson ACC, Sunnucks P, Hales DFH. Microevolution, low clonal diversity and genetic affinities of parthenogenetic *Sitobion* aphids in New Zealand. Mol Ecol. 1999; 8: 1655–1666. PMID: 10583829

23. Llewellyn KS, Loxdale HD, Harrington R, Brookes CP, Clark SJ, Sunnucks P. Migration and genetic structure of the grain aphid (*Sitobion avenae*) in Britain related to climate and clonal fluctuation as revealed using microsatellites. Mol Ecol. 2003; 12: 21–34. PMID: 12492875

24. Papura D, Simon JC, Halkett F, Delmotte F, Le Gallic JF, Dedryver CA. Predominance of sexual reproduction in Romanian populations of the aphid *Sitobion avenae* inferred from phenotypic and genetic structure. Heredity. 2003; 90: 397–404. doi: 10.1038/sj.hdy.6800262 PMID: 12714986

25. Simon JC, Leterme N, Delmotte F, Martin O, Estoup A. Isolation and characterization of microsatellite loci in the aphid species, *Rhopalosiphum padi*. Mol Ecol Notes. 2001; 1: 4–5.

26. Duan XL, Qiao XF, Peng X, Zuo YY, Liu GM, Chen MH. Amplifying stability of microsatellite loci in different *Rhopalosiphum padi* (Linnaeus) populations in China. J Plant Protect. 2015; 42: 297–303.

27. Zane L, Bargelloni L, Patarnello T. Strategies for microsatellite isolation: a review. Mol Ecol. 2002; 11: 1–16. PMID: 11903900

28. Wilson ACC, Massonnet B, Simon JC, Prunier-Leterme N, Dolatti L, Llewellyn KS, et al. Cross-species amplification of microsatellite loci in aphids: assessment and application. Mol Ecol Notes. 2004; 4: 104–109.

29. Graveley BR, Brooks AN, Carlson J, Duff MO, Landolin JM, Yang L, et al. The developmental transcriptome of *Drosophila melanogaster*. Nature. 2011; 471: 473–479. doi: 10.1038/nature09715 PMID: 21179090

30. Chen X, Li J, Xiao S, Liu X. *De novo* assembly and characterization of foot transcriptome and microsatellite marker development for *Paphia textile*. Gene. 2016; 576: 537–543. doi: 10.1016/j.gene.2015.11.001 PMID: 26546834

31. Hou R, Bao ZM, Wang S, Su HL, Li Y, Du HX, Hu JJ, Wang S, Hu XL. Transcriptome sequencing and *de novo* analysis for Yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. Plos One. 2011; 6: e21560. doi: 10.1371/journal.pone.0021560 PMID: 21720557

32. Artigaud S, Thorne MA, Richard J, Lavaud R, Jean F, Flye-Sainte-Marie J, et al. Deep sequencing of the mantle transcriptome of the great scallop *Pecten maximus*. Marine Genomics. 2014; 15: 3–4. doi: 10.1016/j.margen.2014.03.006 PMID: 24731930

33. Uliano-Silva M, Americo JA, Brindeiro R, Dondero F, Prosdocimi F, de Freitas Rebelo M. Gene discovery through transcriptome sequencing for the invasive mussel *Limnoperna fortunei*. PLoS ONE. 2014; 9: e102973. doi: 10.1371/journal.pone.0102973 PMID: 25047650

34. Deng YW, Lei QN, Tian QL, Xie SH, Du XD, Li JH, et al. *De novo* assembly, gene annotation, and simple sequence repeat marker development using Illumina paired-end transcriptome sequences in the pearl oyster *Pinctada maxima*. Biosci Biotech Bioch. 2014; 78: 1685–1692.

35. Shiel BP, Hall NE, Cooke IR, Robinson NA, Strugnell JM. *De novo* characterisation of the greenlip abalone transcriptome (*Haliotis laevigata*) with a focus on the heat shock protein 70 (HSP70) family. Marine Biotech. 2014; 17: 23–32.

36. Zhang LL, Li L, Zhu YB, Guo XM. Transcriptome analysis reveals a rich gene set related to innate immunity in the eastern oyster (*Crassostrea virginica*). Mar Biotech. 2014; 16: 17–33.

37. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions, and mutational mechanisms: a review. Mol Ecol. 2002; 11: 2453–2465. PMID: 12453231

38. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet. 2002; 30: 194–200. doi: 10.1038/ng822 PMID: 11799393

39. Liu GH, Xu MJ, Song HQ, Wang CR, Zhu XQ. *De novo* assembly and characterization of the transcriptome of the pancreatic fluke *Eurytrema pancreaticum* (Trematoda: Dicrocoeliidae) using Illumina paired-end sequencing. Gene. 2016; 576: 333–338. doi: 10.1016/j.gene.2015.10.045 PMID: 26494161

40. Zhang LY, Bernard M, Leroy P, Feuillet C, Sourdille P. High transferability of bread wheat EST-derived SSRs to other cereals. Theor Appl Genet. 2005; 111: 677–687. doi: 10.1007/s00122-005-2041-5 PMID: 16034582

41. Liang XQ, Chen XP, Hong YB, Liu HY, Zhou GY, Li SX, et al. Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and Arachis wild species. BMC Plant Biol. 2009; 9: 35. doi: 10.1186/1471-2229-9-35 PMID: 19309524

42. Huang DN, Zhang YQ, Jin MD, Li HK, Song ZP, Wang YG, et al. Characterization and high cross-species transferability of microsatellite markers from the floral transcriptome of *Aspidistra saxicola* (Asparagaceae). Mol Ecol Resour. 2014; 14: 569–577. doi: 10.1111/1755-0998.12197 PMID: 24286608

43. Schwarz D, Robertson HM, Feder JL, Varala K, Hudson ME, Ragland GJ, et al. Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. BMC Genomics. 2009; 10: 1.

44. Bai X, Zhang W, Orantes L, Jun TH, Mittapalli O, Mian MR, et al. Combining next-generation sequencing strategies for rapid molecular resource development from an invasive aphid species, *Aphis glycines*. PLoS ONE. 2010; 5: e11370. doi: 10.1371/journal.pone.0011370 PMID: 20614011

45. Cook N, Aziz N, Hedley PE, Morris J, Milne L, Karley AJ, et al. Transcriptome sequencing of an ecologically important graminivorous sawfly: a resource for marker development. Conserv Genet Res. 2011; 3: 789–795.

46. Margam VM, Coates BS, Bayles DO, Hellmich RL, Agunbiade T, Seufferheld MJ, et al. Transcriptome sequencing, and rapid development and application of SNP markers for the legume pod borer *Maruca vitrata* (Lepidoptera: Crambidae). PLoS ONE. 2011; 6: e21388. doi: 10.1371/journal.pone.0021388 PMID: 21754987

47. Pascual L, Jakubowska AK, Blanca JM, Canizares J, Ferre J, Gloeckner G, et al. The transcriptome of *Spodoptera exigua* larvae exposed to different types of microbes. Insect Biochem Mole. 2012; 42: 557–570.

48. Li RQ, Zhu HM, Ruan J, Qian WB, Fang XD, Shi ZB, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20: 265–272. doi: 10.1101/gr.097261.109 PMID: 20019144

49. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389–3402. PMID: 9254694

50. Schuelke M. An economic method for the fluorescent labeling of PCR fragments. Nat biotechnol. 2000; 18: 233–234. doi: 10.1038/72708 PMID: 10657137

51. Goudet J. FSTAT Version 2.9.3.2 for windows: a computer program to calculate F-statistics; 2002. Available from: http://www2.unil.ch/popgen/softwares/fstat.

52. Park SDE. Trypanotolerance in West African cattle and the population genetic effects of selection. Ph. D. Thesis, University of Dublin. 2001.

53. Yeh FC, Yang R, Boyle TJ, Ye Z, Xiyan JM. PopGene32, Microsoft Windows-based freeware for population genetic analysis, version 1.32; 2000 [cited 1 December 2000]. Available from: https://sites.ualberta.ca/~fyeh/popgene.html.

54. Rousset F. Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour. 2008; 8: 103–106. doi: 10.1111/j.1471-8286.2007.01931.x PMID: 21585727

55. Ryman N, Palm S. POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. Mol Ecol Notes. 2006; 6: 600–602.

56. Weber JL. Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. Genomics. 1990; 7: 524–30. PMID: 1974878

57. Wang ZY, Fang BP, Chen JY, Zhang XJ, Luo ZX, Huang LF, et al. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). BMC Genomics. 2010; 11: 726. doi: 10.1186/1471-2164-11-726 PMID: 21182800

58. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, et al. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics. 2008; 9: 312. doi: 10.1186/1471-2164-9-312 PMID: 18590545

59. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, et al. Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics. 2009; 10: 219. doi: 10.1186/1471-2164-10-219 PMID: 19435504

60. Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS. *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics. 2010; 11: 400. doi: 10.1186/1471-2164-11-400 PMID: 20573269

61. Chen EH, Wei DD, Shen GM, Yuan GR, Bai PP, Wang JJ. *De novo* characterization of the *Dialeurodes citri* transcriptome: mining genes involved in stress resistance and simple sequence repeats (SSRs) discovery. Insect Mol Biol. 2014; 23: 52–66. doi: 10.1111/imb.12060 PMID: 24164346

62. Rispe C, Legeai F, Papura D, Bretaudeau A, Hudaverdian S, Le Trionnaire G, et al. *De novo* transcriptome assembly of the *Grapevine phylloxera* allows identification of genes differentially expressed between leaf-and root-feeding forms. BMC Genomics. 2016; 17: 1.

63. Liu MY, Qiao GR, Jiang J, Yang HQ, Xie LH, Xie JZ, et al. Transcriptome sequencing and *de novo* analysis for ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. Plos One. 2012; 7: e46766. doi: 10.1371/journal.pone.0046766 PMID: 23056442

64. Ostrander EA, Jong PM, Rine J, Duyk G. Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. P Natl Acad Sci USA. 1992; 89: 3419–3423.

65. Blair MW, Pedraza F, Buendia HF, Gaitán-Solís E, Beebe SE, Gepts P, et al. Development of a genome-wide anchored microsatellite map for common bean (*Phaseolus vulgaris* L.). Theor Appl Genet. 2003; 107: 1362–1374. doi: 10.1007/s00122-003-1398-6 PMID: 14504741

66. Karaiskou N, Buggiotti L, Leder E, Primmer CR. High degree of transferability of 86 newly developed zebra finch EST-linked microsatellite markers in 8 bird species. J Heredity. 2008; 99: 688–693.

67. Zheng X, Pan C, Diao Y, You Y, Yang C, Hu Z. Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). BMC Genomics. 2013; 14: 490. doi: 10.1186/1471-2164-14-490 PMID: 23870214

68. Archak S, Meduri E, Kumar PS, Nagaraju J. InSatDb: a microsatellite database of fully sequenced insect genomes. Nucleic Acids Res. 2007; 35: D36–D39. doi: 10.1093/nar/gkl778 PMID: 17082205

69. Yuan Y, Zhang LF, Wu GX, Zhu JY. High-throughput discovery microsatellites in *Tomicus yunnanensis* (Coleoptera: Scolytinae). J Environ Entomol. 2014; 36: 166–170.

70. Wei DD, Shi JX, Zhang XX, Chen SC, Wei D, Wang JJ. Analysis of microsatellite loci from *Bactrocera dorsalis* based on transcriptome dataset. Chinese J Appl Ecol. 2014; 25: 1799–1805.

71. Luo M, Zhang H, Bin SY, Lin JT. High-throughput discovery of SSR genetic markers in the mealybug, *Phenacoccus solenopsis* (Hemiptera: Pseudococcidae), from its transcriptome database. Acta Entomol Sin. 2014; 57: 395–400.

72. Comeault AA, Sommers M, Schwander T, Buerkle C, Farkas TE, Nosil P, et al. *De novo* characterization of the *Timema cristinae* transcriptome facilitates marker discovery and inference of genetic divergence. Mol Ecol Resour. 2012; 12: 549–561. doi: 10.1111/j.1755-0998.2012.03121.x PMID: 22339780

73. Wang B, Ekblom R, Castoe TA, Jones EP, Kozma R, Bongcam-Rudloff E, et al. Transcriptome sequencing of black grouse (*Tetrao tetrix*) for immune gene discovery and microsatellite development. Open Biol. 2012; 2: 120054. doi: 10.1098/rsob.120054 PMID: 22724064

74. Varshney RK, Sigmund R, Börner A, Korzun V, Stein N, Sorrells ME, et al. Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Sci. 2005. 168: 195–202.

75. Tang Y, Luo H, Nie K. Analysis of microsatellites from *Schistosoma mansoni* ESTs. Chinese J Prev Vet Medicine. 2007; 29: 629–633.

76. Zhu JY, Yang P, Zhang Z, Wu GX, Yang B. Transcriptomic immune response of *Tenebrio molitor* pupae to parasitization by *Scleroderma guani*. PLoS ONE. 2013; 8: e54411. doi: 10.1371/journal.pone.0054411 PMID: 23342153

77. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res. 2000; 10: 967–981. PMID: 10899146

78. Meglécz E, Nève G, Biffin E, Gardner MG. Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model eukaryote species. PLoS ONE. 2012; 7: e40861. doi: 10.1371/journal.pone.0040861 PMID: 22815847

79. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol. 2001; 18: 1161–1167. PMID: 11420357

80. Karaoglu H, Lee CMY, Meye W. Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol. 2005; 22: 639–649. doi: 10.1093/molbev/msi057 PMID: 15563717

81. Kim TS, Booth JG, Gauch HG, Sun Q, Park J, Lee YH, et al. Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics. 2008; 9: 31. doi: 10.1186/1471-2164-9-31 PMID: 18215294

82. Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, et al. Genome-wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium*. PLoS ONE. 2011; 6: e21298. doi: 10.1371/journal.pone.0021298 PMID: 21713003

83. Castagnone-Sereno P, Danchin EGJ, Deleury E, Guillemaud T, Malausa T, Abad P. Genome-wide survey and analysis of microsatellites in nematodes, with a focus on the plant-parasitic species *Meloidogyne incognita*. BMC Genomics. 2010; 11: 598. doi: 10.1186/1471-2164-11-598 PMID: 20973953

84. Bird A. CpG-rich islands and the function of DNA methylation. Nature. 1986; 321: 209–213. doi: 10.1038/321209a0 PMID: 2423876

85. Schorderet D, Gartler S. Analysis of CpG suppression in methylated and nonmethylated species. P Natl Acad Sci USA. 1992; 89: 957–961.

86. Durand J, Bodénès C, Chancerel E, Frigerio JJM, Vendramin G, Sebastiani F, et al. A fast and cost-effective approach to develop and map EST-SSR markers: Oak as a case study. BMC Genomics. 2010; 11: 570. doi: 10.1186/1471-2164-11-570 PMID: 20950475

87. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980; 32: 314. PMID: 6247908