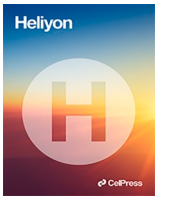
Heliyon 9 (2023) e12735

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Genome survey sequencing and mining of genome-wide microsatellite markers in yellow-billed babbler (*Turdoides affinis*)

Trisha Mondal [a,1], Prateek Dey [a,**,1], Divya Kumari [a], Swapna Devi Ray [a], Goldin Quadros [b], Venkata Hanumat Sastry Kochiganti [c], Ram Pratap Singh [a,*]

[a] *Department of Life Science, Central University of South Bihar, Gaya, 824236, India*
[b] *Wetland Ecology Division, Sálim Ali Centre for Ornithology and Natural History, Anaikatty, Coimbatore, 641108, Tamil Nadu, India*
[c] *National Institute of Animal Nutrition and Physiology, Bengaluru, 560030, India*

ARTICLE INFO

Keywords:
WGS
Microsatellites
Co-operative breeding
Turdoides affinis

ABSTRACT

Turdoides affinis is a species of group dwelling old world passerine of family Leiothrichidae. Unavailability of genome-wide sequence and species-specific molecular markers have hindered comprehensive understanding of cooperative breeding behaviour in *T. affinis*. Therefore, we generated genome-wide microsatellite markers through whole genome short read sequencing of *T. affinis*. A total of 68.8 gigabytes of paired-end raw data were sequenced containing 195,067,054 reads. Total sequenced reads spanned a coverage of 17X with genome size of 1.18 Gb. A large number of microsatellite markers (265,297) were mined in the *T. affinis* genome using Krait, and 50 most informative markers were identified and validated further. *In-silico* PCR results validated 47 markers. Of these 47 markers, five were randomly selected and validated *in-vitro* in twelve individuals of *T. affinis*. Genotyping data on these five loci estimated observed heterozygosity (H_0) and expected heterozygosity (H_e) ratios between 0.333 – 0.833 and 0.851–0.906, respectively. Effective allele size ranged from 6.698 to 10.667, inbreeding coefficient of the population ranged from 0.080 to 0.631 and null allele frequency was calculated at 0.055 to 0.303. Polymorphic information content of all the five loci varied between 0.850 and 0.906. Probabilities of exclusion and identity across 5 loci was estimated to be 0.95 and 0.0036, respectively. All the loci showed significant adherence to Hardy-Weinberg equilibrium. The microsatellite markers reported in this study will facilitate future population genetics studies on *T. affinis* and other congeneric species.

1. Introduction

Yellow billed babbler (*Turdoides affinis*) is an old world insectivore passerine of family Leiothrichidae [1,2]. Traditionally *T. affinis* was placed in the mega babbler family of Timilidae (275 species in 50 genera) [3], although recent taxonomic revisions have incorporated *T. affinis* in Leiothrichidae family. *T. affinis* prefers grassy shrub lands and strives sympatrically with various other babbler/allied species in the same habitat. Babbler and allied species are mostly sedentary with limited migration and characterized by

* Corresponding author.
** Corresponding author.
E-mail addresses: pratikdey23@gmail.com (P. Dey), rampratap@cusb.ac.in (R.P. Singh).
[1] Contributed equally.

https://doi.org/10.1016/j.heliyon.2022.e12735
Received 26 March 2022; Received in revised form 17 December 2022; Accepted 26 December 2022
Available online 3 January 2023

close ecological, morphological and behavioural similitude with other congeneric passerines [2,4,5]. Such an adaption leads to high sympatry amongst babblers, which in turn makes taxonomic and phylogenetic studies on babbler tediously vexing. Further with high level of sympatry, incongruences and non-phyletic clustering in babbler family is rampant and a universal systematic conclusion on babbler phylogeny is hard to reach [5,6]. For instance, position of *T. affinis* in either *Turdoides* or *Argya* genus within babbler phylogenetic tree is also debated and more in-depth studies are a must to confirm the same [7]. In the most recent and exhaustive phylogeny of 452 babbler species, Cai and workers [8] have placed *T. affinis* in *Argya* genus (sister to *Turdoides* genus) within the family Leiothrichidae.

From an evolutionary point of view although babblers are similar in life history traits and ecology, subtle differences in behaviour and physiology amongst its members has attracted attention of ecologists and evolutionary biologists alike. For instance *T. affinis* unlike most members of Leiothrichidae family display 'cooperative breeding' behaviour, a system where multiple individuals show care and parent-like qualities towards the offspring of a single nest/brood [9]. Though the phenomenon of cooperative breeding has been reported in about 9% of bird species worldwide (mostly in tropical climate), such a behaviour still puzzle biologists [10,11]. Explanation regarding the evolution of cooperative breeding in birds encompasses multiple scientific theories and studies have hinted at no straight or singular hypothesis for explanation of such behaviour [11]. Only sporadic attempts have been made to study cooperative breeding in *T. affinis* [12] without elucidating any deep insights to the context which is attributed to unavailability of species-specific molecular markers. Often cooperative breeding groups are apparently family groups in which cooperation is among relatives. It has been postulated that individuals of cooperatively breeding group can benefit from helping (genetic) relatives, thereby enhancing their own fitness through kin selection [13,14]. However, the Kin-selection hypothesis has never been tested in *T. affinis* due to the unavailability of molecular markers.

Since the early 1990s microsatellites has emerged as an indispensable tool for studying parentage and kinship patterns in birds [15]. Microsatellites being highly polymorphic and spread across multiple loci helped revolutionise parentage studies in natural populations of birds [15,16]. However, utilization of microsatellites for studying parentage pattern and kinship relation in a species, commands substantial initial investment in terms of generation of whole genome sequence, identification of microsatellite loci, designing of locus-specific primers, and subsequent in-vitro analysis [15]. Markers developed in model organisms or distantly related bird species, may not be robust enough for efficient relay of genetic information in *T. affinis*. Advances in massively parallel Next Generation Sequencing (NGS) technology over the last 20 years has made it possible to overcome such constraints with more efficient and powerful approaches. NGS can not only be used to generate whole genomes of non-model organisms but also extract thousands of

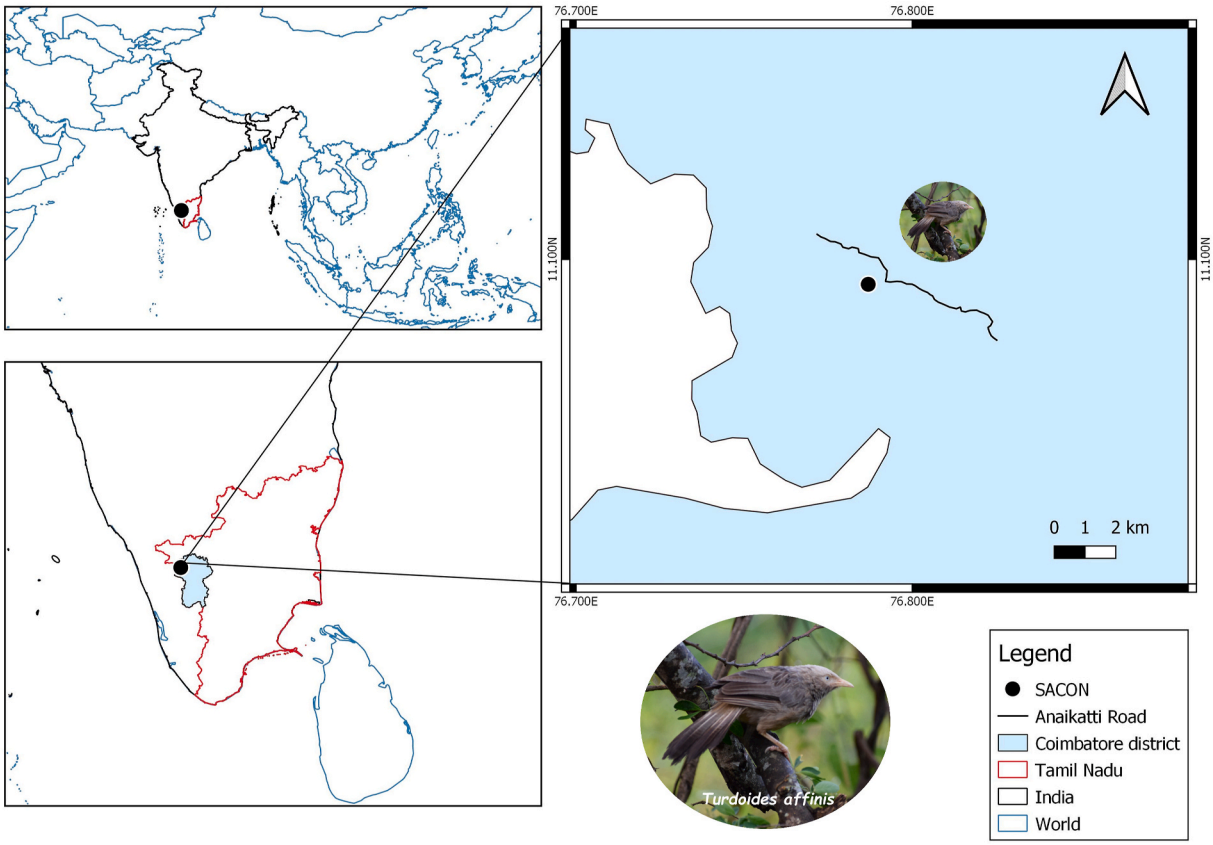


Fig. 1. Map of T. affinis sampling region. The location marked as 'Anaikatti road' on the map is the site of road-kill collection.

microsatellites or Short Sequence Repeats (SSRs) from these newly sequenced genomes [17,18]. Whole genome sequence of an organism provides a deep understanding of evolutionary characteristics of its repeat elements as well as the microsatellite motifs identified from them paints an unbiased picture into its *de-novo* mutation events [17,19]. Microsatellites are generally considered an extremely efficient molecular marker for population genetics studies due to their genetic co-dominance, dispersal throughout the genome, high polymorphism, reproducibility, transferability amongst individuals and ease of automatic allele detection [20–24]. Further, with the advent of 'Seq-to-SSR' approach, single/multiple libraries of the target species, is sufficient to generate thousands of microsatellite motifs from a single sequencing run [18,24,25]. Also, the 'Seq-to-SSR' approach is much preferred technique in low microsatellite abundant genomes such as birds, bats and corals [18,25].

Hence, we aimed to generate genome-wide microsatellite markers for *T. affinis* using short read sequencing, which may be informative for population and kinship genetic studies. Therefore we sequenced whole genome of a single individual of *T. affinis* through Illumina Next Seq 550 to generate millions of paired end reads, mined microsatellite motifs from the reads, predicted fifty most probable putative/polymorphic microsatellite loci through in-silico tools and validated five loci in the wet lab.

2. Materials and methods

2.1. Study area and sample collection

The samples of *T. affinis* used in this study were obtained as fresh road-kill specimens from Anaikatty Hills (Fig. 1). Anaikatty Hills (11.1048° N 76.7683° E), are a part of Western Ghats, one amongst the 35 biodiversity hotspots of the world [26]. As part of this high biodiverse region, Anaikatty hills harbour at least three species of group dwelling babblers [1]. Anaikatty hills are characterized by relatively undisturbed natural habitat for bird species including *T. affinis*. As population genetics indices are susceptible to various evolutionary and ecological processes, Anaikatty hills offer undisturbed natural sanctuary to its resident bird populations.

The collected samples were transported immediately to the lab for further processing under favourable conditions. Permission for collection of road-kill birds was obtained from Tamil Nadu Forest Department (Ref. No.WL5 (A)/2219/2018; Permit No. 14/2018) for this study. A total twelve road-kill samples were collected. Muscle tissue was sampled from the specimens and stored at −20 °C in DESS buffer (20% DMSO, 0.25 M tetra-sodium EDTA, Sodium Chloride till saturation, pH 7.5). Lysis buffer (10 mM Tris-pH 8.0, 10 mM EDTA-pH 8.0, and 100 mM NaCl) along with 40 μl of 20% SDS and 40 μl of Proteinase K was used to digest about 25 mg of the sampled tissues. DNA was extracted from the digested lysate using modified Phenol, Chloroform and Isoamyl alcohol method [27]. The quality of the isolated DNA was assessed on 1% agarose gel and quantified using spectrophotometer (DeNovix, USA) and Qubit 4 Fluorometer (ThermoFisher Scientific, USA).

2.2. Genome survey sequencing

Amongst the twelve samples of *T. affinis* collected, a select male specimen (sample no. 7) was used for generation of paired-end genomic libraries using TruSeq DNA PCR-Free library preparation kit (Illumina Inc., USA). About 1100 ng of the isolated DNA was used to generate paired-end genomic library of insert size 350 (2 × 150) base pairs. The DNA was fragmented using focused ultrasonicator (Covaris M220, USA) to a desired length as per the protocol recommended by the manufacturer. Following TruSeq DNA PCR-Free library preparation kit subsequent clean-up of the fragmented DNA, blunt-end creation and adapter ligation was performed [28]. The mean peak size of the generated library was checked using Fragment Analyzer AATI 5200 (Agilent, USA). QIAseq Library Quant Assay Kit (Qiagen N.V., Germany) was used to quantify the library and the selected library was normalized to 4 pico-moles before sequencing. The library was sequenced using NextSeq550 (Illumina Inc., USA) and at the end of the sequencing run, high-quality paired-end reads were obtained.

2.3. De novo genome assembly

FastQC was used to evaluate the quality of the raw sequences (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The raw data were de-multiplexed and trimming of adapters was performed using *bcl2fastq* software (Illumina Inc., USA). Reads with Phred (Q) score of 30 or above were only selected using *Seqtk* software for further downstream analysis [29]. Structural characteristics of the sequenced data were estimated using *KmerGenie* and *GCE* from the cleaned reads [30,31]. K-mer analysis on peak depth and best predicted K-mer selection was performed using *KmerGenie*. Depth of the genomes was assessed from raw reads using *GCE*. Further K-mer counting and graphical analysis of K-mer distribution was also done through *Jellyfish* and *GenomeScope* [32,33].

Subsequently high quality clean reads were utilized for de-novo assembly using *SOAPdenovo2* and *SPAdes* [34,35]. *SOAPdenovo2* employed *de Bruijn* graph method using an optimal Kmer depth as predicted by *KmerGenie* to assemble reads into large contigs/scaffolds. Similarly *SPAdes* used a multi Kmer approach employing *de Bruijn* graph method to assemble the cleaned reads into large contigs/scaffolds. Both *SOAPdenovo2* and *SPAdes* were executed using modified commands to obtain the best possible assembly. The assemblies were evaluated and compared for various parameters using QUAST-LG [36]. The completeness of the assemblies was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) by comparing the assembled *T. affinis* genome to single-copy orthologs of genes in Passeriformes order present in OrthoDB database (www.orthodb.org) [37].

2.4. Genome-wide microsatellite mining and primer design

Initially *PAL_FINDER* programme was employed to extract microsatellite motifs from raw sequence reads, which resulted low yield [18]. Therefore, the assembled genome through *SPAdes* was used for microsatellite motif identification using *Krait* [38]. *Krait* was employed to search all potential microsatellite motifs in the *T. affinis* de novo assembly. The search parameters used for detection of di-, tri-, tetra-, penta-, and hexa-nucleotide motifs were set at a minimum of 6, 5, 5, 5, 5 repeats respectively. *Primer3* embedded in *Krait* was used to design primers from flanking regions around the identified microsatellite motifs.

2.5. Primer screening, microsatellite marker validation and polymorphic detection

Default criterion in *Primer3* was used to design primers with a 100 base pair flanking region around the identified motifs. From the total primer pairs generated, 50 most polymorphic/putative microsatellite loci were selected for validation (*in-silico* and *in-vitro)*. The primer pair outputs were subjected to the following filters: (i) high number of repeats (a minimum of 10–20bp), these microsatellites were considered more polymorphic [21,25]; (ii) perfect microsatellites (no stray base composition within a repeat stretch present/devoid of composite microsatellites) were only selected, these are reported to be very stable and putative in the genome [25,39]; (iii) design of primers was done in such a way to avoid secondary structure and 3' end to target was flanked by 100 bp, the size of the product is selected between 200 and 300 bp and annealing temperature between 59°C-61 °C. All these conditions were considered to increase the stability of the primer and helped in amplifying the desired region containing the SSRs easily. These 50 select most putative primer pairs were validated *in-silico* using *FastPCR* software [40]. Of these 50 putative primers, five primers were selected randomly for *in-vitro* validation. Twelve *T. affinis* specimen collected opportunistically as road-kill was utilized for initial primer validation and genetic diversity analysis. PCR amplification was performed in 25 μl volume reactions using 12.5 μl Qiagen Type-it microsatellite (Qiagen N.V., Germany) PCR kit master mix, 1 μl forward and reverse primer (10 nM conc. each), 30 ng template DNA, 2.5 μl Q-solution (provided in Type-it kit) and 5 μl nuclease free water. The microsatellites were amplified under suitable conditions (Supplementary 1). All the PCR products were visualized under UV Gel documentation system in a 2.5% agarose gel.

Subsequently to identify peak size and density, the PCR products were run on capillary electrophoresis system Agilent Fragment Analyzer (AATI 5200). To separate the PCR amplicons 'dsDNA 905 (1-500bp)' kit (Agilent, USA) along with 55 cm AATI 5200 capillary (Agilent, USA) was used. The dsDNA 905 kit specified a DNA sizing range of 35–500 bp, with ±5% accuracy and low loading concentration up to 0.5 ng/μl, whereas the 55 cm capillary specified a separation resolution of 1–3 bp when used with the same kit. At first, the PCR products were standardised to a 20 ng/μl concentration using Qubit 4 Fluorometer. The standardised PCR products were diluted to three different concentrations (1:10, 1:5 and 1:2) with nuclease free water, amongst which the dilution concentration of 1:2 was found to be most suitable for further usage. Finally, 2 μl of diluted sample was added to 20 μl of diluent buffer (provided with the kit) and run on AATI 5200 multiple times for peak scoring and genotyping. An aliquot of the dsDNA marker was also incorporated with each run for size calibration. A peak in a PCR product was only scored when at least two attempts estimated exactly the same peak size

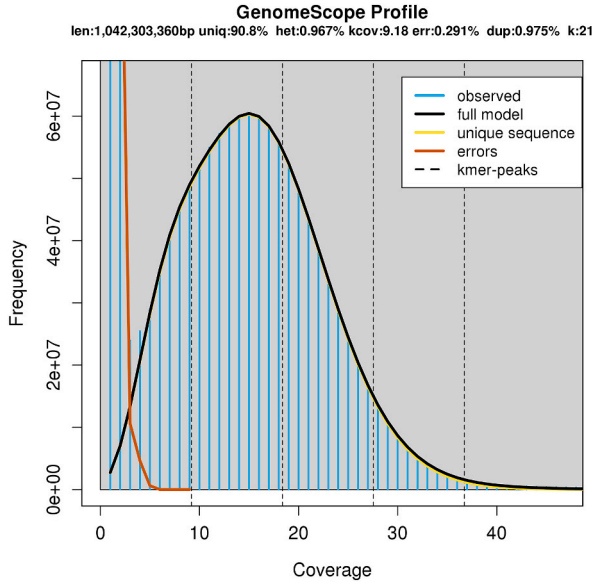


Fig. 2. Pattern of K-mer distribution of T. affinis sequence reads. K-mer distribution: blue bars; Modelled distribution without the error K-mers (red line): black line; Maximum K-mer coverage specified in the model: yellow line. Len: estimated total genome length from pair end reads; Uniq: non-repetitive portion of the genome (useful for assembly); Het: heterozygosity rate of reads; Kcov: mean K-mer coverage for heterozygous bases; Err: error rate in reads; Dup: duplication rate. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Comparison of *T. affinis* assembly (*SOAPdenovo2* and *SPAdes*) metrics and quality with similar short read assemblies of *A. rosecollis* [49]; *M. undulatus*, *G. gallus*, *T. guttata*, *F. peregrinus* and *N. notabilis* [46]; *A. vittata* [52] and *A. macao* [53].

Assembly	*T. affinis* (*SOAPdenovo2* assembly)	*T. affinis* (*SPAdes* assembly)	*Agapornis roseicollis*	*Melopsittacus undulatus*	*Gallus gallus*	*Taeniopygia guttata*	*Amazona vittata*	*Ara macao*	*Falco peregrinus*	*Nestor notabilis*
Genome size (Giga bases)	1.13	1.18	1.1	1.2	1.21	1.2	1.58	1.2	1.2	1.14
Sequencing depth	Calculated from raw reads	17X	100X	160X	50.6X	6X	27X	16X	105X	32X
No. of contigs	847,340	299,202	–	–	–	–	–	–	–	–
Largest contig	69,610	322,963	–	–	–	–	–	–	–	–
GC (%)	41.9	42.06	–	–	–	–	–	–	–	–
N50 (contigs)	2.0 K	14.2 K	5.4 K	55.6 K	2800 K	38.5 K	6.9 K	6.3 K	28.5 K	16 K
N75	906	4920	–	–	–	–	–	–	–	–
L50	126,376	18,417	–	–	–	–	–	–	–	–
L75	340,682	50,614	–	–	–	–	–	–	–	–
No. of N's per 100 per kilo base pairs	4896.39	1223.44	–	–	–	–	–	–	–	–

and the area under the peak was more than 10% [41]. Following the above mentioned criteria, the peaks were identified for genotyping purpose.

2.6. Genetic diversity analysis

All the twelve *T. affinis* individuals were collectively considered to be a single population. From the microsatellite scores of fragment analysis, the allele sizes were estimated and a comprehensive genotype table was prepared. GenAlEx 6.51b2 was used to calculate no. of Effective Alleles, no. of Alleles, Fixation Index (or In-breeding coefficient), Expected Heterozygosity, Observed Heterozygosity, deviation from Hardy-Weinberg equilibrium and probabilities of exclusion (PoE) as well as identity (PoI) [42]. The Polymorphic Information Content (PIC) value for each loci was calculated manually following the formula: $1 - \sum_{i=1}^{n} P(i^2)$; where P is the frequency of ith allele [43]. The null allele frequencies were estimated using software FreeNA [44]. The aim of such an analysis with limited sample size was done to validate identified microsatellite markers and test the feasibility of our designed work flow.

3. Results and discussion

3.1. Genome size prediction and characteristics of paired end reads

About 68.8 gigabytes of paired-end raw data were generated using ~350 base pair inserts containing 195,067,054 (2 × 97,533,527) reads. The sequence read length ranged between 110 and 151 base pairs and GC content of the raw reads was calculated at 45%. *KmerGenie* estimated best K size at 55 for the cleaned reads. Genome size predicted by *KmerGenie* was estimated to be 1,180,249,703 (1.18) Gb (Supplementary 2). The depth was estimated at ~17X from the raw reads using *GCE*. *Jellyfish* and *GenomeScope* calculated heterozygosity, duplication rate and repeat ratio of the cleaned reads at 0.97%, 0.97% and 10.2%, respectively with an estimated genome size of 1.1 Gb (Fig. 2).

Low heterozygosity (0.97%) and repeat ratio (10.2%) of the raw reads aid assemblers like *SOAPDenovo2* or *SPAdes* to assemble high quality genome assembly with relative ease [45]. The genome size of *T. affinis* reported in this study is found similar to that of other bird species reported by previous workers [18,46–49]. For instance, the *de-novo* genome assemblies of 45 bird species sequenced by Zhang and workers ranged from 1.05 Gb to 1.26 Gb in size [46].

3.2. Genome sequencing assembly

Evaluation of both the assemblies on various parameters suggested the assembly generated through *SPAdes* was the best (Table 1, Fig. 3). *SPAdes* using the cleaned reads generated 299,202 contigs containing 1087735178 base pairs. The largest contig length of *SPAdes* assembly was estimated at 322,963 base pairs as compared to 69,610 base pairs of *SOAPdenovo2* assembly. The N50 value of *SPAdes* assembly was estimated at 14,247 (14.2 K) as opposed to 2025 of *SOAPdenovo2* assembly. Missing nucleotides in the *SPAdes*

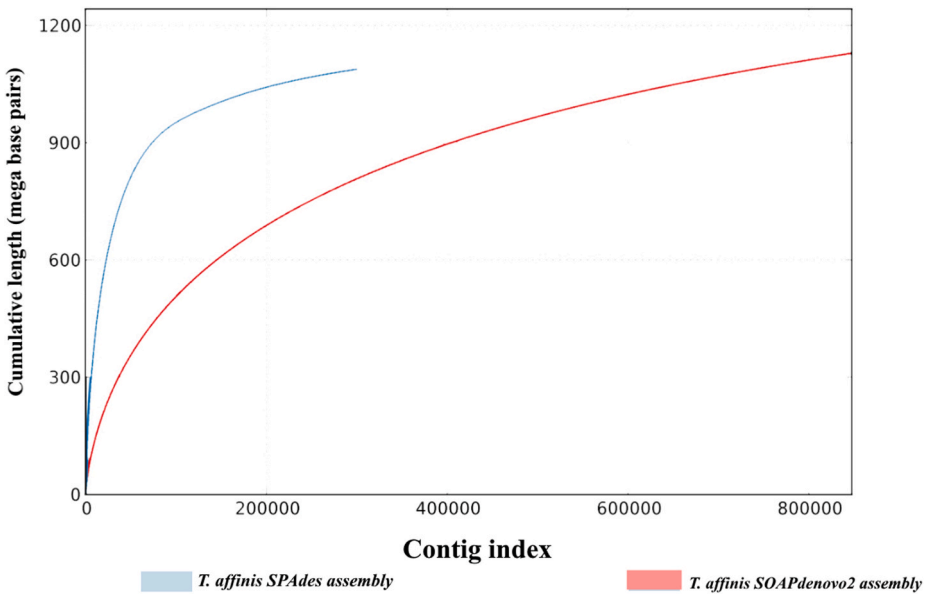


Fig. 3. Comparison of contig in both the assemblies of T. affinis assembly (SOAPdenovo2 and SPAdes). The contig length (on Y-axis) for each contig index (on X-axis) shows that SPAdes assembly has fewer contig fragments but larger size of the contig, creating a more cohesive assembly with fewer errors.

Table 2
Comparison of *T. affinis* assembly (*SOAPdenovo2* and *SPAdes*) metrics and quality with similar short read assemblies of *A. rosecollis* [49]; *M. undulatus*, *G. gallus*, *T. guttata*, *F. peregrinus* and *N. notabilis* [46]; *A. vittata* [52] and *A. macao* [53] using BUSCO.

Assembly	*T. affinis* (*SOAPdenovo2* assembly)	*T. affinis* (*SPAdes* assembly)	*Agapornis roseicollis*	*Melopsittacus undulatus*	*Gallus gallus*	*Taeniopygia guttata*	*Amazona vittata*	*Ara macao*	*Falco peregrinus*	*Nestor notabilis*
Complete (%)	18.3	44.7	85.2	81.2	88.8	81.9	59.5	45.2	88.2	70.3
Missing (%)	74	45.9	7.5	14.2	10.5	5.6	16.1	30	7.5	14.8
Fragmented (%)	7.7	9.4	7.3	4.6	0.7	12.5	24.4	24.8	4.3	14.9

assembly was estimated at 1223.44 counts per 100 kilo base pairs of the assembly as opposed to 4896.39 counts per 100 kilo base pairs in *SOAPdenovo2* assembly. The completeness of both the assemblies was estimated using BUSCO. *SPAdes* assembly of *T. affinis* showed 44.7% of all orthologous gene sets in Passeriformes order were present whereas *SOAPdenovo2* assembly showed a presence of only 18.7% genes (Table 2, Fig. 4). *SOAPdenovo2* assembly showed 8025 missing genes, as opposed to 4965 missing genes in *SPAdes* assembly. Overall comparison of both the assemblies through QUAST-LG and BUSCO favoured downstream analysis of *SPAdes* assembly for microsatellite motif identification on all indices.

The newly assembled *T. affinis* genome (*SPAdes* assembly) was compared on various indices with selected previously published avian genomes (Tables 1 and 2). GC content of previously reported avian genomes ranged from 39% to 42% [48], similar to the GC percentage of *SPAdes* assembly of *T. affinis* estimated at 42.06%. Contig N50 values of previously reported avian genomes ranged from 4.6 K to 55 K [46,48,50], similar to *T. affinis* genome sequenced in this study (14.5 K). Previous studies on avian genomes reported BUSCO of genome assemblies scores between 45.2 and 88.8% [49,50]. Low BUSCO score of *T. affinis* genome (*SPAdes* assembly; 44.7%) in this study may be attributed to exclusive usage of short reads for sequence assembly. Previous reports noted that employing large insert libraries during library preparation increased the quality indices during genome assembly [46,48].

3.3. Microsatellite marker discovery

A total of 197,505 potential microsatellite motifs were identified using *PAL_FINDER* from raw sequence reads (Supplementary 3). Avian genomes are characterized by low SSRs densities [18,46]. Hence, we assumed motif identification from raw reads may result in loss of many microsatellite loci as corroborated by previous studies on *Coragyps atratus* genome [47].

Subsequently *T. affinis* genome generated through *SPAdes* assembler was used to identify microsatellite motifs through *Krait*, which resulted into 25% more microsatellite discovery. A total of 265,297 microsatellite motifs were identified by *Krait*, averaging around 23.52 base pairs in length and covering 0.49% of the assembly sequences (Table 3, Fig. 5). Microsatellite distribution frequency in the sequenced genome was calculated at 208.74 microsatellites per mega base (Mb) pairs of the assembly. Ignoring very few mononucleotides, the mined microsatellite motifs contained 174,020 (65.59%) dinucleotide, 28,304 (10.67%) trinucleotide, 45,321 (17.08%) tetranucleotide, 13,973 (5.27%) pentanucleotide and 3670 (1.38%) hexanucleotide (Table 3, Fig. 5). The repeats of AA, AAAG, CC, AC, AT, AG, AAT, AGG, AACGG, and AAC were the most abundant repeats respectively; interspersed throughout the genome (Fig. 5). Previous studies on avian genomes identified 90,346–282,728 microsatellite motifs covering 0.13%–0.49% of their respective assemblies [48]. Further relative abundance of microsatellite motifs from assembled avian genomes were estimated between 80.9 and 256.9 microsatellites per Mb of their respective assemblies [48]. The number and abundance of microsatellite motifs mined from *T. affinis SPAdes* assembly is similar to previously published reports on avian genomes. Huang and co-workers [48] observed AT, AC, AG, AAT, and AAAG repeats were the most abundant pattern in previously reported avian genomes very similar to most abundant repeats identified in *T. affinis* genome. Although exact microsatellite motif abundance changes from one species to another, similarity in most abundant motifs in di-, tri- and tetra-nucleotide repeats could still be observed across all avian genomes

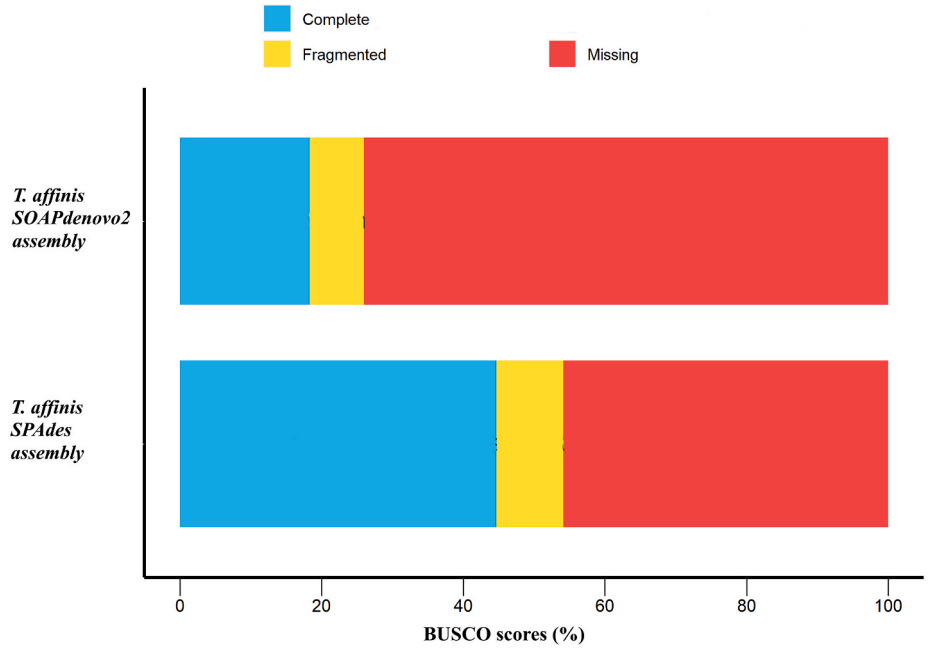


Fig. 4. The BUSCO assessment pipeline was applied to both the T. affinis assembly (SOAPdenovo2 and SPAdes). With higher number of complete gene sets, SPAdes assembly was estimated to be employed for further downstream analysis.

Table 3
Distribution and frequency of different microsatellite motifs mined from *T. affinis* assembly (*SPAdes*).

Type of microsatellite repeats	Counts	Length (base pairs)	Percent (%)	Average Length (base pairs)	Relative Abundance (loci/Mega base pairs)	Relative Density (loci/Mega base pairs)
Dinucleotides	174,020	3,143,296	65.59	18.06	136.92	2473.16
Trinucleotides	28,304	536,982	10.67	18.97	22.27	422.5
Tetranucleotides	45,321	1,799,576	17.08	39.71	35.66	1415.91
Pentanucleotides	13,973	583,975	5.27	41.79	10.99	459.47
Hexanucleotides	3670	174,810	1.38	47.63	2.89	137.54

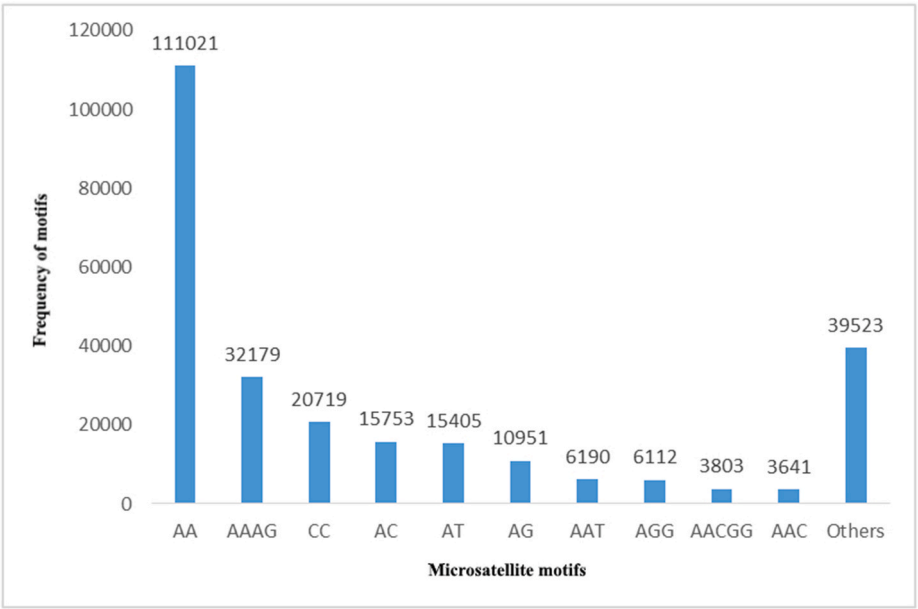


Fig. 5. Comparative characterization of the 10 most abundant microsatellite repeats interspersed throughout the T. affinis genome.

discussed in this study.

3.4. *In-silico and in-vitro validation of microsatellite markers and genetic diversity analysis*

A set of 50 most polymorphic/putative primers were selected for *in-silico* validation (Supplementary 4). Except 3 primer pairs, remaining 47 primer pairs displayed successful amplification when subjected to *in-silico* PCR through *FASTPCR* software (Supplementary 5). Of these 47 pairs, a set of five primers (Supplementary_1) was randomly selected and used for microsatellite loci validation though PCR in twelve samples of *T. affinis*. All five primer pairs displayed successful amplification in all *T. affinis samples* (Supplementary 6). The amplicons varied from 170 to 500 base pairs. The allele sizes also varied within the same range as confirmed by fragment analysis (Supplementary 7).

Observed heterozygosity (H_0) and expected heterozygosity (H_e) ratios range between 0.333 – 0.833 and 0.851–0.906, respectively

Table 4
Attributes of *T. affinis* population (12 individual) investigated in this study through genotyping of 5 putative microsatellite loci. The summary statistics are as detailed: E_{allele}: No. of Effective Alleles = 1/(Sum pi^2), where pi is the frequency of the ^{i}th allele for the population & Sum pi^2 is the sum of the squared population allele frequencies; F_{IS} - Fixation Index (or In-breeding coefficient) = (He - Ho)/He = 1 - (Ho/He); H_e - Expected Heterozygosity; H_o - Observed Heterozygosity; NA_F – Null allele frequency; and p_{HWE} – p value of Hardy-Weinberg equilibrium; PIC: polymorphic information content for each loci.

Primer name	E_{allele}	F_{IS}	H_e	H_o	NA_F	p_{HWE}	PIC
Di_1	10.286	0.538	0.903	0.417	0.25733	0.001	0.902
Tri_3	9.600	0.442	0.896	0.500	0.21197	0.002	0.895
Tetra_1	10.286	0.631	0.903	0.333	0.30348	0.000	0.902
Penta_1	10.667	0.080	0.906	0.833	0.05556	0.032	0.906
Hexa_3	6.698	0.412	0.851	0.500	0.17283	0.000	0.850

(Table 4). Allele numbers varied from 12 to 15 for the five loci, with effective allele size ranged from 6.698 to 10.667. The inbreeding coefficient (F_{IS}) of the population ranged from 0.080 to 0.631 (Table 4). The Null allele frequency was estimated between 0.055 and 0.303. PIC values for each loci ranged from 0.850 to 0.906 (Table 4). Average probability of exclusion (PoE) and probability of identity (PoI) across increasing locus combinations were calculated at 0.95 and 0.003, respectively (Supplementary 8). High PIC value of all the five primers (average 0.89) analysed in this study indicates very high polymorphism at the loci, making them suitable for utilization in various population genetic studies (42). The results of PoE and PoI report high exclusion probabilities (0.95) and low probability of identity (0.0036) indicating these loci may be optimum for individual identification and parentage analysis [51]. All the loci showed significant adherence to Hardy-Weinberg equilibrium indicating no drift. Based on *in-vitro* validation results of five *in-silico* validated microsatellite markers, we presume that all 47 microsatellite markers identified in this study are highly informative, and would be highly useful for future population genetics studies on *T. affinis* to identify drivers of cooperative breeding behaviour.

4. Conclusions

Overall, we successfully decoded the whole genome of *T. affinis*, identified a large number of microsatellite markers, validated them *in-silico* and also in the wet lab using a limited number of samples. These microsatellite markers will be of immense help for future studies in identifying the drivers of cooperative breeding behaviour in *T. affinis*. Preliminary population genetics information will benefit future studies on natural selection, gene flow, genetic drift and breeding systems of *T. affinis* species. In addition, the whole genome data would be highly useful for elucidating evolutionary history and the phylogenetic position of this species in the tree of life. Further, the genome sequence generated here can facilitate comprehensive genetic mapping, candidate gene identification and alleviate the content of genomic information of *T. affinis*.

Author contribution statement

Ram Pratap Singh, Sastry Venkata Hanumat Kochiganti, and Goldin Quadros: Conceived and designed the experiments.
Prateek Dey, Trisha Mondal, Divya Kumari, and Swapna Devi Ray: Performed the experiments.
Prateek Dey, Trisha Mondal, Divya Kumari: Analysed and interpreted the data.
Ram Pratap Singh, GQ: Contributed reagents, materials, analysis tools or data.
Trisha Mondal, Prateek Dey, Ram Pratap Singh, Divya Kumari, Swapna Devi Ray, Sastry Venkata Hanumat Kochiganti, and Goldin Quadros: Wrote the paper.

Funding statement

Dr. Ram Pratap Singh was supported by Science and Engineering Research Board [CRG/2020/002439].

Data availability statement

Data associated with this study has been deposited at NCBI GenBank under the accession number BioProject, BioSample and SRR numbers are PRJNA811562, SAMN26343563, and SRR18188812 respectively.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Permissions: Dr. P. Pramod, Senior Principal Scientist, Sálim Ali Centre for Ornithology and Natural History (SACON) for permission regarding *T. affinis* samples stored at SACON.
The research work was funded by the Science and Engineering Research Board (SERB), Government of India (CRG/2020/002439). TM was supported with fellowship (F.82-44/2020 (SA-III)) by the University Grant Commission (UGC).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.heliyon.2022.e12735.

References

[1] R. Grimmett, C. Inskipp, T. Inskipp, Birds of the Indian Subcontinent: India, Pakistan, Sri Lanka, Nepal, Bhutan, Bangladesh and the Maldives, Bloomsbury Publishing, 2016.

[2] A. Cibois, M.V. Kalyakin, H. Lian-Xian, E. Pasquet, Molecular phylogenetics of babblers (Timaliidae): revaluation of the genera Yuhina and Stachyris, J. Avian Biol. 33 (4) (2002) 380–390.
[3] Dickinson EC, Bahr N, Dowsett R, Pearson D, Remsen V, Roselaar CS, Schodde D. The Howard and Moore Complete Checklist of Birds of the World.
[4] M. Gelang, A. Cibois, E. Pasquet, U. Olsson, P. Alström, P.G. Ericson, Phylogeny of babblers (Aves, Passeriformes): major lineages, family limits and classification, Zool. Scripta 38 (3) (2009) 225–236.
[5] R.G. Moyle, M.J. Andersen, C.H. Oliveros, F.D. Steinheimer, S. Reddy, Phylogeny and biogeography of the core babblers (Aves: timaliidae), Syst. Biol. 61 (4) (2012) 631–651.
[6] A. Cibois, M. Gelang, P. Alström, E. Pasquet, J. Fjeldså, P.G. Ericson, U. Olsson, Comprehensive phylogeny of the laughingthrushes and allies (Aves, Leiothrichidae) and a proposal for a revised taxonomy, Zool. Scripta 47 (4) (2018) 428–440.
[7] B. Campbell, E. Lack (Eds.), A Dictionary of Birds, A&C Black, 2011.
[8] T. Cai, A. Cibois, P. Alström, R.G. Moyle, J.D. Kennedy, S. Shao, R. Zhang, M. Irestedt, P.G. Ericson, M. Gelang, Y. Qu, Near-complete phylogeny and taxonomic revision of the world's babblers (Aves: Passeriformes), Mol. Phylogenet. Evol. 130 (2019) 346–356.
[9] B.J. Hatchwell, J. Komdeur, Ecological constraints, life history traits and the evolution of cooperative breeding, Anim. Behav. 59 (6) (2000) 1079–1086.
[10] A. Cockburn, Prevalence of different modes of parental care in birds, Proc. Biol. Sci. 273 (1592) (2006) 1375–1383.
[11] B.J. Hatchwell, The evolution of cooperative breeding in birds: kinship, dispersal and life history, Phil. Trans. Biol. Sci. 364 (1533) (2009) 3217–3227.
[12] A.F. Skutch, Helpers at Birds' Nests: a Worldwide Survey of Cooperative Breeding and Related Behavior, University of Iowa Press, 1999.
[13] C. Riehl, Evolutionary routes to non-kin cooperative breeding in birds, Proc. Biol. Sci. 280 (1772) (2013), 20132245.
[14] D.F. Westneat, M.S. Webster, Molecular analysis of kinship in birds: interesting questions and useful techniques, in: Molecular Ecology and Evolution: Approaches and Applications, Birkhäuser, Basel, 1994, pp. 91–126.
[15] S.P. Flanagan, A.G. Jones, The future of parentage analysis: from microsatellites to SNPs and beyond, Mol. Ecol. 28 (3) (2019) 544–567.
[16] L. Brouwer, S.C. Griffith, Extra-pair paternity in birds, Mol. Ecol. 28 (22) (2019) 4864–4882.
[17] J.W. Davey, P.A. Hohenlohe, P.D. Etter, J.Q. Boone, J.M. Catchen, M.L. Blaxter, Genome-wide genetic marker discovery and genotyping using next-generation sequencing, Nat. Rev. Genet. 12 (7) (2011) 499–510.
[18] T.A. Castoe, A.W. Poole, A.J. De Koning, K.L. Jones, D.F. Tomback, S.J. Oyler-McCance, J.A. Fike, S.L. Lance, J.W. Streicher, E.N. Smith, D.D. Pollock, Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake, PLoS One 7 (2) (2012), e30953.
[19] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (6) (2004) 435–445.
[20] N.B. Freimer, M. Slatkin, Microsatellites: evolution and mutational processes, Variation in the Human Genome 197 (1996) 51–67.
[21] C.R. Primmer, H. Ellegren, Patterns of molecular evolution in avian microsatellites, Mol. Biol. Evol. 15 (8) (1998) 997–1008.
[22] C. Duran, N. Appleby, D. Edwards, J. Batley, Molecular genetic markers: discovery, applications, data storage and visualisation, Curr. Bioinf. 4 (1) (2009) 16–27.
[23] K.A. Selkoe, R.J. Toonen, Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers, Ecol. Lett. 9 (5) (2006) 615–629.
[24] I. Fernandez-Silva, J. Whitney, B. Wainwright, K.R. Andrews, H. Ylitalo-Ward, B.W. Bowen, R.J. Toonen, E. Goetze, S.A. Karl, Microsatellites for next-generation ecologists: a post-sequencing bioinformatics pipeline, PLoS One 8 (2) (2013), e55990.
[25] M.G. Gardner, A.J. Fitch, T. Bertozzi, A.J. Lowe, Rise of the machines–recommendations for ecologists when using next generation sequencing for microsatellite development, Molecular Ecology Resources 11 (6) (2011) 1093–1101.
[26] R.F. Noss, W.J. Platt, B.A. Sorrie, A.S. Weakley, D.B. Means, J. Costanza, R.K. Peet, How global biodiversity hotspots may go unrecognized: lessons from the North American Coastal Plain, Divers. Distrib. 21 (2) (2015) 236–244.
[27] J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning: a Laboratory Manual, Cold spring harbor laboratory press, 1989.
[28] P. Dey, S.K. Sharma, I. Sarkar, S.D. Ray, P. Pramod, V.H. Kochiganti, G. Quadros, S.S. Rathore, V. Singh, R.P. Singh, Complete mitogenome of endemic plum-headed parakeet Psittacula cyanocephala–characterization and phylogenetic analysis, PLoS One 16 (4) (2021), e0241098.
[29] W. Shen, S. Le, Y. Li, F. Hu, SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation, PLoS One 11 (10) (2016), e0163962.
[30] R. Chikhi, P. Medvedev, Informed and Automated K-Mer Size Selection for Genome Assembly, HiTSeq, 2013.
[31] B. Liu, Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, W. Fan, Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects, 2013 arXiv preprint arXiv:1308.2012.
[32] G.W. Vurture, F.J. Sedlazeck, M. Nattestad, C.J. Underwood, H. Fang, J. Gurtowski, M.C. Schatz, GenomeScope: fast reference-free genome profiling from short reads, Bioinformatics 33 (14) (2017) 2202–2204.
[33] G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics 27 (6) (2011) 764–770.
[34] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience 1 (1) (2012), 2047-17X.
[35] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477.
[36] A. Mikheenko, A. Prjibelski, V. Saveliev, D. Antipov, A. Gurevich, Versatile genome assembly evaluation with QUAST-LG, Bioinformatics 34 (13) (2018) i142–i150.
[37] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (19) (2015) 3210–3212.
[38] L. Du, C. Zhang, Q. Liu, X. Zhang, B. Yue, Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design, Bioinformatics 34 (4) (2018) 681–683.
[39] M. Liu, X. Wang, L. Ma, L. Cao, H. Liu, D. Pu, S. Wei, Genome-wide developed microsatellites reveal a weak population differentiation in the hoverfly Eupeodes corollae (Diptera: syrphidae) across China, PLoS One 14 (9) (2019), e0215888.
[40] R. Kalendar, D. Lee, A.H. Schulman, FastPCR software for PCR, in silico PCR, and oligonucleotide assembly and analysis, in: DNA Cloning and Assembly Methods, Humana Press, Totowa, NJ, 2014, pp. 271–302.
[41] A. Bastías, F. Correa, P. Rojas, R. Almada, C. Munoz, B. Sagredo, Identification and characterization of microsatellite loci in maqui (Aristotelia chilensis [Molina] Stunz) using next-generation sequencing (NGS), PLoS One 11 (7) (2016), e0159825.
[42] R. Peakall, P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update, Bioinformatics 28 (2012) 2537–2539.
[43] D. Botstein, R.L. White, M. Skolnick, R.W. Davis, Construction of a genetic linkage map in man using restriction fragment length polymorphisms, Am. J. Hum. Genet. 32 (3) (1980) 314.
[44] M.P. Chapuis, M. Lecoq, Y. Michalakis, A. Loiseau, G.A. Sword, S. Piry, A. Estoup, Do outbreaks affect genetic population structure? A worldwide survey in Locusta migratoria, a pest plagued by microsatellite null alleles, Mol. Ecol. 17 (16) (2008) 3640–3653.
[45] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, Genome Res. 24 (8) (2014) 1384–1395.
[46] G. Zhang, C. Li, Q. Li, B. Li, D.M. Larkin, C. Lee, J.F. Storz, A. Antunes, M.J. Greenwold, R.W. Meredith, A. Ödeen, Comparative genomics reveals insights into avian genome evolution and adaptation, Science 346 (6215) (2014) 1311–1320.
[47] D.J. Wostenberg, J.A. Fike, S.J. Oyler-McCance, M.L. Avery, A.J. Piaggio, Development of microsatellite loci for two New World vultures (Cathartidae), BMC Res. Notes 12 (1) (2019) 1–6.
[48] J. Huang, W. Li, Z. Jian, B. Yue, Y. Yan, Genome-wide distribution and organization of microsatellites in six species of birds, Biochem. Systemat. Ecol. 67 (2016) 95–102.
[49] H. Van der Zwan, F. Van der Westhuizen, C. Visser, R. Van der Sluis, Draft de novo genome sequence of Agapornis roseicollis for application in avian breeding, Anim. Biotechnol. 29 (4) (2018) 241–246.

[50] A.L. Ducrest, S. Neuenschwander, E. Schmid-Siegert, M. Pagni, C. Train, D. Dylus, Y. Nevers, A. Warwick Vesztrocy, L.M. San-Jose, M. Dupasquier, C. Dessimoz, New genome assembly of the barn owl (*Tyto alba alba*), Ecol. Evol. 10 (5) (2020) 2284–2298.

[51] V. Costa, J. Pérez-González, P. Santos, P. Fernández-Llario, J. Carranza, A. Zsolnai, I. Anton, J. Buzgó, G. Varga, N. Monteiro, A. Beja-Pereira, Microsatellite markers for identification and parentage analysis in the European wild boar (Sus scrofa), BMC Res. Notes 5 (1) (2012) 1–6.

[52] T.K. Oleksyk, J.F. Pombert, D. Siu, A. Mazo-Vargas, B. Ramos, W. Guiblet, Y. Afanador, C.T. Ruiz-Rodriguez, M.L. Nickerson, D.M. Logue, M. Dean, A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education, GigaScience 1 (1) (2012), 2047-17X.

[53] C.M. Seabury, S.E. Dowd, P.M. Seabury, T. Raudsepp, D.J. Brightsmith, P. Liboriussen, Y. Halley, C.A. Fisher, E. Owens, G. Viswanathan, I.R. Tizard, A multi-platform draft de novo genome assembly and comparative analysis for the scarlet macaw (*Ara macao*), PLoS One 8 (5) (2013), e62415.