Thesis for the Master of *Science*

Non-coding transcriptome maps across twenty tissues of Korean black chicken, Yeonsan Ogye

Hyosun Hong

Graduate School of Hanyang University

February 2018

Thesis for Master of *Science*

Non-coding transcriptome maps across twenty tissues of Korean black chicken, Yeonsan Ogye

Thesis Supervisor: Jin-Wu Nam

A Thesis submitted to the graduate school of
Hanyang University in partial fulfillment of the requirements
for the degree of *Master of Science*

Hyosun Hong

February 2018

Department of Life science
Graduate School of Hanyang University

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Non-coding transcriptome maps across twenty tissues of Korean black chicken, Yeonsan Ogye

Hyosun Hong

Department of Life Science

Graduate School of

Hanyang University

The Yeonsan Ogye (Ogye) is the rare chicken breed populated in Korean peninsula, which has a unique black-color appearance of entire body including feather, skin, comb, eyes, shank, and claws. Although some protein-coding genes related to the unique feature have been examined, none of non-coding elements were globally investigated. In this study, high-throughput RNA sequencing (RNA-seq) and reduced representation bisulfite sequencing (RRBS) were performed to construct whole non-coding transcriptome maps across twenty different tissues of Ogye. The resulting maps included 6900 long non-coding RNA (lncRNA) genes (9529 transcripts) comprising 1290 known and 5610 novel lncRNA genes. Comparing to lncRNAs

previously annotated in *galllus gallus red junglefowl*, the considerable number were either fragments of protein-coding genes or not expressed in Ogye tissues. Newly annotated Ogye lncRNA genes showed a tissue-specific expression and a simple gene structures constituting with 2 or 3 exons, as previously reported. Systematic analyses of sequencing data and other genomic data demonstrated that about 39% tissue-specific lncRNAs displayed functional evidences. Particularly, HSF2-associated lncRNAs were discovered as ones functionally linked to protein-coding genes specifically expressed in black skins (skin, shank, and comb), tended to be more syntenically conserved in mammals, and were differentially expressed in black tissues against white tissues. Our findings and resulting maps provide not only a comprehensive catalogue of lncRNAs but also a set of functional lncRNAs that will facilitate understanding non-coding genome regulating unique phenotypes and future use of genomic-breeding of chicken.

Chapter 1. Introduction

The Yeonsan Ogye (Ogye) chicken is one of the rarest breeds of *Gallus gallus domesticus*. Domesticated in the Korean peninsula, it probably originated from the Indonesian Ayam Cemani black chicken, which populates tropical, high-temperature areas (Dharmayanthi et al. 2017). Ogye shares common features-such as black plumage, skin, shank, and fascia-with Ayam Cemani (Dharmayanthi et al. 2017), although it has a smaller comb and shorter legs. Silkie fowl (Silkie), one of the most popular black-bone chickens, also has black skin but has white or varied color plumage (Dorshorst et al. 2011). Several genes involved in Silkie skin hyperpigmentation have been reported in previous studies (Shinomiya et al. 2012; Li et al. 2011; Dorshorst et al. 2011). Recently, transcriptomes from Chinese native black chickens were compared with those from white chickens to globally identify hyperpigmentation-related genes (Zhang et al. 2015). However, studies of the molecular mechanisms and pathways related to black chicken hyperpigmentation have been restricted to coding genes.

A major part of the non-coding transcriptome corresponds to long non-coding RNAs (lncRNAs), which originate from intergenic, intervening, or antisense-overlapping regions of protein-coding genes (Ponting, Oliver, and Reik 2009; Simon et al. 2013; Morris and Mattick 2014). lncRNAs are

defined as transcripts longer than 200nt and are mostly untranslated because they lack an open reading frame; however, they interact with RNA binding proteins and have diverse intrinsic RNA functions (Wang et al. 2008; Hellwig and Bass 2008; Ferre, Colantoni, and Helmer-Citterich 2016). They tend to be localized to subcellular areas, particularly the nucleus, and often interact with heterochromatin remodelers and DNA methylation regulators to regulate gene expression at the epigenetic level. For instance, DNMT1-associated colon cancer repressed lncRNA-1 (DACOR1) is localized to genomic sites, known to be differentially methylated, and regulates methylation at least 50 CpG sites by recruiting DNMT1 in colon cancers (Merry et al. 2015).

lncRNAs are also known to regulate gene expression at other levels: transcriptional, post-transcriptional, translational, and post-translational (Hirota et al. 2008; Wang et al. 2008; Zalfa et al. 2003; Hellwig and Bass 2008; Liu et al. 2012). They regulate distant genes by modulating the recruitment of transcription factors (TFs) to target genes. Only a few lncRNAs, however, have been experimentally validated as functional; most candidates remain unvalidated. In particular, some lncRNAs have been shown to regulate the expression of neighboring genes in a *cis*-acting manner (Ulitsky and Bartel 2013; Garding et al. 2013; Zhang et al. 2012; Jariwala and Sarkar 2016; Sahu, Singhal, and Chinnaiyan

2015). Enhancer-associated lncRNAs (eRNAs) are a well-known group in this class that regulate the expression of downstream genes. Knockdown of eRNAs reduces target gene expression, suggesting their function as *cis*-acting elements (Kim, Hemberg, Gray, Costa, Bear, Wu, Harmin, Laptewicz, Barbara-Haley, and Kuersten 2010; Marques et al. 2013; De Santa et al. 2010). eRNA regulatory roles are known to be achieved via several mechanisms: trapping transcription factors, directing chromatin roofing, and inducing DNA methylation (Orom et al. 2010; Lai et al. 2013; Wang et al. 2008; Feng et al. 2006; Bertani et al. 2011; Dimitrova et al. 2014). On the other hand, lncRNAs that associate with post-transcriptional regulators control target splicing and stability. For instance, antisense lncRNA from the *FGFR2* locus promotes cell-type specific alternative splicing of FGFR2 by interacting with polycomb complex (Gonzalez et al. 2015).

Despite their regulatory roles, only a few lncRNAs are highly conserved across vertebrates (Mercer et al. 2008). lncRNAs generally exhibit either poor conservation at the nucleotide level or conservation in a short region only, particularly compared to protein-coding genes (Derrien et al. 2012; Cabili et al. 2011; Quinn and Chang 2016). Although sequence conservation is often likely to indicate related function, sometimes it is difficult to detect conservation across multiple genome sequences because of technical challenges. lncRNAs, however, appear to be

syntenically conserved with protein-coding genes, which suggests that lncRNAs could have evolutionarily conserved roles in similar genomic contexts (Djebali et al. 2012; Ulitsky et al. 2011a; Li and Chang 2014). A zebrafish lncRNA, linc-*oip5*, which has a short region of sequence conservation with mammalian orthologs in the last exon, also exhibits preserved genomic architecture in its size and arrangement of exons; furthermore, linc-*oip5* loss of function disrupts zebrafish embryonic development, which can be rescued by the mammalian orthologs (Ulitsky et al. 2011b). Thus, examining the genomic context and/or short regions of conservation in a lncRNA may be necessary for understanding lncRNA function.

lncRNA expression signatures also provide hints about lncRNA functional roles at the cellular level. Global lncRNA profiling demonstrated that lncRNAs generally exhibit lower expression than protein-coding genes (Nam and Bartel 2012; Derrien et al. 2012; Pauli et al. 2012) but tend to be uniquely or specifically expressed in distinct tissues, developmental stages, conditions, or disease states (Iyer et al. 2015; Derrien et al. 2012; Cabili et al. 2011; Mercer et al. 2008; Dinger et al. 2008; Faghihi et al. 2008; Nam and Bartel 2012). For instance, one lncRNA, SAMMSON, is specifically expressed in melanoma cells during melanogenesis and is known to regulate the process at the epigenetic level (Leucci et al. 2016). In

addition, large-scale analyses of lncRNA and protein-coding gene co-expression led to the finding that a considerable number of paired genes are actually co-regulated by common TFs (Ghosh et al. 2015; Guttman et al. 2009). Often common TF binding motifs have been discovered in the promoters of the co-expressed lncRNA and protein-coding genes, suggesting that the co-regulated genes could share functional roles (Pang et al. 2009; Liao et al. 2011). Thus, to predict lncRNA biological functions, co-expression networks of lncRNAs and protein-coding genes from large scale transcriptomic data have been constructed and used for the inference of function (Lv et al. 2016; Wang et al. 2016; Cogill and Wang 2014).

Although genome and transcriptome maps of livestock animals, such as rainbow trout, cow, goat, and chicken (Weikard, Hadlich, and Kuehn 2013; Billerey et al. 2014; Li et al. 2012; Ren et al. 2016; Al-Tobasei, Paneru, and Salem 2016), have been recently constructed, only a few non-coding transcriptome studies have been done in those genomes. To date, 9,681 lncRNAs have been annotated in the red jungle fowl *Gallus gallus* genome, but these studies have been limited to a few tissues and many lncRNAs seem to be missing. Thus, a comprehensive non-coding transcriptome map of Ogye will help us understand phenotypic similarities and differences between Ogye and *Gallus gallus*.

Chapter 2. Comprehensive non-coding transcriptome maps of Ogye

Section 1. A comprehensive lncRNA catalogue of Ogye

To construct an Ogye transcriptome map, total RNA samples were collected from twenty tissue samples from 8-month-old Ogye (Figure 1A) and, in total, about 1.5 billion RNA-seq reads (843 million single-end reads and 638 million paired-end reads) were analyzed (Figure 1B). Pooled single- and paired-end RNA-seq reads of each tissue were mapped to the Ogye draft genome (Genome; https://www.ncbi.nlm.nih.gov/genome/; PRJNA412424) using STAR (ver 2.4.2) (Dobin et al. 2013), and subjected to transcriptome assembly using Cufflinks (ver 2.1.1) (Trapnell et al. 2010), leading to the construction of transcriptome maps for twenty tissues. The resulting maps were combined using Cuffmerge (ver 1.0.0) and, in total, 206,084 transcripts from 103,405 loci were reconstructed in the Ogye genome. In the unified Ogye transcriptome map, in addition to 15,766 protein-coding genes, 1290 known (3266 transcripts) and 5610 novel (6263 transcripts) lncRNA genes were confidently annotated using our lncRNA annotation pipeline, adopted from our previous study (You, Yoon, and Nam 2017) (Figure 2A). Compared to previously annotated chicken lncRNAs from *Gallus*

gallus, only 34% were redetected in the Ogye lncRNA catalogue. In fact, the remainder were mainly either fragments of protein-coding genes in which exon junctions were missed during transcriptome assembly or not expressed in all twenty Ogye tissues (Table 1). Consistent with other species (Al-Tobasei, Paneru, and Salem 2016; Billerey et al. 2014; Weikard, Hadlich, and Kuehn 2013; Pauli et al. 2012), the median gene length and the median exon number of Ogye lncRNAs were less than those of protein-coding genes (Figure 2B-C).

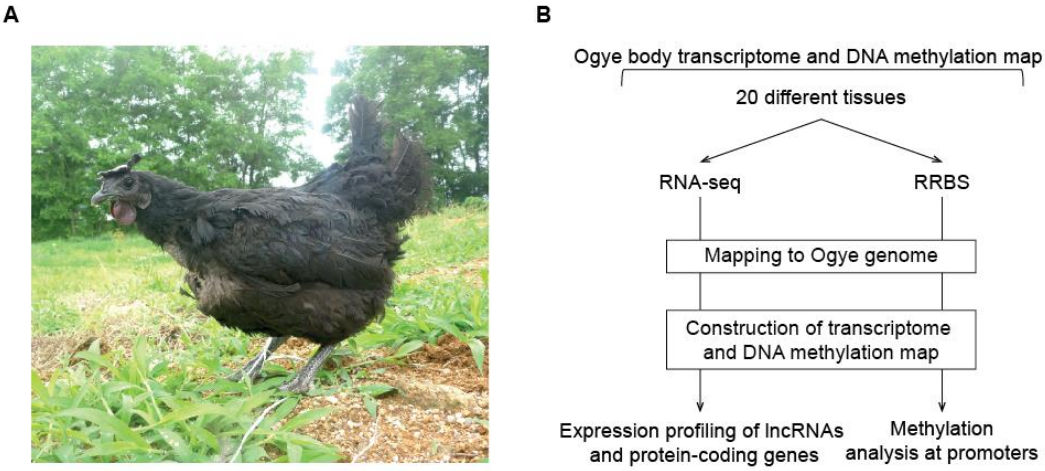

Figure 1. Yeonsan Ogye and study design (A) Yeonsan Ogye (B) A schematic flow for the analyses of coding and non-coding transcriptomes and DNA methylation from twenty different tissues.

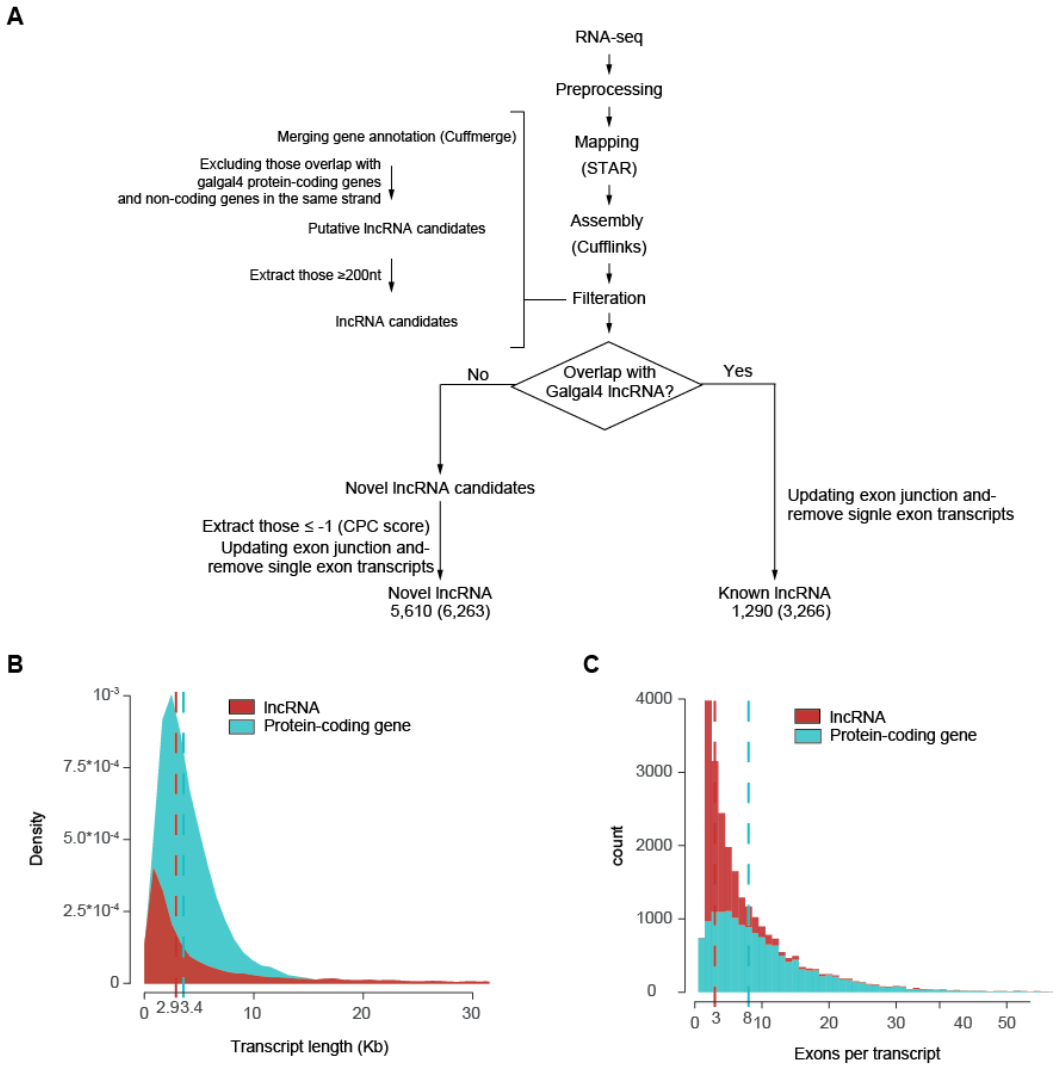

Figure 2. Annotation of Ogye lncRNA genes. (A) A pipeline for lncRNA annotations. (B) Distribution of transcript length (red for lncRNAs and cyan for protein-coding genes). The vertical dotted lines indicate the median. (C) Distribution of exon number per transcript. Otherwise, as in (B).

Table 1. Ogye and Galgal4 lncRNAs

Categories		count
Galgal4 lncRNAs overlapping with Ogye lncRNA	-	1,311
Galgal4 lncRNAs non-overlapping with Ogye lncRNA	protein-coding gene fragment	3,848
	not expressed in 20 tissues	3,407
	missed in Ogye	276

To profile the expression of protein-coding and lncRNA genes across tissues, fragments per kilobase of exons per million mapped reads (FPKM) were measured for transcripts using RSEM (v1.2.25) (Li and Dewey 2011). 6,565 lncRNA genes were expressed with FPKM ≥ 1 in at least one tissue, whereas 13,765 protein-coding genes were. As previously reported (Greco, Gorospe, and Martelli 2015; Cech and Steitz 2014; Cabili et al. 2011), Ogye lncRNAs generally displayed a tissue-specific expression pattern and some lncRNAs were solely expressed in a single tissue, although a few displayed ubiquitous expression across tissues. Tissue-specific genes with a four-fold higher maximum expression value than the mean value over twenty tissues were depicted on the genome using a Circos plot (Figure 3, green track). About 75% of lncRNA genes (5191 loci) were tissue-specific, a significantly higher proportion than that of protein-coding genes (45%; Figure 4A; $P < 2.2e^{-16}$; Wilcoxon rank sum test). The fractions of lncRNAs that were tissue-specific ranged from 2.4 % (Fascia) to 12.5% (Kidney), much higher percentages than those of protein-coding genes, which ranged from 0.4% (Fascia) to 4.2% (Kidney) (Figure 4B). Hierarchical clustering

of commonly expressed lncRNA genes among tissues using the PHYLIP package (ver 3.6) (Felsenstein 1989) defined functionally and histologically-related tissue clusters well. In particular, 2,317 lncRNAs were specifically expressed in the comb, skin, and shank, which are black tissues in Ogye (Figure 4B). Only 780 lncRNAs were ubiquitously expressed across all tissues (Figure 4B).

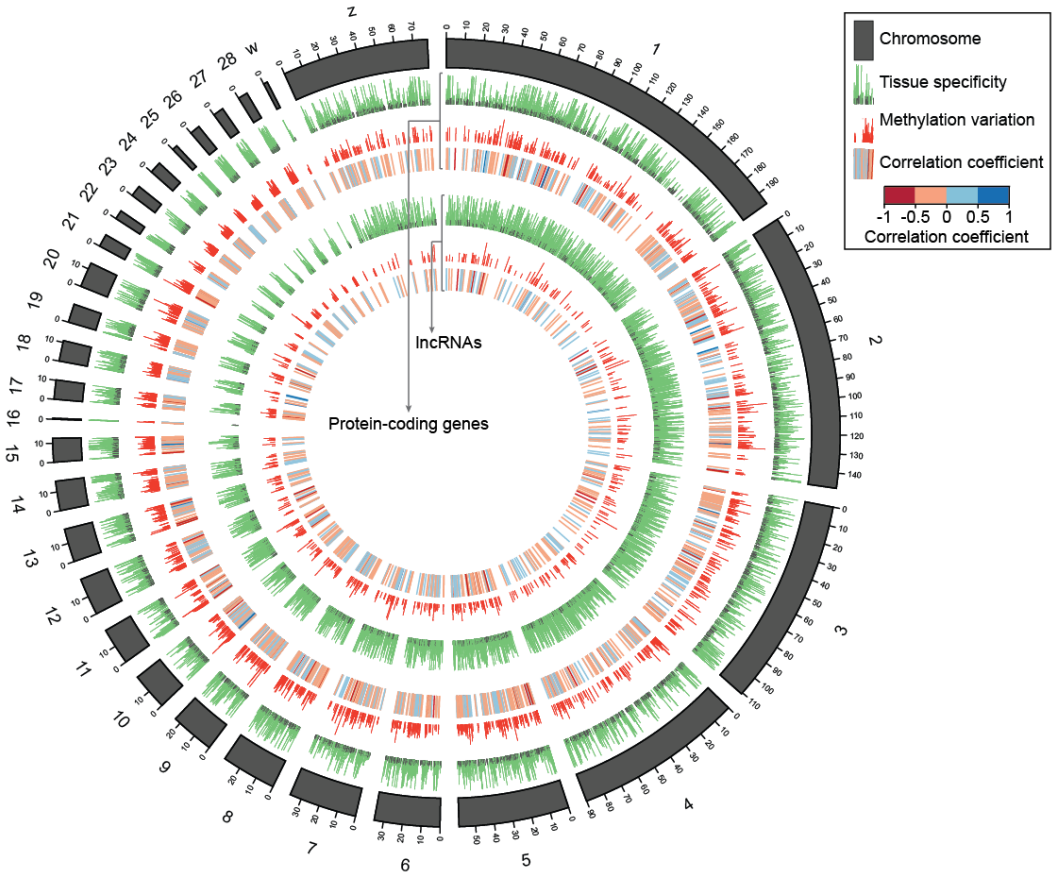

Figure 3. Comprehensive coding and non-coding transcriptome maps of Yeonsan Ogye. Circos plot illustrating the expression variability (green bars) of lncRNA and protein-coding genes, the methylation variability (red bars) at tissue-specific, differentially methylated CpG sites in the promoters, and the correlation coefficients between expression and methylation levels across chromosomes (heatmaps).

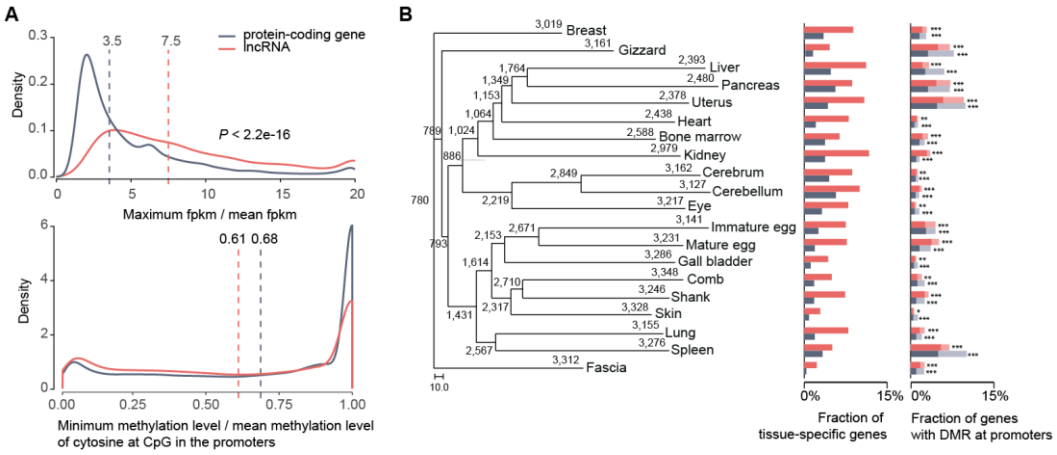

Figure 4. Tissue specificity of lncRNAs and protein-coding genes (A) Shown are the distributions of the maximum versus mean expression values of lncRNA (red line) and protein-coding genes (black line) across tissues (top), and the distributions of the maximum versus mean methylation levels of each cytosine in the promoter of lncRNAs (red line) and protein-coding genes (black line) (bottom). The vertical dotted lines indicate the median value of the respective distribution (black for protein-coding genes and red for lncRNAs). (B) Numbers of commonly or uniquely expressed lncRNAs across tissues are shown in the phylogenetic tree of tissues. The numbers at the leaf nodes indicate lncRNAs expressed in the indicated tissue (FPKM ≥ 1) and the numbers at the internal nodes indicate those commonly expressed in the indicated tissues. Of the expressed genes in a certain tissue, the fraction of the tissue-specific genes (red for lncRNA and black for protein-coding genes) and the fraction of genes with a differentially methylated region (DMR) in the promoters are indicated as bar graphs. Of the genes with a DMR, tissue-specific genes (dark) and others (light) were distinguished. The scale bar represents 10.0, which is the unit of 120 differentially expressed genes across tissues.

Section 2. Tissue-specific DNA methylation landscape of Ogye genome

To correlate the tissue-specific lncRNA expression with the epigenetic status of a respective tissue, reduced representation bisulfite sequencing (RRBS) data from twenty tissue samples were used (GEO; https://www.ncbi.nlm.nih.gov/geo/; GSE104355). RRBS reads were mapped to the Ogye draft genome. The DNA methylation signals (C to T changes in CpGs) across chromosomes were, then, calculated using Bismark in each sample (version 0.7.0) (Krueger and Andrews 2011). A significant correlation (nominal $P \leq 0.05$) between the expression levels and the methylation signals in the region 2kb upstream of genes across twenty tissues was demonstrated along with a variation of the signals (Figure 3). The variability was measured as the relative standard deviation. Of lncRNAs and protein-coding genes with tissue-specific differentially methylated CpG sites (tDMC) that include ≥ five reads with C to T changes in the promoter region in ≥ 10 tissues, 6.4% of the lncRNAs and 9.3% of the protein-coding genes displayed a significant negative correlation (nominal $P \leq 0.05$) between their promoter methylation levels and their expression levels, percentages that were significantly higher than those of random-pair controls (Figure 5A; $P = 1.30 \times 10^{-6}$ for lncRNAs; $P = 7.93$

X 10^{-36} for protein-coding genes; Fisher's exact test). However, only about 3% of genes showed a positive correlation between their expression and methylation signals, which is comparable or less than the control (Figure 5B; P = 0.87 for lncRNAs; P = 0.013 for protein-coding genes). Collectively, these results show that CpG methylation in the promoters represses the expression of target genes.

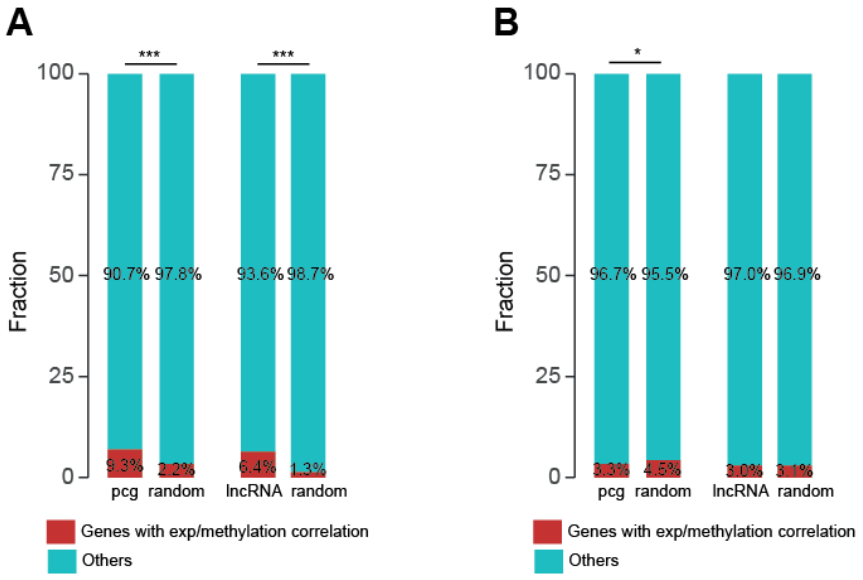

Figure 5. The proportion of genes with correlation between the methylation level and their expression. (A) The proportion of genes (protein-coding genes (left) and lncRNAs (right)) with a significant negative correlation (red) between the methylation level in their promoters and their expression values is shown. (B) The proportion of genes with a significant positive correlation between the methylation level and their expression values is shown. Otherwise, as in (A).

Chapter 3. Co-expression analyses of lncRNAs specify tissue-specific functional cluster

Section 1. Tissue-specific expression signatures of lncRNAs

As lncRNAs tend to be specifically expressed in a tissue or in related tissues, they could be better factors for defining genomic characteristics of tissues than protein-coding genes. To prove this idea, principle component analyses (PCA) were performed with tissue-specific lncRNAs and protein-coding genes (Figure 6). As expected, the 1st, 2nd, and 3rd PCs of lncRNAs enabled us to predict the majority of variances, and better discerned distantly-related tissues and functionally and histologically-related tissues (i.e., black tissues and brain tissues) (Figure 6A) than those of protein-coding genes (Figure 6B).

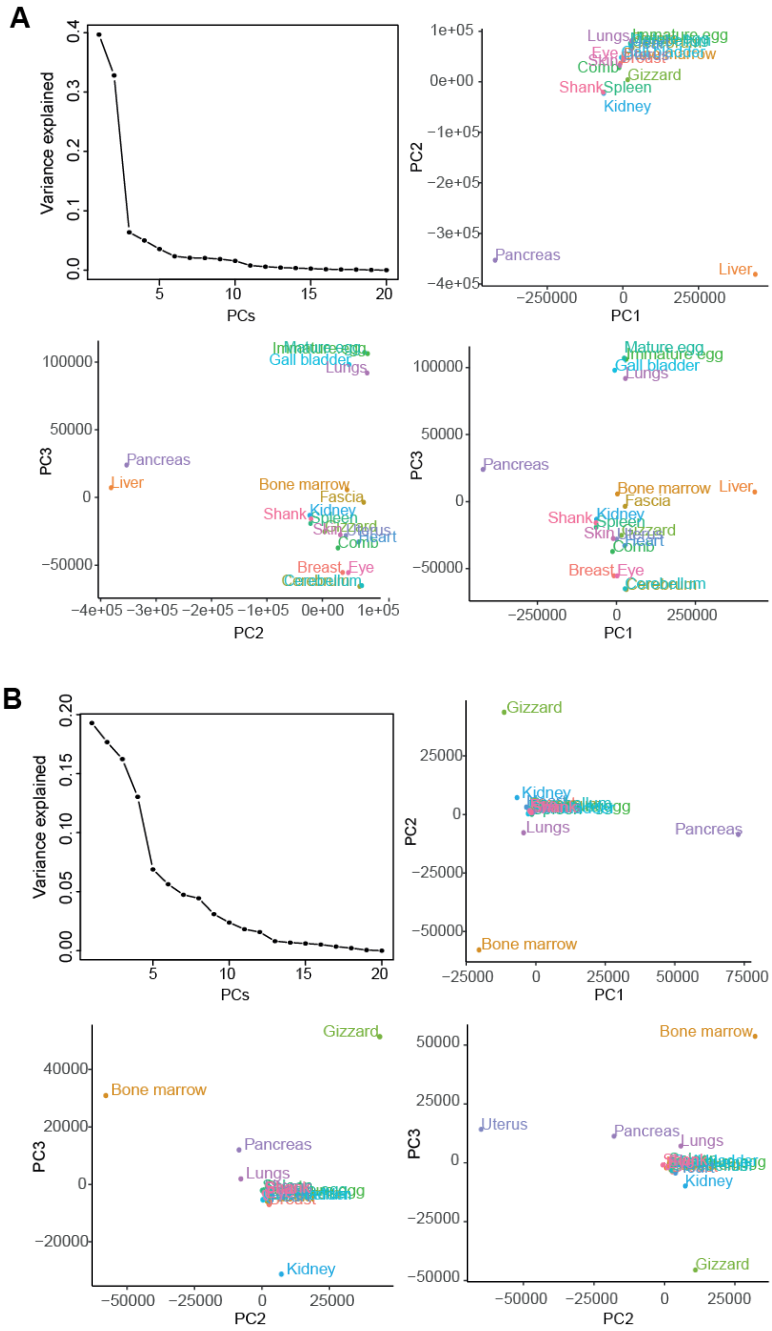

Figure 6. Principle component analysis of protein-coding and lncRNA genes. (A) Principal component analysis (PCA) using tissue-specific lncRNAs. PCs explaining the variances are indicated with the amount of the contribution in the left-top plot. PCA plots with PC1, PC2, and PC3 were demonstrated in a pairwise manner. Each tissue is indicated on the PCA plot with a specific color. (B) PCA using tissue-specific protein-coding genes. Otherwise, as in (A).

To identify functional clusters of lncRNAs, pairwise correlation coefficients between tissue-specific lncRNAs were calculated and the co-expression patterns across 20 tissues were clustered, defining 16 co-expression clusters (Figure 7A). As expected, each co-expression cluster was defined as a functional group, highly expressed in a certain tissue (kidney, eye, pancreas, uterus, mature egg, immature egg, breast, heart, liver, lung, gall bladder, gizzard, bone marrow, or spleen) or related tissues (brain and black tissues). In particular, the largest co-expression cluster, the brain-specific group, included 930 co-expressed lncRNAs, highly expressed in cerebrum and cerebellum. The second largest cluster, the black tissue-specific group, included 479 co-expressed lncRNAs, highly expressed in fascia, comb, skin, and shank (Figure 7A). Clusters of related tissues also display distinct sub-modules corresponding to each tissue. For instance, lncRNA clusters specific to black tissues displayed sub-clusters including sub-cluster 1 specific to shank and sub-cluster 2 specific to comb, although the sub-clusters shared skin-specific expression (Figure 7B).

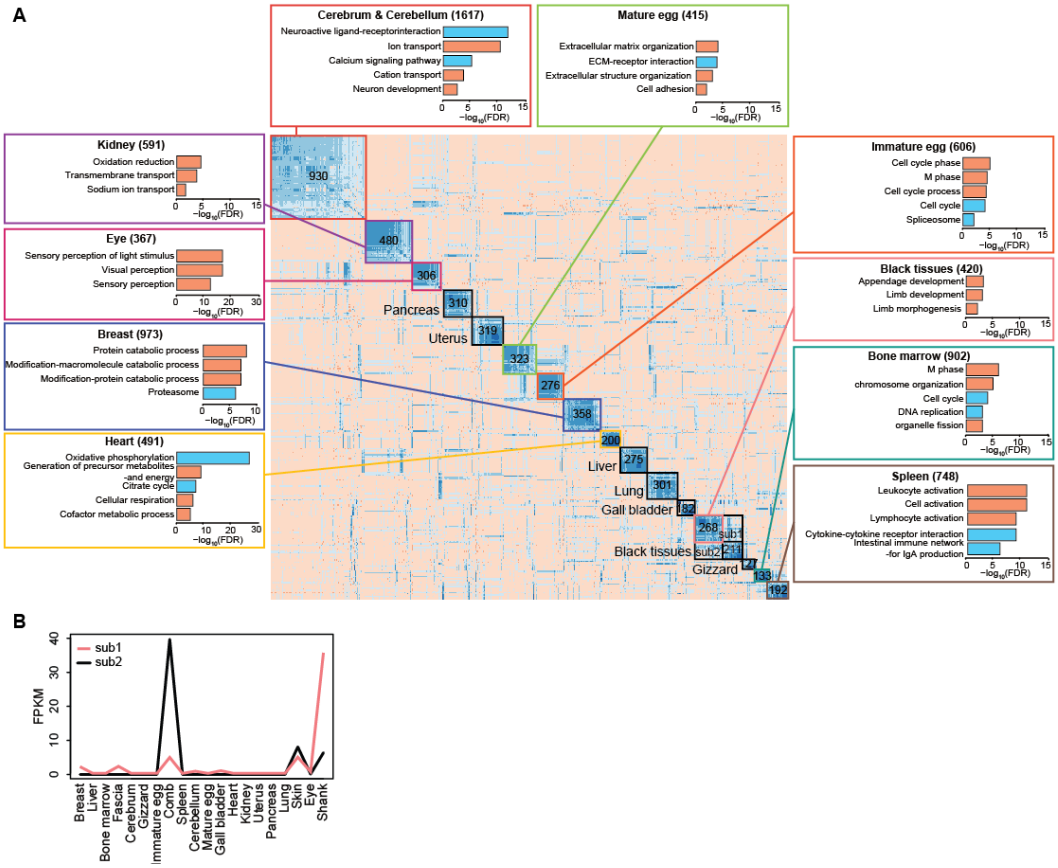

Figure 7. Co-expression clusters of lncRNAs and functional annotations. (A) Co-expression clustering of lncRNAs across twenty tissues defines sixteen clusters and two sub-clusters specific to a tissue or a set of similar tissues. The boxes outlined in a color indicate clusters that have significant GO biological processes (orange bars) or KEGG pathway terms (cyan bars) associated with the protein-coding genes co-expressed with lncRNAs in the respective cluster. The significant enrichment of terms was tested using the hypergeometric test and adjusted by FDR, indicated with a logarithmic scale on the X-axis in the box. Clusters outlined in black are those that had neither a significant association with any GO term nor any co-expressed protein-coding genes. (B) Expression patterns of sub-clusters (sub 1 for shank, 2 for comb) in the black tissue cluster.

Section 2. Functional annotations of lncRNA co-expression clusters

The functional role of each co-expressed lncRNA cluster can be indirectly evident by those of significantly co-expressed mRNAs (Lv et al. 2016; Wang et al. 2016; Cogill and Wang 2014). Thus, exclusively co-expressed mRNAs to each lncRNA cluster were identified with following criteria: mean Pearson's correlation ($\bar{r}$) $\geq$ 0.5 with members within a cluster and the differences between the corresponding $\bar{r}$ and the mean correlation ($\bar{r}_i$) with all other groups $\geq$ 0.3, and subsequently subjected to the gene ontology (GO) analyses using DAVID (Huang da, Sherman, and Lempicki 2009) (Figure 7A). Particularly, 1617 mRNAs exclusively correlated to brain-specific lncRNA group (930 lncRNAs) were identified and had brain-function specific terms, such as neuroactive ligand-receptor interaction (q = 2.18 X 10^{-12}; False discovery rate, FDR correction). By contrast, 748 mRNAs exclusively correlated to spleen-specific lncRNAs were identified and have immune-related terms, such as leukocyte activation (q = 2.37X 10^{-12}). Likewise, 10 out of 16 co-expression clusters of lncRNAs had functional evidences with significantly enriched GO terms and KEGG pathway (Figure 7A).

Chapter 4. Coherent expression models of lncRNA and mRNA

Section 1. Tissue-specific, co-expressed groups of lncRNAs and mRNAs

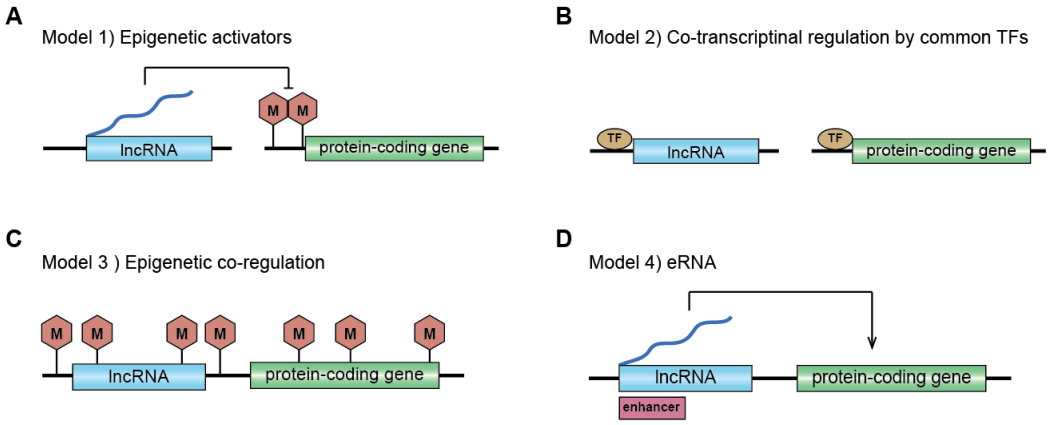

Figure 8. Co-regulational models of lncRNA and protein-coding genes. (A) lncRNAs as epigenetic activators that suppress the methylation level in the promoter of protein-coding genes. (B) Transcriptional co-regulation of lncRNA and protein-coding genes by common TFs. (C) Epigenetic co-regulation of neighboring lncRNA and protein-coding genes. (D) eRNAs that activate the expression of neighboring protein-coding genes.

The coherent expression of two different RNA classes could be in part the outcome of either active regulation by lncRNAs in *cis* and *trans*, or co-regulation by common regulators, such as TFs or epigenetic regulators, in *cis* and *trans* (Figure 8). Regulation of gene expression by lncRNAs often involves engagement with chromatin remodelers, such as polycomb repressive complexes (PRCs) that mediate the suppression of target mRNA

expression (Wu et al. 2013; Zhao et al. 2008) or demethylases that open the chromatin structure to enhance the expression of target mRNAs (Chalei et al. 2014; Di Ruscio et al. 2013) (Figure 8A). Remote co-expression of lncRNAs and mRNAs can be also regulated by common TFs (Ghosh et al. 2015; Guttman et al. 2009) (Figure 8B). Co-expressed genes tend to have common TF binding motifs in their promoters. However, *cis*-regulation of mRNA expression by lncRNAs is known to be associated with common epigenetic factors (Figure 8C) or enhancers (Figure 8D).

Section 2. lncRNA as epigenetic activators

To find lncRNAs that act as epigenetic activators that reduce methylation levels, lncRNAs with expression levels that are significantly negatively correlated with the methylation level in the promoters of co-expressed protein-coding genes (nominal $P \leq 0.01$) were examined in each co-expression cluster. In this case, the lncRNAs are thought to reduce the methylation level in the promoters of the co-expressed protein-coding genes. Of the lncRNAs in clusters, the expression of 15.0%~72.9% displayed significantly negative correlation with methylation levels in the promoters of co-expressed protein-coding genes, which were compared to those of random protein-coding gene cohorts (Figure 9A). Clusters specific to brain, kidney, mature egg, breast, heart and spleen included significantly more lncRNAs with a significant correlation than did the

random controls (P = 0.026~ 7.71 X 10^{-13}) but this was not true for the black tissue cluster. To identify DNA methylation activators with more confidence, we also examined whether the expression and methylation of the co-expressed coding genes were correlated (nominal $P \leq 0.01$). 820 lncRNAs in the clusters were identified as confident DNA methylation activator candidates (Figure 9B). Genes encoding lncRNAs that act as DNA methylation regulators of protein-coding genes were mostly 100kb apart, and only five were within 100kb from target genes, suggesting that lncRNAs that function as epigenetic activators mostly play their roles in *trans*-form rather than *cis*-form.

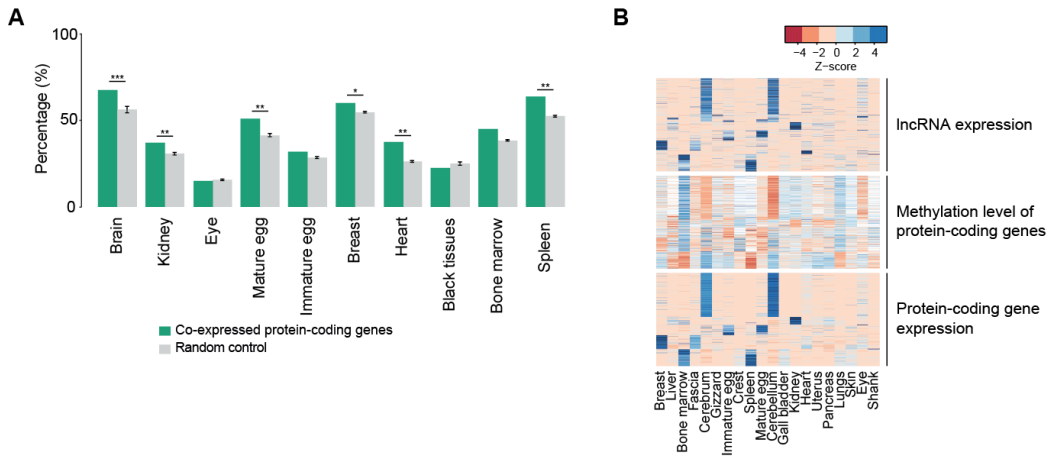

Figure 9. lncRNAs as epigenetic activators. (A) The proportions of lncRNAs with expression levels that are correlated with the methylation level in the promoter of co-expressed protein-coding genes (dark green) in each cluster are shown in bar graphs. The numbers were compared to the mean methylation level of randomly selected protein-coding genes. To test the significance of the enrichment of lncRNAs as epigenetic activator candidates, 1000 number-matched random cohorts were compared to the original numbers (* $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$). (B) lncRNAs as epigenetic activators whose expression levels are negatively correlated with the methylation level in the promoters of protein-coding genes, which in turn are negatively correlated with the level of protein-coding gene expression, as shown in heatmaps. The key indicates the z-score range of the expression values. White indicates N.A.

Section 3. Transcriptional regulation by common transcription factors

To identify co-expressed pairs of lncRNAs and mRNAs regulated by common TFs, TF binding sites (TFBSs) enriched in the promoters of the co-expressed genes were examined. For this analysis, sequences 2kb upstream of the co-expressed genes were extracted and enriched sequence motifs were identified using the multiple expectation-maximization for motif elicitation (MEME) suite (Machanick and Bailey 2011). The resulting motifs were subjected to analysis by the TOMTOM program (Gupta et al. 2007) to annotate TFBSs based on TRANSFAC database v3.2 (Wingender et al. 1997). As a result, 14 common TFs that have significantly abundant binding sites in the promoters of lncRNA and protein-coding genes were detected (Figure 8B; corresponding to model 2). To discern TFs available in chicken genomes, PANTHER (Mi et al. 2010; Thomas et al. 2003) was used to examine whether there are chicken orthologs of the TFs and whether the orthologs are expressed in the corresponding tissues (FPKM $\geq$ 1). Finally, five TFs, including HSF2 and SP1, were identified as candidates (Figure 10A). HSF2 and SP1 binding sites were more recurrently detected across tissues than others and were significantly enriched in the promoters of 478 lncRNAs and 634 protein-coding genes. Although the binding motifs were slightly degenerated from the annotated motifs, the HSF2 motifs were similar in the promoters of lncRNA genes and protein-coding genes (Figure 10B).

To examine further whether the respective TFs actually affect the expression of lncRNAs and protein-coding genes, the correlation between the expression of each TF and co-expressed genes in each cluster was examined. Interestingly, HSF2 expression had a strong positive correlation with expression of genes in black tissues but not in other tissues (Figure 10C). The expression pattern for each of the five lncRNAs and protein-coding genes that were highly correlated with that of HSF2 was specific for skin, shank, and comb compared to other tissues (Figure 10D). Thus, HSF2 is a promising candidate for regulating the black tissue-specific expression of lncRNAs and protein coding genes. Taken together, our data indicate that of a total of 3466 lncRNA in ten clusters, 615 (17.74%) appear to be co-regulated with co-expressed protein-coding genes by common TFs, such as HSF2.

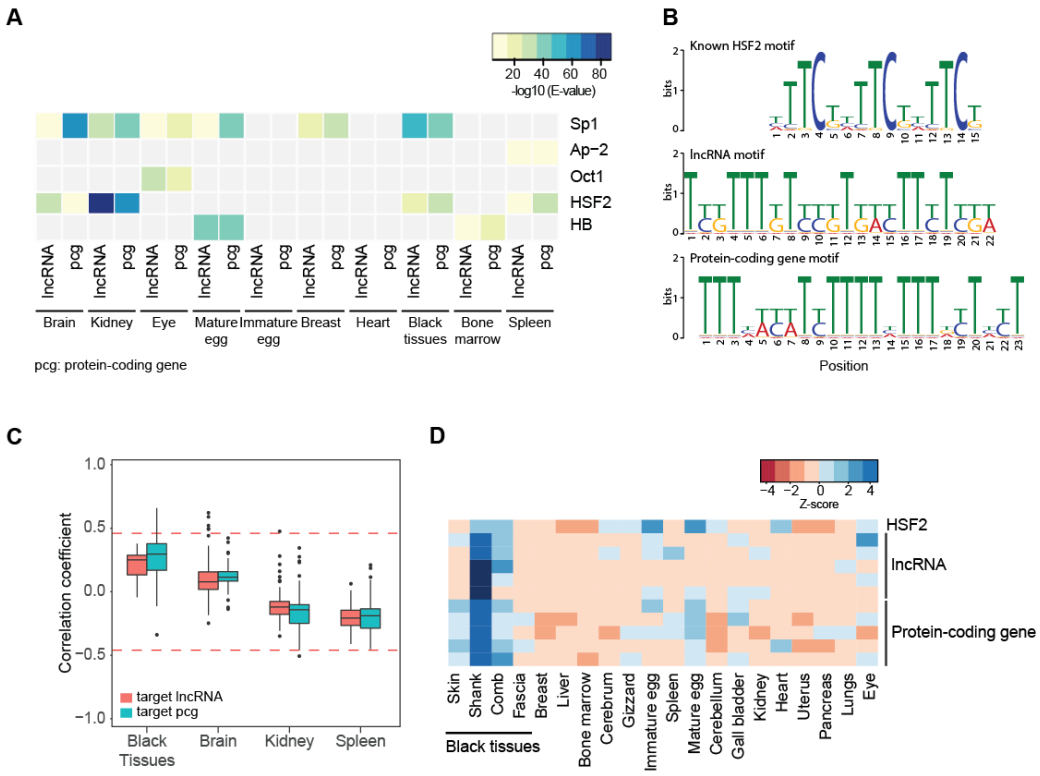

Figure 10. Co-transcriptional regulation of lncRNA and protein-coding genes by common TFs. (A) TFs (Sp1, Ap-2, Oct1, HSF2, and HB) with binding motifs that are significantly co-enriched in the promoters of lncRNAs in a tissue-specific cluster and their co-expressed protein-coding genes are shown in the heatmap. The TFs are expressed in the indicated tissues. The significance of the motif enrichment was tested using MEME and E values are presented with color codes (blue: more significance, yellow: less significance) in the key. PCG indicates protein-coding gene. (B) The HSF2 binding motif. A known motif is shown in the top panel, a motif in lncRNA promoters is shown in the middle panel, and a motif in protein-coding gene promoters is shown in the bottom panel. (C) The expression correlation between co-regulated genes (red boxes for lncRNAs and green boxes for protein-coding genes) and HSF2 across tissues. Red lines indicate the significance level of the correlation coefficient ($P \leq 0.05$). (D) Expression pattern of HSF2 and its target genes that have the top 5 correlations with HSF2.

Section 4. Coherent expression of neighboring lncRNA and protein-coding genes

Previous studies showed that lncRNAs and their neighboring protein-coding genes are highly correlated in their expression across tissues and developmental stages (Nam and Bartel 2012; Ulitsky et al. 2011a). To examine how the co-expressed lncRNAs and mRNAs in our study are co-localized in chromosomes, lncRNAs from each group were first classified based on the closest distances (≤10kb, ≤100kb, >100kb, and other chromosomes) from the significantly co-expressed protein-coding genes (nominal $P \leq 0.01$; Pearson's correlation) (Figure 11A). Genes encoding co-expressed pairs of lncRNAs and mRNAs are significantly proximally co-localized within 10kb (Figure 11A left; $P \leq 0.05$, Fisher's exact test), compared to random controls (Figure 11A right) but not those of lncRNAs and mRNAs in the range of 10~100kb or in the 100kb outside. Overall, 2 ~ 15 % of the co-expressed pairs in the clusters tended to be proximally co-regulated within 10kb.

To examine how neighboring lncRNAs and protein-coding genes are tissue-specifically co-regulated, the pairs within 10kb were classified into three categories on the basis of their relative orientations (head-to-tail, tail-to-tail, or head-to-head). The correlation coefficients of the pairs in each category were compared to those of lncRNA and random protein-coding gene controls from tissue-specific gene sets (Figure 11B) or from

ubiquitously expressed gene sets (Figure 11C). Both neighboring lncRNA and protein-coding gene pairs displayed significantly greater correlation than did random controls, regardless of the category, in both sets (Figure 11C). The correlations were also compared to those of neighboring protein-coding gene pairs. Whereas the correlations of the ubiquitously expressed, neighboring lncRNAs and protein-coding genes were significantly lower than those of ubiquitously expressed neighboring protein-coding gene pairs in the head-to-tail and head-to-head categories (Figure 11C), the correlation coefficients of the tissue-specific pairs were slightly yet insignificantly higher than those of neighboring protein-coding gene pairs (Figure 11B).

To dissect factors that affect the co-regulation of tissue-specific neighboring lncRNA and protein-coding gene pairs, the pairs with a high correlation ($P \leq 0.05$) between the methylation levels of their promoters (methylation-related group – model 3) and those with no correlation (methylation-unrelated group) were divided. Tissue-specific neighboring lncRNA and protein-coding gene pairs showed no more expression correlation than did neighboring protein-coding genes in the methylation-related group (Figure 11D; $P = 0.71$, Wilcoxon rank sum test), whereas they did show a significantly higher correlation in the methylation-unrelated group (Figure 11E; $P \leq 0.001$ for head-to-tail, $P \leq 0.05$ for head-to-head,

Wilcoxon rank sum test), which suggests that neighboring lncRNAs and protein-coding genes in the methylation-unrelated group have a regulatory interaction between them.

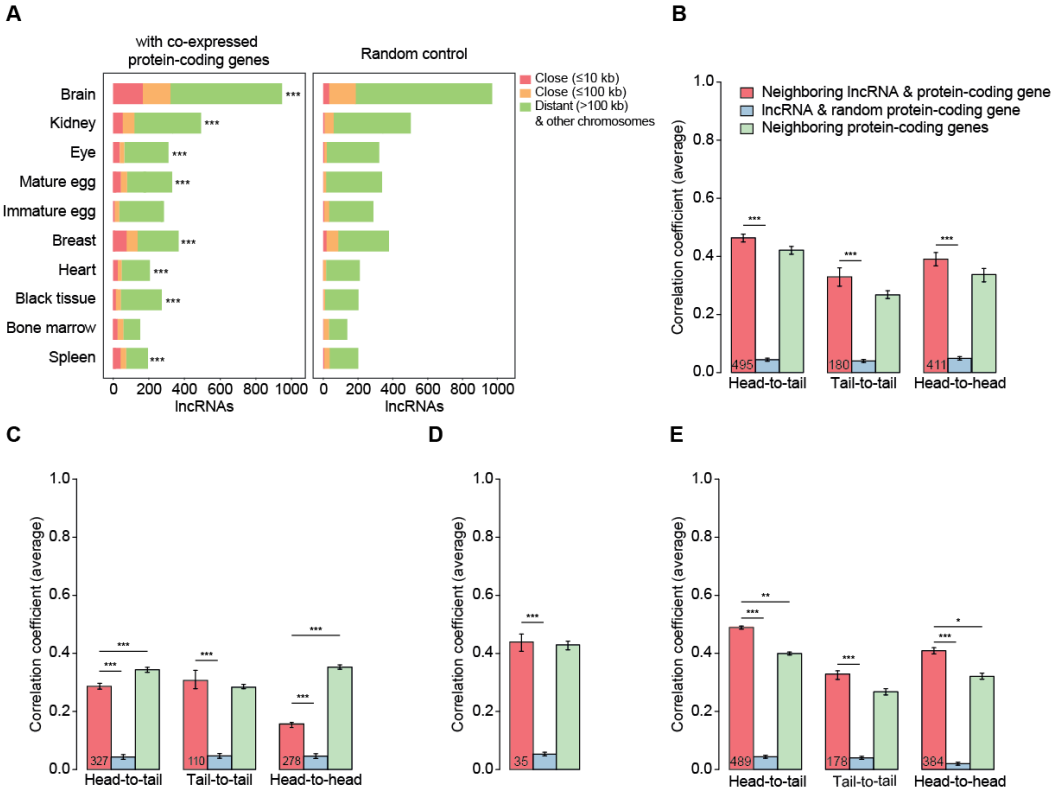

Figure 11. Co-regulation of neighboring lncRNA and protein-coding genes. (A) Shown are the numbers of lncRNAs, classified by the distance from the closest protein-coding gene (red for the ≤ 10 kb group, orange for the ≤ 100 kb group, and green for the > 100kb or on another chromosome group) (left). *, **, and *** indicate $P \leq 0.05$, $P \leq 0.01$, and $P \leq 0.001$, respectively. (B) The average correlation coefficients of tissue-specific lncRNA and protein-coding gene pairs in close neighborhoods (≤10 kb) are shown based on their relative orientations (head-to-tail, tail-to-tail, and head-to-head) (red bars). The average correlation coefficients of random pairs are also shown (blue bars) and those of tissue-specific protein-coding gene pairs in close neighborhoods (≤10 kb) are shown with green bars. *, **, and *** indicate $P \leq 0.05$, $P \leq 0.01$, and $P \leq 0.001$, respectively. Error bars indicate the standard error. The number in the bars indicates the number of analyzed pairs. (C) Average correlation coefficients of ubiquitously expressed lncRNAs and neighboring protein-coding genes. Otherwise, as in (B) (D) The average correlation coefficients of neighboring lncRNA and protein-coding genes with similar methylation levels in their promoters (methylation-related) are shown in bar graphs. Otherwise, as in (B). (E) The average correlation coefficients of tissue-specific lncRNA and protein-coding genes (methylation-unrelated), except for those of (C). Otherwise, as in (B).

Section 5. Enhancer-associated RNA-mediated gene regulator

Previous studies showed that lncRNAs associated with enhancers could regulate their neighboring protein-coding genes (Li, Notani, and Rosenfeld 2016). Genomic association between lncRNAs and enhancers, detected in embryonic developmental stages in the chicken (Seki et al. 2017), revealed that lncRNAs in the methylation-unrelated group are more significantly associated with enhancers than those in the other group (Figure 12A; $P = 2.72 \times 10^{-6}$; Fisher's exact test). As a result, 136 head-to-tail lncRNAs, 67 tail-to-tail lncRNAs and 124 head-to-head lncRNAs were considered as enhancer-associated lncRNA candidates (eRNAs). The eRNAs (corresponding to model 4) had a greater correlation with neighboring protein-coding genes only in the head-to-tail group (Figure 12B), whereas non-eRNAs displayed a greater correlation in the head-to-head orientation, which could allow sharing of common promoters (Figure 12C). A few eRNAs were discovered to have strong bi-directional transcriptional activity (Figure 12D), as previously reported (Kim, Hemberg, Gray, Costa, Bear, Wu, Harmin, Laptewicz, Barbara-Haley, Kuersten, et al. 2010; Wang et al. 2011)

Next, to identify TFs binding to genomic regions that transcribe eRNAs, TF binding sites detected from all the genomic regions associated with enhancers were profiled and were compared to those of TFs detected from the enhancers specific to a certain tissue (Figure 12E). Oct1 and HSF2

binding sites were significantly localized in eRNAs specific to black tissues ($P < 3.09 \times 10^{-5}$ for Oct1; $P < 3.11 \times 10^{-7}$ for HSF2; binomial test). Besides the TFs specific to black tissues, GR, YY1, RAP1 and GATA1, and HSF3 binding sites were localized in eRNAs specific to heart, eye, spleen and bone marrow, respectively (Figure 12E). Interestingly, HSF2 was a common TF candidate for co-regulating lncRNAs and protein-coding genes at a distance.

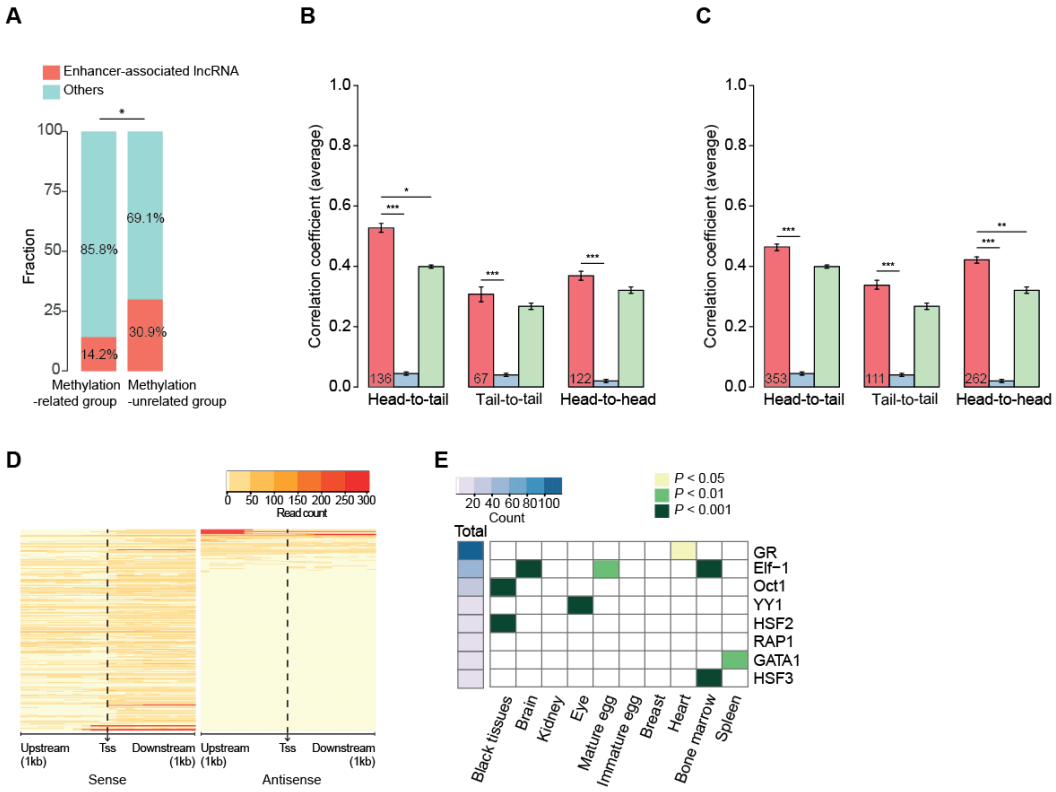

Figure 12. Co-regulation of neighboring eRNA and protein-coding genes. (A) The proportion of eRNAs (red) in the methylation-related group (Figure 9D) and -unrelated group (Figure 9E). The significance of the difference in the fraction of eRNAs was tested using Fisher's exact test. ** indicates $P \leq 0.01$. (B) The average correlation coefficients of tissue-specific eRNAs. Otherwise, as in Figure 9B. (C) The average correlation coefficients of tissue-specific lncRNAs not associated with enhancers. Otherwise, as in Figure 9B. (D) The read counts are indicated with color codes (described in the key) in the sense (left) and antisense (right) strands based on the relative position from the eRNA TSS. Yellow indicate no read. (E) TF binding motifs significantly associated with the eRNAs. The total count of the indicated TF binding sites in eRNAs is indicated in the heatmap (left) and the significance of the association over the total background is indicated with color-coded *P* values across tissues. The significance of a specific TF binding motif was tested using a binomial test in each tissue.

Chapter 5. Black skin-specific conserved lncRNAs

As mentioned earlier, unlike other chicken breeds, both the plumage and skin of the Ogye are black. To identify lncRNAs potentially functionally related to this trait, lncRNAs specifically co-expressed in black tissues (Figure 7A) were further investigated by comparing to those in non-black skin of other chicken breeds. Of 479 lncRNAs specific to black tissues, 47 were significantly two-fold up- (29) or down-regulated (18) in Ogye black skin, compared to those in brown leghorn skin (Figure 13A; FDR < 0.05).

To find functionally conserved lncRNAs, the 47 differentially expressed lncRNAs were examined for synteny and sequence conservation in human and mouse genomes. Synteny conservation considers whether orthologs of a certain lncRNA's neighboring genes are positionally conserved in these mammalian genomes (Figure 13B). As a result of this analysis, about 10% of lncRNAs were found to be syntenically conserved in both the human and mouse genomes and about 25% were syntenically conserved in at least one genome (Figure 13C), percentages that are comparable to those of the protein-coding genes (Figure 13D). However, sequence similarity analyses by the BLAST showed that only 6% of the syntenically conserved lncRNAs had conserved sequences relative to sequences in either the human or mouse

genomes (Figure 13C), which is much lower than that of protein-coding genes (56%). Taken together, our data showed that 16 lncRNAs were syntenically or sequentially conserved and differentially expressed in black tissue (Figure 13E). Of the 16 lncRNAs that have evidence of black tissue-specific function, four, including eRNAs, were associated with HSF2 binding motifs, whereas of the 104 that have synteny and sequence conservation but are not differentially expressed in black tissues, only one was associated with HSF2. The presence of HSF2 binding motifs appears to be significantly related to black tissue-specific expression (Figure 13F; $P \leq 0.0008$, Fisher's exact test).

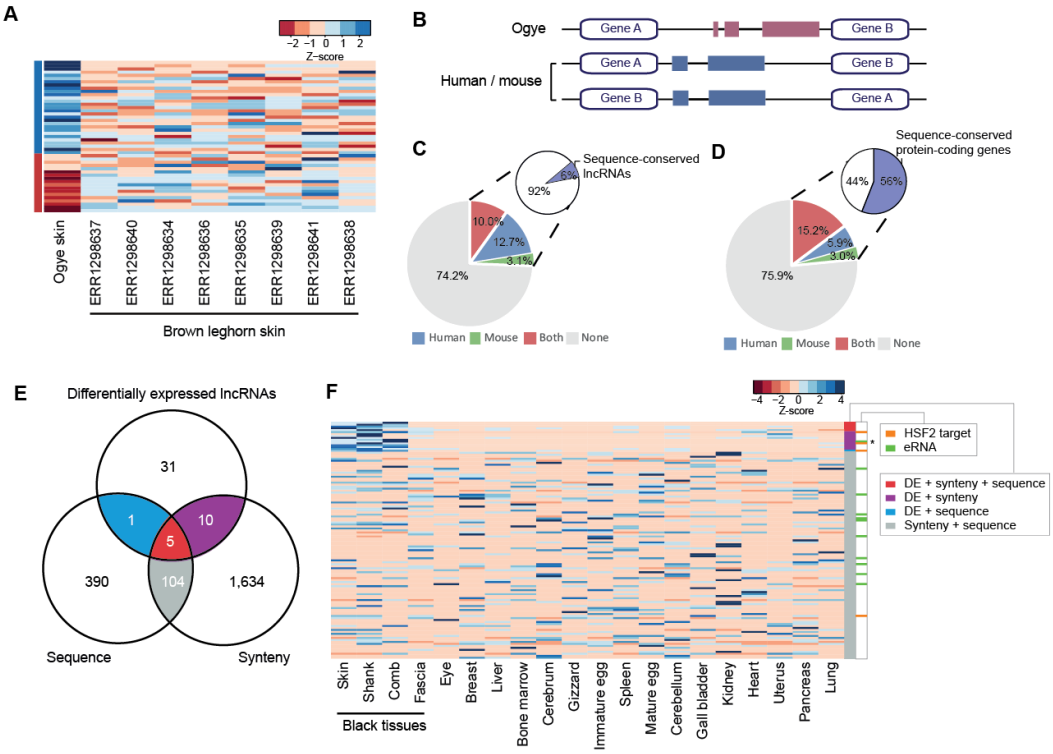

Figure 13. Black tissue-specific lncRNAs with sequence and synteny conservation. (A) The expression patterns of differentially expressed lncRNAs in Ogye skin, compared to brown leghorn skin samples. Expression levels are indicated with a color-coded Z-score (red for low and blue for high expression) as shown in the key. (B) A cartoon showing a lncRNA that is syntenically conserved with up- and down-stream protein-coding genes in the human and/or mouse genome. (C) The fraction of lncRNAs with syntenic conservation in the human (blue), mouse (green) or both (red) genomes is shown in the pie chart. Of the syntenically conserved lncRNAs, the fraction of lncRNAs with sequence conservation (purple) in the human or mouse genome is indicated in the secondary pie charts. (D) The fraction of protein-coding genes with synteny conservation is indicated in the pie chart. Otherwise, as in (C). (E) The numbers of differentially expressed lncRNAs in black skin with evidence of sequence and synteny conservation are indicated in a Venn diagram. (F) Evidence for differential expression (DE) + synteny + sequence (red), DE + synteny conservation (purple), or DE + sequence conservation (blue) for 16 black-skin specific lncRNAs is shown in a heatmap. 104 non-specific lncRNAs with evidence of sequence + synteny conservation are indicated in gray. The co-regulation models associated with a certain lncRNA are indicated to the left with color codes (orange for HSF2 binding and green for eRNAs). * indicates the eRNA associated with HSF2. The expression level is indicated with a color-coded z-score, as shown in the key.

For instance, linc-THEM184c is significantly up-regulated in black tissue (Figure 14B), its locus is syntenically conserved with neighboring genes, TMEM184C and EDNRA, in both human and mouse genomes, and its promoter includes a HSF2 binding motif (Figure 14A). In addition, the protein-coding genes that are co-expressed with this lncRNA are enriched for GO terms that are functionally relevant for black skin: keratinocyte differentiation, angiogenesis, and ECM-receptor-interaction (Figure 14C). Among the co-expressed genes, 31 have HSF2 binding sites in their promoters (Figure 14A).

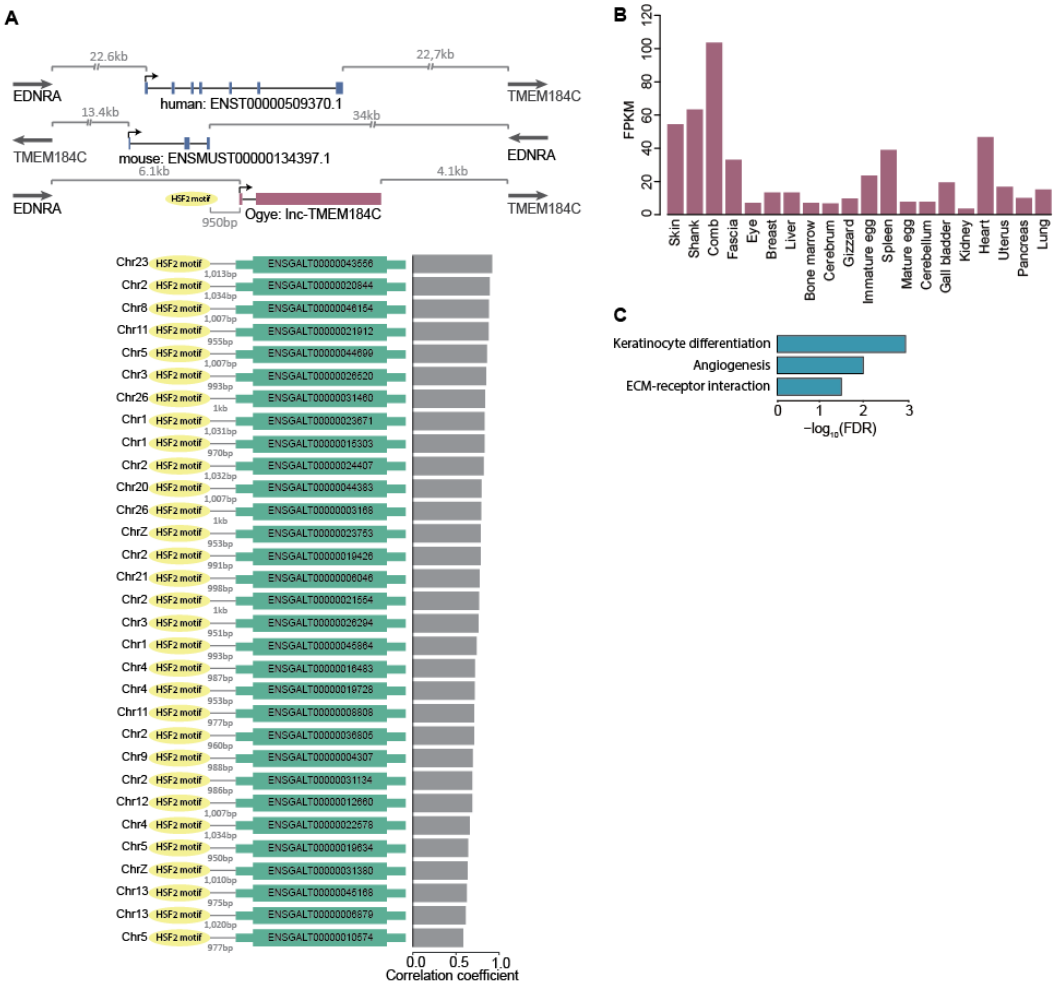

Figure 14. An example of a black skin-specific lncRNA with synteny conservation, which is transcriptionally regulated by HSF2. (A) Ogye lncRNA (lnc-TMEM184C) with synteny conservation in human and mouse genomes (top). The lncRNA has an HSF2 binding motif in its promoter; this motif is also present in the promoters of protein-coding genes with correlated expression (below). Gray bar plots indicate the expression correlation between the lncRNA and the protein-coding genes. (B) The lnc-TMEM184C expression pattern across 20 tissues. (C) GO terms that are significantly associated with the protein-coding genes that are co-expressed with lnc-TMEM184C.

As another example, black-tissue specific linc-FAM204A is syntenically conserved with the RAB11FIP2 and FAM204A genes in the human and mouse genomes (Figure 15A). This lncRNA was highly expressed in black tissues including the skin, shank, and comb but had no expression in other tissues except for the eye (Figure 15B). The co-expressed protein-coding genes are enriched for functionally relevant GO terms melanogenesis, ECM-receptor interaction, and Wnt signaling (Figure 15C). Interestingly, the human and Ogye lncRNA orthologs share a conserved sequence of 389 nt, which includes multiple miRNA 7-mer target sites (Figure 15A).

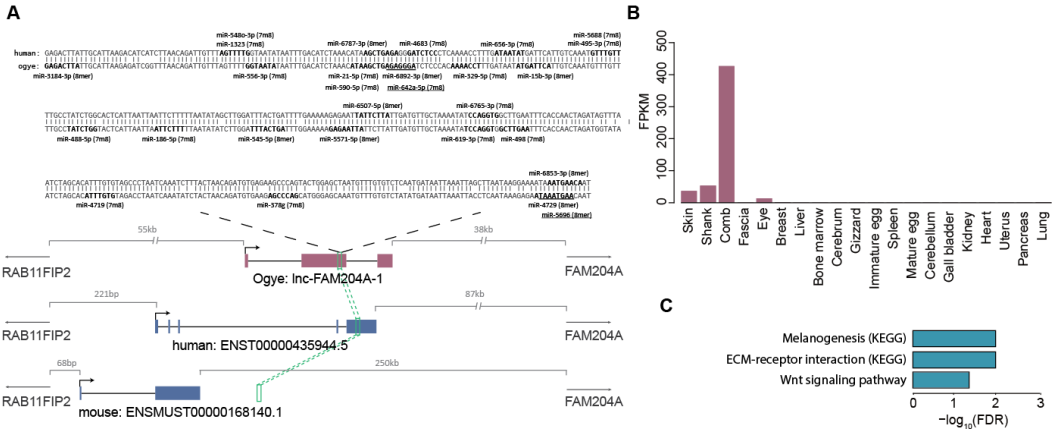

Figure 15. An example of a black skin-specific lncRNA with synteny and sequence conservation. (A) An example of an Ogye lncRNA (linc-FAM204A) that contains a sequence that is conserved in the exonic region (green box) of human lncRNA ENST00000435944.5 but not in the corresponding mouse gene. It is also syntenically conserved with sequences in both the mouse and human genome. 7-mer mRNA target sites in the conserved region are indicated in the sequences (top). (B) Expression pattern of linc-FAM204A. (C) GO terms that are significantly associated with protein-coding genes that are co-expressed with linc-FAM204A.

Discussion

In this study, 6900 multiple-exon lncRNAs were identified from twenty tissues of Ogye; about 18% had been previously annotated in *Galllus gallus red junglefowl*. The remainder of the previously annotated lncRNAs were mostly not expressed in Ogye or were false annotations, suggesting that the current chicken lncRNA annotations should be reconstructed more carefully. Our Ogye lncRNAs resembled previously annotated lncRNAs in mammals in their genomic characteristics, including transcript length, exon number, and tissue-specific expression pattern, providing evidence for the accuracy of the new annotations. Hence, the Ogye lncRNA catalogue may help us to improve lncRNA annotations in the chicken reference genome.

The majority of lncRNAs showed a tissue-specific expression pattern, defining functionally coherent co-expression clusters. The tissue-specific expression and the coherent expression of lncRNA genes with other protein-coding genes could be attributed to common epigenetic and transcriptional regulation. In fact, of the lncRNAs in clusters, 39.3% had evidence associating them with at least one model (Figure 16A); most commonly, these involved lncRNAs that act as epigenetic activators of protein-coding gene expression and common TFs that bind to the lncRNA and protein-coding gene promoters (Figure 16B). Interestingly, 126 lncRNAs had evidence supporting both the epigenetic activator and TF models (Table

2). 79 lncRNAs had functional evidence supporting their identity as eRNAs. Although lncRNAs are known to be mostly involved in epigenetic repression of genes, our study intentionally focused on lncRNAs as epigenetic activators by correlating the level of lncRNAs and the methylation in target gene promoters. Furthermore, because only a subset of CpG sites are sometimes related to the chromatin state and transcriptional activity of target genes, averaging CpG methylation signals in the promoter might underestimate the fraction of epigenetically activating lncRNAs in our study.

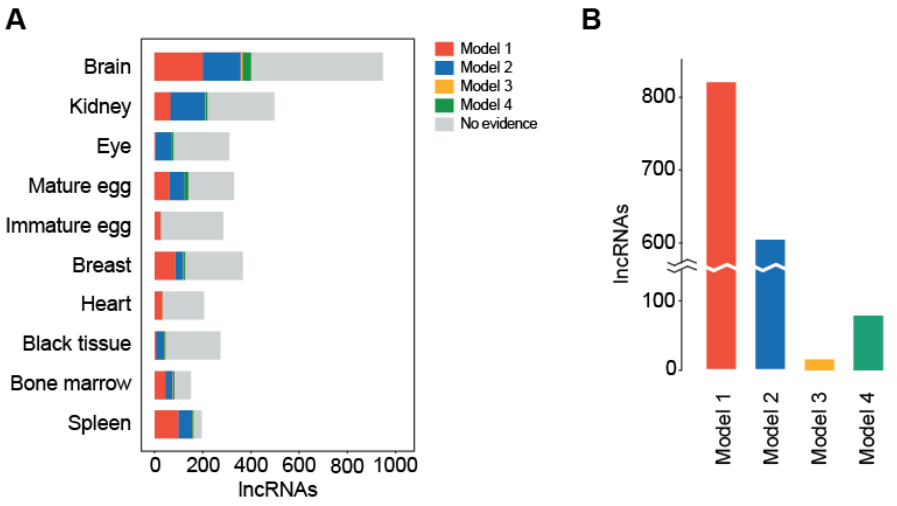

Figure 16. Proportions of lncRNAs that are explained by each functional model. (A) The numbers of lncRNAs, associated with specific clusters, with characteristics that are explained by the different co-regulation models are indicated in the stacked bar graphs (red for epigenetic activator (model1); blue for co-transcriptional regulation by TFs (model 2); yellow for epigenetic co-regulation (model 3); green for eRNA-mediated regulation (model 4); grey for no associated model. (B) The numbers of lncRNAs from all clusters with characteristics that are explained by each model.

Table 2. lncRNAs that are supported by more than two functional models

Model	count
model1 & model2 & model4	5
model1 & model2	126
model1 & model3	5
model1 & model4	15
model2 & model3	3
model2 & model4	8

lncRNA and protein-coding genes co-expressed in black tissues had HSF2 binding sites in their promoters and were specifically correlated with the level of HSF2 across tissues, supporting that the genes are co-regulated by HSF2. Moreover, enhancers that included HSF2 binding sites were associated with eRNAs specific to black tissue, indicating that HSF2 is the most likely regulator of black tissue-specific expression. Because the ancestor of Ogye appears to have originated in the rainforest, it makes sense that heat shock-related factors could be involved in melanogenesis and hyper-pigmentation processes, which would help avoid the absorption of too much heat. One of the black skin-specific lncRNAs, lnc-THMEM184c, is most abundantly expressed in comb, and HSF2 appears to co-regulate lnc-THMEM184c and its co-expressed protein-coding genes, which are related to keratinocyte differentiation and ECM-receptor interaction (Figure 14).

In addition, several previous studies that also focused on animal coat color showed that the color can be determined by the amount and type of

melanin produced and released by melanocytes present in the skin (Ito, Wakamatsu, and Ozeki 2000; Ito and Wakamatsu 2008). Melanin is produced by melanosomes, large organelles in melanocytes, in a process called melanogenesis. Wnt signaling has a regulatory role in the melanogenesis pathway and is also required for the developmental process that leads to melanocyte differentiation from neural crest cells (Dunn et al. 2000; Guo et al. 2016). One of the candidate lncRNAs related to the process is linc-FAM204A, whose co-expressed protein-coding genes are associated with GO terms melanogenesis, ECM-receptor interaction, and Wnt signaling pathway (Figure 15C). linc-FAM204A, which contains a short-conserved motif, is broadly preserved in mammalian genomes, including the human, rhesus macaque, mouse, dog, and elephant genomes. Among these orthologs, the human ortholog is known as CASC2, and is suppressed in lung, colorectal, renal and other cancers by miR-21-5p targeting via the conserved 7-mer site (Figure 15A).

Taken together, these results indicate that coding and non-coding RNAs functionally relevant to black and other tissues could help explain unique genomic and functional characteristics of a Korean domestic chicken breed, Yeonsan Ogye. Additionally, these findings could provide unprecedented

insight for future studies with industrial and agricultural applications, as well as for scientific analysis of chicken genomes.

Materials and Methods

Acquisition and care of Yeonsan Ogye

Yeonsan Ogye chicken (object number: 02127), obtained from the Animal Genetic Resource Research Center of the National Institute of Animal Science (Namwon, Korea), was used in the study. The care and experimental use of Ogye was reviewed and approved by the Institutional Animal Care and Use Committee of the National Institute of Animal Science (IACUC No.: 2014-080). Ogye management, treatment, and sample collection and further analysis of all raw data were performed at the National Institute of Animal Science.

Preparation of RNA-seq libraries

Total RNAs were extracted from twenty Ogye tissues using 80% EtOH and TRIzol. The RNA concentration was checked by Quant-IT RiboGreen (Invitrogen, Carlsbad, USA). To assess the integrity of the total RNA, samples were run on a TapeStation RNA screentape (Agilent, Waldbronn, Germany). Only high quality RNA samples (RIN $\geq$ 7.0) were used for RNA-seq library construction. Each library was independently prepared with 300ng of total RNA using an Illumina TruSeq Stranded Total RNA Sample Prep Kit (Illumina, San Diego, CA, USA). The rRNA in the total RNA was depleted using a Ribo-Zero kit. After rRNA depletion, the remaining RNA

was purified, fragmented and primed for cDNA synthesis. The cleaved RNA fragments were copied into the first cDNA strand using reverse transcriptase and random hexamers. This step was followed by second strand cDNA synthesis using DNA Polymerase I, RNase H and dUTP. The resulting cDNA fragments then underwent an end repair process, the addition of a single 'A' base, after which adapters were ligated. The products were purified and enriched with PCR to create the final cDNA library. The libraries were quantified using qPCR according to the qPCR Quantification Protocol Guide (KAPA Library Quantificatoin kits for Illumina Sequecing platforms) and qualified using the TapeStation D1000 ScreenTape assay (Agilent Technologies, Waldbronn, Germany).

Preparation of RRBS libraries

Preparation of reduced representation bisulfite sequencing (RRBS) libraries was done following Illumina's RRBS protocol. 5ug of genomic DNA that had been digested with the restriction enzyme MspI and purified with a QIAquick PCR purification kit (QIAGEN, Hilden, Germany) was used for library preparation, which was done using a TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, USA). Eluted DNA fragments were end-repaired, extended on the 3′ end with an 'A', and ligated with Truseq adapters. After ligation had been assessed, the products, which ranged

from 175 to 225bp in length (insert DNA of 55-105 bp plus adaptors of 120 bp), were excised from a 2%(w/v) Low Range Ultra Agarose gel (Biorad, Hercules, USA) and purified using the QIAquick gel extraction protocol. The purified DNA underwent bisulfite conversion using an EpiTect Bisulfite Kit (Qiagen, 59104). The bisulfite-converted DNA libraries were amplified by PCR (four cycles) using PfuTurbo Cx DNA polymerase (Agilent, 600410). The final product was then quantified using qPCR and qualified using the Agilent Technologies 2200 TapeStation assay (Agilent, Waldbronn, Germany). The final product was sequenced using the HiSeq™ 2500 platform (Illumina, San Diego, USA).

Annotations of protein-coding and lncRNA genes

To annotate protein-coding genes in the Ogye genome, *Gallus gallus* (red junglefowl) protein-coding genes downloaded from Ensembl biomart (release 81; http://www.ensembl.org/biomart) were mapped onto the Ogye draft genome v1.0 using GMAP (v2015-07-23)(Wu and Watanabe 2005). Genes that had greater than 90% coverage and identity were selected as Ogye protein-coding genes. As a result, 14,264 protein-coding genes were subjected to further analysis.

For lncRNA gene annotations, RNA-seq data from twenty different tissues (Breast, Liver, Bone marrow, Fascia, Cerebrum, Gizzard, Immature egg,

Comb, Spleen, Mature egg, Cerebellum, Gall bladder, Kidney, Heart, Uterus, Pancreas, Lung, Skin, Eye, and Shank) were produced in both single end and paired-end types. Sequences were preprocessed to filter nucleotides with low quality scores using FASTQC (v 0.10.1) (Andrews) and were mapped to the Ogye draft genome using STAR (v2.4.2)(Dobin et al. 2013) with the options '--runMode alignReads --alignIntronMin 67 --alignIntronMax 36873 -outReadsUnmapped Fastx -outFilterMismatch_ Nmax 999 --outFilterMismatchNoverLmax 0.02 -alignMatesGapMax 1000000 -outSAMtyp_ e BAM SortedByCoordinate --outFilterMultimapNmax 5 --outWigType wiggle -outWigStran_ d Stranded --outWigNorm RPM'. Initial transcriptome assemblies from twenty tissues were performed with Cufflinks (ver 2.1.1)(Trapnell et al. 2010) with the parameter '--library-type fr-firststrand' and the resulting assemblies were combined using Cuffmerge (ver 1.0.0)(Trapnell et al. 2010) with the default option. In total, 206,084 transcripts from 103,405 loci were annotated in the Ogye genome. To distinguish lncRNAs from other biotypes of RNAs, such as mRNAs, tRNAs, rRNAs, snoRNAs, miRNAs, and other small non-coding RNAs, these sequences were downloaded from ENSEMBL biomart and aligned with the resulting transcript sequences. Any transcripts overlapping at least 1nt with known RNAs were excluded. Of the remainder, those of less than 200nt in length and within 200 bp of protein-coding genes on the same strand were further

excluded to avoid the inclusion of fragmented RNAs. 54,760 lncRNA candidate loci (60257 transcripts) were retained and compared with a chicken lncRNA annotation of NONCODE (v2016) (Zhao et al. 2016). Of the candidates, 2094 loci (5215 transcripts) overlapped with previously annotated chicken lncRNAs. 52,666 non-overlapping loci (55,042 transcripts) were further examined to determine whether they had coding potential using coding potential calculator (CPC) scores (Kong et al. 2007). Those with a score greater than -1 were filtered out, and the remainder (14,108 novel lncRNA candidate loci without coding potential) were subjected to the next step. Because many candidates still appeared to be fragmented, those with a single exon but with neighboring candidates within 36,873bp, which is the intron length of the 99^{th} percentile, were re-examined using both exon-junction reads consistently presented over twenty tissues and the maximum entropy score (Yeo and Burge 2004), as done in our previous study (You, Yoon, and Nam 2017). If there were at least two junction reads spanning two neighboring transcripts or if the entropy score was greater than 4.66 in the interspace, two candidates were reconnected, and those with a single exon were discarded. In the final version, 9529 transcripts from 6900 lncRNA loci (5610 novel and 1290 known) were annotated as lncRNAs.

DNA methylation profiling

RRBS reads with a low quality score (*Phred quality score* < 20) were discarded using FastQC (v0.10.1). The remaining reads were aligned to the Ogye draft genome (v1.0) using Bismark (Krueger and Andrews 2011). The methylation level of each cytosine in a CpG region was calculated using Bismark methylation extractor. Tissue-specific, differentially methylated CpG sites (tDMC), covered with at least five reads in a promoter, were considered for downstream analysis. A tissue specific site is defined as one in which its mean methylation across tissues is at least four time greater than the signal in a certain tissue. A promoter region is defined as the region 2 kb upstream of the 5' end of genes.

Expression profiling

The expression values (FPKM) of lncRNA and protein-coding genes were estimated using RSEM (v1.2.25) in each tissue. The values across tissues were normalized using the quantile normalization method. In all downstream analyses, lncRNA or protein-coding genes with FPKM ≥ 1 in at least one tissue were used. lncRNAs for which the maximum expression value across twenty tissues was at least four-fold higher than the mean value were considered to exhibit tissue-specific expression. In total, 5,191 (75%)

lncRNAs were considered to be tissue-specific across twenty different tissues.

Hierarchical clustering of expressed lncRNAs across tissues

To perform hierarchical clustering of commonly expressed lncRNA genes among tissues, the list of expressed lncRNAs in each tissue was used as a input vector for phylogenetic clustering. The clustering was done using the PHYLIP package. lncRNAs with FPKM $\geq$ 1 in a certain tissue were considered to be expressed in a certain tissue. As two tissues share more common genes, they become more closely clustered.

Clustering of co-expressed lncRNAs

Hierarchical clustering was performed to search for expression clusters of lncRNAs across twenty tissues using Pearson's correlation coefficient metrics. Clusters in which more than 80% of their members are most highly expressed in the same or related tissues (brain and black tissues) were regarded as tissue-specific. Sub-clusters in the brain and black tissue clusters were further defined with the same criterion mentioned above.

Defining coding genes co-expressed with lncRNAs in a cluster

Protein-coding genes with a high mean correlation with lncRNAs in a cluster (Pearson's correlation ≥ 0.5), but for which the mean correlation to the cluster is at least 0.3 greater than those of other clusters, were assigned to the co-expressed set of the cluster. Each set of mRNAs was used to perform gene ontology (GO) term and pathway enrichment analyses using DAVID (Huang da, Sherman, and Lempicki 2009). Terms were only selected when the false discovery rate (FDR) q value was ≤ 0.05.

Correlation of the methylation level of neighboring lncRNA and protein-coding genes.

The methylation levels at CpG sites in the promoters of neighboring lncRNA and protein-coding genes were correlated with each other over twenty tissues (using Pearson's correlation coefficients). Only tissues in which a certain position had sufficient read coverage (at least five) were considered for measuring the correlation. If the nominal P value was ≤ 0.05, then the pair of lncRNA and protein-coding genes was considered as having a significantly correlated interaction.

Correlating the expression level of lncRNAs with the methylation level of protein-coding genes

To identify lncRNAs as potential epigenetic activators, the expression of lncRNAs and the methylation at CpG sites in the promoters of protein-coding genes were correlated over twenty tissues using a non-parametric correlation method (Spearman's correlation). Only pairs of lncRNA and protein-coding genes exhibiting a nominal P value $\leq$ 0.01 were considered as having a significantly correlated interaction. Of the resulting pairs, if the protein-coding mRNAs had a significant correlation (nominal P value $\leq$ 0.01) between their expression level and the methylation level in their promoter, its paired lncRNA was regarded as an epigenetic activator.

Prediction of TFBSs

To identify enriched TFBSs in the promoters of the co-expressed lncRNAs in each tissue-specific cluster and in the promoters of the co-expressed protein-coding genes within the cluster, the promoter sequences were examined using the MEME suite (V4.9.0). Motifs that exhibit an E-value $\leq$ 1×10^{-5} were selected as enriched motifs, associated with the corresponding tissue. The resulting motifs were searched for in the Transfac database (Wingender et al. 1997) using TomTom (Gupta et al. 2007). As a result, 14 TFBSs significantly enriched in a certain tissue or in a

set of similar tissues were detected ($P \leq 0.01$), of which 6 had associated TF orthologs (SP1, HEN1, HSF2, HB, AP-2, Oct1) encoded in the genome. However, HEN1 was not expressed in a corresponding tissue (FPKM ≤1). In addition, to confirm TFs related to enhancers, enhancer sequences were compared with the resulting TFBSs.

Identification of enhancer regions

To annotate enhancer regions in the Ogye draft genome, annotation files including all enhancers in the *Gallus gallus* (*red junglefowl*) genome were downloaded from the NCBI gene expression omnibus (GEO, GSE75480). Enhancer sequences extracted using our in-house script were aligned to the Ogye draft genome using BLAST (-p blastn). Regions that significantly matched the original enhancers (E-value ≤ 1 X 10^{-5}) and with high coverage of more than 80% were annotated as Ogye enhancers.

Transcriptional activity of eRNAs

To examine bi-directional transcriptional activity of eRNAs, total mapped reads in the range spanning 1kb upstream to 1 kb downstream of the eRNA transcription start site (TSS) were re-examined on both forward and reverse strands.

Correlation of expression between neighboring lncRNA and protein-coding genes

Pairs consisting of a lncRNA and its closest neighboring protein-coding gene within 10kb were classified into three groups based on their genomic orientations: head-to-head (can be divergently overlapped), head-to-tail (including only independent lncRNAs with evidence of a TSS and cleavage and polyadenylation site; otherwise, these lncRNAs must be at least 1kb apart from each other), and tail-to-tail (can be convergently overlapped). The correlation of the expression of these pairs was calculated over twenty tissues using Pearson's correlation method. The average correlation coefficient values and their standard errors were calculated in the respective groups. As a random control, the expression of 1000 random pairs of lncRNA and protein-coding genes were correlated using the same method. As another control, number-matched pairs of neighboring protein-coding genes were also correlated with each other.

Synteny and sequence conservation

To examine the conservation of synteny of a lncRNA, its closest downstream and upstream neighboring protein-coding genes in the Ogye genome were matched to their orthologous genes in the mouse and human genomes. If a lncRNA is located between the two orthologous genes, regardless of

direction, that lncRNA was regarded as syntenically conserved. GENCODE lncRNA annotations (v25 for human and vM11 for mouse) were analyzed for this study. To check for sequence conservation, Ogye lncRNA sequences were aligned to lncRNA sequences from other species, intronic sequences, and their flanking sequences (up to 4 Mb) using BLAST. For a significant match, an E-value 1 X 10^{-6} was used as a cutoff.

Analysis of lncRNA differential expression

To identify lncRNAs that are differentially expressed between Ogye and Brown leghorn skin tissues, Brown leghorn skin RNA-seq data were downloaded from the NCBI SRA (ERR1298635, ERR1298636, ERR1298637, ERR1298638, ERR1298639, ERR1298640, and ERR1298641). Reads were mapped to the *Gallus gallus* Galgal4 reference genome using Bowtie (V1.0.0), and the average mismatch rates were estimated across read positions. If the mismatch rate was greater than 0.1 at a certain position, sequences on high mismatch side of the position were trimmed using seqtk (https://github.com/lh3/seqtk), and then sickle was used with the default option for read quality control. Preprocessed reads from RNA-seq data were mapped onto the chicken Galgal4 reference genome using STAR. The read counts of lncRNAs were performed using HTSeq (v0.6.0) and the differential expression analysis was performed using DESeq (Anders and

Huber 2010). Genes with a greater than two-fold difference in expression and a FDR q value $\leq$ 0.05 were considered to be differentially expressed.

References

Al-Tobasei, Rafet, Bam Paneru, and Mohamed Salem. 2016. 'Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout', *PLoS ONE*, 11: e0148940.

Anders, S., and W. Huber. 2010. 'Differential expression analysis for sequence count data', *Genome Biol*, 11: R106.

Andrews, Simon. 'FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ '.

Bertani, S., S. Sauer, E. Bolotin, and F. Sauer. 2011. 'The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin', *Mol Cell*, 43: 1040-6.

Billerey, Coline, Mekki Boussaha, Diane Esquerré, Emmanuelle Rebours, Anis Djari, Cédric Meersseman, Christophe Klopp, Daniel Gautheret, and Dominique Rocha. 2014. 'Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing', *BMC Genomics*, 15: 499.

Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. 2011. 'Integrative annotation of human large

intergenic noncoding RNAs reveals global properties and specific subclasses', *Genes Dev*, 25: 1915-27.

Cech, T. R., and J. A. Steitz. 2014. 'The noncoding RNA revolution-trashing old rules to forge new ones', *Cell*, 157: 77-94.

Chalei, Vladislava, Stephen N. Sansom, Lesheng Kong, Sheena Lee, Juan F. Montiel, Keith W. Vance, and Chris P. Ponting. 2014. 'The long non-coding RNA Dali is an epigenetic regulator of neural differentiation', *Elife*, 3: e04530.

Cogill, S. B., and L. Wang. 2014. 'Co-expression Network Analysis of Human lncRNAs and Cancer Genes', *Cancer Inform*, 13: 49-59.

De Santa, F., I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C. L. Wei, and G. Natoli. 2010. 'A large fraction of extragenic RNA pol II transcription sites overlap enhancers', *PloS Biol*, 8: e1000384.

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo. 2012. 'The GENCODE v7 catalog of human long noncoding RNAs:

analysis of their gene structure, evolution, and expression', *Genome Res*, 22: 1775-89.

Dharmayanthi, Anik Budhi, Yohei Terai, Sri Sulandari, M. Syamsul Arifin Zein, Toyoko Akiyama, and Yoko Satta. 2017. 'The origin and evolution of fibromelanosis in domesticated chickens: Genomic comparison of Indonesian Cemani and Chinese Silkie breeds', *PLoS ONE*, 12: e0173147.

Di Ruscio, Annalisa, Alexander K. Ebralidze, Touati Benoukraf, Giovanni Amabile, Loyal A. Goff, Jolyon Terragni, Maria Eugenia Figueroa, Lorena Lobo De Figueiredo Pontes, Meritxell Alberich-Jorda, Pu Zhang, Mengchu Wu, Francesco D/'Alo, Ari Melnick, Giuseppe Leone, Konstantin K. Ebralidze, Sriharsa Pradhan, John L. Rinn, and Daniel G. Tenen. 2013. 'DNMT1-interacting RNAs block gene-specific DNA methylation', *Nature*, 503: 371-76.

Dimitrova, N., J. R. Zamudio, R. M. Jong, D. Soukup, R. Resnick, K. Sarma, A. J. Ward, A. Raj, J. T. Lee, P. A. Sharp, and T. Jacks. 2014. 'LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint', *Mol Cell*, 54: 777-90.

Dinger, M. E., P. P. Amaral, T. R. Mercer, K. C. Pang, S. J. Bruce, B. B. Gardiner, M. E. Askarian-Amiri, K. Ru, G. Solda, C. Simons, S. M.

Sunkin, M. L. Crowe, S. M. Grimmond, A. C. Perkins, and J. S. Mattick. 2008. 'Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation', *Genome Res*, 18: 1433-45.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras. 2012. 'Landscape of transcription in human cells', *Nature*, 489: 101-8.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29: 15-21.

Dorshorst, Ben, Anna-Maja Molin, Carl-Johan Rubin, Anna M. Johansson, Lina Strömstedt, Manh-Hung Pham, Chih-Feng Chen, Finn Hallböök, Chris Ashwell, and Leif Andersson. 2011. 'A Complex Genomic Rearrangement Involving the Endothelin 3 Locus Causes Dermal Hyperpigmentation in the Chicken', *PLOS Genetics*, 7: e1002412.

Dunn, Karen J, Bart O Williams, Yi Li, and William J Pavan. 2000. 'Neural crest-directed gene transfer demonstrates Wnt1 role in melanocyte expansion and differentiation during mouse development', *Proceedings of the National Academy of Sciences*, 97: 10050-55.

Faghihi, M. A., F. Modarresi, A. M. Khalil, D. E. Wood, B. G. Sahagan, T. E. Morgan, C. E. Finch, G. St Laurent, 3rd, P. J. Kenny, and C. Wahlestedt. 2008. 'Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase', *Nat Med*, 14: 723-30.

Felsenstein, Joseph. 1989. 'PHYLIP - Phylogeny Inference Package (Version 3.2)', *Cladistics*, 5: 164-66.

Feng, J., C. Bi, B. S. Clark, R. Mady, P. Shah, and J. D. Kohtz. 2006. 'The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator', *Genes Dev*, 20: 1470-84.

Ferre, F., A. Colantoni, and M. Helmer-Citterich. 2016. 'Revealing protein-lncRNA interaction', *Brief Bioinform*, 17: 106-16.

Garding, Angela, Nupur Bhattacharya, Rainer Claus, Melanie Ruppel, Cordula Tschuch, Katharina Filarsky, Irina Idler, Manuela Zucknick, Maïwen Caudron-Herger, and Christopher Oakes. 2013. 'Epigenetic upregulation of lncRNAs at 13q14. 3 in leukemia is linked to the In Cis downregulation of a gene cluster that targets NF-kB', *PLoS Genet*, 9: e1003373.

Ghosh, Sourav, Satish Sati, Shantanu Sengupta, and Vinod Scaria. 2015. 'Distinct patterns of epigenetic marks and transcription factor binding sites across promoters of sense-intronic long noncoding RNAs', *Journal of genetics*, 94: 17-25.

Gonzalez, I., R. Munita, E. Agirre, T. A. Dittmer, K. Gysling, T. Misteli, and R. F. Luco. 2015. 'A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature', *Nat Struct Mol Biol*, 22: 370-6.

Greco, S., M. Gorospe, and F. Martelli. 2015. 'Noncoding RNA in age-related cardiovascular diseases', *J Mol Cell Cardiol*, 83: 142-55.

Guo, Haiying, Yizhan Xing, Yingxin Liu, Yan Luo, Fang Deng, Tian Yang, Ke Yang, and Yuhong Li. 2016. 'Wnt/β-catenin signaling pathway activates melanocyte stem cells in vitro and in vivo', *Journal of Dermatological Science*, 83: 45-51.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. 2007. 'Quantifying similarity between motifs', *Genome Biology*, 8: R24.

Guttman, Mitchell, Ido Amit, Manuel Garber, Courtney French, Michael F. Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W. Carey, John P. Cassady, Moran N. Cabili, Rudolf Jaenisch, Tarjei S. Mikkelsen, Tyler Jacks, Nir Hacohen, Bradley E. Bernstein, Manolis Kellis, Aviv Regev, John L. Rinn, and Eric S. Lander. 2009. 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, 458: 223-27.

Hellwig, S., and B. L. Bass. 2008. 'A starvation-induced noncoding RNA modulates expression of Dicer-regulated genes', *Proc Natl Acad Sci U S A*, 105: 12897-902.

Hirota, K., T. Miyoshi, K. Kugou, C. S. Hoffman, T. Shibata, and K. Ohta. 2008. 'Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs', *Nature*, 456: 130-4.

Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009. 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat Protoc*, 4: 44-57.

Ito, S., and K. Wakamatsu. 2008. 'Chemistry of mixed melanogenesis--pivotal roles of dopaquinone', *Photochem Photobiol*, 84: 582-92.

Ito, S., K. Wakamatsu, and H. Ozeki. 2000. 'Chemical analysis of melanins and its application to the study of the regulation of melanogenesis', *Pigment Cell Res*, 13 Suppl 8: 103-9.

Iyer, M. K., Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, A. Poliakov, X. Cao, S. M. Dhanasekaran, Y. M. Wu, D. R. Robinson, D. G. Beer, F. Y. Feng, H. K. Iyer, and A. M. Chinnaiyan. 2015. 'The landscape of long noncoding RNAs in the human transcriptome', *Nat Genet*, 47: 199-208.

Jariwala, N., and D. Sarkar. 2016. 'Emerging role of lncRNA in cancer: a potential avenue in molecular medicine', *Ann Transl Med*, 4: 286.

Kim, Tae-Kyung, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley,

and Scott Kuersten. 2010. 'Widespread transcription at neuronal activity-regulated enhancers', *Nature*, 465: 182-87.

Kim, Tae-Kyung, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F. Worley, Gabriel Kreiman, and Michael E. Greenberg. 2010. 'Widespread transcription at neuronal activity-regulated enhancers', *Nature*, 465: 182-87.

Kong, Lei, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. 2007. 'CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine', *Nucleic Acids Research*, 35: W345-W49.

Krueger, Felix, and Simon R. Andrews. 2011. 'Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics*, 27: 1571-72.

Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel, and R. Shiekhattar. 2013. 'Activating RNAs associate with Mediator to enhance chromatin architecture and transcription', *Nature*, 494: 497-501.

Leucci, Eleonora, Roberto Vendramin, Marco Spinazzi, Patrick Laurette, Mark Fiers, Jasper Wouters, Enrico Radaelli, Sven Eyckerman, Carina

Leonelli, Katrien Vanderheyden, Aljosja Rogiers, Els Hermans, Pieter Baatsen, Stein Aerts, Frederic Amant, Stefan Van Aelst, Joost van den Oord, Bart de Strooper, Irwin Davidson, Denis L. J. Lafontaine, Kris Gevaert, Jo Vandesompele, Pieter Mestdagh, and Jean-Christophe Marine. 2016. 'Melanoma addiction to the long non-coding RNA SAMMSON', *Nature*, 531: 518-22.

Li, Bo, and Colin N Dewey. 2011. 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC bioinformatics*, 12: 323.

Li, Lingjie, and Howard Y. Chang. 2014. 'Physiological roles of long noncoding RNAs: insight from knockout mice', *Trends in Cell Biology*, 24: 594-602.

Li, Tingting, Suya Wang, Rimao Wu, Xueya Zhou, Dahai Zhu, and Yong Zhang. 2012. 'Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing', *Genomics*, 99: 292-98.

Li, Wenbo, Dimple Notani, and Michael G. Rosenfeld. 2016. 'Enhancers as non-coding RNA transcription units: recent insights and future perspectives', *Nat Rev Genet*, 17: 207-23.

Li, Yulin, Xuping Zhu, Liu Yang, Junying Li, Zhengxing Lian, Ning Li, and Xuemei Deng. 2011. 'Expression and network analysis of genes

related to melanocyte development in the Silky Fowl and White Leghorn embryos', *Molecular Biology Reports*, 38: 1433-41.

Liao, Q., C. Liu, X. Yuan, S. Kang, R. Miao, H. Xiao, G. Zhao, H. Luo, D. Bu, H. Zhao, G. Skogerbo, Z. Wu, and Y. Zhao. 2011. 'Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network', *Nucleic Acids Res*, 39: 3864-78.

Liu, X., D. Li, W. Zhang, M. Guo, and Q. Zhan. 2012. 'Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay', *EMBO J*, 31: 4415-27.

Lv, Yuanda, Zhikai Liang, Min Ge, Weicong Qi, Tifu Zhang, Feng Lin, Zhaohua Peng, and Han Zhao. 2016. 'Genome-wide identification and functional prediction of nitrogen-responsive intergenic and intronic long non-coding RNAs in maize (Zea mays L.)', *BMC Genomics*, 17: 350.

Machanick, Philip, and Timothy L. Bailey. 2011. 'MEME-ChIP: motif analysis of large DNA datasets', *Bioinformatics*, 27: 1696-97.

Marques, Ana C., Jim Hughes, Bryony Graham, Monika S. Kowalczyk, Doug R. Higgs, and Chris P. Ponting. 2013. 'Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs', *Genome Biology*, 14: R131.

Mercer, T. R., M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick. 2008. 'Specific expression of long noncoding RNAs in the mouse brain', *Proc Natl Acad Sci U S A*, 105: 716-21.

Merry, C. R., M. E. Forrest, J. N. Sabers, L. Beard, X. H. Gao, M. Hatzoglou, M. W. Jackson, Z. Wang, S. D. Markowitz, and A. M. Khalil. 2015. 'DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer', *Hum Mol Genet*, 24: 6240-53.

Mi, Huaiyu, Qing Dong, Anushya Muruganujan, Pascale Gaudet, Suzanna Lewis, and Paul D. Thomas. 2010. 'PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium', *Nucleic Acids Research*, 38: D204-D10.

Morris, Kevin V., and John S. Mattick. 2014. 'The rise of regulatory RNA', *Nat Rev Genet*, 15: 423-37.

Nam, J. W., and D. P. Bartel. 2012. 'Long noncoding RNAs in C. elegans', *Genome Res*, 22: 2529-40.

Orom, U. A., T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytnicki, C. Notredame, Q. Huang, R. Guigo, and R. Shiekhattar. 2010. 'Long noncoding RNAs with enhancer-like function in human cells', *Cell*, 143: 46-58.

Pang, K. C., M. E. Dinger, T. R. Mercer, L. Malquori, S. M. Grimmond, W. Chen, and J. S. Mattick. 2009. 'Genome-wide identification of long noncoding RNAs in CD8+ T cells', *J Immunol*, 182: 7738-48.

Pauli, A., E. Valen, M. F. Lin, M. Garber, N. L. Vastenhouw, J. Z. Levin, L. Fan, A. Sandelin, J. L. Rinn, A. Regev, and A. F. Schier. 2012. 'Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis', *Genome Res*, 22: 577-91.

Ponting, Chris P., Peter L. Oliver, and Wolf Reik. 2009. 'Evolution and Functions of Long Noncoding RNAs', *Cell*, 136: 629-41.

Quinn, Jeffrey J., and Howard Y. Chang. 2016. 'Unique features of long non-coding RNA biogenesis and function', *Nat Rev Genet*, 17: 47-62.

Ren, Hangxing, Gaofu Wang, Lei Chen, Jing Jiang, Liangjia Liu, Nianfu Li, Jinhong Zhao, Xiaoyan Sun, and Peng Zhou. 2016. 'Genome-wide analysis of long non-coding RNAs at early stage of skin pigmentation in goats (Capra hircus)', *BMC Genomics*, 17: 67.

Sahu, Anirban, Udit Singhal, and Arul M Chinnaiyan. 2015. 'Long noncoding RNAs in cancer: from function to translation', *Trends in cancer*, 1: 93-109.

Seki, R., C. Li, Q. Fang, S. Hayashi, S. Egawa, J. Hu, L. Xu, H. Pan, M. Kondo, T. Sato, H. Matsubara, N. Kamiyama, K. Kitajima, D. Saito, Y. Liu, M. T. Gilbert, Q. Zhou, X. Xu, T. Shiroishi, N. Irie, K.

Tamura, and G. Zhang. 2017. 'Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features', *Nat Commun*, 8: 14229.

Shinomiya, Ai, Yasunari Kayashima, Keiji Kinoshita, Makoto Mizutani, Takao Namikawa, Yoichi Matsuda, and Toyoko Akiyama. 2012. 'Gene duplication of endothelin 3 is closely correlated with the hyperpigmentation of the internal organs (Fibromelanosis) in silky chickens', *Genetics*, 190: 627-38.

Simon, Matthew D., Stefan F. Pinter, Rui Fang, Kavitha Sarma, Michael Rutenberg-Schoenberg, Sarah K. Bowman, Barry A. Kesner, Verena K. Maier, Robert E. Kingston, and Jeannie T. Lee. 2013. 'High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation', *Nature*, 504: 465-69.

Thomas, Paul D, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. 2003. 'PANTHER: a library of protein families and subfamilies indexed by function', *Genome research*, 13: 2129-41.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. 2010. 'Transcript assembly and quantification by RNA-Seq reveals unannotated

transcripts and isoform switching during cell differentiation', *Nat Biotechnol*, 28: 511-5.

Ulitsky, I., and D. P. Bartel. 2013. 'lincRNAs: genomics, evolution, and mechanisms', *Cell*, 154: 26-46.

Ulitsky, I., A. Shkumatava, C. H. Jan, H. Sive, and D. P. Bartel. 2011a. 'Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution', *Cell*, 147: 1537-50.

Ulitsky, Igor, Alena Shkumatava, Calvin H Jan, Hazel Sive, and David P Bartel. 2011b. 'Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution', *Cell*, 147: 1537-50.

Wang, Dong, Ivan Garcia-Bassets, Chris Benner, Wenbo Li, Xue Su, Yiming Zhou, Jinsong Qiu, Wen Liu, Minna U. Kaikkonen, Kenneth A. Ohgi, Christopher K. Glass, Michael G. Rosenfeld, and Xiang-Dong Fu. 2011. 'Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA', *Nature*, 474: 390-94.

Wang, X., S. Arai, X. Song, D. Reichart, K. Du, G. Pascual, P. Tempst, M. G. Rosenfeld, C. K. Glass, and R. Kurokawa. 2008. 'Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription', *Nature*, 454: 126-30.

Wang, Yueying, Songyi Xue, Xiaoran Liu, Huan Liu, Tao Hu, Xiaotian Qiu, Jinlong Zhang, and Minggang Lei. 2016. 'Analyses of long non-coding RNA and mRNA profiling using RNA sequencing during the pre-implantation phases in pig endometrium', *Scientific Reports*, 6.

Weikard, Rosemarie, Frieder Hadlich, and Christa Kuehn. 2013. 'Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing', *BMC Genomics*, 14: 789.

Wingender, E., A. E. Kel, O. V. Kel, H. Karas, T. Heinemeyer, P. Dietze, R. Knuppel, A. G. Romaschenko, and N. A. Kolchanov. 1997. 'TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation', *Nucleic Acids Res*, 25: 265-8.

Wu, Liang, Pierre Murat, Dijana Matak-Vinkovic, Adele Murrell, and Shankar Balasubramanian. 2013. 'Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins', *Biochemistry*, 52: 9519-27.

Wu, Thomas D, and Colin K Watanabe. 2005. 'GMAP: a genomic mapping and alignment program for mRNA and EST sequences', *Bioinformatics*, 21: 1859-75.

Yeo, G., and C. B. Burge. 2004. 'Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals', *J Comput Biol*, 11: 377-94.

You, B. H., S. H. Yoon, and J. W. Nam. 2017. 'High-confidence coding and noncoding transcriptome maps', *Genome Res*, 27: 1050-62.

Zalfa, F., M. Giorgi, B. Primerano, A. Moro, A. Di Penta, S. Reis, B. Oostra, and C. Bagni. 2003. 'The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses', *Cell*, 112: 317-27.

Zhang, Bin, Gayatri Arun, Yuntao S Mao, Zsolt Lazar, Gene Hung, Gourab Bhattacharjee, Xiaokun Xiao, Carmen J Booth, Jie Wu, and Chaolin Zhang. 2012. 'The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult', *Cell Reports*, 2: 111-23.

Zhang, Jianqin, Fuzhu Liu, Junting Cao, and Xiaolin Liu. 2015. 'Skin Transcriptome Profiles Associated with Skin Color in Chickens', *PLoS ONE*, 10: e0127301.

Zhao, J., B. K. Sun, J. A. Erwin, J. J. Song, and J. T. Lee. 2008. 'Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome', *Science*, 322: 750-6.

Zhao, Yi, Hui Li, Shuangsang Fang, Yue Kang, Wei wu, Yajing Hao, Ziyang Li, Dechao Bu, Ninghui Sun, Michael Q. Zhang, and Runsheng Chen. 2016. 'NONCODE 2016: an informative and valuable data source of long non-coding RNAs', *Nucleic Acids Research*, 44: D203-D08.

국문요지

연산 오계는 깃털, 피부, 빗, 눈, 정강이, 발톱 등 몸 전체가 검은색 외관을 가진 한국의 재래 닭이다. 닭의 검은색 외관에 대한 기존 연구들은 단백질 번역 유전자 (Protein-coding gene)에 제한적이고, 비 번역 유전자 (long non-coding RNA; lncRNA)와 관련된 연구는 진행된 바가 거의 없다. 본 연구에서는 연산 오계의 20개 조직에서 생산된 RNA sequencing (RNA-seq)과 reduced representation bisulfite sequencing (RRBS)를 이용하여 비 번역 전사체 지도를 작성하였다. 오계의 비 번역 전사체 지도는 1290개의 알려진 lncRNA와 5610개의 새로운 lncRNA를 포함한 6900개의 lncRNA로 구성되어 있으며, 이미 알려진 galllus gallus red junglefowl의 lncRNA의 상당수가 단백질 번역 유전자 조각이거나 오계의 20개 조직에서 발현되지 않음을 보였다. 본 연구에서 동정한 오계 lncRNA의 상당 수가 조직 특이적인 발현 양상을 보였고, 조직 특이적인 lncRNA의 약 39%가 기능적 증거를 보였다. 특히, HSF2에 의해 조절되는 lncRNA가 검은 피부 조직에 특이적으로 발현하는 단백질 번역 유전자와 기능적으로 연관되어 있었고, 포유류에서 synteny가 보존되는 경향이 있었다. 또한, 흰 피부 조직과 비교 하였을 때 검은 피부 조직에서 차별적으로 발현되는 것을 확인 하였다.

따라서, 본 연구는 종합적인 lncRNA catalogue를 제공해 줄 뿐만 아니라, 독특한 표현형을 조절하는 비 번역 유전체를 이해하는데 도움을 줄 것이다.