GigaScience

PacBio assembly with Hi-C mapping generates an improved, chromosome-level goose genome

--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00133
Full Title:	PacBio assembly with Hi-C mapping generates an improved, chromosome-level goose genome
Article Type:	Data Note
Funding Information:	National Key R & D Program of China (2018YFD0500403) — Prof. Mingzhou Li National Natural Science Foundation of China (U19A2036) — Prof. Mingzhou Li National Natural Science Foundation of China (31872335) — Not applicable National Natural Science Foundation of China (31772576) — Prof. Mingzhou Li National Natural Science Foundation of China (31802044) — Dr. Yan Li China Postdoctoral Science Foundation (2018M643514) — Dr. Yan Li
Abstract:	Background The domestic goose is an economically important and scientifically valuable waterfowl; however, a lack of high-quality genomic data has hindered research concerning its genome, genetics, and breeding. As domestic geese breeds derive from both the swan goose (Anser cygnoides) and the graylag goose (Anser anser), we selected a female Tianfu goose (A. anser × A. cygnoides) for genome sequencing. We generated a high-quality goose genome assembly by adopting a hybrid de novo assembly approach that combined PacBio single-molecule real-time sequencing, high-throughput chromatin conformation capture mapping, and Illumina short-read sequencing. Findings We generated a 1.11 Gb goose genome with contig and scaffold N50 values of 1.85 Mb and 33.12 Mb, respectively. The assembly contains 39 chromosomes (2n = 78) accounting for ca. 88.36% of the goose genome. Compared with previous goose assemblies, our assembly has more continuity, completeness, and accuracy; the annotation of core eukaryotic genes and universal single-copy orthologs has also been improved. We have identified 17,568 protein-coding genes (PCGs) and a repeat content of 8.67% (96.57 Mb) in this genome assembly. We also explored the spatial organization of chromatin and gene expression in the goose genome, in terms of inter-chromosomal interaction patterns, compartments, topologically associating domains, and promoter-enhancer interactions. Conclusions We present the first chromosome-level assembly of the goose genome. This will be a valuable resource for future genetic and genomic studies on geese.
Corresponding Author:	Mingzhou Li, Ph.D. Sichuan Agricultural University Chengdu, Sichuan CHINA
Corresponding Author Secondary Information:	

Corresponding Author's Institution:	Sichuan Agricultural University
Corresponding Author's Secondary Institution:	
First Author:	Yan Li
First Author Secondary Information:	
Order of Authors:	Yan Li
	Guangliang Gao
	Yu Lin
	Silu Hu
	Yi Luo
	Guosong Wang
	Long Jin
	Qigui Wang
	Jiwen Wang
	Qianzi Tang
	Mingzhou Li, Ph.D.
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly	Yes

encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

PacBio assembly with Hi-C mapping generates an improved, chromosome-level goose genome

Yan Li[1, †], Guangliang Gao[1, 2, †, *], Yu Lin[1, †], Silu Hu[1], Yi Luo[1], Guosong Wang[1,3], Long Jin[1], Qigui Wang[2], Jiwen Wang[1], Qianzi Tang[1], Mingzhou Li[1, *]

[1]Institute of Animal Genetics and Breeding, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China;

[2]Institute of Poultry Science, Chongqing Academy of Animal Sciences, Chongqing 402460, China;

[3]Department of Animal Science, Texas A&M University, College Station 77843, United States of America

[†] These authors contributed equally to this paper.

[*] Corresponding author(s): Guangliang Gao: guanglianggaocq@hotmail.com; Mingzhou Li: mingzhou.li@sicau.edu.cn.

Abstract

Background:

The domestic goose is an economically important and scientifically valuable waterfowl; however, a lack of high-quality genomic data has hindered research concerning its genome, genetics, and breeding. As domestic geese breeds derive from both the swan goose (*Anser cygnoides*) and the graylag goose (*Anser anser*), we selected a female Tianfu goose (*A. anser* × *A. cygnoides*) for genome sequencing. We generated a high-quality goose genome assembly by adopting a hybrid *de novo* assembly approach that combined PacBio single-molecule real-time sequencing, high-throughput chromatin conformation capture mapping, and Illumina short-read sequencing.

Findings:

We generated a 1.11 Gb goose genome with contig and scaffold N50 values of 1.85 Mb and 33.12

Mb, respectively. The assembly contains 39 chromosomes (2n = 78) accounting for ca. 88.36% of the goose genome. Compared with previous goose assemblies, our assembly has more continuity, completeness, and accuracy; the annotation of core eukaryotic genes and universal single-copy orthologs has also been improved. We have identified 17,568 protein-coding genes (PCGs) and a repeat content of 8.67% (96.57 Mb) in this genome assembly. We also explored the spatial organization of chromatin and gene expression in the goose genome, in terms of inter-chromosomal interaction patterns, compartments, topologically associating domains, and promoter-enhancer interactions.

Conclusions:

We present the first chromosome-level assembly of the goose genome. This will be a valuable resource for future genetic and genomic studies on geese.

Key Words: goose genome, chromosome-length assembly, hybrid *de novo* assembly approaches, annotation, Pacbio, Hi-C

Data description

Context

The goose is a member of the family Anatidae and is an economically important waterfowl with distinctive characters. Domesticated geese derive from the swan goose (*Anser cygnoides*) and the graylag goose (*Anser anser*)[1], and approximately 6,000 years of artificial selection have led to significant alterations in their body size, reproductive performance, egg production, feather color, and other features[2]. Currently, more than 181 domesticated breeds are reared globally to supply meat, eggs, and valuable byproducts (feathers, fatty liver) for human consumption[2,3,4]. The domestic goose is also well suited to sustainable production practices because fiber can form part of its diet, which then lessens competition for human food[5]. Its excellent disease resistance and behavioral patterns also allow for large-scale

farming and easy management[6]. Interestingly, despite the liver weight of goose increasing 5–10 times after two to three weeks of overfeeding, the amount of fat in hepatic cells (and other biomedical parameters) returns to normal levels when overfeeding ceases. This suggests that the goose liver could provide a novel animal model for the study of human non-alcoholic fatty liver disease[6].

The goose was one of the earliest animals to be domesticated[2,7], and wide-ranging genomic and breeding research has been conducted to study its domestication process and the unique morphological and physiological features of these animals. For example, recently published goose genome sequences have been assembled into scaffolds using short reads from the Illumina platform[8,9]; however, the genetic basis of the fatty liver of goose and their selective breeding remains largely unknown. To address such issues, a high-quality genome sequence is required. Currently, there are many advantages to using hybrid *de novo* assembly approaches to improve the quality of genome assemblies. This is because short, accurate reads from the Illumina platform can be combined with the longer, less accurate reads generated by the single-molecule real-time (SMRT) sequencing platform[10]. With Hi-C, linking information can then be ordered and oriented into scaffolds, after which assembly errors can be identified and corrected[11]. This approach has been applied to improve the genome assemblies of many species, including humans[12], goats[13], rockfish[14], *Aedes aegypti*[11], and barley[15].

Here, we have generated a high-quality goose assembly with chromosome-length scaffolds by adopting a hybrid *de novo* assembly approach using a combination of short reads from the Illumina platform, long reads from the PacBio platform, and Hi-C-based chromatin interaction maps. Our draft goose genome comprises fewer scaffolds than currently available goose genome assemblies, and these scaffolds are of a higher-quality and are more continuous and accurate. Our new genome assembly thus provides a valuable resource for exploring the molecular basis of the morphological and physiological

features of the goose, and will facilitate further genomic, genetic, and breeding studies of this domesticated waterfowl.

Methods

a) Sample collection and sequencing

We extracted genomic DNA from the liver tissue of a healthy adult female (136 days old) from the Tianfu goose maternal line, which was provided by the Experimental Farm of Waterfowl Breeding of Sichuan Agricultural University (Chengdu, Sichuan, China; **Figure 1**). We then carried out single-molecule real-time DNA sequencing of ca. 20-kb inserts using the PacBio Sequel platform. This yielded approximately 84.31 Gb of high-quality sequence data that were used to initially assemble the genome (**Table 1**). Next, 149.70 Gb of high-quality sequence data were generated from a 350-bp insert size Hi-C library, as previously reported[13]. Finally, 350-bp paired-end libraries constructed from the same genomic DNA were sequenced on the Illumina HiSeq platform, producing a further 181.52 Gb of sequence data. In total, we obtained approximately 415.53 Gb of high-quality sequencing data (ca. 324.63× coverage) for our draft assembly of the goose genome (**Table 1**).

b) *De novo* assembly of the goose genome

The size of the goose genome was estimated by k-mer distribution analysis to be 1.28 Gb. To assemble the genome, we first performed an initial assembly with the PacBio long-reads alone, using Falcon[16] software. We used the pbsmrtpipe pipeline of the smrtlink software to correct this assembly sequence, which resulted in a draft assembly with a contig N50 of 1.72 Mb (**Table S1**). Next, we used the single-molecule sequence reads to scaffold these contigs and fill gaps, using SSPACE-Long[17] and PBJelly[18], respectively. Pilon[19] software was then used to map the short reads to the assembly and correct sequence errors (**Table S1**). Most of these scaffolds were assembled into 39 chromosomes when the Hi-

C reads were aligned using Lachesis[20] software (**Table S2, Figure S1**); this is consistent with the number of goose chromosomes (2n = 78) reported in previous studies[21]. With these methods we generated a high-quality goose assembly with a contig N50 of 1.85 Mb and scaffold N50 of 33.12 Mb (**Table 2**). The average GC content is 42.15% and the total genome size is 1.11 Gb, which is consistent with previous studies[8,9] and suggests that our goose assembly is reliable.

c) Repeat sequence and gene annotation

De novo methods and homology-based approaches were used to annotate the repeat content of the goose genome. First, we used *ab initio*-prediction software, including LTR-finder[22], RepeatMolder[23], and RepeatScout[24], to perform *de novo* annotation of the genome. For homology-based predictions, we identified repeat regions across species in published RepBase sequences[25] using RepeatMasker[26] and RepeatProteinMask[27] software. Combined with these results, the repeat region of the goose genome was further predicted with RepeatMasker software. From these analyses, we identified 92.11 Mb of repetitive DNA (**Table S3**) accounting for 8.67% of our assembly, which is much higher than has been reported in previous studies[8,9]. Long interspersed nuclear elements (LINEs) were the most abundant repeat element identified, representing 6.83% of the genome. The proportion of LINE repetitive sequences identified in this study was also higher than has been reported in two previous goose genome assemblies (**Table S3**).

We performed PCGs annotation by combining *ab initio*-based, homology-based, and RNA-sequencing-based prediction methods. First, GenScan[28], Geneid[29], and Augustus[30] were used for *ab initio*-based predictions. Next, we selected six high-quality genomes, namely *Homo sapiens* (GCF_000001405.39), *Mus musculus* (GCF_000001635.26), *Gallus gallus* (GCF_000002315.6), *Anas platyrhynchos* (GCF_003850225.1), *Meleagris gallopavo* (GCF_000146605.3), and *Taeniopygia guttata* (GCF_003957565.1), to use for homology-based annotation of our goose draft genome using

TBLASTN[31] and GeneWise[32] software. We found 8,255 common orthologous groups across these seven species (**Figure S2**). To optimize genome annotation, total RNA was extracted from 11 samples (abdominal fat, brain, duodenum, heart, liver, lung, muscular stomach, ovary, pancreas, pectoral muscle, and spleen) taken from the same individual whose DNA was used for the draft genome assembly. We pooled equal amounts of the total RNA from each of the 11 tissues and then performed RNA-seq on this pooled sample using the Illumina platform. After filtering, these data were used to annotate protein-coding regions of the genome assembly using Trinity[33] and TopHat[34]. Finally, the predictions from each method described above were integrated using EVM[35]; overall, 17,568 PCGs were predicted (**Table 3, Figure 2**). To identify long noncoding RNAs (lncRNAs), the goose genome reads were aligned by STAR[36] and subjected to Cufflinks[37] and TACO[38] for assembly and filtering. CPC2[39] was then applied to perform coding potential analysis, and PfamScan[40] was used to check for domain hits against Pfam31-A[41]. After removing all likely domains, 3,287 lncRNAs and 542 transcripts of uncertain coding potential were identified.

Data validation and quality control

a) Assessment of genome assembly completeness

Our assembly has more scaffolds and fewer contigs, and significantly improved contig and scaffold N50 values, than the goose genome assemblies presented in two previous studies (**Figure 3**). Moreover, we have annotated more repeat and coding sequence regions than these previous studies (**Table 3**), which suggests that we have generated an improved genome assembly and annotation. The 39 chromosomes described in our study account for 88.36% of the assembled genome and are longer than those previously reported[8,9], again indicating that our draft goose genome represents a significant improvement on previous work. The GC content of our genome assembly is 42% and the size of the genome is 1.11 Gb

(**Table 2**). This is comparable to the sizes reported for the two previously constructed goose genomes[8,9] and is characteristic of avian genomes[42]. We also mapped short-insert paired-end reads (350 bp) to our draft goose genome and obtained mapping and coverage rates of 97.25% and 99.71%, respectively. Finally, we downloaded 19 wild goose resequencing[43] datasets from public databases and mapped them to our assembly, and to the two earlier draft goose genomes. We found that the mapping rate of our assembly was higher than that of the previously assembled genomes (**Table S4**), indicating that it is more contiguous. Taken together, these results demonstrate the improvements made by our study in the assembly and annotation of the goose genome, in comparison to previous studies[8,9].

To evaluate the completeness of our draft genome, we determined the number of conserved eukaryotic and universal genes present in our assembly by applying the core eukaryotic genes mapping approach software (CEGMA) and using a set of benchmarking universal single-copy orthologs (BUSCO). We found that 211 of the 248 (85.08%) core eukaryotic genes and 2,586 (97%) of the universal single-copy orthologs were assembled in our genome. Compared with previous studies, this suggests that our genome assembly is more complete than previous drafts of the goose genome[8,9].

To explore the hypothesis that the leptin gene was lost from goose[8], we downloaded leptin sequences from avian and mammal genomes to use as reference sequences in BLASTP searches of our newly assembled goose genome. We found no sequences similar to leptin in our draft assembly. Furthermore, although the human genome region that contains the leptin gene (chromosome 7, 126.0 to 129.4 Mb) aligned with the goose genome, we did not find a sequence similar to the leptin gene in this region. These results confirm the previous finding that the leptin gene is not present in the goose genome[8].

b) Phylogenetic tree and lineage-specific gene families

Using OrthoMCL[44], 16,157 orthologous gene families across 17 species (ostrich, duck, goose,

chicken, turkey, saker, red-legged seriema, African crowned crane, pelican, little egret, crested ibis, cormorant, great crested grebe, pigeon, woodpecker, zebra finch, and lizard) were identified. Based on 2,389 shared single-copy ortholog gene clusters, we constructed a maximum likelihood phylogenetic tree using the RAxML software[45]. This revealed that goose and duck diverged about 31.60 million years ago (Mya), which is comparable to the divergence time of chicken and turkey (32.33 Mya; **Figure S3**). We also noted that lineage-specific genes in the goose genome were significantly enriched for olfactory receptor activity (GO:0004984, $p = 3.85\times10^{-24}$), G protein-coupled receptor activity (GO:0004930, $p = 6.67\times10^{-13}$), and integral component of membrane (GO:0016021, $p = 0.01$; **Table S5**). As a migratory bird, the goose is adapted for long-distance migration, which exposes them to a diversity of food as they seek out ideal habitats. We propose that such influences could strengthen the interactions between odorants and the receptors of the olfactory mucosa, and could underlie receptor family evolution in the goose genome.

c) Expansion and contraction of gene families

The expansions and contractions of gene clusters in the goose genome were identified in comparison to nine other avian genomes using the CAFE program[46]. We found 839 expanded gene families (**Table S6**) and 2,193 contracted gene families (**Table S7**). Interestingly, the expanded gene families were mainly enriched for olfactory receptor activity (GO:0004984, $p = 8.58\times10^{-51}$), G protein-coupled receptor activity (GO:0004930, $p = 5.81\times10^{-25}$), and integral component of membrane (GO:0016021, $p = 3.20\times10^{-6}$), which is consistent with the results from our analysis of lineage-specific genes (**Table S5**). This further confirms that the migratory adaptations of the goose are reflected by unique characteristics in the goose genome that contrast with those of nonmigratory birds. Other expanded gene families were enriched for ATPase-coupled transmembrane transporter activity (GO:0042626, $p = 1.96\times10^{-06}$),

NAD(P)+-protein-arginine ADP-ribosyl transferase activity (GO:0003956, $p = 3.20\times10^{-04}$), ATPase activity (GO:0016887, $p = 8.28\times10^{-05}$), and aspartic-type endopeptidase activity (GO:0004190, $p = 9.63\times10^{-06}$; **Table S6**), while gene families contracted in the goose were significantly enriched for transmembrane transport (GO:0055085, $p = 8.30\times10^{-04}$), ion channel activity (GO:0005216, $p = 1.87\times10^{-9}$), ion transmembrane transport (GO:0034220, $p = 5.30\times10^{-6}$), and ATPase-coupled intramembrane lipid transporter activity (GO:0140326, $p = 8.60\times10^{-10}$; **Table S7**). As these pathways are related to ATP utilization, ATP production, and energy regulation, these data support a previous finding that goose energy metabolism is different to that in other avian species[47]. This feature of the goose is likely related to its migratory habits and artificial selection—the goose is unique among migratory birds because of its large body size, which requires much energy for long-distance, high altitude flying[48].

d) Genes under positive selection

We identified 52 positively selected genes (PSGs) in the goose genome based on orthologous genes from the 17 species above, using a branch-site model and F3x4 codon frequencies in Codeml (**Table S8**). Some of these PSGs, such as GCH1[49], MDH2[50], and OGFOD2[51] are involved in hypobaric hypoxia and hypoxic sensing. The viral transcription-related genes RPL7A[52], SNW1[53], and POU2F3[54] are also under positive selection in the goose, indicating that disease resistance may help the goose adapt to high altitude migration, and to harsh, changeable environments[55, 56].

e) Initial characterization of the three-dimensional organization of goose genome

We analyzed the inter-chromosomal interaction pattern[57], compartments[58, 59], topologically associating domains (TADs)[60], and promoter-enhancer interactions (PEI)[61] of the goose genome. The matrix resolution of our Hi-C experiment reached ~2 Kb (defined as the smallest locus size such that 80%

of loci have at least 1,000 contacts), which was adequate for subsequent analyses of the chromatin architecture. Our results showed that the whole inter-chromosomal interaction pattern was distinguished by two clusters, that is, short chromosomes and longer chromosomes, which suggests that goose chromosomes tend to interact with one another on the basis of size (**Figure 4**). As for the identification of A and B compartments, which represent relatively active and inactive chromatin states, respectively, we found that the number of transcriptional start sites (TSSs) in each 100 Kb bin was significantly correlated with PC1 values (R = 0.39, $p = 2.2\times10^{-16}$; **Figure S5**), and that the transcripts per kilobase millions (TPMs) of PCGs located in A compartments were significantly higher than those in B compartments ($p = 2.2\times10^{-16}$; **Figure S6, Table S9**). We identified 734 TADs across the goose assembly, accounting for 80% of the genome (**Figure S7, Table S10**). The mean and median sizes of the TADs were 1.21 Mb and 1.00 Mb, respectively. We also observed that the TSSs of PCGs were enriched in TAD-boundary regions (**Figure S8**). After filtering for interaction distances lower than 20 Kb, we identified 13,017 PEIs (**Table S11**) and found that gene expression levels positively correlated with the number of PEIs (**Figure S9**). This is suggestive of additive effects of enhancers on target-gene transcription levels.

Availability of supporting data

The goose assembled draft genome sequence is available at National Center for Biotechnology Information (NCBI) GenBank through the accession number WTSS00000000; The high-quality Hi-C data are available through the NCBI Sequence Read Archive (SRA) database under accession number SRR10483522. The high-quality PacBio long-read sequencing data have been deposited in the NCBI SRA (SRR10483521). The high-quality Illumina short-read sequencing data are available through NCBI SRA accession number: SRR10483516, SRR10483517, SRR10483518 and SRR10483520. The

transcriptome data are available through the NCBI SRR10483519.

List of abbreviations

(1) Anser anser：A. anser;

(2) Anser cygnoides：A. cygnoides;

(3) BUSCO: Benchmarking Universal Single-Copy Orthologs;

(4) CHMP1B: charged multivesicular body protein 1B;

(5) CEGMA: Core Eukaryotic Genes Mapping Approach software;

(6) GCH1: GTP cyclohydrolase 1;

(7) Hi-C, Chromosome conformation capture;

(8) IVNS1ABP: influenza virus NS1A binding protein;

(9) LINEs: Long interspersed nuclear elements;

(10) LncRNAs: long noncoding RNAs;

(11) OGFOD2: 2-oxoglutarate and iron dependent oxygenase domain containing 2

(12) MDH257: malate dehydrogenase 2

(13) PCGs: protein coding genes

(14) PEI: promoter-enhancer interactions;

(15) PSGs: positively selected genes;

(16) SMRT: single-molecule real-time;

(17) TADs: topological associated domains;

(18) TPMs: transcripts per kilobase millions.

Ethics approval

All animal experiments were approved and reviewed by Animal Care and Use Committee Institutional of Sichuan Agricultural University (Approval No. DKY-B20121406) and the Ministry of Science and Technology of the People's Republic of China (Approval No. 2006–398).

Competing interests

The authors declare no competing interest.

Acknowledgments

This work was supported by grants from the National Key R & D Program of China (2018YFD0500403), the National Natural Science Foundation of China (U19A2036, 31872335, 31772576 and 31802044) and the China Postdoctoral Science Foundation (2018M643514).

Author contributions

Mingzhou Li, Guangliang Gao designed and supervised the project. Yan Li, Yu Lin, Qianzi Tang, Silu Hu performed bioinformatics analyses. Jiwen Wang, Yan Li and Yi Luo contributed to collect the samples. Mingzhou Li, Qigui Wang, Guangliang Gao, Yi Luo and Long Jin were involved in the data analyses and wrote the manuscript.

References

1. Shi XW, Wang JW, Zeng FT, et al. Mitochondrial DNA cleavage patterns distinguish independent origin of Chinese domestic geese and western domestic geese. Biochem Genet. 2006; 44(5-6) : 237-245.
2. Kozák J. Variations of geese under domestication. World's Poult Sci J. 2019; 75(2): 247-260.
3. Goluch-Koniuszy Z, Haraf G. Geese for slaughter and wild geese as a source of selected mineral elements in a diet. J Elementol. 2018; 23: 1343-1360.
4. Janan J, Tóth P, Hutas I, et al. Effects of dietary micronutrient supplementation on the reproductive traits of laying geese. Acta Fytotech Zootech. 2015; 18(1) : 6-9.
5. Zhang Y, Sha Z, Guan F, et al. Impacts of geese on weed communities in corn production systems

and associated economic benefits. Biol Control. 2016. 99: 47-52.

6. Wang G, Jin L, Li Y, et al. Transcriptomic analysis between Normal and high-intake feeding geese provides insight into adipose deposition and susceptibility to fatty liver in migratory birds. BMC genomics. 2019; 20(1): 372.

7. Honka J, Heino M, Kvist L, et al. Over a thousand years of evolutionary history of domestic geese from Russian archaeological sites, analysed using ancient DNA. Genes. 2018; 9(7): 367.

8. Lu L, Chen, Y, Wang Z, et al. The goose genome sequence leads to insights into the evolution of waterfowl and susceptibility to fatty liver. Genome Biol. 2015; 16(1): 89.

9. Gao G, Zhao X, Li Q, et al. Genome and metagenome analyses reveal adaptive evolution of the host and interaction with the gut microbiota in the goose. Sci Rep. 2016; 6: 32961.

10. Schadt E, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010; 19(R2): R227-R240.

11. Dudchenko O, Batra SS, Omer AD, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017; 356(6333): 92-95.

12. Pendleton M, Sebra R, Pang AWC, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015; 12(8): 780–786.

13. Bickhart DM, Rosen BD, Koren S, et al. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nat Genet. 2017; 49(4): 643.

14. Liu Q, Wang X, Xiao Y, et al. Sequencing of the black rockfish chromosomal genome provides insight into sperm storage in the female ovary. DNA Research, 2019. 26(6):453–464,

15. Mascher M, Gundlach H, Himmelbach A, et al. A chromosome conformation capture ordered

sequence of the barley genome. Nature. 2017; 544(7651): 427

16. Chin CS, Peluso P, Sedlazeck FJ et al. Phased diploid genome assembly with single molecule real-time sequencing. Nat Methods. 2016;13:1050.

17. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinf. 2014; 15(1): 211.

18. English AC, Richards S, Han Y, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012; 7(11): e47768.

19. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. Plos One. 2014; 9(11): e112963.

20. Burton JN, Adey A, Patwardhan RP, et al. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol. 2013; 31(12): 1119–1125.

21. Jun X, Tianxing L, Qing C, et al. Karyotypes of Zhedong White Goose and Siji Goose. China Poultry. 2007; 21(9): 27-29.

22. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27(2): 573–580.

23. RepeatMolder software. http://www.repeatmasker.org/RepeatModeler/.

24. Price AL , Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics2005;21(suppl 1):i351–8.

25. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21:i351–8. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015; 6(1):11.

26. Maja TG, Nansheng C. Using RepeatMasker to identify repetitive elements in genomic sequences.

Curr Protoc Bioinf. 2009; 25(1): 4.10.11–14.10.14.

27. Allred DB, Cheng A, Sarikaya M, et al. . Three-dimensional architecture of inorganic nanoarrays electrodeposited through a surface-layer protein mask. Nano Lett. 2008;8(5):1434–8.

28. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997; 268(1): 78–94.

29. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Curr Protoc Bioinf. 2007; 18(1): 4.3.1-4.3.28.

30. Stanke M, Steinkamp R, Waack S. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004; 32(suppl_2): W309–W312.

31. Gertz EM, Yu YK. Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 2006; 4(1): 41.

32. Birney E, Clamp M, Durbin R. Gene Wise and Genomewise. Genome Res. 2004; 14(5): 988–995.

33. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29(7): 644–652.

34. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9): 1105–1111.

35. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008; 9(1): R7.

36. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1): 15–21.

37. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq

reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28(5): 511–515.

38. Niknafs YS, Pandian B, Iyer HK, et al. TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat Methods. 2017; 14(1): 68.

39. Kang YJ, Yang DC, Kong L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017; 45(W1): W12-W16.

40. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42(D1): D222–D230.

41. Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. Nucleic Acids Res. 2004; 32 (suppl_1): D138–D141.

42. Zhang G, Li C, Li Q, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014; 346(6215): 1311-1320.

43. Ottenburghs J, Megens HJ, Kraus RH, et al. A history of hybrids? Genomic patterns of introgression in the True Geese. BMC Evol Biol. 2017; 17(1): 201.

44. Fischer S, Brunk BP, Chen F, et al. Using OrthoMCL to assign proteins to OrthoMCL- DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinf. 2011; 35(1): 6-12.

45. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9): 1312–1313.

46. Bie T, Cristianini N, Demuth J. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006; 22(10): 1269-1271.

47. Józefiak DA, Rutkowski A, Martin SA. Carbohydrate fermentation in the avian ceca: a review. Anim Feed Sci Technol. 2004; 113(1-4): 1-15.

48. Watanabe, YY. Flight mode affects allometry of migration range in birds. Ecol Lett. 2016; 19(8): 907-914.

49. He Y B, Duojizhuoma C Y, Cai-juan D B, et al. GCH1 plays a role in the high-altitude adaptation of Tibetans. Zool Res. 2017; 38(3): 155–162.

50. Lu H, Wang R, Li W, et al. Plasma proteomic study of acute mountain sickness susceptible and resistant individuals. Sci Rep. 2018; 8(1): 1-9.

51. McDonough M A, Loenarz C, Chowdhury R, et al. Structural studies on human 2-oxoglutarate dependent oxygenases. Curr Opin Struct Biol. 2010; 20(6): 659-672.

52. Brahma R, Gurumayum S, Naorem L D, et al. Identification of hub genes and pathways in Zika virus infection using RNA-seq data: a network-based computational approach. Viral Immunol. 2018; 31(4): 321-332.

53. Verma S, De Jesus P, Chanda S K, et al. SNW1, a novel transcriptional regulator of the NF-κB pathway. Mol Cell Biol. 2019; 39(3): e00415-18.

54. Huang Y H, Klingbeil O, He X Y, et al. POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. Gene Dev. 2018; 32(13-14): 915-928.

55. Jones E, Fugger L, Strominger J. MHC class II proteins and disease: a structural perspective. Nat Rev Immunol. 2006; 6(4): 271.

56. Cardona C, Xing Z, Sandrock C. Avian influenza in birds and mammals. Comp Immunol. 2009; 32(4): 255-273.

57. Battulin N, Fishman VS, Mazur AM, et al. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. Genome Biol. 2016; 17(1): 6.

58. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range

interactions reveals folding principles of the human genome. Science. 2009; 326(5950): 289-293.

59. Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Mol Cell. 2017; 67(5): 837-852.

60. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7389): 376-380.

61. Ron G, Globerson Y, Moran D and Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun. 2017; 8(1): 2237.

Table1 Summary of sequencing data for goose genome assembly.

Pair-end libraries	Insert size (bp)	Total data (Gb)	Read length (bp)	Sequence coverage (×)
Illumina reads	350	181.52	150	141.81
Pacbio reads	20,000	84.31		65.86
Hi-C	350	149.70	150	116.95
Total		415.53		324.63

Table2 Comparison of quality metrics of this study and the previous goose genome assemblies.

Genomic features	This study	Lu *et al.*[a]	Gao *et al.*[b]
Estimate of genome size (bp)	1,277,099,016	1,208,661,181	1,198,802,839
Total length of assembled contigs (bp)	1,113,842,245	1,086,838,604	1,100,859,441
Total size of assembled scaffolds (bp)	1,113,913,845	1,122,178,121	1,130,663,797
Number of contigs (>2kb)	2,771	60,979	53,336
Number of scaffolds （>2kb）	2,055	1,050	1,837
Contigs N50 (bp)	1,849,874	27,602	35,032
Scaffolds N50 (bp)	33,116,532	5,202,740	5,103,766
Longest contig (bp)	10,766,871	201,281	399,111
Longest scaffold (bp)	70,896,740	24,051,356	20,207,557
GC content (%)	42.15	38.00	41.68
Number of gene model	17,568	16,150	16,288
Repeats share in genome (%)	8.67	6.33	6.90

[a] From the ref. 8. [b] From the ref. 9.

Table 3 A comparative summary of predicted genes within each goose genome assembly.

Property	This study	Lu *et al.*[a]	Gao *et al.*[b]
Total genes length (bp)	326,863,440	439,289,059	500,923,091
Genes percentage of genome (%)	29.34	39.25	44.31
Total exons number	152,392	158,713	167,532
Average exons per gene	8.67	10.92	10.29
Total exons length (bp)	26,883,354	25,763,242	26,157,477
Exons percentage of genome (%)	2.41	2.31	2.31
Average exons length (bp)	176.41	162.33	156.13
Average introns length (bp)	2224.97	2867.48	3139.07

[a] From the ref. 8. [b] From the ref. 9.

Figure 1 A picture of a female adult goose used for genome sequencing

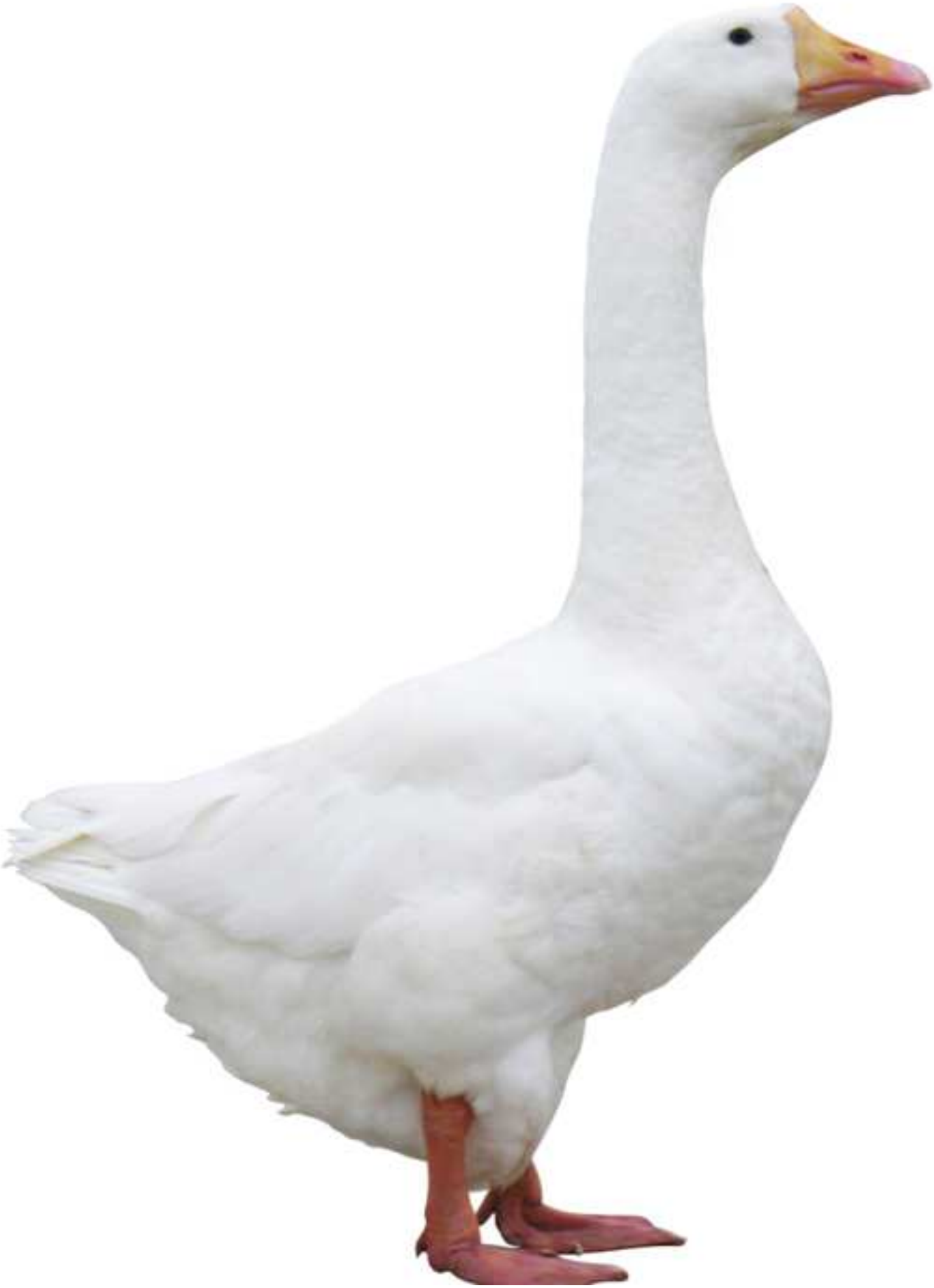

Figure 2 The gene density distribution in the goose genome

Click here to download Figure Figure 2.tif

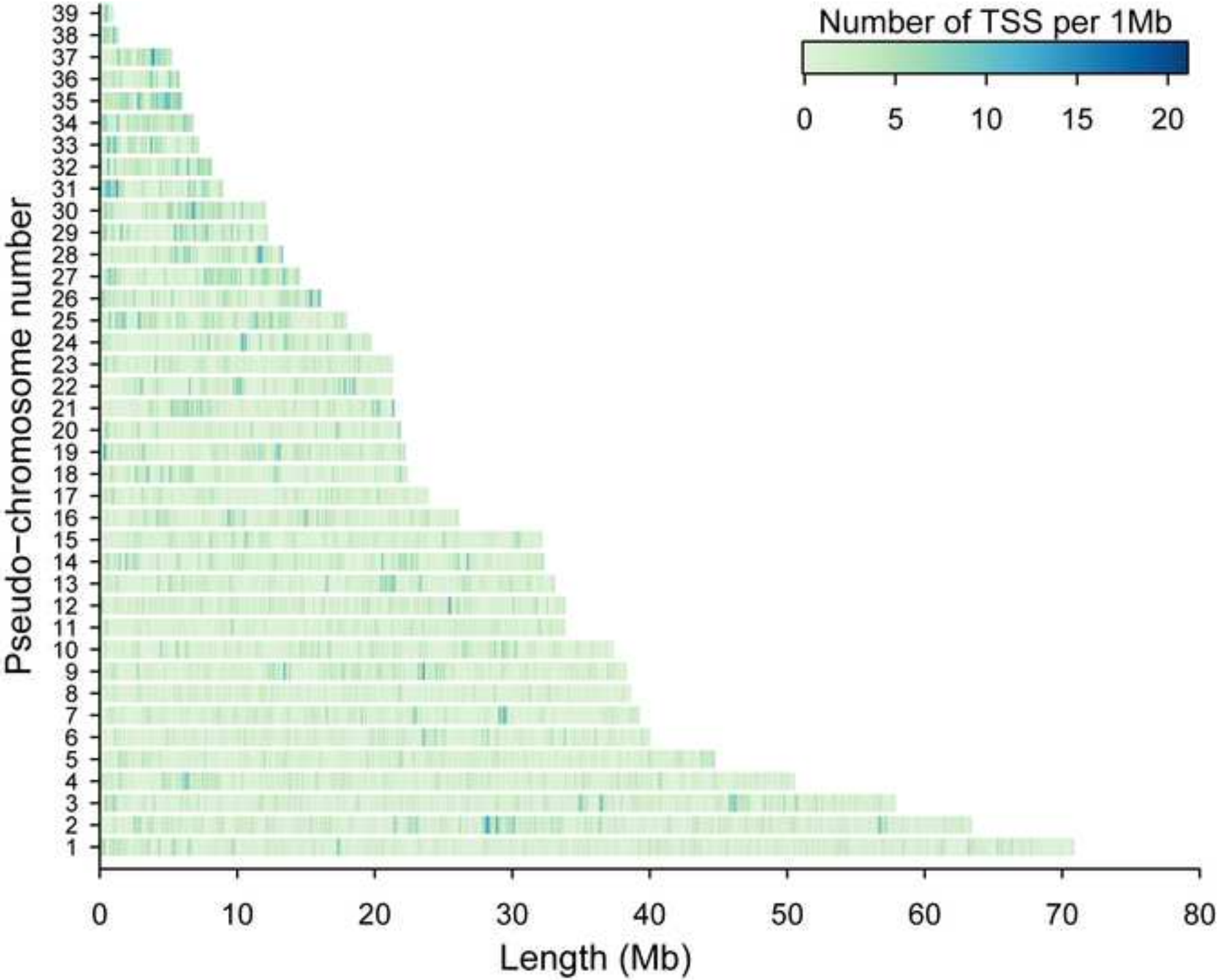

Figure 3 Comparison of the distribution and coverage of the scaffolds for the assembly with previous goose genome assemblies.

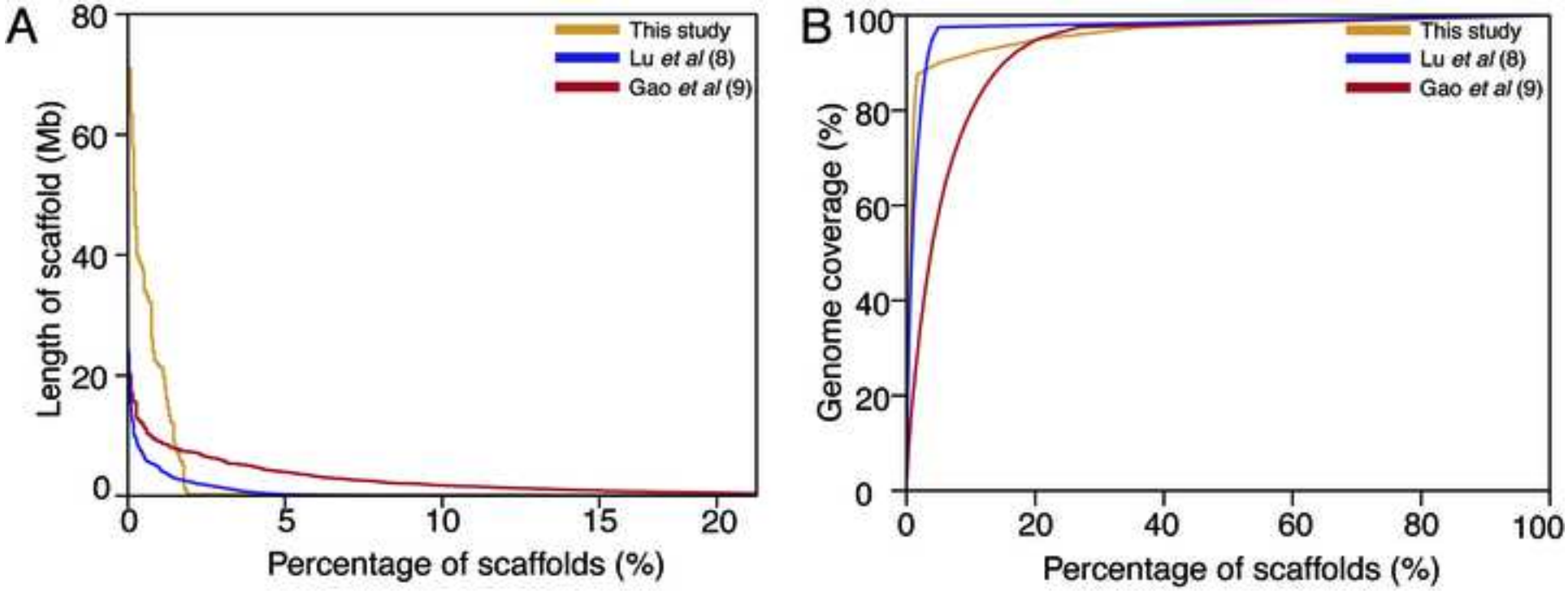

Figure 4 The dendrogram of inter-pseudo-chromosome interaction pattern generated by average linkage

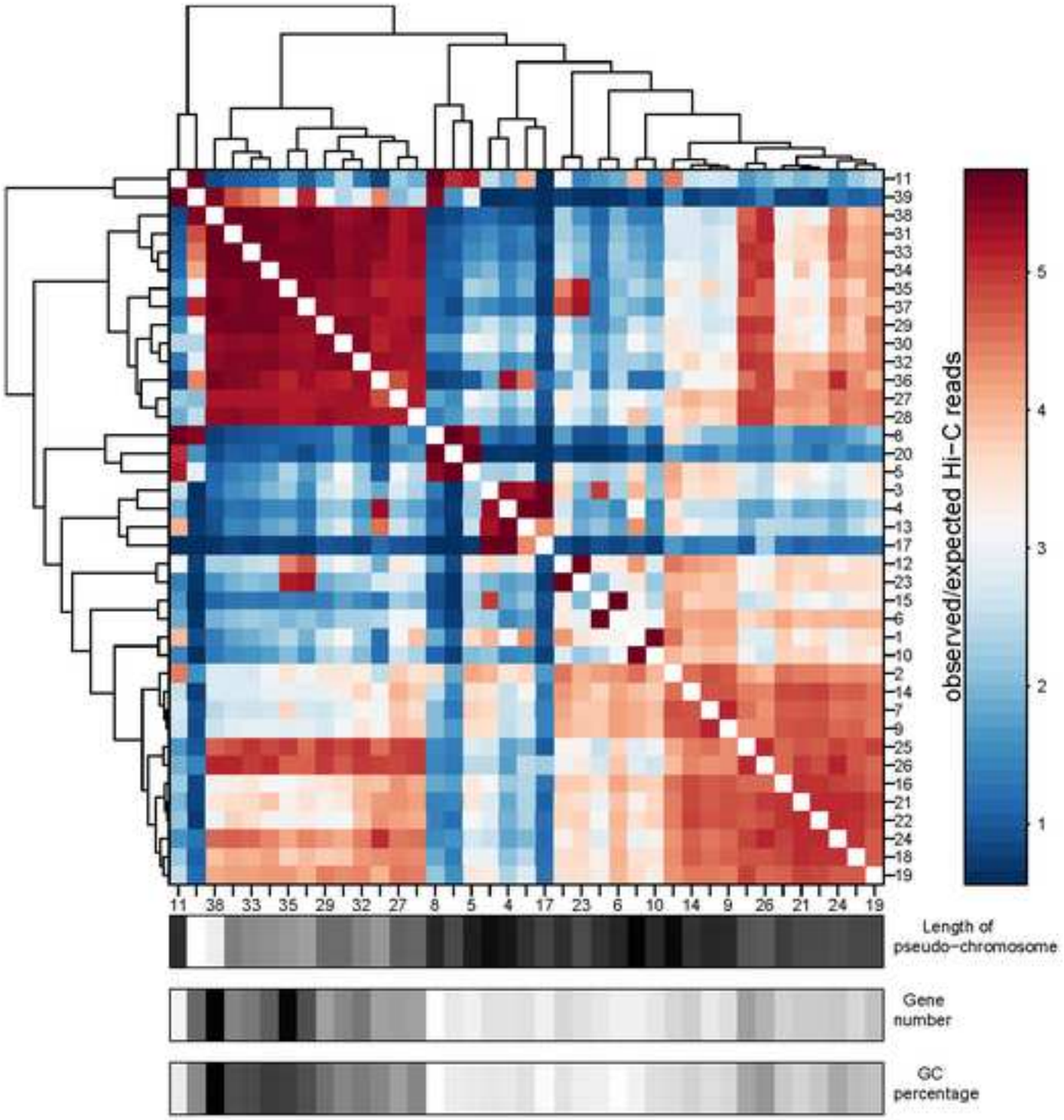

Figure S1 The Hi-C interaction contact heatmap of goose pseudochromosome genome assembly (bin size is 100kb).

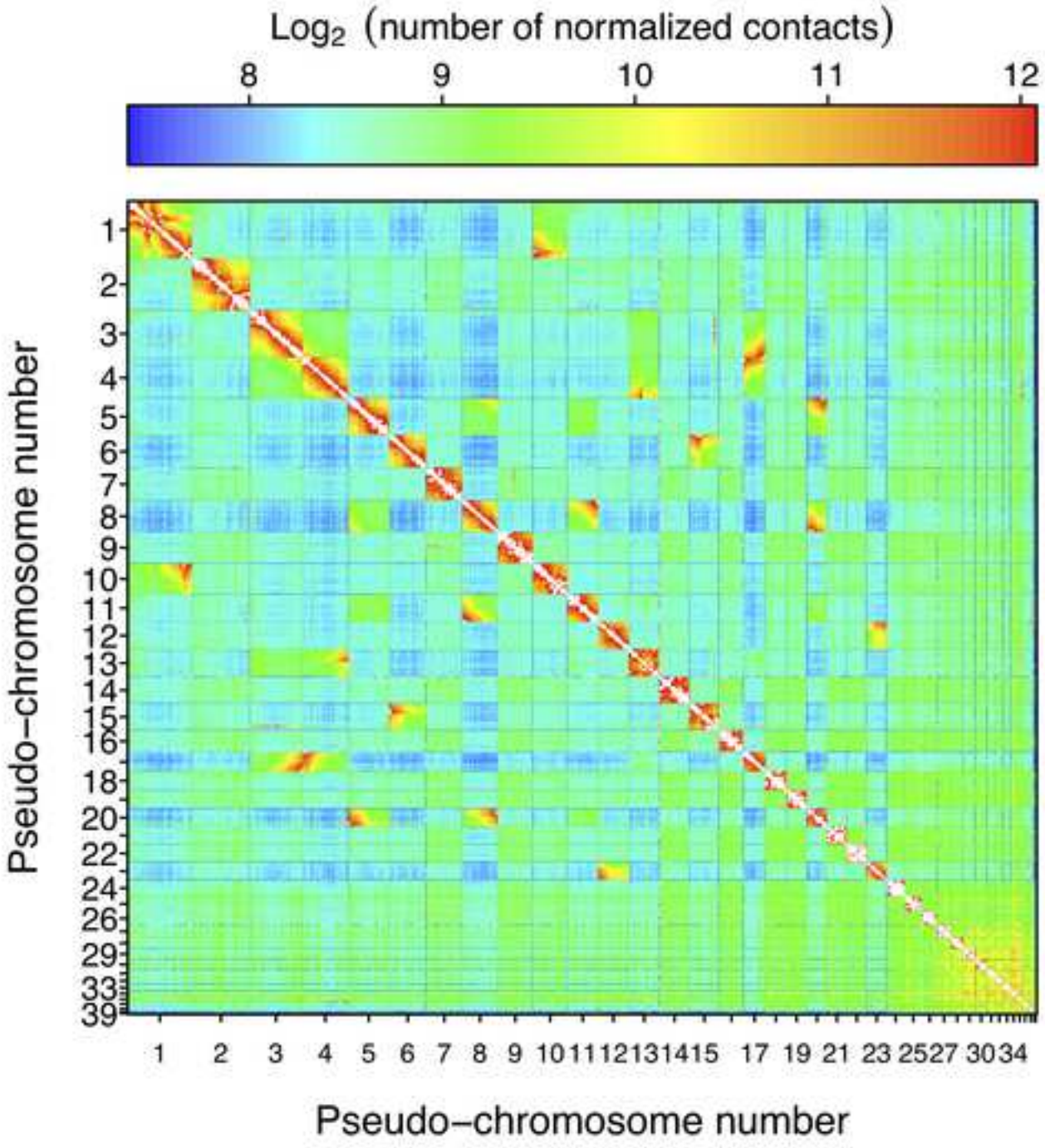

Figure S2 The shared homologous gene families across the six species

Click here to download Figure Figure S2.tiff

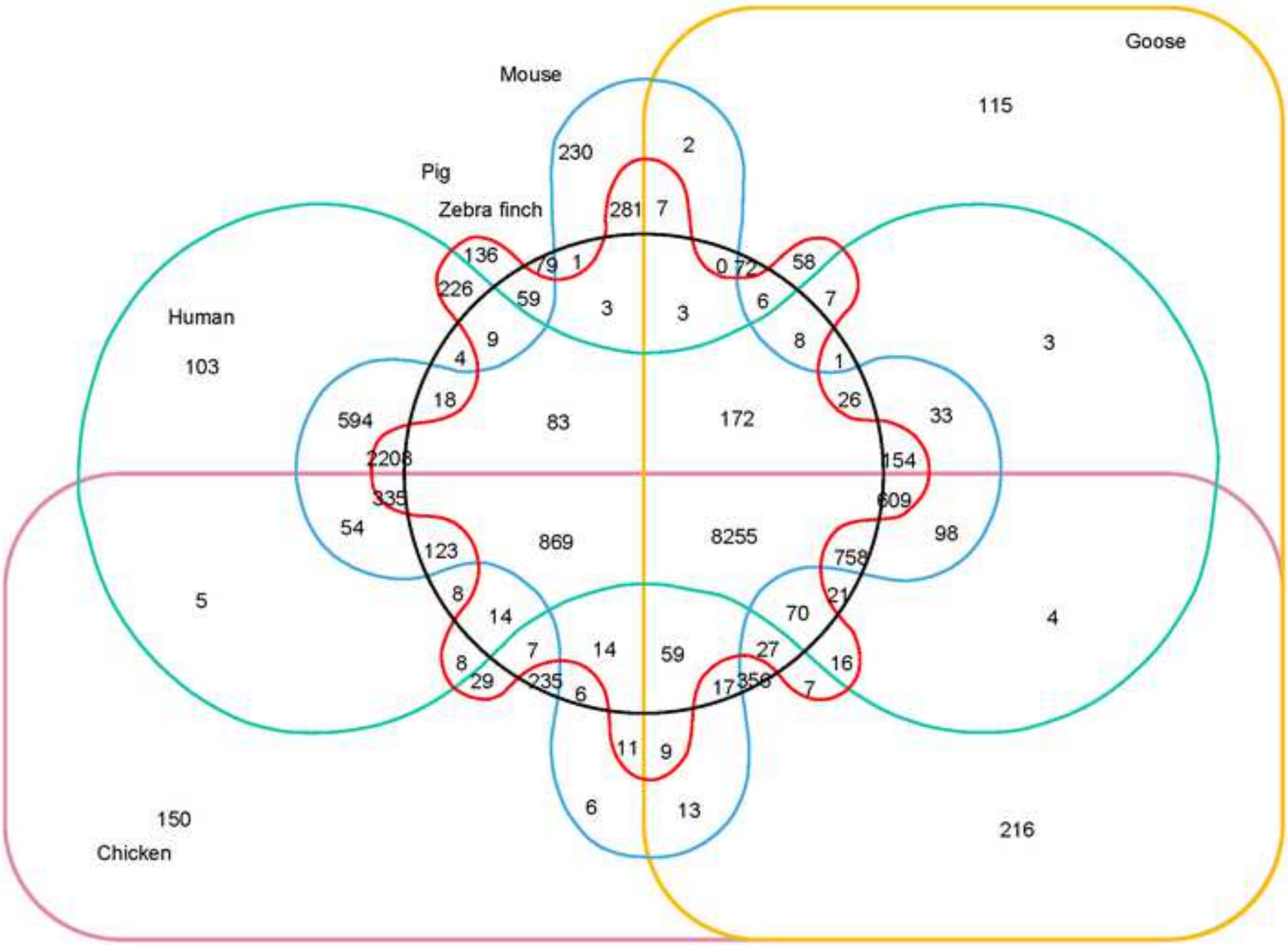

Figure S3 Divergence of time and the expansion, contraction gene families in the seventeen species

Click here to download Figure Figure S3.jpg

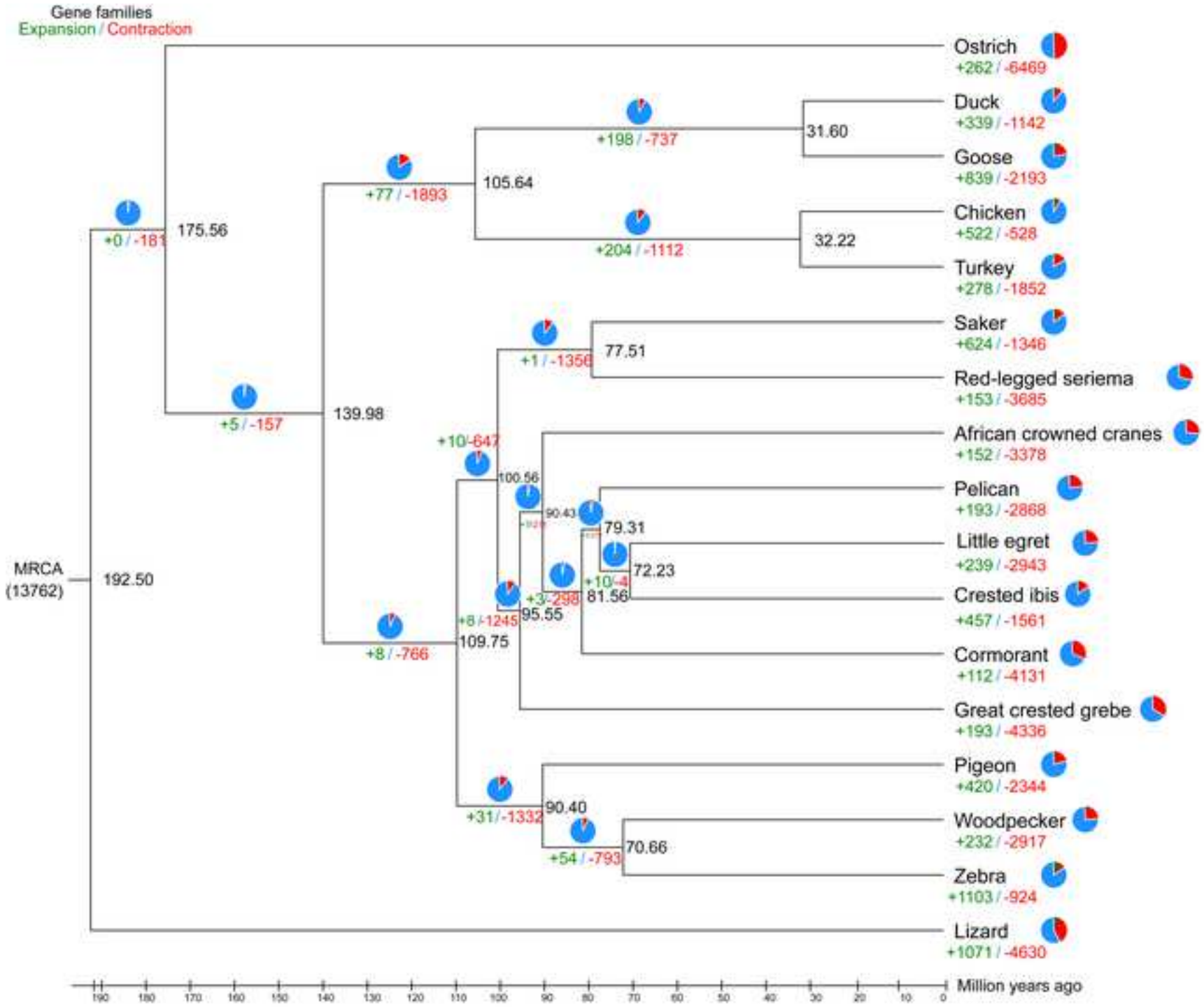

Figure S4 Resolution evaluation of Hi-C matrix.

Click here to download Figure Figure S4.tif

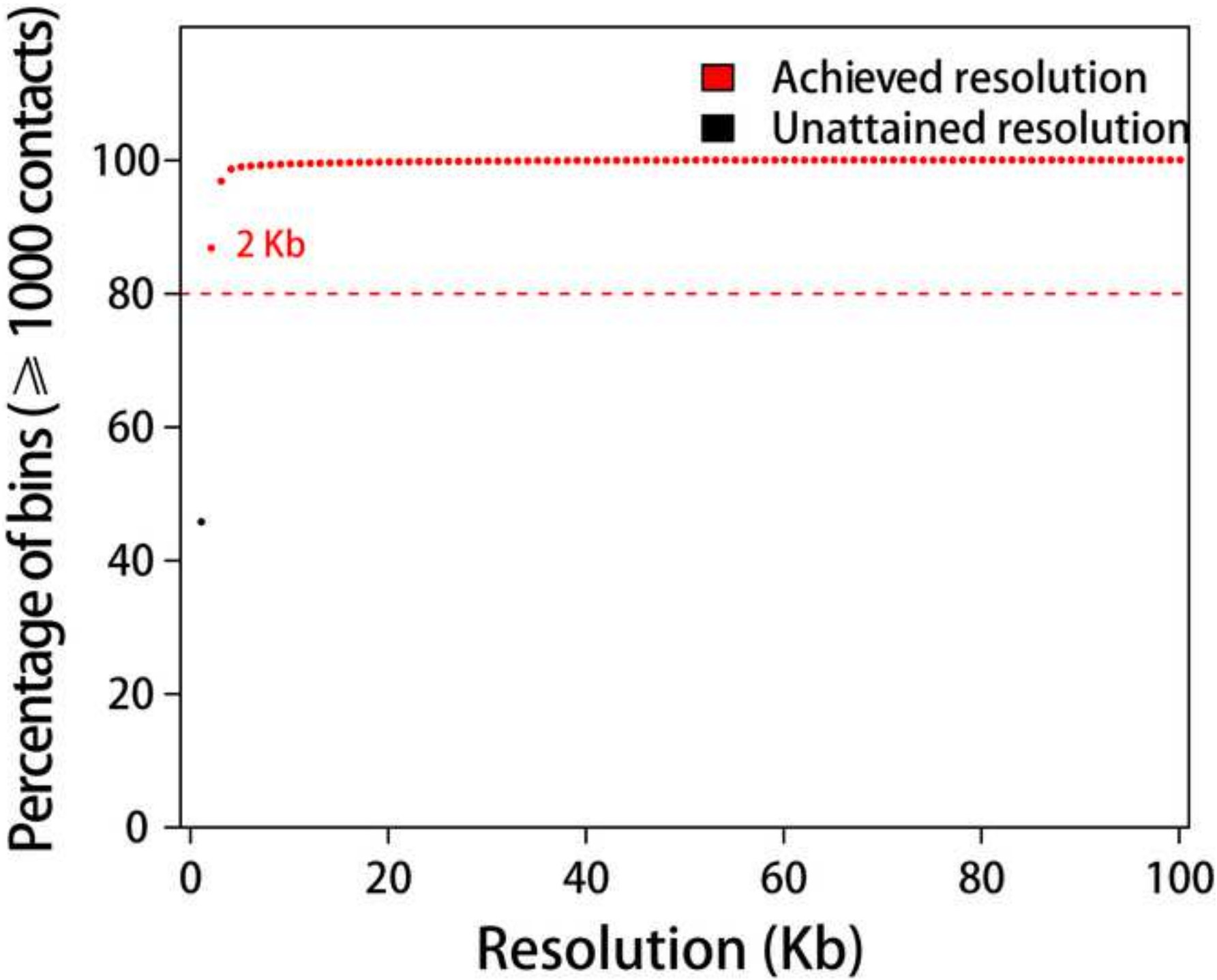

Figure S5 Vioplot of PC1 values in 100 Kb bins with various number of TSSs.

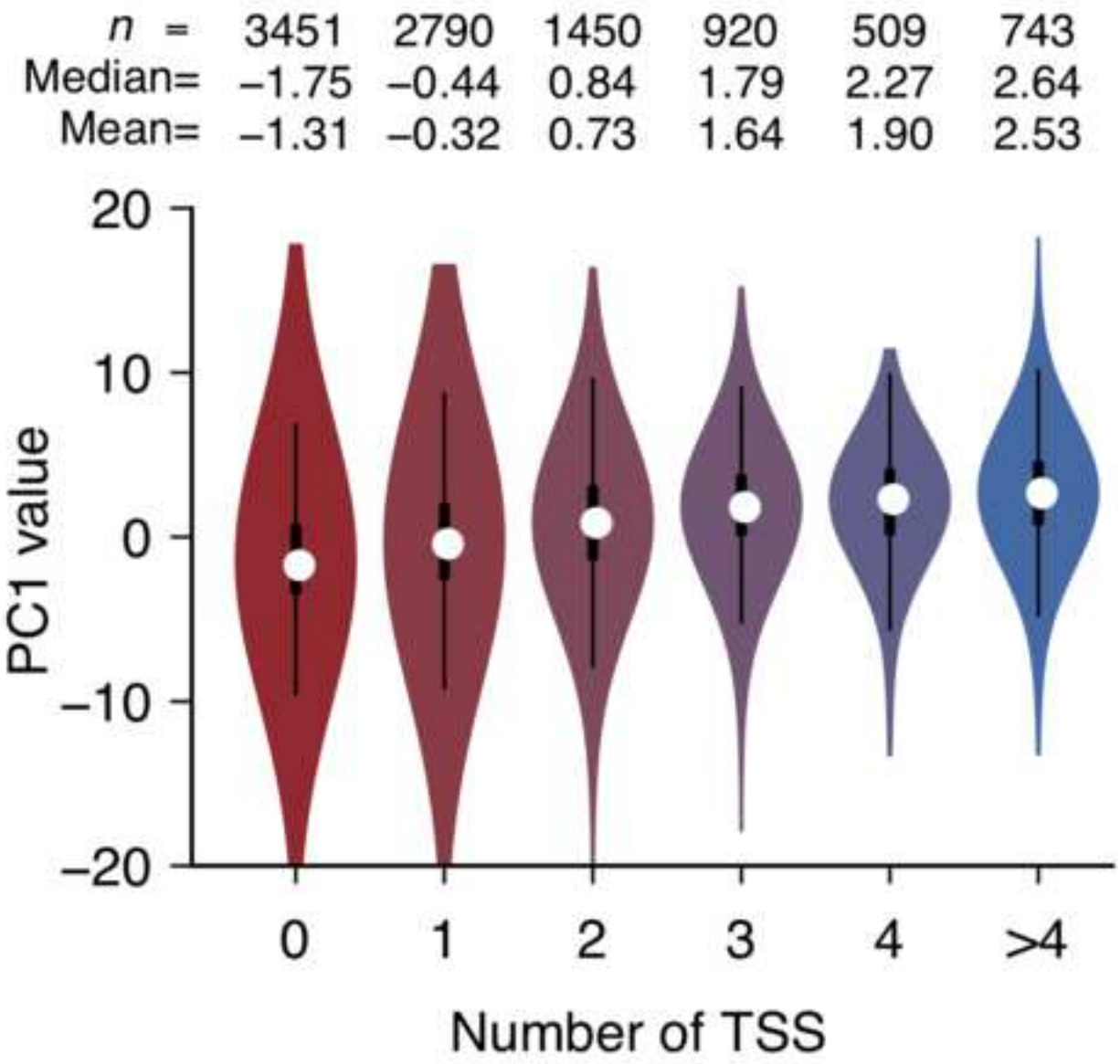

Figure S6 The Genes in compartment A expressed significantly higher than that in compartment B.

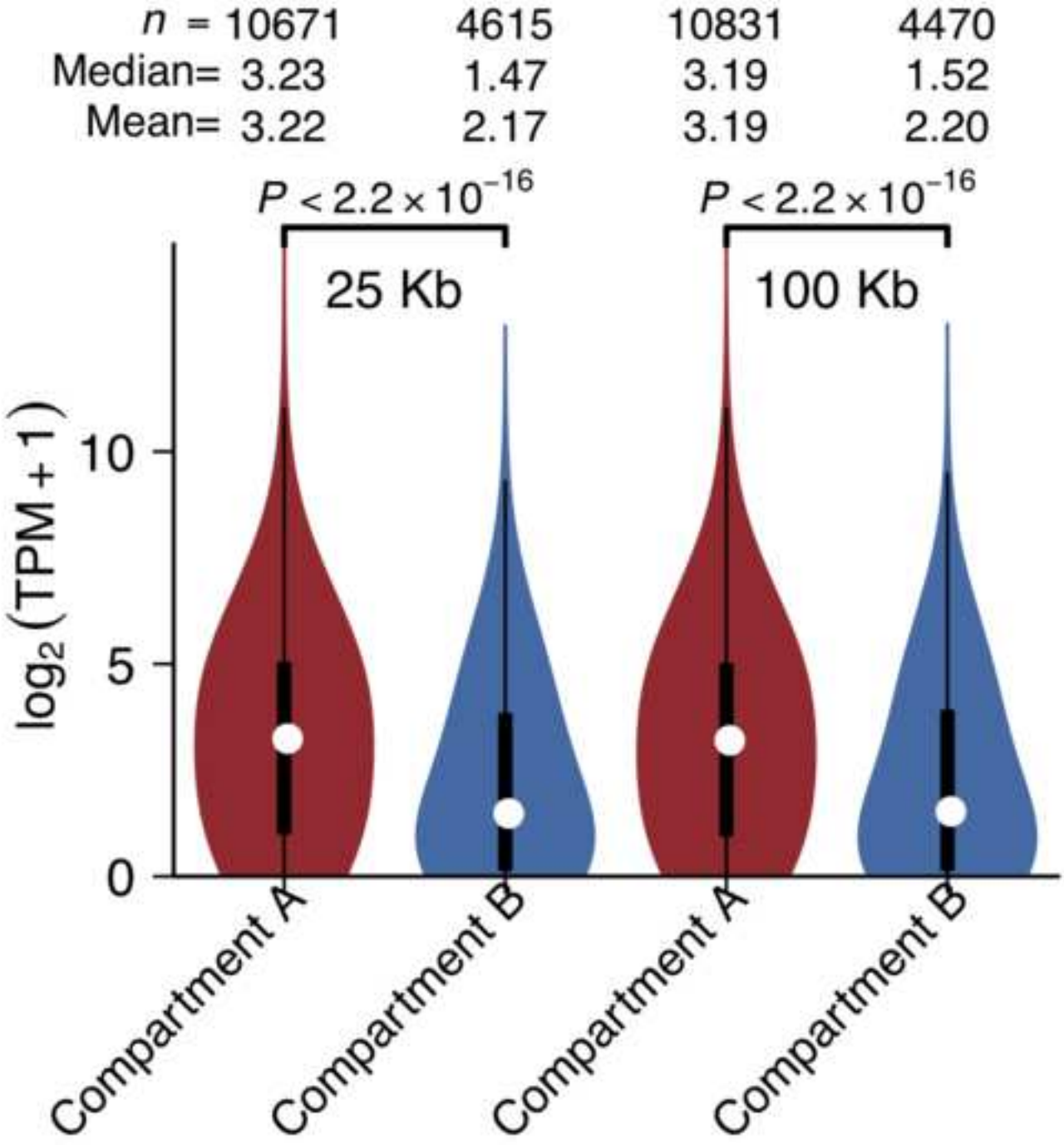

Figure S7 TAD distribution across the goose genome assembly.

Click here to download Figure Figure S7.tif

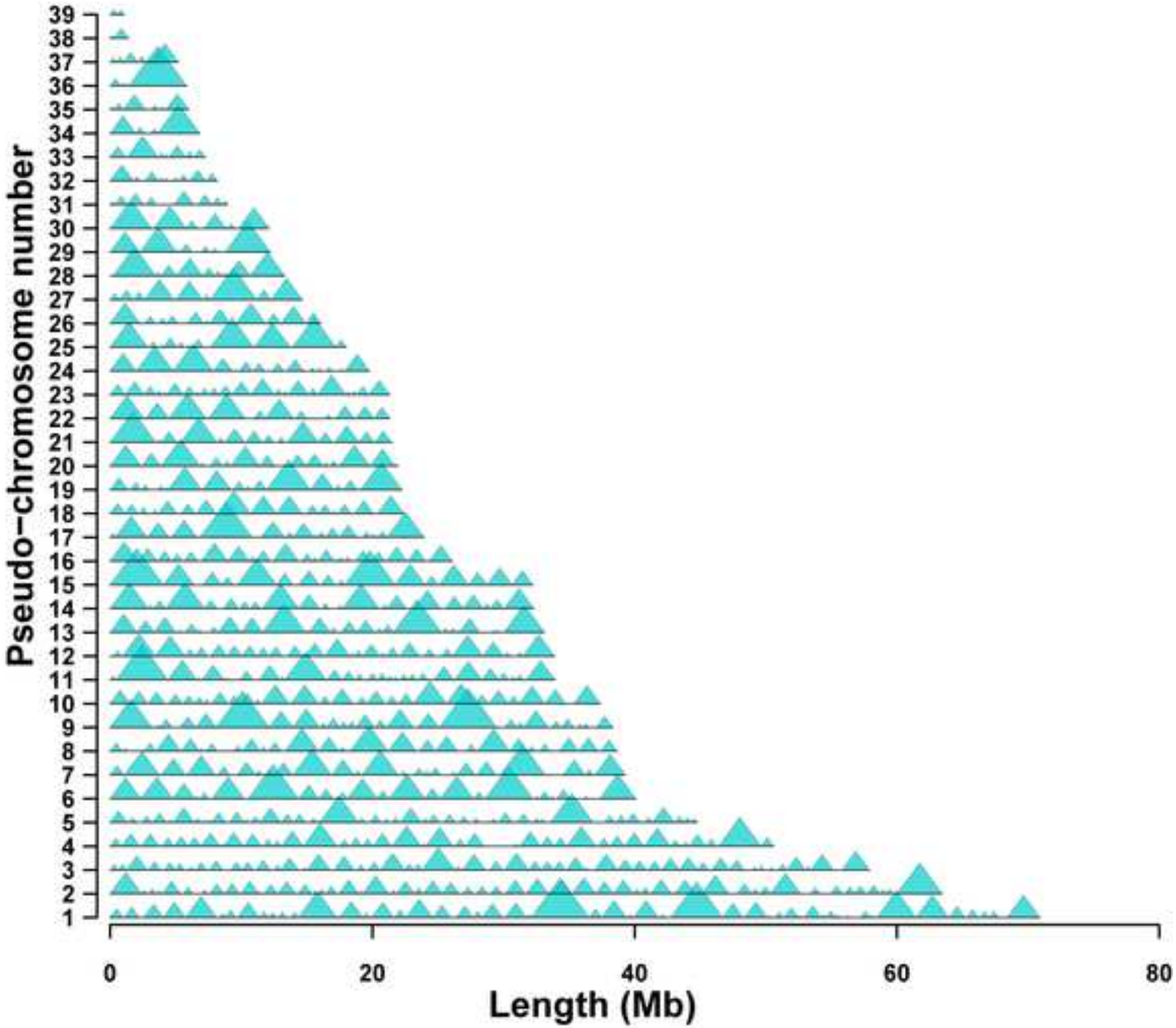

Figure S8 TSSs of genes were enriched in TAD boundary regions.

Click here to download Figure Figure S8.tif

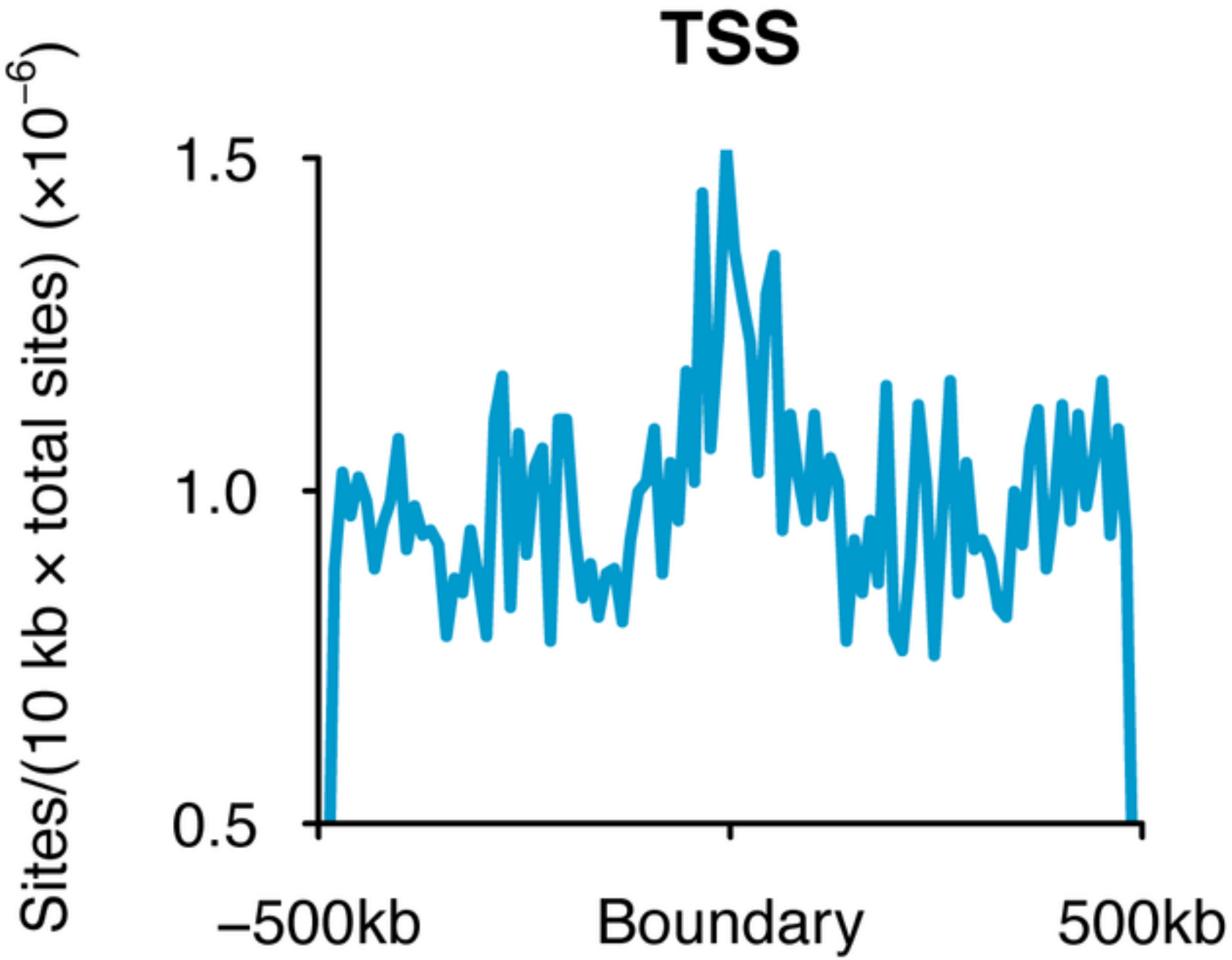

Figure S9 The additive effects of enhancers on target-gene expression levels.

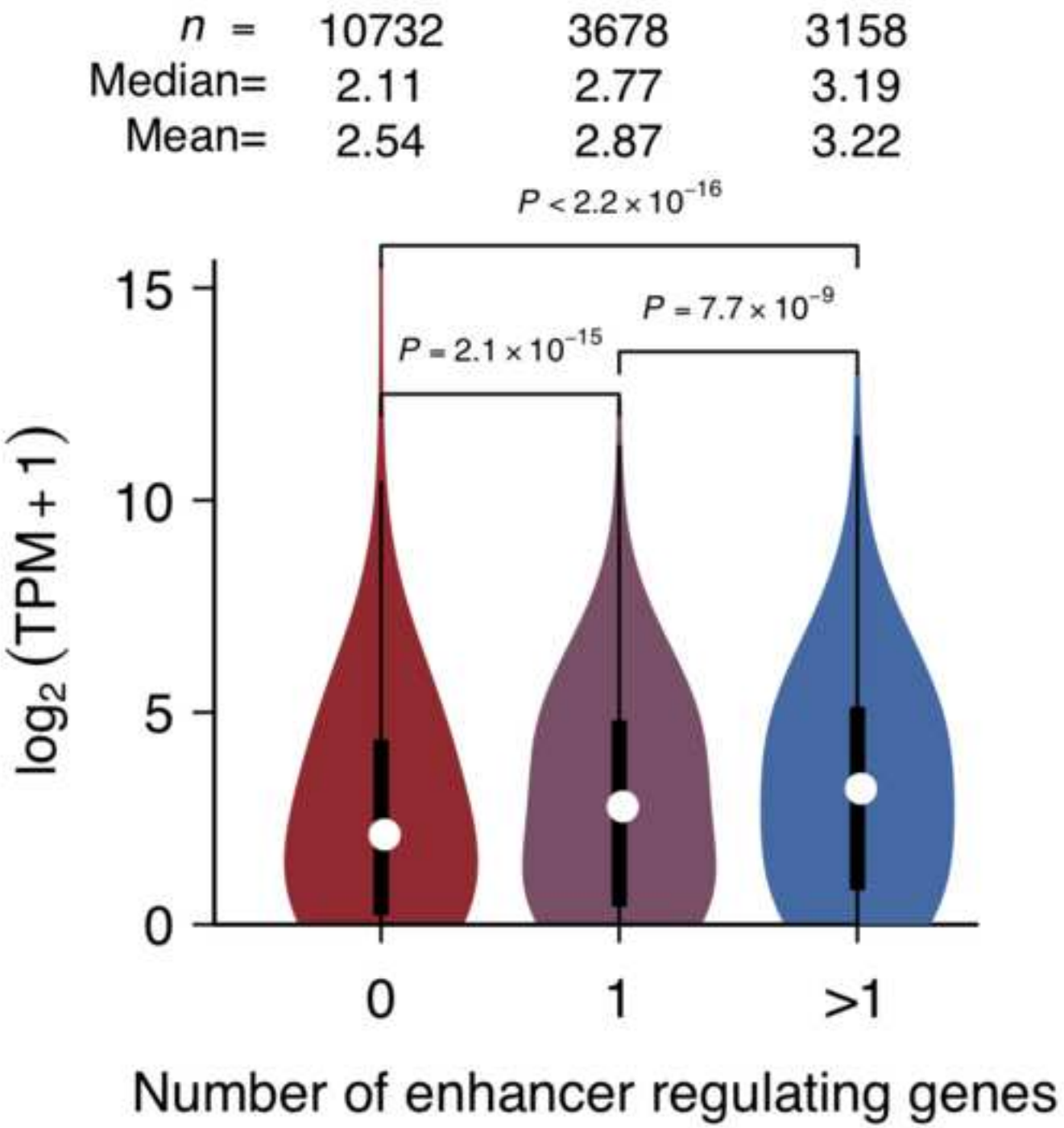

Table S1 Summary of the Pacbio initial assembly and Hi-C reads mapping used for goose genome assembly process.

Table S2 Summary of the length of pseudo-chromosomes in goose genome.

Table S3 A comparative summary of assembled repeat content in this study and previous studies.

Table S4 Summary the map rates of the wild goose resequencing data.

Table S5 Gene ontology (GO) enrichment analysis for the lineage-specific gene annotation in goose genome.

Table S6 Functional gene categories enriched for the goose genome-specific expansion gene family.

Table S7 Functional gene categories enriched for the contraction of genes family in goose genome.

Table S8 Positively selected genes (PSGs) identified in the goose genome.

Table S9 The PC1 values (100 Kb) through Principal Component Analysis (PCA) and A-B index values (25 Kb).

Table S10 TAD in genome coordinates of our goose genome by using method of DI values.

Table S11 Detailed information of promoter-enhancer interactions (PEIs) identified in goose genome.