Characterization of Duck (*Anas platyrhynchos*) Short Tandem Repeat Variation by Population-Scale Genome Resequencing

Wenlei Fan[1,2], Lingyang Xu[1], Hong Cheng[3], Ming Li[3], Hehe Liu[1], Yong Jiang[1], Yuming Guo[2], Zhengkui Zhou[1]* and Shuisheng Hou[1]*

[1] Key Laboratory of Animal (Poultry) Genetics Breeding and Reproduction, Ministry of Agriculture and Rural Affairs, State Key Laboratory of Animal Nutrition, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China, [2] State Key Laboratory of Animal Nutrition, College of Animal Science and Technology, China Agricultural University, Beijing, China, [3] College of Animal Science and Technology, Northwest A&F University, Yangling, China

Short tandem repeats (STRs) are usually associated with genetic diseases and gene regulatory functions, and are also important genetic markers for analysis of evolutionary, genetic diversity and forensic. However, for the majority of STRs in the duck genome, their population genetic properties and functional impacts remain poorly defined. Recent advent of next generation sequencing (NGS) has offered an opportunity for profiling large numbers of polymorphic STRs. Here, we reported a population-scale analysis of STR variation using genome resequencing in mallard and Pekin duck. Our analysis provided the first genome-wide duck STR reference including 198,022 STR loci with motif size of 2–6 base pairs. We observed a relatively uneven distribution of STRs in different genomic regions, which indicates that the occurrence of STRs in duck genome is not random, but undergoes a directional selection pressure. Using genome resequencing data of 23 mallard and 26 Pekin ducks, we successfully identified 89,891 polymorphic STR loci. Intensive analysis of this dataset suggested that shorter repeat motif, longer reference tract length, higher purity, and residing outside of a coding region are all associated with an increase in STR variability. STR genotypes were utilized for population genetic analysis, and the results showed that population structure and divergence patterns among population groups can be efficiently captured. In addition, comparison between Pekin duck and mallard identified 3,122 STRs with extremely divergent allele frequency, which overlapped with a set of genes related to nervous system, energy metabolism and behavior. The evolutionary analysis revealed that the genes containing divergent STRs may play important roles in phenotypic changes during duck domestication. The variation analysis of STRs in population scale provides valuable resource for future study of genetic diversity and genome evolution in duck.

Keywords: short tandem repeat, duck, whole genome resequencing, population genetics, variation

INTRODUCTION

Short tandem repeats (STRs), also known as short sequence repeats(SSRs) or microsatellites, are tandem repeat nucleotides of 1–6 bp in DNA sequences. These sequences are ubiquitously present in eukaryotic and prokaryotic genomes, and occur in both genic and intergenic regions (Toth et al., 2000). They are often highly variable with mutation rates dependent on several factors, including the STR motif length, repeat number, purity and their locations in the genome (Brandstrom and Ellegren, 2008; Payseur et al., 2011). The length polymorphism makes STRs into more informative genetic markers, hence the STRs are used extensively in varied aspects such as genetic diversity analysis (Santos et al., 2016; Ginja et al., 2017), genetic linkage mapping (Rohrer et al., 1996) and parentage analysis (Cherel et al., 2011). Emerging evidence suggests that STRs may play a regulatory role in complex traits of numerous species, including mice, dogs, and humans (Rd and Garner, 2004; Gemayel et al., 2010). STR variations in coding regions directly produce mutant proteins and cause human genetics disorders, such as Huntington disease and Fragile X syndrome (Pearson et al., 2005). The presence of STRs in promoters, 3′-untranslated regions (3′-UTR), 5′-untranslated regions (5′-UTR) and introns might regulate gene expression and function through various mechanisms (Sonay et al., 2015; Bagshaw, 2017). For instance, the copy number variation of the "CCG" trinucleotide repeat identified in the promoter of *Pleomorphic adenoma gene 1* (*PLAG1*), was proved to be the potential causative mutation influencing bovine stature by serving as nuclear factor binding sites and modulating the expression of *PLAG1* (Karim et al., 2011).

Population genetics has played an important role in exploring genetic variations in human and domestic animals (Larson and Burger, 2013). Investigating the population genetics and natural selection of STRs could enable us to understand their mutational mechanisms and functional impacts. Population scale analysis of STR variations in humans and flies has provides a genomic view of factors affecting STR polymorphism (Payseur et al., 2011; Fondon et al., 2012; Willems et al., 2014). On the other hand, population-specific STRs are usually located in candidate regions under selection and are potentially responsible for diverse phenotypes. For example, comparative genomic study of STRs in developmental genes of 92 breeds of dogs revealed that profound evolution in the snout morphology of domestic dog breeds are caused by length mutations of a compound STR in the gene *Runx2* (Rd and Garner, 2004). However, in domestic animals, most studies have limited themselves to only a few genomes or very small panels of STRs. The investigation of diversity and origin of STRs, the characterization of their population-genetic properties, and the determination of their functional impacts are still active areas of research.

The advances of next generation sequencing (NGS) have generated large amounts of sequence data at low costs, providing an opportunity to profile STR variations on a population-scale. However, STRs are substantially more difficult to detect based on short reads produced by NGS platform (Gymrek, 2017).

Therefore, STR variation analysis has been absent in most whole-genome resequencing studies due to the lack of tools. Recently, numerous methods have been developed to identify STR variants in human, including lobSTR (Gymrek et al., 2012), STRviper (Cao et al., 2014), popSTR (Kristmundsdóttir et al., 2017), and HipSTR (Willems et al., 2017). The program lobSTR was a custom algorithm for genotyping STRs, which has been used to characterize the variation of genome-wide STR loci in Phase 1 of the 1000 Genomes Project and the Simons Genome Diversity Project (Willems et al., 2014; Mallick et al., 2016).

Duck (*Anas platyrhynchos*) is one of the most common domestic fowls that is derived from mallard since 500 BC in central China (Qu et al., 2009). The Pekin ducks as the most elite breed has undergone intensive artificial selection since the Ming Dynasty (A.D. 1368-1644). Compared with their wild ancestors, Pekin ducks show many striking changes such as noble entirely white plumage, an extraordinary big body size and excellent performance of egg production. The Pekin duck is also important model organism for the study of lipid metabolism disorders and immune resistance for viruses, due to its low susceptibility to nonalcoholic fatty liver disease and influenza A viruses infection (Gaidet et al., 2008). Several studies have been performed to develop 100s of duck STR markers (Maak et al., 2000; Huang et al., 2005) and these STRs has also been widely used in researches such as genetic diversity, paternity test, genetic map construction and QTL mapping (Huang et al., 2006, 2007; Su et al., 2007; Ren et al., 2009). However, for the majority of STRs in the duck genome, their population genetics and functional impacts remain poorly defined. Fortunately, the availability of duck reference genomic sequence (Huang et al., 2013) and a dozen of duck genome resequencing data, offer us the opportunity for population scale profiling of duck STRs.

In the present study, we characterized the composition and distribution of STRs in duck reference genome and analyzed the variation of these STRs using genome resequencing in Pekin duck and mallard. Next, we attempt to investigate their population-genetic properties and potential role in duck domestication.

MATERIALS AND METHODS

Samples and Sequencing

A total of 23 mallard and 26 Pekin duck resequencing data were used to profile the STR genotypes. Of which, 21 sequence data were generated in our previous work (Zhou et al., 2018) available at Genome Sequence Archive (GSA)[1], and the other 28 previously published data were download from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)[2]. We converted the 28 SRA files to fastq format using SRA toolkit (fastq-dump –split-3). All samples were sequenced to an average coverage of 12-fold (range 10–16-fold) with 125PE or 150PE reads on an Illumina Hi-seq (2500 or X Ten) machine. A summary of sample information was presented in **Additional File 1**.

---

[1]http://gsa.big.ac.cn

[2]https://www.ncbi.nlm.nih.gov/sra

Identification of STRs in Duck Reference Genome

The reference genome of Pekin duck used in this study was BGI duck 1.0 reference (GCA_000355885.1), which have been assembled into chromosome level in our previous work based on an RH map (Rao et al., 2012; Zhou et al., 2018). The chromosome level reference genome was available at www.duckbase.org/Download. Tandem Repeats Finder (TRF) was run on the duck chromosomes with a match weight of 2, a mismatch and indel penalty of 7, an 80% probability of matching and a 10% probability of an indel (Benson, 1999). The output was filtered in order to include only repeats with motif length between 2 and 6 base pairs. We removed STRs that localized to areas that might preclude unique mapping, such as large repeats or transposable elements. Transposons and other repetitive elements were identified using RepeatMasker and the TRF results in or within 20 bases of these regions were removed. We furtherly removed STRs with alignment scores below thresholds suggested by Willems et al., STRs located next to or within 20 bases of another STR (Willems et al., 2014). Finally, we successfully assembled the first duck genome-wide STR reference with 198,022 loci included.

Genotyping STRs in Duck Population

We conducted a comprehensive survey of STR variants using 49 high coverage NGS data from two duck population. LobSTR was applied with default parameters for alignment and STR discovery as previously described (Gymrek et al., 2012). Briefly, we first created a lobSTR reference index based on STR reference using lobstr_index.py script. Then we carried out lobSTR alignment to create the STR alignment bam files, and the final STR variants allelotypes were identified throughout all samples based on the merged alignment file. LobSTR employs an explicit model to enhance accuracy by avoiding stutter noise caused by PCR amplification of a STR locus. The genotyped STR loci were then filtered with the following properties as Gymrek's recommendation (Mallick et al., 2016): mainly based on coverage, call rate (percent of samples with a genotype call for a given locus), and the metrics Q and DISTENDS reported in the VCF file generated by lobSTR.

- Average coverage $< 3 \times$
- Average $-\log10(1-Q) < 0.8$
- Call rate < 0.8
- Reference allele length > 80 bp

After filtering loci we additionally filtered individual calls with:

- Coverage $< 3 \times$
- $-\log10(1-Q) < 0.8$
- Absolute value of DISTENDS score > 20

After filtering, 141,289 loci remained for further analysis.

Genotyping of SNP in Duck Population

The raw reads were mapped to the reference genome with Burrows-Wheeler alignment (BWA aln) (Li and Durbin, 2009) using the default parameters. Then, the reads mapped to the

exact same position on the reference genome were removed with MarkDuplicates in Picard tools to avoid any influence on variant detection. We additionally performed local realignment using GATK (Depristo et al., 2011) to enhance the alignments in regions of insertion-deletion (Indel) polymorphisms. By applying HaplotypeCaller in GATK, we generated a VCF file containing SNPs and short indels. The output was further filtered using VCFtools (version 0.1.15) (Danecek et al., 2011). SNPs that did not pass the following criteria were excluded: (1) the mean sequencing depth (over all included individuals) had to be $>3\times$ and $<30\times$; (2) SNPs had to have a minor allele frequency >0.05 and a max allele frequency <0.99; (3) the maximum missing rate was <0.1; and (4) SNPs had only two alleles.

Primer Design and *in silico* PCR Validation

The 200 bp flanking sequences of each unique STR were extracted for primer design. The Primer3 software[3] were used and parameters were set as follows: primer length of 18–27 nucleotides, melting temperatures of 55–65°C, GC content of 30–70%, and predicted PCR products of 100–300 bp in length. We further used *in silico* PCR analysis to align primer pairs to reference genome and that matched more than one genome location were removed. The software (e-PCR-2.3.12)[4] was used for *in silico* PCR analysis with the following default parameters: 50 bp margin, 2 bp mismatch, 1 bp gap, and 50–1,000 bp product size.

Validating lobSTR Accuracy Using Sanger Sequence and Capillary Electrophoresis

PCR amplifications were carried out in 25 μL reaction mixtures, comprising approximately 50 ng of template DNA, 1.5–2 mm MgCl2 (TaKaRa, Japan), 200 μm of each dNTP, 15 pmol of each primer, and 0.3 U of Ampli *Taq* DNA polymerase (TaKaRa, Japan). Amplifications were performed using the following PCR procedure: an initial denaturation step for 5 min at 95°C, followed by 35 cycles of 95°C for 45 s, 30 s at locus-specific annealing temperature (55–65°C) and 50 s at 72°C, and a final elongation for 10 min at 72°C. The PCR products were subjected to Sanger sequencing to confirm the sequence identity. PCR fragments amplified using fluorescence labeled primer were separated by capillary electrophoresis on ABI 377XL instruments (Applied Biosystems), according to the manufacturer's recommendations. Genotypes were called using GeneMapperV4.1 software (Applied Biosystems). Details of seven STRs selected for PCR amplifications and its primer information can be found in **Additional File 2**.

Correlations of STR Variation With Reference Tract Length and Purity

The heterozygosity of each STR was calculated using the formula $1-\sum^{n}_{i=1}P_i^2$, where P_i is the frequency of allele i and n is the total

---

[3]http://bioinfo.ut.ee/primer3/

[4]ftp://ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR/

TABLE 1 | Distribution and composition of duck STR reference with respect to motif length.

Motif length (bp)	Number of loci	% in reference	Abundance (No./Mb)	Common motifs (% in each category)
Di-	54,347	27.44	49.95	AT(44.81), AC(23.13), CA(14.09)
Tri-	31,711	16.01	29.15	AAT(26.95), AAC(23.40), ATT(9.03)
Tetra-	76,100	38.43	69.95	AAAC(26.91), AAAT(22.52), ATTT(8.21)
Penta-	25,750	13.00	23.67	AAAAC(23.62), AAAAT(10.76), ACAAA(5.98)
Hexa-	10,114	5.11	9.30	AAAAAT(10.61), AAAAAC(7.39), AAAACA(3.64)
Total	198,022	100.00	182.02	

number of alleles. The mannwhitneyu and pearsonr function in python package scipy.stats were used for significant difference and Pearson correlation analysis, respectively. The TRF-reported motif length and purity values were used to categorize the STRs. Repeats were binned by reference tract length, and the mean heterozygosity for each bin was determined to generate plots.

Principal Component Analysis

We carried out principal components analysis (PCA) using EIGENSOFT software (Patterson et al., 2006) with both STRs and SNPs call set. A Tracy–Widom test was used to determine the significance level of the eigenvectors. We used autosome STR loci with heterozygosity greater than 10% that were called in at least 80% of samples. To encode STRs in bi-allelic format, we followed the convention suggested by Patterson et al. (2006) and encoded each STR allele in the frequency range of 5–95% as a separate bi-allelic marker. This gave 199,999 STR "markers" from 43,939 unique STR loci. The figures were then plotted using the first and second principle components with R packages.

Evaluation of Genetic Diversity With Polymorphic STRs

The 10% most heterozygous (13,256 loci) autosomal STR loci were used for genetic diversity analysis. To determine whether the two populations had systematically different heterozygosity at these loci, a paired comparison for each STR locus was performed and the cdf function in the scipy.stats.binom python package was used to calculate the P-value.

Functional Annotation of Divergent STRs Using PANTHER and KEGG Database
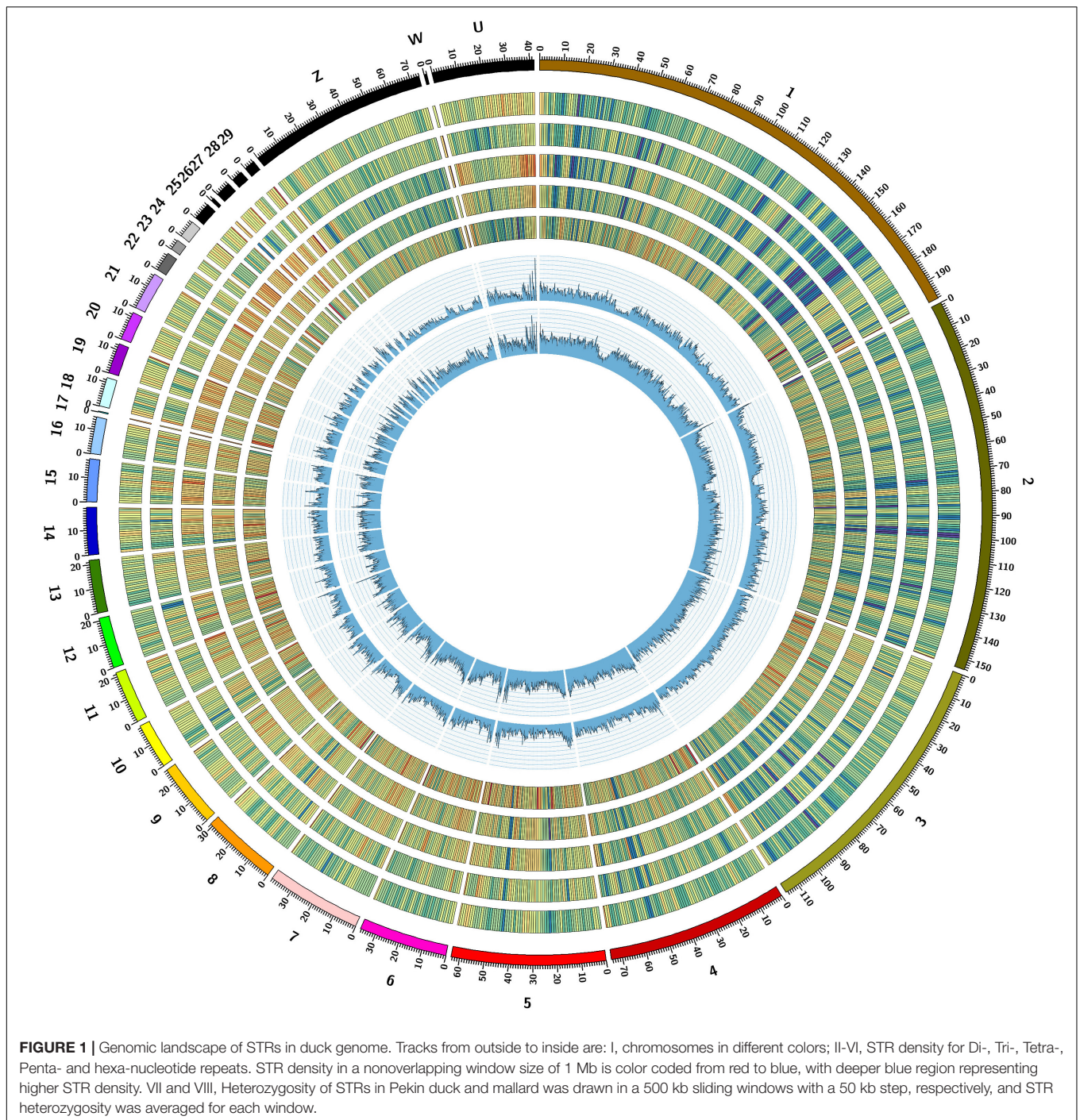
We formally characterized the function genes containing divergent STRs by searching for overrepresented pathways associated with these genes. Firstly, we extracted the protein sequences for the genes containing divergent STRs. Then, we enriched this genes using "Fasta Protein Sequence" by "Gene-list Enrichment" in KOBAS 3.0 (Wu et al., 2006; Xie et al., 2011). And "Gallus gallus (chicken)," "hypergeometric test/Fisher's exact test" and "Benjamini and Hochberg (1995)" were chose as "Species," "Statistical method," and "FDR correction method," respectively. Finally, we obtained 9 significant pathways (Corrected P-value < 0.05).

RESULTS

STR Frequency and Distribution in the Duck Genome

By screening the genome assembly of Pekin duck, we obtained the first duck STR reference including 198,022 STR loci with repeat motif of di-, tri-, tetra-, penta-, and hexa-nucleotide. The total length of all STR sequences covered 2.10% of the draft genome assembly with an average density of 182.0 STRs/Mb (**Table 1** and **Figure 1**). Among different nucleotide types, the STR frequency was negatively correlated with the number of nucleotide, except tetra-nucleotides. Tetra-nucleotides were the most abundant category, accounting for 38.43% of total STRs, followed by di-nucleotides (27.44%), and tri-nucleotides (16.01%). In contrast, penta-nucleotides and hexa-nucleotides were less frequent compared with others (**Table 1**). We furtherly examined the nucleotide composition of each motif type and found that some combinations of nucleotides were more prevalent than others in each class. Our analysis showed that STRs with an [A]nT and [A]nC motif were more frequent in each category. Conversely, GC-rich STR motifs were rare in the duck genome. For example, the AT motif was dramatically overrepresented in di-nucleotide motifs, and it was also the most frequent motif in the entire duck genome, which accounting for 12.30% of the total SSR loci discovered. Similarly, the AAT, AAAC, AAAAC, and AAAAAT were the most abundant repeats types in each class. The top three most abundant STR motifs for each class were listed in **Table 1**, and these STRs represent 82.03, 59.38, 57.64, 40.36, and 21.64% in di-, tri-, tetra-, penta-, and hexa-nucleotide repeats, respectively.

STRs were classified into six categories according to its genomic location and annotation of the duck reference genome, to unveil how it was organized in the duck genome. According to our analysis, the majority of STRs were commonly mapped onto intergenic (122,987 STRs) and intronic (68,457 STRs) regions, which together comprised 96.70% of our STR reference (**Table 2**). There is no apparent difference in STR contents between intergenic regions and introns. Despite low abundance of STRs located in coding region, we found an overrepresentation of Tri- and Hexa-nucleotide repeats in this region (**Figure 2A**). Moreover, the enriched trinucleotide repeats in exonic and 5′-UTR region were mostly rich in G/C, such as TCC, GAG, and GCA. In contrast to coding regions, A/T-rich trinucleotide repeats were enriched in other genomic regions. However, the

FIGURE 1 | Genomic landscape of STRs in duck genome. Tracks from outside to inside are: I, chromosomes in different colors; II-VI, STR density for Di-, Tri-, Tetra-, Penta- and hexa-nucleotide repeats. STR density in a nonoverlapping window size of 1 Mb is color coded from red to blue, with deeper blue region representing higher STR density. VII and VIII, Heterozygosity of STRs in Pekin duck and mallard was drawn in a 500 kb sliding windows with a 50 kb step, respectively, and STR heterozygosity was averaged for each window.

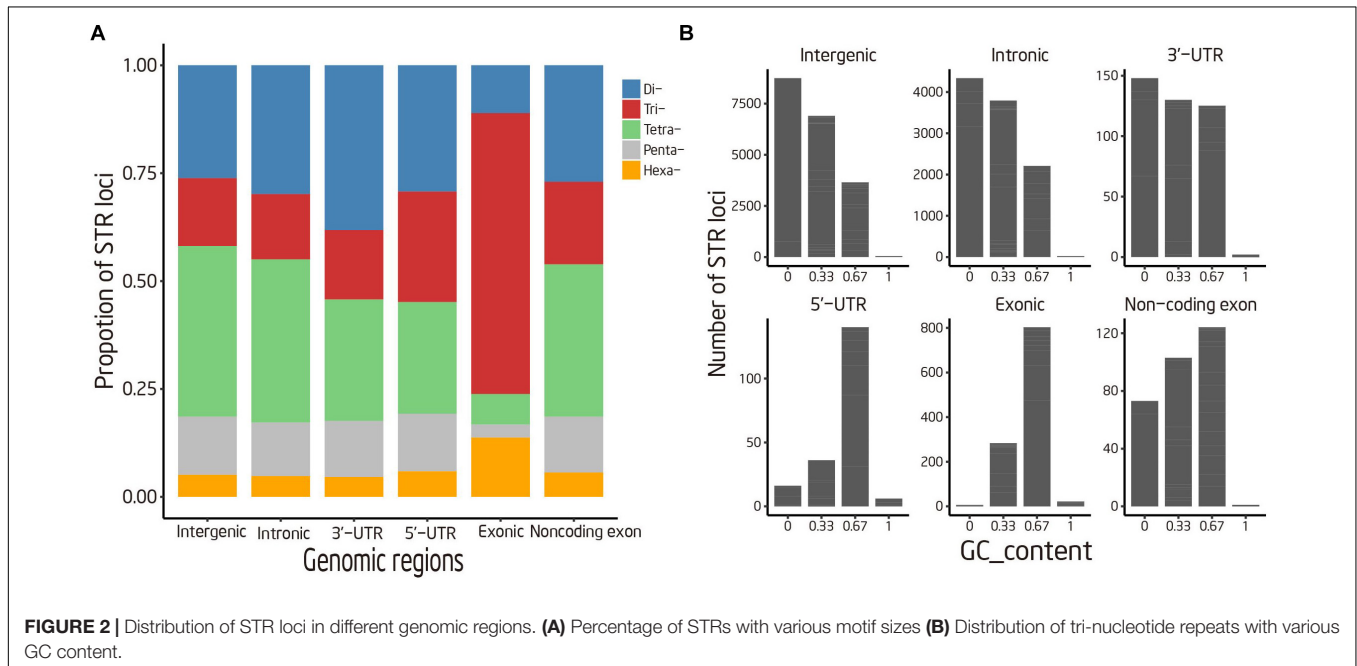trinucleotide repeats containing no "A/T" were quite rare in all genomic regions (**Figure 2B**).

Systematic Profiling of STRs in Duck Populations and Experiment Validation of STR Genotypes

To identify polymorphic STRs in duck populations, whole genome sequence data from 23 mallard and 26 Pekin duck were mapped against reference genome and genotyped using lobSTR. The genomic coverage of sequence data for each animal was around 10∼16× (**Additional File 1**). In total, 187,874 STRs and 9,354,319 calls were detected in our dataset, with an average STR coverage of 4.8∼7.9×for each sample (**Table 3** and **Additional File 1**). We were not able to genotype about 5% of loci in our reference. Most of these loci have allele lengths greater than 80 bp that could not be spanned by Illumina reads or not suitable for the "allelotype" requirement. Furtherly, we subjected the resulting

TABLE 2 | Distribution of duck STR in different genomic regions.

Motif length (bp)	Intergenic	Intronic	3'-UTR	5'-UTR	Exonic	Noncoding exon
Di-	32,136	20,409	962	226	189	425
Tri-	19,339	10,355	405	198	1113	301
Tetra	48,620	25,894	709	200	121	556
Penta	16,560	8,503	328	103	52	204
Hexa-	6,332	3,296	116	46	235	89
Total	122,987	68,457	2,520	773	1,710	1,575

FIGURE 2 | Distribution of STR loci in different genomic regions. **(A)** Percentage of STRs with various motif sizes **(B)** Distribution of tri-nucleotide repeats with various GC content.

genotypes to stringent filtering to ensure high quality calls. After filtering, we obtained 141,289 high quality loci with an average length of 19.8 bp, ranging from 11 to 80 bp. Among these STRs, 25,624, 14,944, 36,302, 10,271, and 2,750 STRs were polymorphic (with at least two allele) with repeat motif of di-, tri-, tetra-, penta-, and hexa-nucleotide, respectively (**Table 3**).

To evaluate lobSTR prediction, we randomly selected 7 loci with motif size of three or four and confirmed the sequence identity of PCR products using Sanger sequencing (**Additional File 3**). The genotypes of this 7 loci were furtherly validated using Capillary electrophoresis, the gold standard for STR genotyping. We observed that the homozygous STRs with a read

TABLE 3 | Overview of STRs genotyped in duck population.

Motif length (bp)	Number genotyped	Number passed filter	Number polymorphic
Di-	52,164	38,519	25,624
Tri-	30,394	22,948	14,944
Tetra-	73,503	58,206	36,302
Penta-	22,881	16,333	10,271
Hexa-	8,932	5,283	2,750
Total	187,874	141,289	89,891

coverage of $\geq 3\times$ could be correctly genotyped by lobSTR. For heterozygous STRs, lobSTR may correctly call one allele and miss the other one due to insufficient reads coverage. We observed for homozygous loci 88.89% (40/45) were correctly called whereas the heterozygous showed a lower correct rates of 62.37% (58/93) (**Additional File 4**).
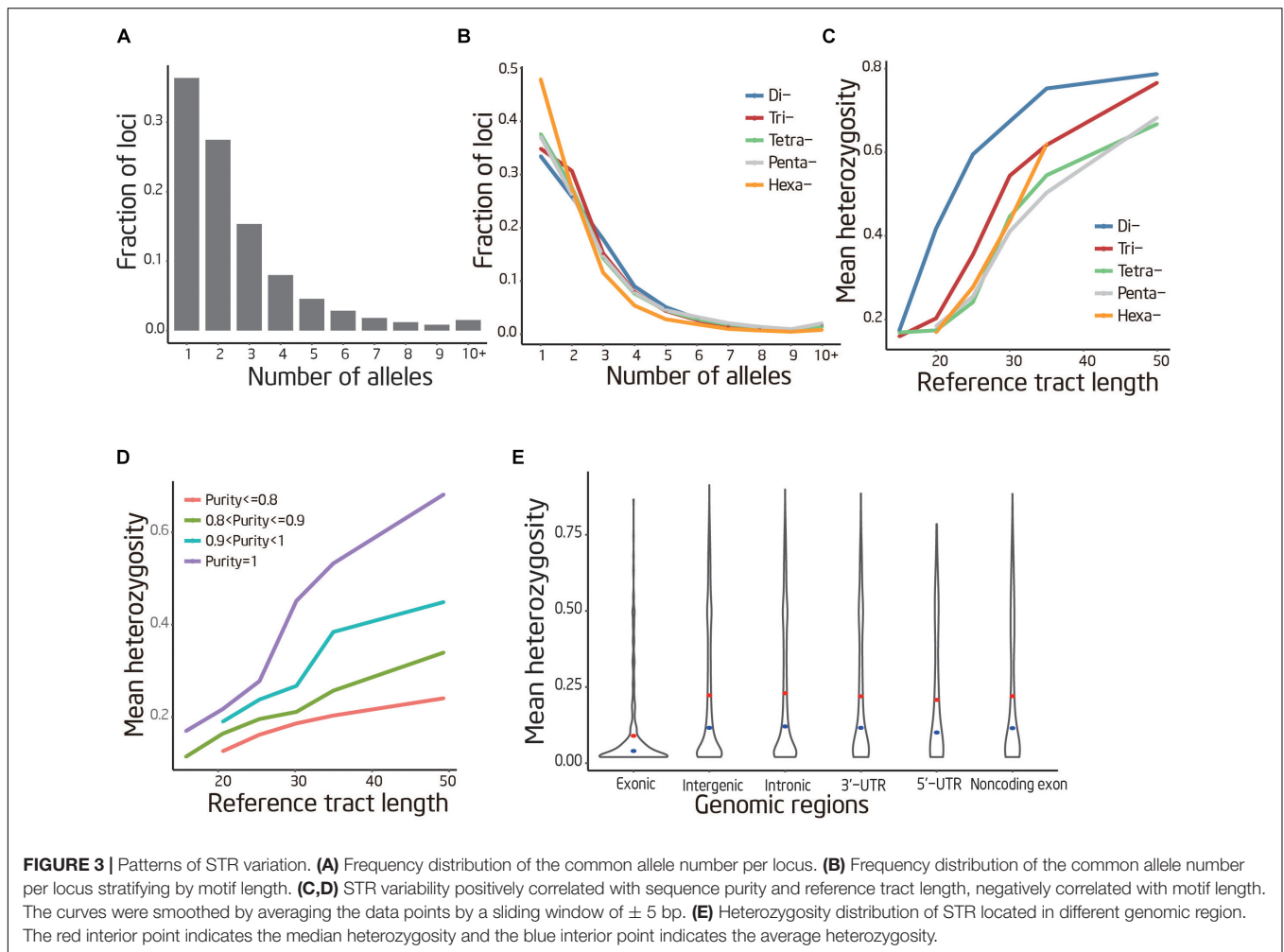
The full catalog of STR variations is publicly available at http://www.duckbase.org/Diversity in VCF format. In addition, to facilitate the access and effective utilization of the duck STR markers, an integrative table was also available at http://www.duckbase.org/Diversity. Details of each STR, such as location information, repeat type, reference copy number, reference tract length, heterozygosity and locus specific primer pairs could be obtained in the integrative table.

Patterns of STR Variation

The overall trends of STR polymorphism were examined using genotypes identified by lobSTR. Of the 141,289 genotyped loci, 36.1% were observed having more than two common alleles, and some loci have more than 20 common alleles. The remaining loci are either fixed across all individuals (36.4%) or have only two common alleles (27.5%) (**Figure 3A**). These patterns changed significantly when stratifying by motif length, with longer motif

FIGURE 3 | Patterns of STR variation. **(A)** Frequency distribution of the common allele number per locus. **(B)** Frequency distribution of the common allele number per locus stratifying by motif length. **(C,D)** STR variability positively correlated with sequence purity and reference tract length, negatively correlated with motif length. The curves were smoothed by averaging the data points by a sliding window of $\pm$ 5 bp. **(E)** Heterozygosity distribution of STR located in different genomic region. The red interior point indicates the median heterozygosity and the blue interior point indicates the average heterozygosity.

lengths showing less variability. For instance, only 33.5% of dinucleotides are fixed compared to 47.9% of Hexa-nucleotide (**Figure 3B**).

To examine the relationships between repeat features and variability, heterozygosity of each STR were used as a metric of variation. The present study revealed that heterozygosity depends strongly on properties and genomic locations of the STRs. Our analysis showed that heterozygosity was positively correlated with STR sequence purity ($r = 0.13$, $P < 10^{-100}$) and reference track length ($r = 0.22$, $P < 10^{-100}$) (**Figures 3C,D**). In addition, the coding STRs demonstrated significantly decreased heterozygosity compared to noncoding STRs (Mann–Whitney U-test; $P < 10^{-10}$) (**Figure 3E**).
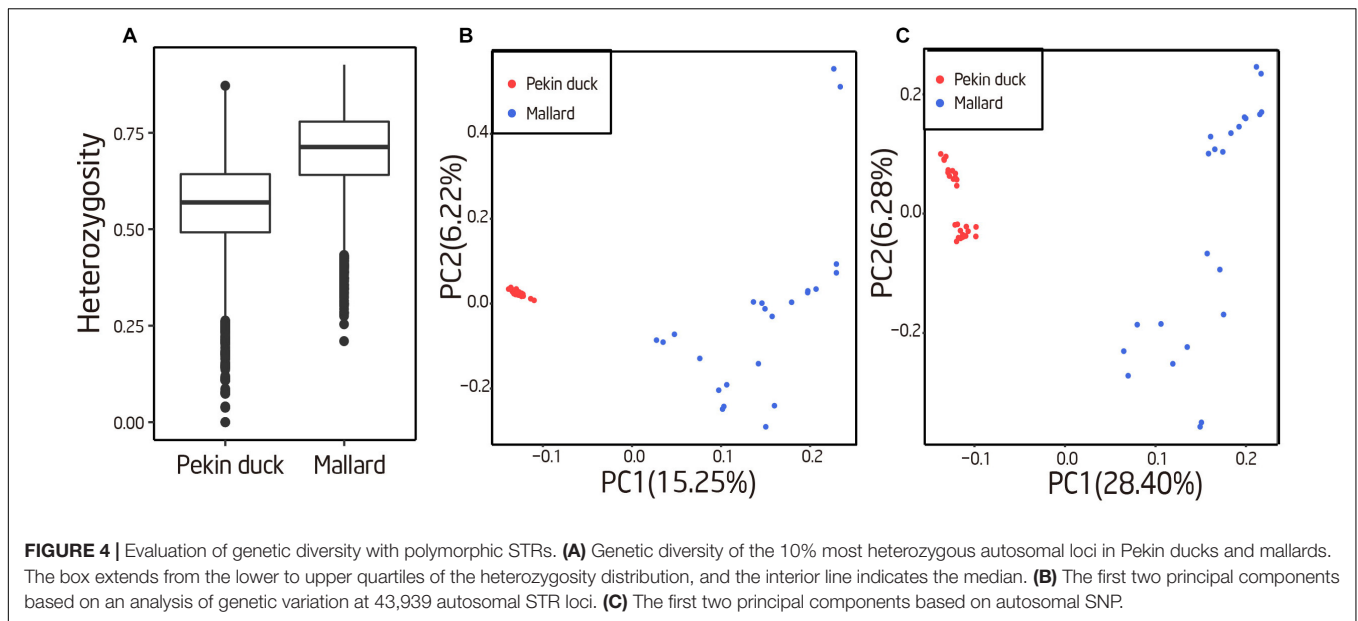
Population Genetic Analyses Using STR Genotype

Population genetic properties of mallard and Pekin ducks were explored using the obtained STR genotypes. We first compared the heterozygosities of the 10% most variable autosomal loci between the two breeds, and we found that mallard have a significant higher heterozygosities than Pekin duck (sign test; $P < 10^{-50}$), suggesting a higher genetic diversity in mallard

(**Figures 1, 4A**). We performed PCA using the autosome STR loci which have heterozygosity greater than 10%. For comparison, we also performed PCA on autosome SNPs of the same individuals. Mallard and Pekin ducks were clearly separated into two groups on the first principle components from both STR and SNP based PCA Tracy–Widom test; $P < 10^{-26}$). Meanwhile, the PCA analysis also revealed higher diversity in mallard than Pekin ducks, which was concordant with the heterozygosities analysis (**Figures 4B,C**). Our results agree with the known genetic diversity and population structure of the two breeds.

Population-Differentiated STRs Between Pekin Duck and Mallard

In this study, we identified 89,891 polymorphic STRs in duck population. To investigate the selection signature involved in these STRs, we checked if any STRs were divergent between mallard and Pekin duck by searching for STRs with high degrees of fixation index (F_{ST}) (**Additional File 5**). In total, we identified 3122 divergent STRs with a significance level of $P < 0.01$ (Z-test). Around 38% of these divergent STR overlap genic elements, i.e., exons, introns, untranslated regions (3'-UTR and 5'-UTR) and noncoding exon, and remaining 62%

FIGURE 4 | Evaluation of genetic diversity with polymorphic STRs. **(A)** Genetic diversity of the 10% most heterozygous autosomal loci in Pekin ducks and mallards. The box extends from the lower to upper quartiles of the heterozygosity distribution, and the interior line indicates the median. **(B)** The first two principal components based on an analysis of genetic variation at 43,939 autosomal STR loci. **(C)** The first two principal components based on autosomal SNP.

are intergenic variants. A total of 740 unique genes were found to containing divergent STRs, and an enrichment test was performed to search for significantly overrepresented KEGG pathways ($P_{FDR} < 0.05$) using these genes. Nine significantly enriched pathways were identified in this analysis including nervous system and energy metabolism related pathways, such as neuroactive ligand–receptor interaction and ABC transporters (**Additional File 6**). The conspicuous cluster related to nervous system and energy metabolism implies that these pathways may be involved in the behavioral alternations of duck domestication.

DISCUSSION

In the last few years, population-scale sequencing projects have made tremendous progress in documenting genetic variation across various vertebrate species. Population scale characterization of STR variation has been reported for humans (Willems et al., 2014), cattle (Xu et al., 2017), and porcine (Liu et al., 2017). Here, we scanned all possible STR loci in duck reference genome and conducted a population-scale analysis of STR variation, providing new insights for further exploration of duck genome.

Our analysis significantly augments the level of knowledge of STR variation in duck population. Before our current study, NCBI has reported only several 100s of STR containing sequence in duck. Our catalog provides the first duck STR reference including 198,022 STR loci with motifs of 2–6 bp. The number of STRs and their density identified in our study was lower than that in human (Willems et al., 2014), which confirmed the prediction that STR abundance is lower in avian than in mammals (Primmer, 1997; Brandstrom and Ellegren, 2008). Without taking of mononucleotide repeats into account, it is generally reported that the most abundant STRs may vary in different organisms, e.g., trinucleotide and dinucleotides

repeats were the most common STRs for insect and mammals, respectively (Li et al., 2002). However, the tetra-nucleotide repeats appeared significantly overrepresented in our catalog, which was in agreement with previous reports in other birds (Huang et al., 2016). In addition, our analysis also confirmed the common view that genomic STRs with GC-rich repeats are rare in vertebrate species, such as human, giant panda, and birds (Willems et al., 2014; Huang et al., 2015; Huang et al., 2016).

The distributions of STRs also vary in different regions of a genome. It is well known that noncoding regions generally contain more abundant STR than coding regions. Despite the relative lowest abundance of STR in exons, a propensity of tri- and hexa-nucleotides in exons were observed in our study, which was in agreement with observation in human (Willems et al., 2014) and giant panda (Huang et al., 2015). These STR loci can vary without introducing frameshift mutations and therefore may be exposed to weaker purifying selection. Moreover, the GC-rich trinucleotide repeats were also found to be enriched in exon and 5′-UTR region. This observation suggests that the occurrence of trinucleotide repeats in exon and 5′-UTR region is not random, but undergoes positive or negative selective pressure. Previous studies have reported that GC-rich trinucleotide repeats are capable of forming stable hairpin or quadruplex structures and are involved in the regulation of transcription (Jasinska et al., 2003; Kozlowski et al., 2010).

Based on the defined STR reference catalog, our population-scale analysis of STR variation obtained genotypes for 187,874 STR loci. After filtering, 89,891 polymorphic STRs were obtained, which could be useful genetic markers for future studies such as parentage analysis, genetic diversity and QTL mapping. It has been reported that STR variability were determined by its motif characteristics in human (Willems et al., 2014) and *Drosophila melanogaster* (Fondon et al., 2012). Our study confirmed the well-known relationship that shorter repeat motif, longer reference tract length, higher purity, and residing outside of a coding

region are all associated with an increase in STR variability (Pemberton et al., 2009; Fondon et al., 2012; Mallick et al., 2016). The overall positive relationship between repeat feature and polymorphism observed here suggests that STR loci with shorter repeat motif and longer repeat length mutate more rapidly. Clearly, for any given total length of a STR loci, interruptions have the consequence of lowering variability, most likely due to the stabilizing effects of unique sequence within tandem array that prevent replication slippage (Petes et al., 1997; Rolfsmeier and Lahue, 2000; Sibly et al., 2003).

It is well known that most domesticated animals have experienced a "domestication bottleneck" with reduced genetic diversity relative to their wild ancestor(s) (Chen et al., 2007; Gibbs et al., 2009). STRs are important molecular markers for studying genetic diversity and population structure (Tadano et al., 2014; Abebe et al., 2015; Santos et al., 2016). The genome-wide STR genotypes were used to evaluate genetic diversity of Pekin duck and its ancestor (mallard). A higher genetic diversity was observed in mallard than Pekin duck, which suggest that the current genetic diversity of the wild population could be important genetic resource for future breeding programs. Previous study has reported that genome-wide STR genotypes could be used to distinguish population structure and divergence patterns among population groups (Sonay et al., 2015). We performed principal component analysis (PCA) based on STR genotypes and found that mallard and Pekin ducks could be clearly separated into two groups on the first principle components. These STR markers could then potentially be used for conservation and breeding programs of Pekin ducks, since their higher polymorphism information content means that a low number of markers can be used.

Pekin ducks have experienced natural and artificial selections for 1000s of years, and these selections has resulted in striking changes in traits such as boy shape, color, behavior, and reproduction (Qu et al., 2009). However, the molecular mechanisms underlying the selection-causing phenotypic changes in Pekin duck remained largely unknown. To reveal the potential contributions of STRs in the process of duck domestication, we carried out a comparative study to identify the STRs with extreme divergence in allele frequency (F_{ST}) between Pekin duck and mallard. A total of 3,122 divergent STRs were obtained, which overlapped with a set of genes related to nervous system, energy metabolism and behavior. Of the significantly enriched pathways, neuroactive ligand–receptor interaction was found to have 30 genes containing divergent STRs. The enriched genes were also including growth hormone receptor (*GHR*) and thyroid stimulating hormone receptor (*TSHR*), which has been well established to play pivotal role in growth and metabolic regulation of chicken. It was also worth noting that *TSHR* was also located in a selected region as our previous reports (Zhou et al., 2018). *GHR* has been reported to contain a deletion in its exon as causative mutation for sex-linked dwarfism. And this mutation has been widely used in breeding of commercial broiler lines to improve the feed conversion rate (Agarwal et al., 1994). *TSHR* has been identified previously as domestication gene for chicken, and related to the absence of the strict regulation of seasonal reproduction found in natural populations (Yoshimura et al., 2003). This results are in accordance with previous reports

that strong selection on behavioral alternations and improve reproduction ability are the common characteristics of animal domestication such as chicken (Rubin et al., 2010), dog (Axelsson et al., 2013), and yak (Qiu et al., 2015).

CONCLUSION

In this study, we analyzed the composition and distribution of STRs in the duck reference genome and investigated the population-genetic properties of these STRs using genome resequencing on 23 mallard and 26 Pekin ducks. We confirmed the well-known relationship that STR polymorphism was closely related to its genomic location, motif length, purity and repeat number. The population genetic analysis suggests that STR genotypes obtained from genome resequencing data could be used to distinguish population structure and divergence patterns among population groups. Our results also revealed that the genes overlapping with divergent STRs perhaps involved in domestication of duck by influencing the nervous system and energy metabolism. Altogether, our study provides valuable resource for future study of genetic diversity and genome evolution in duck.

ETHICS STATEMENT

All experimental procedures with ducks were performed according to the Guidelines for Experimental Animals established by the Ministry of Science and Technology (Beijing, China). Ethical approval on animal survival was given by the animal ethics committee of the Institute of Animal Sciences (IAS), Chinese Academy of Agricultural Sciences (CAAS, Beijing, China) with the following reference number: IASCAAS-AE-03.

DATA AVAILABILITY

The Illumina sequencing data used in this study can be available at (SRA, https://www.ncbi.nlm.nih.gov/sra) and (GSA, http://gsa.big.ac.cn). The accession numbers for each sample are included in the **Additional Files**. Other data sets supporting the results of this article are included within the article and its **Additional Files**.

AUTHOR CONTRIBUTIONS

ZZ and SH conceived and coordinated the study. WF performed the study and wrote the manuscript. WF, LX, HC, and ML carried out the bioinformatics and experiments analyses. LX, HL, YJ, and YG gave advice about concept and revised manuscript. All authors read and approved the final manuscript.

FUNDING

CAAS-Innovation Team Project (ASTIP-2016-IAS-9 and CAAS-XTCX2016010-03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00520/full#supplementary-material

Additional File 1 | Summary of sample information used in this study.

Additional File 2 | Details of seven STRs selected for Capillary electrophoresis and its primer informations.

Additional File 3 | Sanger sequence of PCR products from seven selected STR.

Additional File 4 | Comparison of capillary electrophoresis and lobSTR results for seven Selected STRs.

Additional File 5 | Chromosome-wise deviation in STR constitutions based on FST.

Additional File 6 | Significant KEGG pathways enriched by genes containing divergent STR.

REFERENCES

Abebe, A. S., Mikko, S., and Johansson, A. M. (2015). Genetic diversity of five local Swedish chicken breeds detected by microsatellite markers. *PLoS One* 10:e0120580. doi: 10.1371/journal.pone.0120580

Agarwal, S. K., Cogburn, L. A., and Burnside, J. (1994). Dysfunctional growth hormone receptor in a strain of sex-linked dwarf chicken: evidence for a mutation in the intracellular domain. *J. Endocrinol.* 142, 427–434. doi: 10.1677/joe.0.1420427

Axelsson, E., Ratnakumar, A., Arendt, M. L., Maqbool, K., Webster, M. T., Perloski, M., et al. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495, 360–364. doi: 10.1038/nature11837

Bagshaw, A. T. M. (2017). Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.* 9, 2428–2443. doi: 10.1093/gbe/evx164

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Brandstrom, M., and Ellegren, H. (2008). Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Res.* 18, 881–887. doi: 10.1101/gr.075242.107

Cao, M. D., Tasker, E., Willadsen, K., Imelfort, M., Vishwanathan, S., Sureshkumar, S., et al. (2014). Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res.* 42:e16. doi: 10.1093/nar/gkt1313

Chen, K., Baxter, T., Muir, W. M., Groenen, M. A., and Schook, L. B. (2007). Genetic resources, genome mapping and evolutionary genomics of the pig (*Sus scrofa*). *Int. J. Biol. Sci.* 3, 153–165. doi: 10.7150/ijbs.3.153

Cherel, P., Glenisson, J., and Pires, J. (2011). Tetranucleotide microsatellites contribute to a highly discriminating parentage test panel in pig. *Anim. Genet.* 42, 659–661. doi: 10.1111/j.1365-2052.2011.02187.x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Depristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Fondon, J. W. I., Martin, A., Richards, S., Gibbs, R. A., and Mittelman, D. (2012). Analysis of microsatellite variation in *Drosophila melanogaster* with population-scale genome sequencing. *PLoS One* 7:e33036. doi: 10.1371/journal.pone.0033036

Gaidet, N., Cattoli, G., Hammoumi, S., Newman, S. H., Hagemeijer, W., Takekawa, J. Y., et al. (2008). Evidence of infection by H5N2 highly pathogenic avian influenza viruses in healthy wild waterfowl. *PLoS Pathog.* 4:e1000127. doi: 10.1371/journal.ppat.1000127

Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 44, 445–477. doi: 10.1146/annurev-genet-072610-155046

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., et al. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528–532. doi: 10.1126/science.1167936

Ginja, C., Gama, L. T., Martinez, A., Sevane, N., Martin-Burriel, I., Lanari, M. R., et al. (2017). Genetic diversity and patterns of population structure in Creole goats from the Americas. *Anim. Genet.* 48, 315–329. doi: 10.1111/age.12529

Gymrek, M. (2017). A genomic view of short tandem repeats. *Curr. Opin. Genet. Dev.* 44, 9–16. doi: 10.1016/j.gde.2017.01.012

Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* 22, 1154–1162. doi: 10.1101/gr.135780.111

Huang, J., Li, W. J., Jian, Z. Y., Yue, B. S., and Yan, Y. F. (2016). Genome-wide distribution and organization of microsatellites in six species of birds. *Biochem. Syst. Ecol.* 67, 95–102. doi: 10.1016/j.bse.2016.05.023

Huang, J., Li, Y.-Z., Du, L.-M., Yang, B., Shen, F.-J., Zhang, H.-M., et al. (2015). Genome-wide survey and analysis of microsatellites in giant panda (*Ailuropoda melanoleuca*), with a focus on the applications of a novel microsatellite marker system. *BMC Genomics* 16:61. doi: 10.1186/s12864-015-1268-z

Huang, Y., Haley, C. S., Wu, F., Hu, S., Hao, J., Wu, C., et al. (2007). Genetic mapping of quantitative trait loci affecting carcass and meat quality traits in Beijing ducks (*Anas platyrhynchos*). *Anim. Genet.* 38, 114–119. doi: 10.1111/j.1365-2052.2007.01571.x

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian, W., et al. (2013). The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.* 45, 776–783. doi: 10.1038/ng.2657

Huang, Y., Tu, J., Cheng, X., Tang, B., Hu, X., Liu, Z., et al. (2005). Characterization of 35 novel microsatellite DNA markers from the duck (*Anas platyrhynchos*) genome and cross-amplification in other birds. *Genet. Sel. Evol.* 37, 455–472. doi: 10.1051/gse:2005011

Huang, Y., Zhao, Y., Haley, C. S., Hu, S., Hao, J., Wu, C., et al. (2006). A genetic and cytogenetic map for the duck (*Anas platyrhynchos*). *Genetics* 173, 287–296. doi: 10.1534/genetics.105.053256

Jasinska, A., Michlewski, G., De, M. M., Sobczak, K., Kozlowski, P., Napierala, M., et al. (2003). Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic Acids Res.* 31, 5463–5468. doi: 10.1093/nar/gkg767

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J. A., Baurain, D., et al. (2011). Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat. Genet.* 43, 405–413. doi: 10.1038/ng.814

Kozlowski, P., de Mezer, M., and Krzyzosiak, W. J. (2010). Trinucleotide repeats in human genome and exome. *Nucleic Acids Res.* 38, 4027–4039. doi: 10.1093/nar/gkq127

Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B., and Halldórsson, B. V. (2017). popSTR: population-scale detection of STR variants. *Bioinformatics* 33, 4041–4048. doi: 10.1093/bioinformatics/btw568

Larson, G., and Burger, J. (2013). A population genetics view of animal domestication. *Trends Genet.* 29, 197–205. doi: 10.1016/j.tig.2013.01.003

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, Y. C., Korol, A. B., Fahima, T., Beiles, A., and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11, 2453–2465. doi: 10.1046/j.1365-294X.2002.01643.x

Liu, C., Liu, Y., Zhang, X., Xu, X., and Zhao, S. (2017). Characterization of porcine simple sequence repeat variation on a population scale with genome resequencing data. *Sci. Rep.* 7:2376. doi: 10.1038/s41598-017-02600-8

Maak, S., Neumann, K., Von, L. G., and Gattermann, R. (2000). First seven microsatellites developed for the Peking duck (*Anas platyrhynchos*). *Anim. Genet.* 31:233.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206. doi: 10.1038/nature18964

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Payseur, B. A., Jing, P., and Haasl, R. J. (2011). A genomic portrait of human microsatellite variation. *Mol. Biol. Evol.* 28, 303–312. doi: 10.1093/molbev/msq198

Pearson, C. E., Nichol Edamura, K., and Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6, 729–742. doi: 10.1038/nrg1689

Pemberton, T. J., Sandefur, C. I., Jakobsson, M., and Rosenberg, N. A. (2009). Sequence determinants of human microsatellite variability. *BMC Genomics* 10:612. doi: 10.1186/1471-2164-10-612

Petes, T. D., Greenwell, P. W., and Dominska, M. (1997). Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae. Genetics* 146, 491–498.

Primmer, C. R. (1997). Low frequency of microsatellites in the avian genome. *Genome Res.* 7, 471–482. doi: 10.1101/gr.7.5.471

Qiu, Q., Wang, L., Wang, K., Yang, Y., Ma, T., Wang, Z., et al. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat. Commun.* 6:10283. doi: 10.1038/ncomms10283

Qu, L., Liu, W., Yang, F., Hou, Z., Zheng, J., Xu, G., et al. (2009). Origin and domestication history of Peking ducks delitermined through microsatellite and mitochondrial marker analysis. *Sci. China Ser. C Life Sci.* 52, 1030–1035. doi: 10.1007/s11427-009-0145-x

Rao, M., Morisson, M., Faraut, T., Bardes, S., Feve, K., Labarthe, E., et al. (2012). A duck RH panel and its potential for assisting NGS genome assembly. *BMC Genomics* 13:513. doi: 10.1186/1471-2164-13-513

Rd, F. J., and Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U.S.A.* 101, 18058–18063. doi: 10.1073/pnas.0408118101

Ren, J., Lu, L., Liu, X., Tao, Z., Zhang, C., Wang, D., et al. (2009). Paternity assessment: application on estimation of breeding value in body-weight at first egg trait of egg-laying duck (*Anas platyrhynchos*). *Mol. Biol. Rep.* 36, 2175–2181. doi: 10.1007/s11033-008-9432-z

Rohrer, G. A., Alexander, L. J., Hu, Z., Smith, T. P., Keele, J. W., and Beattie, C. W. (1996). A comprehensive map of the porcine genome. *Genome Res.* 6, 371–391. doi: 10.1101/gr.6.5.371

Rolfsmeier, M. L., and Lahue, R. S. (2000). Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae. Mol. Cell. Biol.* 20, 173–180. doi: 10.1128/MCB.20.1.173-180.2000

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591. doi: 10.1038/nature08832

Santos, C. H., Santana, G. X., Sa Leitao, C. S., Paula-Silva, M. N., and Almeida-Val, V. M. (2016). Loss of genetic diversity in farmed populations of *Colossoma macropomum* estimated by microsatellites. *Anim. Genet.* 47, 373–376. doi: 10.1111/age.12422

Sibly, R. M., Meade, A., Boxall, N., Wilkinson, M. J., Corne, D. W., and Whittaker, J. C. (2003). The structure of interrupted human AC microsatellites. *Mol. Biol. Evol.* 20, 453–459. doi: 10.1093/molbev/msg056

Sonay, T. B., Carvalho, T., Robinson, M. D., Greminger, M. P., Kruetzen, M., Comas, D., et al. (2015). Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res.* 25, 1591–1599. doi: 10.1101/gr.190868.115

Su, Y., Long, R., Chen, G., Wu, X., Xie, K., and Wan, J. (2007). Genetic analysis of six endangered local duck populations in China based on microsatellite markers. *J. Genet. Genomics* 34, 1010–1018. doi: 10.1016/S1673-8527(07)60114-3

Tadano, R., Kinoshita, K., Mizutani, M., and Tsudzuki, M. (2014). Comparison of microsatellite variations between Red Junglefowl and a commercial chicken gene pool. *Poult. Sci.* 93, 318–325. doi: 10.3382/ps.2013-03547

Toth, G., Gaspari, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967–981. doi: 10.1101/gr.10.7.967

Willems, T., Gymrek, M., and Highnam, G. (2014). The landscape of human STR variation. *Genome Res.* 24, 1894–1904. doi: 10.1101/gr.177774.114

Willems, T., Zielinski, D., Jie, Y., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* 14, 590–592. doi: 10.1038/nmeth.4267

Wu, J., Mao, X., Cai, T., Luo, J., and Wei, L. (2006). KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* 34, W720–W724. doi: 10.1093/nar/gkl167

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322. doi: 10.1093/nar/gkr483

Xu, L., Haasl, R. J., Sun, J., Zhou, Y., Bickhart, D. M., Li, J., et al. (2017). Systematic profiling of short tandem repeats in the cattle genome. *Genome Biol. Evol.* 9, 20–31. doi: 10.1093/gbe/evw256

Yoshimura, T., Yasuo, S., Watanabe, M., Iigo, M., Yamamura, T., Hirunagi, K., et al. (2003). Light-induced hormone conversion of T4 to T3 regulates photoperiodic response of gonads in birds. *Nature* 426, 178–181. doi: 10.1038/nature02117

Zhou, Z., Li, M., Cheng, H., Fan, W., Yuan, Z., Gao, Q., et al. (2018). An intercross population study reveals genes associated with body size and plumage color in ducks. *Nat. Commun.* 9:2648. doi: 10.1038/s41467-018-04868-4