# Discovering All Transcriptome Single-Nucleotide Polymorphisms and Scanning for Selection Signatures in Ducks (*Anas platyrhynchos*)

Ruiyi Lin[1,*], Xiaoyong Du[1,2,*], Sixue Peng[1], Liubin Yang[1], Yunlong Ma[1], Yanzhang Gong[1] and Shijun Li[1]

[1]Key Lab of Animal Genetics, Breeding and Reproduction of Ministry of Education, College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, People's Republic of China. [2]College of Informatics, Huazhong Agricultural University, Wuhan, People's Republic of China. *These authors contributed equally to this work.

**Supplementary Issue: RNA: An Expanding View of Function and Evolution**

**ABSTRACT:** The duck is one of the most economically important waterfowl as a source of meat, eggs, and feathers. Characterizing the genetic variation in duck species is an important step toward linking genes or genomic regions with phenotypes. Human-driven selection during duck domestication and subsequent breed formation has likely left detectable signatures in duck genome. In this study, we employed a panel of >1.4 million single-nucleotide polymorphisms (SNPs) identified from the RNA sequencing (RNA-seq) data of 15 duck individuals. The density of the resulting SNPs is significantly positively correlated with the density of genes across the duck genome, which demonstrates that the usage of the RNA-seq data allowed us to enrich variant functional categories, such as coding exons, untranslated regions (UTRs), introns, and downstream/upstream. We performed a complete scan of selection signatures in the ducks using the composite likelihood ratio (CLR) and found 76 candidate regions of selection, many of which harbor genes related to phenotypes relevant to the function of the digestive system and fat metabolism, including TCF7L2, EIF2AK3, ELOVL2, and fatty acid-binding protein family. This study illustrates the potential of population genetic approaches for identifying genomic regions affecting domestication-related phenotypes and further helps to increase the known genetic information about this economically important animal.

**KEYWORDS:** *Anas platyrhynchos*, RNA-seq, SNP, selection signatures

**SUPPLEMENT:** RNA: An Expanding View of Function and Evolution

**CITATION:** Lin et al. Discovering All Transcriptome Single-Nucleotide Polymorphisms and Scanning for Selection Signatures in Ducks (*Anas platyrhynchos*). *Evolutionary Bioinformatics* 2015:11(S1) 67–76 doi: 10.4137/EBO.S21545.

**TYPE:** Original Research

**RECEIVED:** August 07, 2015. **RESUBMITTED:** November 02, 2015. **ACCEPTED FOR PUBLICATION:** November 08, 2015.

**ACADEMIC EDITOR:** Jike Cui, Associate Editor

**PEER REVIEW:** Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1476 words, excluding any confidential comments to the academic editor.

**FUNDING:** This work was supported by the National Nature Scientific Foundation of China (No. 31172201), the National High Technology Research and Development Program of China (863 Program) (No. 2011AA100305), the National Nature Scientific Foundation of China (31402040), and the Fundamental Research Funds for the Central Universities (No. 2662015PY145). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**COMPETING INTERESTS:** Authors disclose no potential conflicts of interest.

**CORRESPONDENCE:** lishijun@mail.hzau.edu.cn

Published by Libertas Academica. Learn more about this journal.

## Introduction

One of the goals of livestock genomics research is to identify the genetic variation responsible for variation in phenotypic traits, particularly those of economic importance. Characterizing the genetic variation is an important step toward linking genes or genomic regions with phenotypes. The duck is one of the most economically important waterfowl as a source of meat, eggs, and feathers.[1] The duck is very important in some regions of the world, mainly in eastern and southern Asia. The total number of slaughtered ducks has increased significantly for several years in the commercially slaughtered poultry.

The completion of the duck genome sequence and recent advances in DNA sequencing (DNA-seq) technology allow for in-depth characterization of the genetic variations present in duck. Duck genome assembly is based on a domesticated individual from the Beijing breed, containing 78,487 scaffolds with an N50 value of 1.2 Mb and a set of 19,144 genes. The size of the duck genome assembly is 1.1 billion bases, and the heterozygosity rate was estimated to be 0.26%.[1] The single-nucleotide polymorphisms (SNPs) identified in the duck represent an essential step for future improvement of economically important traits through genetic association studies. Certain SNP markers associated with reproductive traits in ducks, especially hatchability, were identified, such as lysozyme,[2] ovomucoid gene,[3] and COLX gene.[4] However, to date, very limited SNPs have been identified within the whole duck genome assembly.

Transcriptome analysis has rapidly been shaped by next-generation sequencing technologies, as the benefits of RNA sequencing (RNA-seq) were acknowledged. The direct sequencing of cDNA libraries in RNA-seq allows for the discovery of new genes, transcripts, alternative splice junctions, fused sequences, and novel RNAs. RNA-seq has extensively been used to study the profile of gene expression, allelic difference in expression, transcriptome characterization, RNA–protein interactions, and alternative splicing.[5] However, very few studies have reported on the viability of SNP detection using RNA-seq.[6–10] Accurate mapping of junction reads to their genomic origins is crucial for avoiding mismatches that are interpreted as false SNPs. However, to date, very few computational pipelines for SNP calling have the ability to map reads in a splice-aware manner,[11] which poses a challenge to accurate SNP detection in RNA-seq data rather than that in DNA-seq data (whole-genome resequencing or exome data).

The complete intraspecific genome can be applied to inspect variants that have been subject to positive selection in the recent past. In theory, a beneficial variant that has been under the pressure of selection will generate distinct patterns in the respective region of the genome, such as reduction in variability and increase in linkage disequilibrium. Accordingly, the beneficial loci can be detected by examining the SNP patterns in intraspecific genome alignments. A number of statistical methods based on different demographic models have been proposed in the recent decades, such as Tajima's D,[12] Hudson–Kreitman–Aguadé test,[13] Fay and Wu's H test,[14] fixation index,[15] composite likelihood ratio (CLR),[16] extended haplotype homozygosity,[17] and integrated haplotype score.[18] Recently, the technical development of sequencing and SNP chips has provided us with high-density markers, enabling the identification of genome-wide selection signatures. Nielsen et al.[19] introduced two major modifications to the CLR method[16] for detecting selective sweeps in whole-genome data, which were based on the site frequency spectrum (SFS) method and implemented in the software, SweepFinder. Based on the code of SweepFinder, Pavlidis et al.[20] developed a faster and more accurate selective sweep detector, termed SweeD. A neutral SFS can be obtained in SweeD without the need to compute the empirical average SFS for the genome. If successful, the detection of selection signatures can provide a straightforward insight into the mechanisms of artificial selection and further uncover the causal genes related to the phenotypic variation.

Here, we describe the transcriptome sequencing of 15 ducks from distinct breeds for the purpose of identifying and annotating the novel forms of genetic variation in ducks using highly accurate SNP calling methods. By analyzing the nucleotide diversity using sequencing data, we aim to identify genomic regions exhibiting the signatures of selection and positional candidate genes reported in proximity to the genomic positions, showing the most significant indications of selection, and to gain a further insight into the genome-wide footprints of duck selection. In addition, the functions associated with the genes putative under selection regions were investigated by gene set enrichment analysis on gene ontology (GO) annotations.

## Materials and Methods

**Experimental animals.** In this experiment, four individuals from the Baigai breed, two from the Ma breed, one from Liancheng White, and one from Longsheng Green were sampled. The RNA-seq data of seven HBK–SPF individuals from Peking duck breed were retrieved from NCBI SRA database. Three feather bulbs from the same individual were pooled as one sample. All research involving animals were conducted according to the regulation (No. 5 proclaim of the Standing Committee of Hubei People's Congress) approved by the Standing Committee of Hubei People's Congress, and the ethics committee of Huazhong Agricultural University, P. R. China.

**Preparation of Illumina libraries and sequence analysis.** Feather bulbs were put into 2-mL tubes containing 1-mL TRIzol reagent (Invitrogen). A magnetic bead homogenizer was used to homogenize the tissue and TRIzol. The quality and quantity of RNA samples were detected by Spectrophotometer ND-1000 (NanoDrop) and denaturing agarose gel electrophoresis. All RNA samples were treated with DNAse I for later use. RNA-seq libraries were constructed using the mRNA-Seq Prep Kit (Illumina) and then sequenced using the paired-end sequencing module of the Illumina HiSeq 2000 platform (100 bp at each end). Low-quality reads were trimmed by trimmomatic toolkit with default options.[21]

**SNP calling.** The scaffold sequences and transcriptome sequences of the duck genome were retrieved from the Ensembl database. We used two approaches to identify SNP: (1) GATK–DNA-seq[22] combined SNPiR pipeline[11] and (2) the recently developed GATK–RNA-seq.

In the first approach, we used the Burrows–Wheeler Aligner[23] to map RNA-seq reads against both the reference genome and the transcriptome. We mapped each of the paired-end reads separately to the reference genome using the commands "bwa aln fastqfile" and "bwa samse -n4." We use BWA to map the data to transcriptome with commands "bwa aln fastqfile1," "bwa aln fastqfile2," and "bwa sampe." For the mapped reads, we used IndelRealigner (default), TableRecalibration (default), and UnifiedGenotyper ("stand_call_conf = 0," "stand_emit_conf = 0," and "output mode = EMIT_ALL_CONFIDENT_SITES"), successively, from GATK for DNA-seq[22] to local realignment, base score recalibration, and candidate SNP calling nearly as suggested by "Best Practices Workflows" in the manual of GenomeAnalysisTK–2.8–1. We called variants with these loose criteria in GATK, which allowed a high sensitivity of SNPiR pipeline[11] to filter the candidate SNPs. We required candidate SNP call with the quality of $Q > 20$. These filters in the SNPiR pipeline included the removal of false calls in duplicated regions, mismatch sites at 5′ read ends, sites in repetitive regions according to RepeatMasker annotation, intronic sites within 4 bp of splice junctions, and sites in homopolymer runs of >5 bp. Then, BLAT[24] was used to map all reads against the duck genome to support unique mapping of the SNPs. In the second approach, we applied a method recently developed in GTAK 3.0 specifically for calling variants in RNA-seq (GATK–RNA-seq). In brief, the key modifications made to GATK–RNA-seq focus on handling splice junctions correctly, which involves specific mapping by STAR aligner[25] and the procedures, including SplitNCigarReads (with parameters: -RMQF 255 -RMQT 60 -U ALLOW_N_CIGAR_READS), Recalibration (default parameters), HaplotypeCaller (with parameters: -dontUseSoftClippedBases -stand_call_conf 20.0 -stand_emit_conf 20.0), and HaplotypeCaller (with parameters: -window 35 -cluster 3 -filterName FS -filter "FS > 30.0" -filterName QD -filter "QD < 2.0"). We merged the resulting variants in VCF4.0 format from the two SNP calling approaches. The consensus variants, with the SNP quality of $Q < 20$, with minor allele frequency of <0.05, or with >50% missing genotypes within the sampled individuals, were filtered.

The gene-based analysis of SnpEff software[26] with standard settings was used to functionally annotate the putative SNPs. For each SNP, the location (exonic, intronic, intergenic, 5′UTR, 3′UTR, splice acceptor or donor site, downstream or upstream) and the functional annotation (nonsynonymous, synonymous) were determined based on the duck reference genome.[1] Gene annotations used in this analysis were taken from the Ensembl database (BGI_duck_1.0.75). For genome-wide detection of positive selection signatures, we converted the genome coordinates of SNPs from scaffold level to duck chromosomes using information from duck Radiation Hybrid (RH) genome maps (unpublished), which was built from the duck RH panel.[27]

**Detecting positive selection.** We first conducted the admixture analysis using NGSadmix[28] to roughly estimate the introgression among different populations. Then, we performed the CLR test using the information from allele frequencies to identify completed sweeps, and the statistic was calculated for a nonoverlapping sliding window of 100 kb across each chromosome respectively. In this process, the SFS of the entire chromosome was considered as the background SFS to calculate the composite likelihood of a recently completed selective sweep in each window. To obtain the empirical distributions of CLR, a neutral sequence equal to each chromosome in length was simulated 10,000 times using ms[29] with the given epoch demography of the instantaneous size change followed by exponential growth (with parameters: ms 30 10000 -t 80 -G 6.93 -eG 0.2 0.0 -eN 0.3 0.5). To convert from the ms parameter, we assumed a mutation rate of $10^{-8}$ per site per generation and considered a segment 100 kb pairs long. Then, a threshold value of significant CLR for each chromosome is determined at $P \leq$ 10E-4.

The GO terms were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID).[30] First, we retrieved the Ensembl IDs of the genes, which were considered to be overlapping if their positions were contained inside the candidate regions of selection. The DAVID was used to analyze enrichment in the GO terms using 257 human orthologs from the 329 duck genes. The GO terms with $P$-values < 0.05 and with a false discovery rate of <25% were used for further analysis in our study.

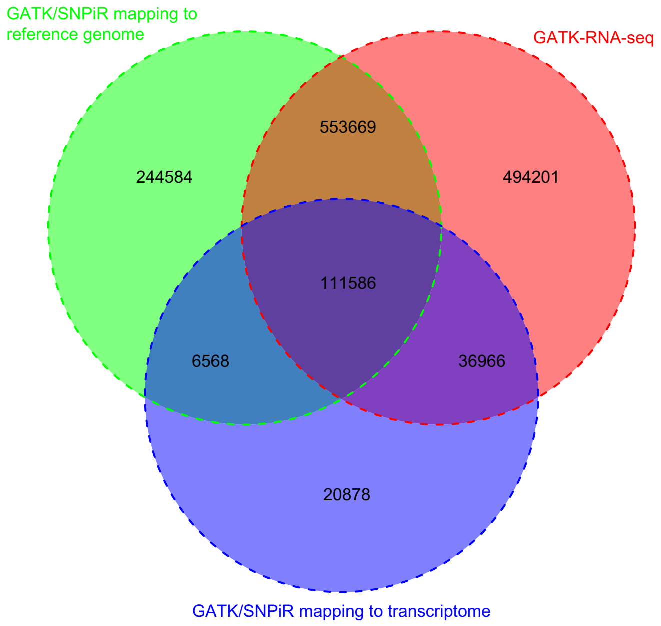

**Figure 1.** Comparison of SNPs identified by the two approaches. We detected 916,407 SNPs by mapping to the reference genome (green) and 175,998 SNPs were called by mapping to the transcriptome (purple), which composed 974,251 unique variants using GATK–DNA-seq combined with the SNPiR pipeline. Besides, we applied the approach of GATK–RNA-seq and then detected 1,196,422 variants (red).

## Results and Discussions

We identified SNPs in the transcriptome of 15 samples using the GATK–DNA-seq[22] combined with the SNPiR pipeline.[11] In total, we were able to detect 916,407 variants by mapping to the reference genome, and 175,998 variants were called by mapping to the transcriptome, yielding 974,251 unique variants in total (Fig. 1). Utilizing the transcriptome for mapping short reads only provides an additional 57,844 variants, mainly because the splice junctions from gene models annotated in the reference duck genome are currently limited. Thus, we applied an approach recently developed in GATK–RNA-seq. Then, 1,196,422 variants were detected in the RNA-seq of 15 samples according to the Best Practices Workflows of GATK–RNA-seq (Fig. 1). We merged the resulting variants from the two SNP calling approaches to obtain 1,468,452 unique SNPs in the duck genome (Fig. 1). The resulting SNPs were deposited into the Database of Short Genetic Variation build 145 (http://www.ncbi.nlm.nih.gov/snp/) with NCBI Submitted SNP (ss) number (1939971667–1947221053). Our result exhibited a transition-to-transversion (ts/tv) ratio of 2.06 for the entire duck genome, which is similar to the overall ts/tv ratio of 2.0–2.1 for the entire human genome.[31] Previous studies have established an expected higher ts/tv ratio (3–4) for human coding regions.[32] We found an estimated ts/tv ratio of 3.51 for duck exonic regions in our study, a good reflection of the genomic variation in transcribed regions.

To date, a specific challenge to call a variant in RNA-seq data is to quantify accurately a different number of mapped

**Table 1.** Summary of SNPs in ducks.

| CATEGORY | COUNT | PERCENT (%) | NOTE |
|---|---|---|---|
| Sample size | N = 15 | – | SPLICE_SITE_REGION' means that a variant is within 2 bp of a splice junction.<br>'SPLICE_SITE_ACCEPTOR' means that the variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).<br>'SPLICE_SITE_DONOR' means that the variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).<br>'UPSTREAM/DOWNSTREAM' means that a variant overlaps with the 1 kb region upstream/ downstream of the gene end site.' Number of effects is larger than the number of SNPs because a variant is annotated for two or more effects. |
| SNP | 1,468,452 | – | |
| UPSTREAM | 269,271 | 11.48 | |
| UTR_5_PRIME | 2,778 | 0.12 | |
| Exonic | 190,444 | 8.12 | |
| Non_coding_exon | 1,207 | 0.05 | |
| Frameshift | 885 | 0.04 | |
| NON_SYNONYMOUS | 57,628 | 2.46 | |
| Synonymous | 130,575 | 5.57 | |
| Nonsyn/Syn ratio (ω) | 0.44 | – | |
| INTRON | 552,382 | 23.56 | |
| SPLICE_SITE_REGION | 115,458 | 4.92 | |
| SPLICE_SITE_DONOR | 34,861 | 1.49 | |
| SPLICE_SITE_ACCEPTOR | 13,864 | 0.59 | |
| UTR_3_PRIME | 30,477 | 1.30 | |
| DOWNSTREAM | 507,849 | 21.66 | |
| INTERGENIC | 627,198 | 26.75 | |
| Number of effects | 2,344,610 | 100 | |

reads to the reference genome and then to determine whether a variant exists. Accurate variant calling could be hampered by (1) highly similar regions in the genome, (2) the library construction for mRNA-seq and the inability to map reads in a splice-aware manner, and (3) the RNA-editing and allele-specific expression, which could mimic SNPs or bias allele frequencies.[33] These hindrances, especially the last one, may have resulted in false positive SNP calling in our study. However, there are not yet any other sources of duck SNP data (like resequencing or exome data) to validate our result variants.

The admixture proportions were estimated from genotype likelihoods using NGSadmix.[28] A graphic representation of cluster structure analysis is depicted in Figure 2. There had already been a significant differentiation between Baigai White, Ma–Liancheng White, and Peking duck populations, while Longsheng Green had shown admixture with all three of those populations.

**Enrichment of variants in functional categories.** After SNP calling, we assigned a functional class to each SNP and provided several fields of information describing the affected transcripts and proteins, if applicable. The RNA-seq data allowed us to enrich SNP variants than whole-genome sequencing in coding exons (8.12%), untranslated regions (UTRs; 1.42%), upstream/downstream (33.41%), and introns (23.56%). The annotated SNPs identified in this study can serve as useful genetic tools and as candidates in searches for phenotype-altering DNA and transcript differences. The resulting SNPs were highly abundant in these four categories (Table 1). Although the libraries for mRNA-seq are supposed to be enriched for mRNAs via poly(A)+ selection, a certain amount of immature pre-mRNA, which carries the introns, usually infiltrates into the libraries. Introns compose a much larger fraction of the duck genome than exonic regions. Those two facts mainly explain why so many SNPs exist in introns than in exonic regions. Most of these variant calls in intronic and intergenic regions are usually of low quality because of low coverage of short reads (Fig. 3). Only a small fraction, 26.75%, of SNPs fell into intergenic regions, which compose 65.46% of duck genome.

RNA degrades quickly, so poly(A)+ selection may catch the 3′ end of an mRNA that is already breaking down, which means that some portion of the 5′ end has already been lost. As a result, mRNA-seq coverage when polyA selection protocols have been used in our study is deeper at the 3′ end of genes and comparatively sparse at the 5′ end. For example, of 165 genes, as shown in Figure 3, there is ~10 × coverage at the 3′ end (3′UTR and last intron on the far right), trailing off to ~3 × by the time you get to the 5′ end (5′UTR and first intron on the far left).

**Genome-wide scanning for selection signatures.** We applied the CLR test to scan for genomic regions showing the signals of recent selection in the duck genome.[20] The regional SFS was compared with the background SFS to calculate CLR, which indicates the likelihood of a signal at each window of 100 kb in length across the whole genome. Significance was determined by the threshold value deriving from the empirical distribution of neutral scenarios simulated in the software ms.[29] A total of 126 windows

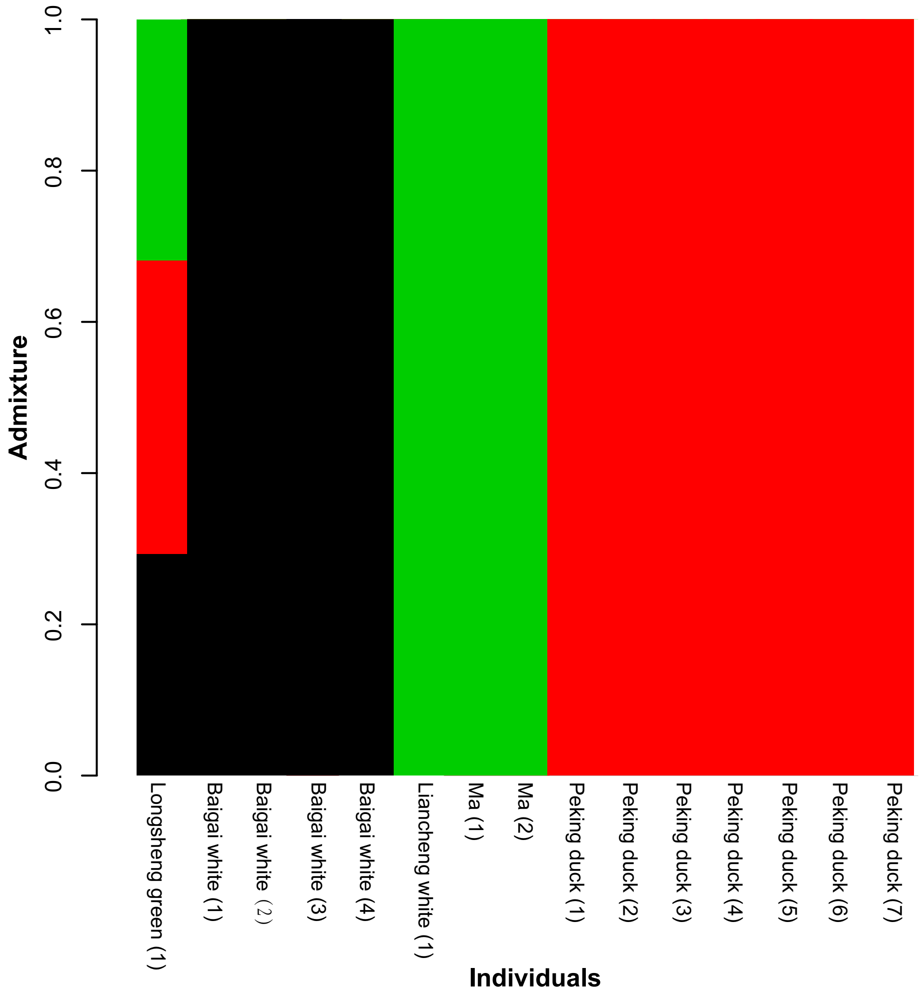

**Figure 2.** Admixture plot presenting genetic structure for 15 duck individuals. The length of each colored segment represents the proportion of the individual's genome from $K = 3$ ancestral populations. The single color bar in Baigai White, Ma–Liancheng White, and Peking duck means that there is no admixture.

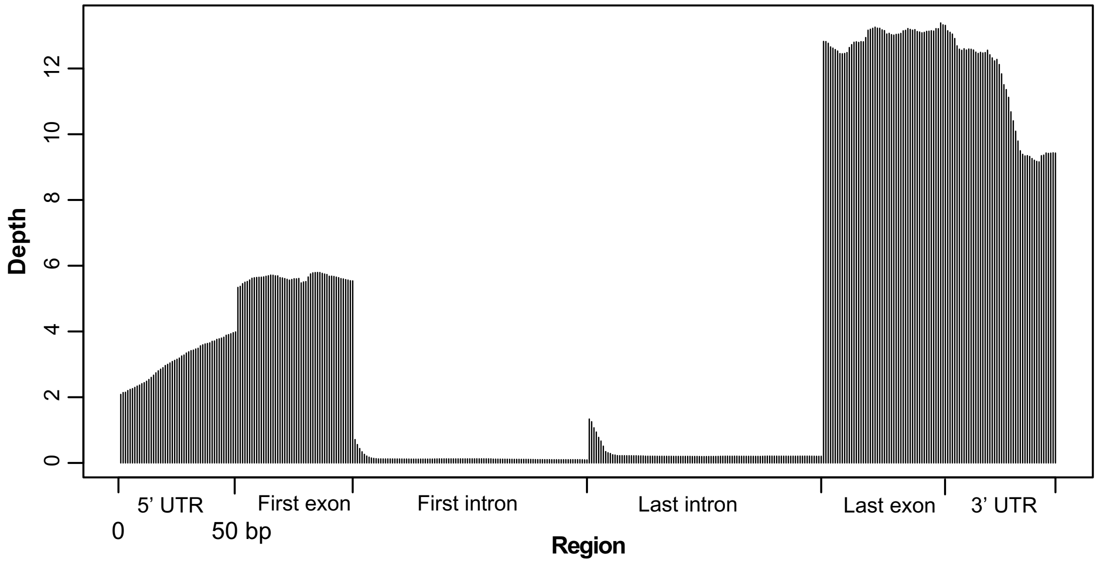

**Figure 3.** Average read depths on different elements (UTR, exon, intron) of 165 genes randomly selected from the duck transcriptome.
The ~10 × coverage at the 3′ end (far right) trailing off to ~3 × by the time you get to the 5′ end (far left) indicated that in some cases, a portion of the 5′ end of a transcript has already been lost in the preparation of RNA-seq. The lower coverage at the 5′ end of transcript resulted in less SNPs calling at the 5′ end than that in the 3′ end. The much lower density of SNP calls in intronic regions and the lower quality of these SNPs are mainly because of very low coverage of short reads in RNA-seq data.

**Table 2.** Summary of significant CRs ($P \leq 0.01$) and distribution of SNPs in five duck populations.

| CHR | WINDOWS (N)[1] | CR (N)[2] | CR SNPS (N)[3] | CHR SNPS (N)[4] | CR LENGTH (KB) | CHR LENGTH (MBP) | CR GENE (N) | CHR GENES (N) |
|---|---|---|---|---|---|---|---|---|
| chr1 | 29 | 11 | 3,682 | 186,609 | 4,200 | 198.28 | 43 | 2,124 |
| chr2 | 17 | 12 | 1,890 | 121,342 | 3,000 | 154.285 | 38 | 1,365 |
| chr3 | 10 | 9 | 2,684 | 103,474 | 1,900 | 115.727 | 24 | 1,207 |
| chr4 | 17 | 9 | 2,573 | 59,448 | 2,900 | 74.523 | 46 | 758 |
| chr5 | 6 | 5 | 1,249 | 75,816 | 1,100 | 63.518 | 20 | 947 |
| chr6 | 4 | 3 | 1,614 | 38,315 | 800 | 36.433 | 13 | 514 |
| chr7 | 2 | 2 | 694 | 41,748 | 400 | 39.268 | 7 | 527 |
| chr8 | 3 | 2 | 386 | 44,361 | 500 | 31.228 | 4 | 522 |
| chr9 | 5 | 3 | 904 | 41,473 | 800 | 26.143 | 15 | 448 |
| chr10 | 0 | 0 | 0 | 28,568 | 0 | 18.705 | 0 | 320 |
| chr11 | 1 | 1 | 181 | 37,742 | 200 | 21.689 | 2 | 419 |
| chr12 | 0 | 0 | 0 | 29,961 | 0 | 20.949 | 0 | 350 |
| chr13 | 0 | 0 | 0 | 35,703 | 0 | 21.836 | 0 | 338 |
| chr14 | 4 | 4 | 2,292 | 36,637 | 800 | 19.493 | 16 | 345 |
| chr15 | 0 | 0 | 0 | 36,406 | 0 | 17.612 | 0 | 430 |
| chr16 | 2 | 1 | 448 | 35,585 | 300 | 15.016 | 13 | 374 |
| chr17 | 0 | 0 | 0 | 4,888 | 0 | 0.387 | 0 | 39 |
| chr18 | 1 | 1 | 624 | 32,263 | 200 | 11.812 | 7 | 308 |
| chr19 | 3 | 3 | 984 | 27,947 | 600 | 12.468 | 11 | 318 |
| chr20 | 0 | 0 | 0 | 32,858 | 0 | 11.803 | 0 | 343 |
| chr21 | 4 | 3 | 1,226 | 25,870 | 801 | 15.674 | 20 | 346 |
| chr22 | 0 | 0 | 0 | 26,388 | 0 | 7.939 | 0 | 251 |
| chr23 | 0 | 0 | 0 | 9,916 | 0 | 4.482 | 0 | 111 |
| chr24 | 0 | 0 | 0 | 25,486 | 0 | 7.225 | 0 | 240 |
| chr25 | 0 | 0 | 0 | 15,503 | 0 | 7.33 | 0 | 175 |
| chr26 | 0 | 0 | 0 | 7,475 | 0 | 1.284 | 0 | 80 |
| chr27 | 0 | 0 | 0 | 27,757 | 0 | 6.462 | 0 | 257 |
| chr28 | 0 | 0 | 0 | 19,044 | 0 | 4.768 | 0 | 201 |
| chr29 | 1 | 1 | 629 | 21,146 | 200 | 4.454 | 11 | 198 |
| chrW | 1 | 1 | 188 | 1,491 | 200 | 2.089 | 10 | 40 |
| chrZ | 16 | 5 | 947 | 32,634 | 2,300 | 74.036 | 29 | 735 |
| Total | 126 | 76 | 23,195 | 1,263,854 | 21201 | 1046.918 | 329 | 14,630 |

**Notes:** [1]Windows of size 100 kb with $P < 10E\text{-}4$. [2]Distinct CRs with $P$-value $< 10$ E-4. [3]Total number of SNPs forming significant CRs. [4]Total number of SNPs used in the chromosome.

had a $P$-value of <10E-4, indicating that the CLR of these windows surpassed every CLR obtained from the distribution of 10,000 neutral simulations. As some of the 126 windows were adjacent to each other, these correspond to 76 distinct core regions (CRs) where the observed CLR value significantly exceeded neutral simulations. The results are presented in Table 1, which also includes the number of candidate regions and SNPs putative under selection for each chromosome. As shown in Figure 3, the transcriptome sequencing has significantly lower sequence coverage in noncoding regions, which certainly affects accurate variant calls in these regions compared with coding regions, and may bias the quality of CLR scan. We found that the DAF (absolute allele frequency) of coding SNPs is significantly larger than that of noncoding SNPs (one-tailed $t$-test, $P < 2.2E\text{-}16$), indicating that the SFS in coding and noncoding regions differs strongly. However, if regions with less callable reads were excluded, power to detect a selective sweep through the CLR could be lowered. We extracted all the 227,948 SNPs in UTR/exon regions and performed a CLR scan. The sweep result from the data of UTR/exon SNPs was comparable to the sweep result from that of whole SNPs.

We investigated genes in the candidate regions of selection, and corresponding genes were identified by comparing

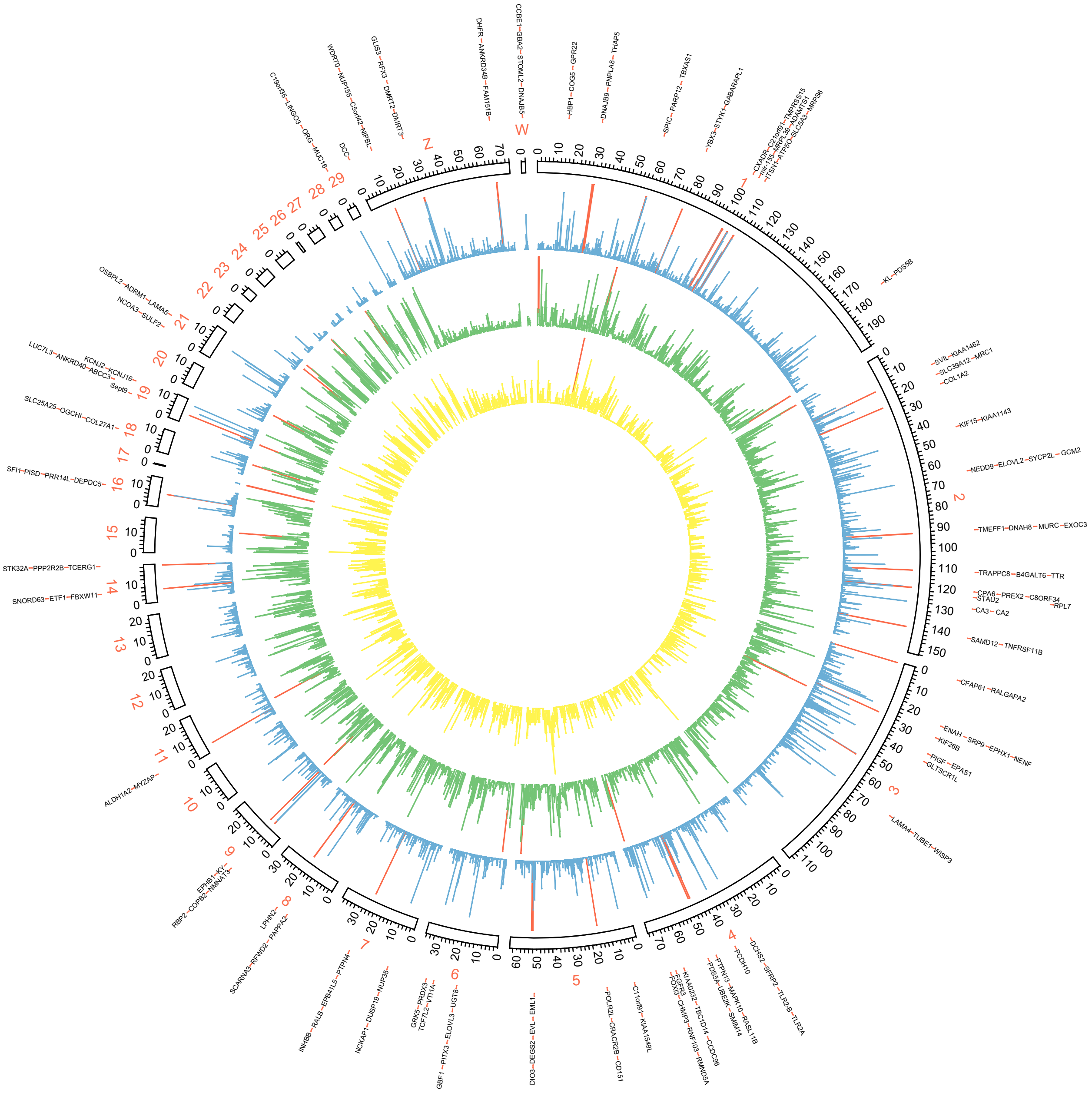

**Figure 4.** Circos plot of the global distribution of genes, SNP variants, and signature of selective sweep along with the genome. The circles from outside to inside illustrate gene density (yellow), SNP density (green), and CLR values (blue). The genes located in regions with significant strong sweep signatures are presenting as outliers. High values in each layout (gene density > 10/100 kb, SNP density > 1000/100 kb, and CLR value > 30) were marked in red histograms.

their genomic locations with the available annotation of the duck genome. In total, 329 genes were located on the 76 distinct CRs (Table 2 and Fig. 4). The region with the largest CLR in the duck genome is located on chromosome 1 (Figs. 4 and 5). This region consists of eight adjacent 100 kb windows with *P*-value < 10E-5. There are several protein coding genes in this top-scoring 1 Mb region (chr1: 28.7 Mb–29.7 Mb), including the protein coding genes DNAJB9 (ENSAPLG00000007950), THAP5 (ENSAPLG00000008047), AVPR2(ENSAPLG00000008090), PNPLA8(ENSAPLG00000008096),NRCAM(ENSAPLG00000008656), and CNTN1 (ENSAPLG00000010381). The region contains 13 nonsynonymous SNPs that passed the quality filtering: one coding change (KB743050.1:385586 | His > Arg) in DNAJB9, one coding change (KB743050.1:396307 | Thr > Ile) in THAP5, two coding changes (KB743050.1:607259 | Phe > Val and KB743050.1:620545 | Asn > Ser) in NRCAM, and nine coding changes in PNPLA8. The PNPLA8 gene encodes a member of the patatin-like phospholipase domain-containing protein family, and the product of PNPLA8 is a calcium-independent phospholipase. Mutations in PNPLA8 are associated with mitochondrial myopathy with lactic acidosis.[34]

**Involved genes and biological processes under selection.** We then sought to investigate the functions associated

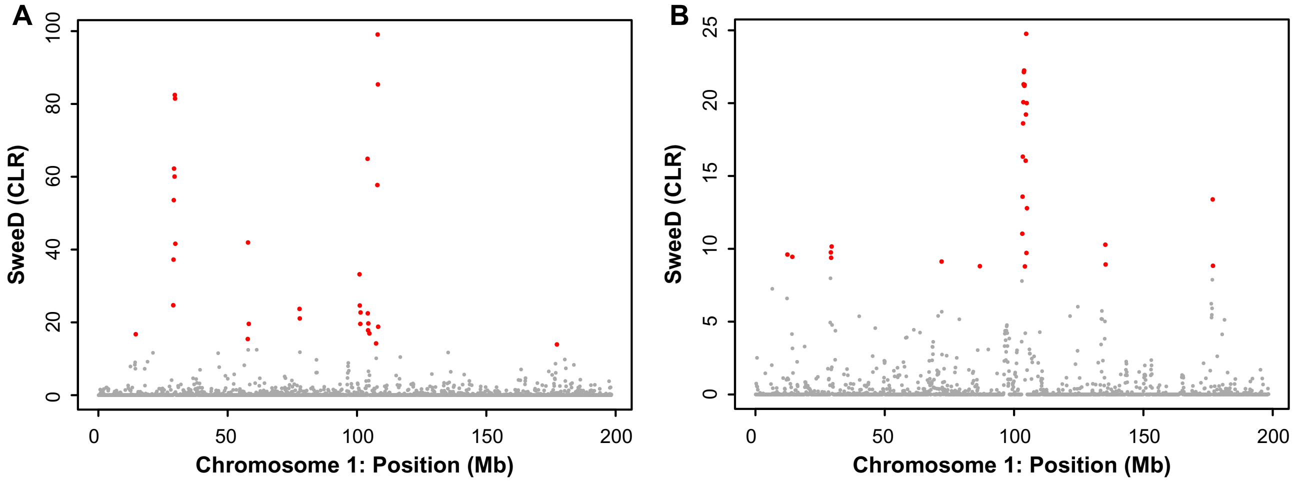

**Figure 5.** The result of selective sweep for the entire chromosome 1. (**A**) The scan of CLR for the sweep signal at each window of 100 kb in length across chromosome 1. We found outlier genomic regions with significant strong sweep signatures at a threshold of 10E-4 (shown in red). (**B**) The scan of CLR for the sweep signal at each window of 100 kb in length across the across chromosome 1 using only SNP data from the exon/UTR regions.

with the genes putative under the 76 selected regions. The analysis of overrepresented annotations and pathways was performed using DAVID with those 329 genes.[30] If the *P*-value was <0.05 for GO annotation, it was considered significant. We identified 21 enriched GO categories ($P < 0.05$; Table 3). The top three terms with the lowest *P*-values were positive regulation of binding (GO:0051099, $P = 5.4$E-3), pancreas development (GO:0031016, $P = 1.4$E-2), and intracellular transport (GO:0046907, $P = 1.5$E-2). There are two kinds of pancreatic secretion function: one is exocrine secretion, which mainly secretes digestive enzymes; the other is endocrine secretion, which secretes the islet A and B cells. The poultry pancreas is one of the important digestive glands. Amylase, protease, and lipase in the small intestine, which are the most important enzymes of the digestive system, are mainly derived from the pancreas.[35] We found that the EIF2AK3 gene is connected with the development of pancreas. Studies also have shown that EIF2AK3 kinase activity is essential for normal development of the islet B cell.[36] In humans, mutations in the EIF2AK3 gene will result in neonatal or early infancy type 1 diabetes.[37] There are five nonsynonymous SNPs in the EIF2AK3 gene (Gly > Arg at the position of the 223th amino acid, Cys > Trp at position 233, Ile > Leu at position 981, Gln > Arg at position 1005, and Lys > Arg at position 1006). Evidence of more effective selection in birds than that in humans can be seen from the previous observations of a higher proportion of nonsynonymous substitutions in birds.[38,39] These specific SNPs in the gene suggest potential functional changes that would be useful for further genetic study of the gene in birds.

Furthermore, the significant GO terms also included "fatty acid biosynthetic process (GO:0006633)" and "regulation of insulin secretion (GO:0050796)." As the fat traits in duck are related to meat quality and nutritive value, the fatty acid biosynthetic process is essential for duck breeding. The ELOVL2 gene is involved in the fatty acid biosynthetic process. Duck ELOVL2 may have the ability to convert α-linolenic acid to docosahexaenoic acid,[40] which is an important omega-3 fatty acid for humans. A previous study has reported that insulin plays an important role in the metabolism of glucose and lipid in poultry.[41] As insulin secretion is associated with pancreas development, the TCF7L2 gene takes part in pancreas development and regulation of insulin secretion. One nonsynonymous SNP was also found in the TCF7L2 gene (Val > Ala at the position of the 145th amino acid). ΔN-Tcf7l2 transgenic mice, which lack the N-terminal β-catenin-binding domain, show impaired glucose tolerance with insulin secretion decreased.[42] Long-chain fatty acid transporter activity (GO:0005324, $P = 5.9$E-2) and fatty acid transporter activity (GO:0015245, $P = 8.7$E-2) were found to be significant in molecular function. The FABP1 and FABP2 genes, which were involved in those two terms, encode the fatty acid-binding protein (FABP) found in the liver or in the intestines. It is thought that the roles of FABPs include fatty acid uptake, transport, and metabolism.[43] We speculated that reasonable regulation of fat in domestic ducks was to cater to the requirements of humans. A previous study has reported that a number of genes in the significant CRs under selection may be important for controlling abdominal fatness in domestic chicken lines.[44]

The current annotation of the duck genome has a limited availability of GO terms, which decreases the sensitivity of the analysis. Therefore, we could provide only suggestive evidences for the overrepresented annotations affected by positive selection.

In conclusion, we achieved high accuracy of SNP calling and enrichment of variants in functional categories using the RNA-seq data. This study provides a genome-wide map of selection signatures in duck genomes and yields insight into the mechanisms of selection in duck breeding. Our results show that genes related to the function of the digestive system and lipid metabolism may also experience positive selection.

**Table 3.** The enriched biological process of GO analysis.

| TERM | GENES | *P*-VALUE |
|---|---|---|
| GO:0051099~positive regulation of binding | AMH, NCOA3, HIPK2, JAK2, PRDX3, EIF2AK3 | 0.0054 |
| GO:0031016~pancreas development | ALDH1A2, HNF1A, EIF2AK3, TCF7L2 | 0.0139 |
| GO:0046907~intracellular transport | ENAH, GRPEL1, NPEPL1, NUP155, VTI1A, COPB2, APP, COG5, GBF1, ZFYVE16, LYST, STX16, GNAS, JAK2, ATP5O, SRP9, HSPA9, TOB1 | 0.0149 |
| GO:0043388~positive regulation of DNA binding | AMH, NCOA3, HIPK2, JAK2, PRDX3 | 0.0186 |
| GO:0050796~regulation of insulin secretion | INHBB, HNF1A, JAK2, TCF7L2 | 0.0200 |
| GO:0001655~urogenital system development | AMH, ALDH1A2, LAMA5, ADAMTS1, NID1, CA2 | 0.0219 |
| GO:0007018~microtubule-based movement | APP, KIF15, TUBE1, TUBB1, DNAH8, KIF26B | 0.0242 |
| GO:0007167~enzyme linked receptor protein signaling pathway | WFIKKN2, FGFR3, KL, CD8B, ZFYVE16, HIPK2, COL1A2, JAK2, EIF2AK3, EPHB1, TOB1 | 0.0274 |
| GO:0002763~positive regulation of myeloid leukocyte differentiation | GNAS, CA2, RUNX1 | 0.0274 |
| GO:0006633~fatty acid biosynthetic process | TBXAS1, HNF1A, ELOVL3, ELOVL2, DEGS2 | 0.0277 |
| GO:0015031~protein transport | EIF4ENIF1, GRPEL1, NPEPL1, NUP155, VTI1A, COPB2, COG5, RNF103, SCFD2, ZFYVE16, LYST, STX16, EXOC3, GNAS, JAK2, NUP35, SRP9, HSPA9, TOB1 | 0.0281 |
| GO:0002791~regulation of peptide secretion | INHBB, HNF1A, JAK2, TCF7L2 | 0.0288 |
| GO:0006891~intra-Golgi vesicle-mediated transport | COPB2, COG5, STX16 | 0.0304 |
| GO:0045184~establishment of protein localization | EIF4ENIF1, GRPEL1, NPEPL1, NUP155, VTI1A, COPB2, COG5, RNF103, SCFD2, ZFYVE16, LYST, STX16, EXOC3, GNAS, JAK2, NUP35, SRP9, HSPA9, TOB1 | 0.0304 |
| GO:0008104~protein localization | EIF4ENIF1, NPEPL1, GRPEL1, HNF1A, AMN, NUP155, VTI1A, COPB2, COG5, RNF103, SCFD2, ZFYVE16, LYST, STX16, EXOC3, GNAS, JAK2, NUP35, SRP9, HSPA9, TOB1 | 0.0310 |
| GO:0015718~monocarboxylic acid transport | HNF1A, ABCC3, FABP1, FABP2 | 0.0357 |
| GO:0016050~vesicle organization | COPB2, GBF1, LYST, ZFYVE16 | 0.0394 |
| GO:0006754~ATP biosynthetic process | ATP5E, AK3, ATP10D, ATP5O, ATP5J | 0.0403 |
| GO:0043523~regulation of neuron apoptosis | NMNAT3, HIPK2, JAK2, PRDX3, ITSN1 | 0.0417 |
| GO:0006917~induction of apoptosis | DCC, APP, RASGRF2, LYST, HIPK2, NAIF1, JAK2, ITSN1, PDCD6, TRAF3 | 0.0436 |
| GO:0012502~induction of programmed cell death | DCC, APP, RASGRF2, LYST, HIPK2, NAIF1, JAK2, ITSN1, PDCD6, TRAF3 | 0.0442 |

## Acknowledgments

We thank Man Rao (INRA, Toulouse, France) for providing the chromosome sequences of duck genome.

## Author Contributions

Conceived and designed the experiments: XD, SL. Analyzed the data: XD, RL. Contributed to the writing of the manuscript: RL, XD, YM, SP, LY, SL. Agreed with manuscript results and conclusions: SL, YG. All the authors reviewed and approved the final manuscript.

## REFERENCES

1. Huang Y, Li Y, Burt DW, et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat Genet*. 2013;45(7):776–83.
2. Huang H-L, Cheng Y-S. A novel minisequencing single-nucleotide polymorphism marker of the lysozyme gene detects high hatchability of Tsaiya ducks (*Anas platyrhynchos*). *Theriogenology*. 2014;82(8):1113–20.
3. Huang HL, Cheng YS, Huang CW, Huang MC, Hsu WH. A novel genetic marker of the ovomucoid gene associated with hatchability in Tsaiya ducks (*Anas platyrhynchos*). *Anim Genet*. 2011;42(4):421–7.
4. Chang M-T, Cheng Y-S, Huang M-C. A novel non-synonymous SNP of the COLX gene and its association with duck reproductive traits. *Mol Cell Probes*. 2012;26(5):204–7.
5. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*. 2011;12(2):87–98.
6. Heap GA, Yang JH, Downes K, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet*. 2010;19(1):122–34.
7. Salem M, Vallejo RL, Leeds TD, et al. RNA-Seq identifies SNP markers for growth traits in rainbow trout. *PLoS One*. 2012;7(5):e36264.
8. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*. 2010;464(7289):773–7.
9. Yang SS, Tu ZJ, Cheung F, et al. Using RNA-seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC Genomics*. 2011;12:199.
10. Quinn EM, Cormican P, Kenny EM, et al. Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One*. 2013;8(3):e58815.
11. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet*. 2013;93(4):641–51.
12. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123(3):585–95.
13. Hudson RR, Kreitman M, Aguade M. A test of neutral molecular evolution based on nucleotide data. *Genetics*. 1987;116(1):153–9.
14. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. *Genetics*. 2000;155(3):1405–13.

15. Wright S. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*. 1965;19:395–420.
16. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*. 2002;160(2):765–77.
17. Sabeti PC, Reich DE, Higgins JM, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002;419(6909):832–7.
18. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. 2006;4(3):e72.
19. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. *Genome Res*. 2005;15(11):1566–75.
20. Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol*. 2013;30(9):2224–34.
21. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.
22. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
24. Kent WJ. BLAT – the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–64.
25. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
26. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92.
27. Rao M, Morisson M, Faraut T, et al. A duck RH panel and its potential for assisting NGS genome assembly. *BMC Genomics*. 2012;13:513.
28. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013;195(3):693–702.
29. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. 2002;18(2):337–8.
30. Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*. 2003;4(5):3.
31. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061–73.
32. Marth GT, Yu F, Indap AR, et al; 1000 Genomes Project. The functional spectrum of low-frequency coding variation. *Genome Biol*. 2011;12(9):R84.
33. Lee J-H, Ang JK, Xiao X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA*. 2013;19(6):725–32.
34. Saunders CJ, Moon SH, Liu X, et al. Loss of function variants in human PNPLA8 encoding calcium-independent phospholipase A2 gamma recapitulate the mitochondriopathy of the homologous null mouse. *Hum Mutat*. 2015;36(3):301–6.
35. Mahagna M, Nir I, Larbier M, Nitsan Z. Effect of age and exogenous amylase and protease on development of the digestive tract, pancreatic enzyme activities and digestibility of nutrients in young meat-type chicks. *Reprod Nutr Dev*. 1995;35(2):201–12.
36. Wang R, McGrath BC, Kopp RF, et al. Insulin secretion and Ca2+ dynamics in β-cells are regulated by PERK (EIF2AK3) in concert with calcineurin. *J Biol Chem*. 2013;288(47):33824–36.
37. Biason-Lauber A, Lang-Muritano M, Vaccaro T, Schoenle EJ. Loss of kinase activity in a patient with Wolcott-Rallison syndrome caused by a novel mutation in the EIF2AK3 gene. *Diabetes*. 2002;51(7):2301–5.
38. Axelsson E, Ellegren H. Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. *Mol Biol Evol*. 2009;26(5):1073–9.
39. Gossmann TI, Santure AW, Sheldon BC, Slate J, Zeng K. Highly variable recombinational landscape modulates efficacy of natural selection in birds. *Genome Biol Evol*. 2014;6(8):2061–75.
40. Gregory MK, James MJ. Functional characterization of the duck and turkey fatty acyl elongase enzymes ELOVL5 and ELOVL2. *J Nutr*. 2014;144(8):1234–9.
41. Zhang W, Kim S, Settlage R, et al. Hypothalamic differences in expression of genes involved in monoamine synthesis and signaling pathways after insulin injection in chickens from lines selected for high and low body weight. *Neurogenetics*. 2015;16(2):133–44.
42. Takamoto I, Kubota N, Nakaya K, et al. TCF7 L2 in mouse pancreatic beta cells plays a crucial role in glucose homeostasis by regulating beta cell mass. *Diabetologia*. 2014;57(3):542–53.
43. Huang H, McIntosh AL, Martin GG, et al. Structural and functional interaction of fatty acids with human liver fatty acid-binding protein (L-FABP) T94A variant. *FEBS J*. 2014;281(9):2266–83.
44. Zhang H, Wang SZ, Wang ZP, et al. A genome-wide scan of selective sweeps in two broiler chicken lines divergently selected for abdominal fat content. *BMC Genomics*. 2012;13:704.