Non-coding Transcriptome Maps Across Twenty Tissues of the Korean Black Chicken, Yeonsan Ogye

Hyosun Hong[1,6], Han-Ha Chai[3,4,6], Kyoungwoo Nam[1], Dajeong Lim[3], Kyung-Tai Lee[3], Yoon Jung Do[3], Chang-Yeon Cho[5] and Jin-Wu Nam[1,2,*]

[1]Department of Life Science, College of Natural Sciences, Hanyang University, Seoul 133791, Republic of Korea

[2]Research Institute for Convergence of Basic Sciences, Hanyang University, Seoul 133791, Republic of Korea

[3]Department of Animal Biotechnology & Environment of National Institute of Animal Science, RDA, Wanju 55365, Republic of Korea

[4]College of Pharmacy, Chonnam National University, Kwangju 61186, Republic of Korea

[5]Animal Genetic Resource Research Center of National Institute of Animal Science, RDA, Namwon 55717, Republic of Korea

[6]These authors contributed equally to this paper.

*Corresponding author

Tel: +82-2-2220-2428, Fax: +82-2-2298-0319, Email: jwnam@hanyang.ac.kr

Abstract

The Yeonsan Ogye (Ogye) is a rare chicken breed that populates the Korean peninsula. The entire body of this bird, including its feathers and skin, has a unique black coloring. Although some protein-coding genes related to this unique feature have been examined, non-coding elements have not been 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 Ogye tissues. The resulting maps included 6900 long non-coding RNA (lncRNA) genes comprising 1290 known and 5610 novel lncRNA genes. Compared to lncRNAs previously annotated in the *Galllus gallus red junglefowl*, a considerable number were either fragments of protein-coding genes or not expressed in Ogye tissues. Newly annotated Ogye lncRNA genes showed tissue-specific expression and simple gene structures containing 2 or 3 exons. Systematic analyses of sequencing data and other genomic data demonstrated that about 39% of the tissue-specific lncRNAs displayed evidence of function. In particular, heat shock transcription factor 2 (HSF2)-associated lncRNAs were discovered to be functionally linked to protein-coding genes that are specifically expressed in black skin tissues, tended to be more syntenically conserved in mammals, and were differentially expressed in black tissues relative to 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 how the non-coding genome regulates unique phenotypes. Furthermore, our results should be of use for future genomic breeding of chickens.

Keywords

lncRNA, Chicken, Black tissues, Tissue-specific genes, Non-coding transcriptome

Background

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 [1]. Ogye shares common features—such as black plumage, skin, shank, and fascia—with Ayam Cemani [1], 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 [2]. Several genes involved in Silkie skin hyperpigmentation have been reported in previous studies [2-4]. Recently, transcriptomes from Chinese native black chickens were compared with those from white chickens to globally identify hyperpigmentation-related genes [5]. 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 [6-8]. 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 [9-11]. 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 [12].

lncRNAs are also known to regulate gene expression at other levels: transcriptional, post-transcriptional, translational, and post-translational [9, 10, 13-15]. 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 [16-20]. 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 [21-23]. eRNA regulatory roles are known to be achieved via several mechanisms: trapping transcription factors, directing chromatin roofing, and inducing DNA methylation [9, 24-28]. 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 [29].

Despite their regulatory roles, only a few lncRNAs are highly conserved across vertebrates [30]. lncRNAs generally exhibit either poor conservation at the nucleotide level or conservation in a short region only, particularly compared to protein-coding genes [31-33]. 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 [34-36]. 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 [37]. 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 [31, 38, 39] but tend to be uniquely or specifically expressed in distinct tissues, developmental stages, conditions, or disease states [30-32, 38, 40-42]. For instance, one lncRNA, SAMMSON, is specifically expressed in melanoma cells during melanogenesis and is known to regulate the process at the epigenetic level [43]. 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 [44, 45]. 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 [46, 47]. 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 [48-50].

Although genome and transcriptome maps of livestock animals, such as rainbow trout, cow, goat, and chicken [51-55], 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.*

Data description

Total RNA samples and Bisulphite-treated DNA samples were collected 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) from 8-month-old Ogye (Fig. 1a). About 1.5 billion RNA-seq reads (843 million single-end reads and 638 million paired-end reads) and 123 million

reduced representation bisulfite sequencing (RRBS) reads were analyzed (Table 1-2). The RNA-seq and RRBS data from this study have been submitted to NCBI Gene Expression Omnibus under SuperSeries accession number GSE104358.

Analyses

A comprehensive Ogye lncRNA catalogue

To construct an Ogye transcriptome map, 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) [56], and subjected to transcriptome assembly using Cufflinks (ver 2.1.1) [57], 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 (see the "Annotation of protein-coding and lncRNA genes" section in the Methods for more details). 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 [58] (Additional file 1: Figure S1a). Compared to previously annotated chicken lncRNAs from *Gallus gallus*, only 34% were redetected in the Ogye lncRNA catalogue. Because approximately 92% previously annotated chicken protein-coding genes were redetected in our Ogye protein-coding gene annotations (data not shown here), the remainder that are missing may not result from genome mis-assembly. 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 (Additional file 1: Figure S1b). Only 276 previously annotated lncRNAs were missed in our Ogye lncRNA annotations. Consistent with other species [39, 51, 52, 55], the median gene

length and the median exon number of Ogye lncRNAs were less than those of protein-coding genes (Additional file 1: Figure S1c-d).

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) [59]. 6,565 lncRNA genes were expressed with FPKM ≥ 1 in at least one tissue, whereas 13,765 protein-coding genes were. As previously reported [32, 60, 61], 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 (Additional file 1: Figure S1e). 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 (Fig. 1c, green track). About 75% of lncRNA genes (5191 loci) were tissue-specific, a significantly higher proportion than that of protein-coding genes (45%; Fig. 1d; $P < 2.2e^{-16}$; 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) (Fig. 1e). Hierarchical clustering of commonly expressed lncRNA genes among tissues using the PHYLIP package (ver 3.6) [62] (see the “Hierarchical clustering of expressed lncRNAs across tissues” section in the Methods for more details) 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 (Fig. 1e). Only 780 lncRNAs were ubiquitously expressed across all tissues (Fig. 1e).

Tissue-specific DNA methylation landscape of the Ogye genome

To correlate tissue-specific lncRNA expression with its epigenetic status in the respective tissue, reduced representation bisulfite sequencing (RRBS) reads were mapped to the Ogye draft genome (Fig. 1b; Table2). The DNA methylation signals (C to T changes in CpGs)

across chromosomes were, then, calculated using Bismark in each sample (version 0.7.0) [63] (see the "DNA methylation profiling" section in the Methods for more details). 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 (Fig. 1c). 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 (Additional file 1: Figure S1f; $P = 1.30 \times 10^{-6}$ for lncRNAs; $P = 7.93 \times 10^{-36}$ 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 (Additional file 1: Figure S1g; $P = 0.87$ 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.

Co-expression analyses of lncRNAs specify tissue-specific functional clusters

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 (Additional file 1: Figure S2). 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) (Additional file 1: Figure S2a) than those of protein-coding genes (Additional file 1: Figure S2b).

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 (Fig. 2). 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) (Additional file 4: Table S3). 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 (Fig. 2). 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 (Additional file 1: Figure S 2c).

The functional role of each co-expressed lncRNA cluster can be indirectly inferred by a set of co-expressed mRNAs [48-50]. Thus, mRNAs that are exclusively co-expressed with each lncRNA cluster were identified with the following criteria: a mean Pearson's correlation ($\bar{r}$) ≥ 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 ≥ 0.3, and were subsequently subjected to gene ontology (GO) analyses using DAVID [64] (Fig. 2; Additional file 5:Table S4). In particular, 1617 mRNAs exclusively correlated to the brain-specific lncRNA group (930 lncRNAs) were identified and had been associated with brain-function specific terms, such as neuroactive ligand-receptor interaction (q = 2.18 X 10^{-12}; False discovery rate, FDR correction). In contrast, 748 mRNAs exclusively correlated to spleen-specific lncRNAs were identified and had been associated with immune-related terms, such as leukocyte activation (q = 2.37X 10^{-12}). Likewise, 10 out of 16 co-expression clusters of

lncRNAs had functional evidence, with significantly enriched GO terms and KEGG pathways (Fig. 2).

lncRNAs as epigenetic activators

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* (Additional file 1: Figure S3). 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 [65, 66] or demethylases that open the chromatin structure to enhance the expression of target mRNAs [67, 68] (Additional file 1: Figure S3a). Remote co-expression of lncRNAs and mRNAs can be also regulated by common TFs [44, 45] (Additional file 1: Figure S3b). 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 (Additional file 1: Figure S3c) or enhancers (Additional file 1: Figure S3d).

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 (Fig. 3a). Clusters specific to brain, kidney, mature egg, breast, heart and spleen included significantly more lncRNAs with a significant correlation than did the random controls ($P = 0.026 \sim 7.71 \times 10^{-13}$)

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 (Fig. 3b). 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.

Transcriptional regulation by common TFs

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 [69] (see the “Prediction of TFBSs” section in the Methods for more details). The resulting motifs were subjected to analysis by the TOMTOM program [70] to annotate TFBSs based on TRANSFAC database v3.2 [71]. As a result, 14 common TFs that have significantly abundant binding sites in the promoters of lncRNA and protein-coding genes were detected (Additional file 1: Figure S4; corresponding to model 2). To discern TFs available in chicken genomes, PANTHER [72, 73] was used to examine whether there are chicken orthologs of the TFs and whether the orthologs are expressed in the corresponding tissues (FPKM ≥ 1). Finally, five TFs, including HSF2 and SP1, were identified as candidates (Fig. 4a). 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 (Figs. 4b).

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 (Fig. 4c). 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 (Fig. 4d). 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.

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 [35, 38]. 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) (Fig. 5a). Genes encoding co-expressed pairs of lncRNAs and mRNAs are significantly proximally co-localized within 10kb (Fig. 5a left; $P \leq 0.05$, Fisher's exact test), compared to random controls (Fig. 5a 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 (Fig. 5b) or from ubiquitously expressed gene sets (Additional file 1: Figure S5a). Both neighboring lncRNA and protein-coding gene pairs displayed significantly greater correlation than did random controls, regardless of the category, in both sets (Fig. 5b; Additional file 1: Figure S5a). 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 (Additional file 1: Figure S5a), the correlation coefficients of the tissue-specific pairs were slightly yet insignificantly higher than those of neighboring protein-coding gene pairs (Fig. 5b).

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 (Fig. 5c; $P = 0.71$, Wilcoxon rank sum test), whereas they did show a significantly higher correlation in the methylation-unrelated group (Fig. 5d; $P \leq 0.001$ 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.

Enhancer-associated RNA-mediated gene regulation

Previous studies showed that lncRNAs associated with enhancers could regulate their neighboring protein-coding genes [74]. Genomic association between lncRNAs and enhancers, detected in embryonic developmental stages in the chicken [75], revealed that

lncRNAs in the methylation-unrelated group are more significantly associated with enhancers than those in the other group (Fig. 5e; $P = 2.72 \times 10^{-6}$; 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; Supplementary Table S5). The eRNAs (corresponding to model 4) had a greater correlation with neighboring protein-coding genes only in the head-to-tail group (Fig. 5f), whereas non-eRNAs displayed a greater correlation in the head-to-head orientation, which could allow sharing of common promoters (Fig. 5g). A few eRNAs were discovered to have strong bi-directional transcriptional activity (Additional file 1: Figure S5b; see the "Transcriptional activity of eRNAs" section in the Methods for more details), as previously reported [76, 77]

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 (Fig. 5h). 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 (Fig. 5h). Interestingly, HSF2 was a common TF candidate for co-regulating lncRNAs and protein-coding genes at a distance (Fig. 5d).

Conserved black skin-specific 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 (Fig. 2) 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

375 those in brown leghorn skin (Fig. 6a; Additional file 7:Table S6; FDR < 0.05).

376 To find functionally conserved lncRNAs, the 47 differentially expressed lncRNAs
377 were examined for synteny and sequence conservation in human and mouse genomes.
378 Synteny conservation considers whether orthologs of a certain lncRNA's neighboring genes
379 are positionally conserved in these mammalian genomes (Fig. 6b). As a result of this
380 analysis, about 10% of lncRNAs were found to be syntenically conserved in both the human
381 and mouse genomes and about 25% were syntenically conserved in at least one genome
382 (Fig. 6c; Additional file 8:Table S7), percentages that are comparable to those of the protein-
383 coding genes (Fig. 6d). However, sequence similarity analyses by the BLAST showed that
384 only 6% of the syntenically conserved lncRNAs had conserved sequences relative to
385 sequences in either the human or mouse genomes (Fig. 6c; Additional file 8:Table S7), which
386 is much lower than that of protein-coding genes (56%). Taken together, our data showed that
387 16 lncRNAs were syntenically or sequentially conserved and differentially expressed in black
388 tissue (Fig. 6e).

389 Of the 16 lncRNAs that have evidence of black tissue-specific function, four,
390 including eRNAs, were associated with HSF2 binding motifs, whereas of the 104 that have
391 synteny and sequence conservation but are not differentially expressed in black tissues, only
392 one was associated with HSF2. The presence of HSF2 binding motifs appears to be
393 significantly related to black tissue-specific expression (Fig 6f; $P \leq 0.0008$, Fisher's exact
394 test). For instance, linc-THEM184c is significantly up-regulated in black tissue (Fig. 7b), its
395 locus is syntenically conserved with neighboring genes, TMEM184C and EDNRA, in both
396 human and mouse genomes, and its promoter includes a HSF2 binding motif (Fig. 7a). In
397 addition, the protein-coding genes that are co-expressed with this lncRNA are enriched for
398 GO terms that are functionally relevant for black skin: keratinocyte differentiation,
399 angiogenesis, and ECM-receptor-interaction (Fig. 7c). Among the co-expressed genes, 31

have HSF2 binding sites in their promoters (Fig. 7a). As another example, black-tissue specific linc-FAM204A is syntenically conserved with the RAB11FIP2 and FAM204A genes in the human and mouse genomes (Additional file 1: Figure S6a). 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 (Additional file 1: Figure S6b). The co-expressed protein-coding genes are enriched for functionally relevant GO terms melanogenesis, ECM-receptor interaction, and Wnt signaling (Additional file 1: Figure S6c). Interestingly, the human and Ogye lncRNA orthologs share a conserved sequence of 389 nt, which includes multiple miRNA 7-mer target sites (Additional file 1: Figure S6a).

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 (Additional file 1: Figure S7a); 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 (Additional file 1: Figure S7b). Interestingly, 126 lncRNAs had evidence supporting both the epigenetic activator and TF models (Additional file 1: Figure S7c). 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.

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 (Fig. 7).

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 [78, 79]. 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 [80, 81]. 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 (Additional file 1: Figure S6c). 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 (Additional file 1: Figure S6a).

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.

Methods

Acquisition and care of Yeonsan Ogye

Yeonsan Ogye chickens (object number: 02127), obtained from the Animal Genetic Resource Research Center of the National Institute of Animal Science (Namwon, Korea), were 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

473 further analysis of all raw data were performed at the National Institute of Animal Science.

474 **Preparation of RNA-seq libraries**

475 Total RNAs were extracted from twenty Ogye tissues using 80% EtOH and TRIzol. The RNA
476 concentration was checked by Quant-IT RiboGreen (Invitrogen, Carlsbad, USA). To assess
477 the integrity of the total RNA, samples were run on a TapeStation RNA screentape (Agilent,
478 Waldbronn, Germany). Only high quality RNA samples (RIN ≥ 7.0) were used for RNA-seq
479 library construction. Each library was independently prepared with 300ng of total RNA using
480 an Illumina TruSeq Stranded Total RNA Sample Prep Kit (Illumina, San Diego, CA, USA).
481 The rRNA in the total RNA was depleted using a Ribo-Zero kit. After rRNA depletion, the
482 remaining RNA was purified, fragmented and primed for cDNA synthesis. The cleaved RNA
483 fragments were copied into the first cDNA strand using reverse transcriptase and random
484 hexamers. This step was followed by second strand cDNA synthesis using DNA Polymerase
485 I, RNase H and dUTP. The resulting cDNA fragments then underwent an end repair process,
486 the addition of a single 'A' base, after which adapters were ligated. The products were
487 purified and enriched with PCR to create the final cDNA library. The libraries were quantified
488 using qPCR according to the qPCR Quantification Protocol Guide (KAPA Library
489 Quantificatoin kits for Illumina Sequecing platforms) and qualified using the TapeStation
490 D1000 ScreenTape assay (Agilent Technologies, Waldbronn, Germany).

491 **Preparation of RRBS libraries**

492 Preparation of reduced representation bisulfite sequencing (RRBS) libraries was done
493 following Illumina's RRBS protocol. 5ug of genomic DNA that had been digested with the
494 restriction enzyme MspI and purified with a QIAquick PCR purification kit (QIAGEN, Hilden,
495 Germany) was used for library preparation, which was done using a TruSeq Nano DNA
496 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)[82]. 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) [83] and were mapped to the Ogye draft genome using STAR (v2.4.2)[56] with the options ‘--runMode alignReads --alignIntronMin 67 --alignIntronMax 36873 –outReadsUnmapped Fastx –outFilterMismatch_Nmax 999 --outFilterMismatchNoverLmax 0.02 -alignMatesGapMax 1000000 –outSAMtyp_

522 e BAM SortedByCoordinate --outFilterMultimapNmax 5 --outWigType wiggle –outWigStran_
523 d Stranded --outWigNorm RPM'. Initial transcriptome assemblies from twenty tissues were
524 performed with Cufflinks (ver 2.1.1)[57] with the parameter '--library-type fr-firststrand' and
525 the resulting assemblies were combined using Cuffmerge (ver 1.0.0)[57] with the default
526 option. In total, 206,084 transcripts from 103,405 loci were annotated in the Ogye genome.
527 To distinguish lncRNAs from other biotypes of RNAs, such as mRNAs, tRNAs, rRNAs,
528 snoRNAs, miRNAs, and other small non-coding RNAs, these sequences were downloaded
529 from ENSEMBL biomart and aligned with the resulting transcript sequences. Any transcripts
530 overlapping at least 1nt with known RNAs were excluded. Of the remainder, those of less
531 than 200nt in length and within 200 bp of protein-coding genes on the same strand were
532 further excluded to avoid the inclusion of fragmented RNAs. 54,760 lncRNA candidate loci
533 (60257 transcripts) were retained and compared with a chicken lncRNA annotation of
534 NONCODE (v2016) [84]. Of the candidates, 2094 loci (5215 transcripts) overlapped with
535 previously annotated chicken lncRNAs. 52,666 non-overlapping loci (55,042 transcripts)
536 were further examined to determine whether they had coding potential using coding potential
537 calculator (CPC) scores [85]. Those with a score greater than -1 were filtered out, and the
538 remainder (14,108 novel lncRNA candidate loci without coding potential) were subjected to
539 the next step. Because many candidates still appeared to be fragmented, those with a single
540 exon but with neighboring candidates within 36,873bp, which is the intron length of the 99th
541 percentile, were re-examined using both exon-junction reads consistently presented over
542 twenty tissues and the maximum entropy score [86], as done in our previous study [58]. If
543 there were at least two junction reads spanning two neighboring transcripts or if the entropy
544 score was greater than 4.66 in the interspace, two candidates were reconnected, and those
545 with a single exon were discarded. In the final version, 9529 transcripts from 6900 lncRNA
546 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 [63]. 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 ≥ 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 $\geq$ 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 [64]. Terms were only selected when the false discovery rate (FDR) q value was $\leq$ 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 $\leq$ 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 ≤ 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 ≤ 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 ≤ 1 X 10^{-5} were selected as enriched motifs, associated with the corresponding tissue. The resulting motifs were searched for in the Transfac database [71] using TomTom [70]. As a result, 14 TFBSs significantly enriched in a certain tissue or in a set of similar tissues were detected (*P* ≤ 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 [87]. Genes with a greater than two-fold difference in expression and a FDR q value ≤ 0.05 were considered to be differentially expressed.

Tables

Table 1. Sequencing and mapping summary of RNA-seq reads

Samples	Paired-end sequencing reads			Single-end sequencing reads		
	Sequencing reads	Mapping rate	SRA accession	Sequencing reads	Mapping rate	SRA accession

Breast	34,893,064	92.05%	SRX3223583	43,294,022	90.70%	SRX3223603
Liver	33,476,266	85.75%	SRX3223584	48,032,813	85.81%	SRX3223604
Bone marrow	30,975,506	85.00%	SRX3223585	40,286,974	87.99%	SRX3223605
Fascia	33,316,764	84.61%	SRX3223586	42,425,452	87.93%	SRX3223606
Cerebrum	30,887,821	89.95%	SRX3223587	46,455,658	92.32%	SRX3223607
Gizzard	31,537,118	84.00%	SRX3223588	38,689,871	85.82%	SRX3223608
Immature egg	32,009,437	87.73%	SRX3223589	32,048,703	87.80%	SRX3223609
Comb	31,936,332	85.34%	SRX3223590	37,985,049	87.76%	SRX3223610
Spleen	28,946,777	89.70%	SRX3223591	38,704,448	89.33%	SRX3223611
Mature egg	30,873,699	91.98%	SRX3223592	40,650,664	92.17%	SRX3223612
Cerebellum	30,798,145	93.53%	SRX3223593	39,940,946	93.34%	SRX3223613
Gall bladder	35,862,229	84.83%	SRX3223594	35,423,339	87.06%	SRX3223614
Kidney	29,953,007	87.25%	SRX3223595	39,894,009	89.99%	SRX3223615
Heart	30,986,431	94.14%	SRX3223596	45,951,338	91.49%	SRX3223616
Uterus	33,444,002	91.89%	SRX3223597	46,650,355	90.63%	SRX3223617
Pancreas	30,595,568	82.52%	SRX3223598	47,361,192	84.35%	SRX3223618
Lung	31,533,498	87.63%	SRX3223599	45,552,982	92.34%	SRX3223619
Skin	34,442,464	82.36%	SRX3223600	41,934,970	84.00%	SRX3223620
Eye	33,006,509	89.21%	SRX3223601	44,044,630	91.82%	SRX3223621
Shank	28,643,334	94.07%	SRX3223602	47,716,995	79.86%	SRX3223622

Table 2. **Sequencing and mapping summary of RRBS reads**

Samples	Sequencing reads	Mapping rate	SRA accession
Breast	6,042,106	68.90%	SRX3223667
Liver	6,744,208	74.20%	SRX3223668
Bone marrow	5,736,011	72.00%	SRX3223669
Fascia	5,720,194	68.90%	SRX3223670
Cerebrum	6,078,989	70.00%	SRX3223671
Gizzard	5,731,878	69.40%	SRX3223672
Immature egg	6,741,258	67.70%	SRX3223673
Comb	5,948,687	72.90%	SRX3223674
Spleen	6,307,517	77.60%	SRX3223675
Mature egg	6,246,607	69.20%	SRX3223676
Cerebellum	6,291,610	68.20%	SRX3223677
Gall bladder	5,738,180	70.10%	SRX3223678
Kidney	5,470,502	68.60%	SRX3223679
Heart	5,462,739	69.40%	SRX3223680
Uterus	6,046,764	67.90%	SRX3223681
Pancreas	7,100,215	70.30%	SRX3223682
Lung	5,640,120	67.60%	SRX3223683
Skin	7,226,309	72.40%	SRX3223684
Eye	6,956,141	71.90%	SRX3223685
Shank	5,924,463	74.20%	SRX3223686

Availability of supporting data and materials

Raw RNA-seq and RRBS data from twenty different Ogye tissues have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/; GSE104351) under SuperSeries accession number GSE104358. All lncRNA catalogs and expression

tables from this study have also been submitted to NCBI GEO (GSE104351) under the same SuperSeries accession number, GSE104358.

Declarations

Consent for publication

Not applicable

Competing interests

Not applicable

Funding

This work was supported by the Program for Agriculture Science & Technology Development (Project No. PJ01045301; PJ01045303) of the Rural Development Administration.

Author's contributions

JWN and HHC conceived and supervised the project. CYC provided materials. DJL, KTL, and YJD performed RNA-seq experiments. HSH performed the analyses. JWN and HSH wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank all members of the BIG lab for helpful comments and discussions.

Figure legends

Figure 1. Comprehensive coding and non-coding transcriptome maps of Yeonsan Ogye. **a.**

Yeonsan Ogye. **b.** A schematic flow for the analyses of coding and non-coding transcriptomes and DNA methylation from twenty different tissues. **c.** A 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). **d.** 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). **e.** 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 $\geq$ 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 and the enrichment of the tissue-specific genes was tested using Fisher's exact test (* $P \leq 1 \times e^{-5}$, ** $P \leq 1 \times e^{-10}$, *** $P \leq 1 \times e^{-20}$). The scale bar represents 10.0, which is the unit of 120 differentially expressed genes across tissues.

Figure 2. Co-expression clusters of lncRNAs and functional annotations. 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. Sub-clusters in the clusters are indicated where appropriate. The number in each cluster indicates the number of lncRNAs in the cluster and the number in the boxes with functional terms indicates the number of co-expressed protein-coding genes.

Figure 3. 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.

Figure 4. 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 ($P \leq 0.05$). **e.** Expression pattern of HSF2 and its target genes that have the top 5 correlations with HSF2.

Figure 5. 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 ≤ 10kb group, orange for the ≤ 100kb group, and green for the > 100kb or on another chromosome group) (left). *, **, and *** indicate $P \leq 0.05$, ≤ 0.01, and ≤ 0.001, respectively. **b.** The average correlation coefficients of tissue-specific lncRNA and protein-coding gene pairs in close neighborhoods (≤ 10kb) 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 (≤ 10kb) are shown with green bars. *, **, and *** indicate $P \leq 0.05$, ≤ 0.01, and ≤ 0.001, respectively. Error bars indicate the standard error. The number in the bars indicates the number of analyzed pairs. **c.** 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**. **d.** The average correlation coefficients of tissue-specific lncRNA and protein-coding genes (methylation-unrelated), except for those of **c**. Otherwise, as in **b**. **e.** The proportion of eRNAs (red) in the methylation-related group (**c**) and -unrelated group (**d**). ** indicates $P \leq 0.01$. **f.** The average correlation coefficients of tissue-specific eRNAs.

Otherwise, as in **b**. **g.** The average correlation coefficients of tissue-specific lncRNAs not associated with enhancers. Otherwise, as in **b**. **h.** 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.

Figure 6. 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.

Figure 7. An example of black skin-specific lncRNAs 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.

Additional files

Additional file 1: Supplementary Figures S1-S7.

Additional file 2: Table S1. Expression levels of all lncRNAs across twenty tissues.

Additional file 3: Table S2. Tissue-specific lncRNAs and ubiquitously expressed lncRNAs.

Additional file 4: Table S3. lncRNAs corresponding to each cluster (16 clusters).

Additional file 5: Table S4. Enriched biological process GO terms and KEGG pathways for protein-coding genes co-expressed with each cluster (FDR ≤ 0.05).**Additional file 6:** Table S5. lncRNAs associated with enhancers and neighboring protein-coding genes in terms of their genomic orientations: head-to-tail, tail-to-tail, head-to-head.

Additional file 7: Table S6. lncRNAs that are differentially expressed between Ogye and Brown leghorn skin.

Additional file 8: Table S7. Synteny-conserved lncRNAs and sequence-conserved lncRNAs.

References

1. Dharmayanthi A.B., Terai Y., Sulandari S., Zein M.S.A., Akiyama T.,Satta Y., The origin and evolution of fibromelanosis in domesticated chickens: Genomic comparison of Indonesian Cemani and Chinese Silkie breeds. PLOS ONE, 2017: 12(4); e0173147.
2. Dorshorst B., Molin A.-M., Rubin C.-J., Johansson A.M., Strömstedt L., Pham M.-H., et al., A Complex Genomic Rearrangement Involving the Endothelin 3 Locus Causes Dermal Hyperpigmentation in the Chicken. PLOS Genetics, 2011: 7(12); e1002412.
3. Shinomiya A., Kayashima Y., Kinoshita K., Mizutani M., Namikawa T., Matsuda Y., et al., Gene duplication of endothelin 3 is closely correlated with the hyperpigmentation of the internal organs (Fibromelanosis) in silky chickens. Genetics, 2012: 190(2); 627-638.
4. Li Y., Zhu X., Yang L., Li J., Lian Z., Li N., et al., Expression and network analysis of genes related to melanocyte development in the Silky Fowl and White Leghorn embryos. Molecular Biology Reports, 2011: 38(2); 1433-1441.
5. Zhang J., Liu F., Cao J.,Liu X., Skin Transcriptome Profiles Associated with Skin Color in Chickens. PLOS ONE, 2015: 10(6); e0127301.
6. Ponting C.P., Oliver P.L.,Reik W., Evolution and Functions of Long Noncoding RNAs. Cell, 2009: 136(4); 629-641.
7. Simon M.D., Pinter S.F., Fang R., Sarma K., Rutenberg-Schoenberg M., Bowman S.K., et al., High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature, 2013: 504(7480); 465-469.
8. Morris K.V.,Mattick J.S., The rise of regulatory RNA. Nat Rev Genet, 2014: 15(6); 423-437.
9. Wang X., Arai S., Song X., Reichart D., Du K., Pascual G., et al., Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature, 2008: 454(7200); 126-30.
10. Hellwig S.,Bass B.L., A starvation-induced noncoding RNA modulates expression of Dicer-regulated genes. Proc Natl Acad Sci U S A, 2008: 105(35); 12897-902.
11. Ferre F., Colantoni A.,Helmer-Citterich M., Revealing protein-lncRNA interaction. Brief Bioinform, 2016: 17(1); 106-16.
12. Merry C.R., Forrest M.E., Sabers J.N., Beard L., Gao X.H., Hatzoglou M., et al., DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer. Hum Mol Genet, 2015: 24(21); 6240-53.
13. Hirota K., Miyoshi T., Kugou K., Hoffman C.S., Shibata T.,Ohta K., Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature, 2008: 456(7218); 130-4.
14. Zalfa F., Giorgi M., Primerano B., Moro A., Di Penta A., Reis S., et al., The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses. Cell, 2003: 112(3); 317-27.
15. Liu X., Li D., Zhang W., Guo M.,Zhan Q., Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. Embo j, 2012: 31(23); 4415-27.
16. Ulitsky I.,Bartel D.P., lincRNAs: genomics, evolution, and mechanisms. Cell, 2013: 154(1); 26-46.
17. Garding A., Bhattacharya N., Claus R., Ruppel M., Tschuch C., Filarsky K., et al., 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, 2013: 9(4); e1003373.
18. Zhang B., Arun G., Mao Y.S., Lazar Z., Hung G., Bhattacharjee G., et al., The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell reports, 2012: 2(1); 111-123.
19. Jariwala N.,Sarkar D., Emerging role of lncRNA in cancer: a potential avenue in molecular medicine. Ann Transl Med, 2016: 4(15); 286.
20. Sahu A., Singhal U.,Chinnaiyan A.M., Long noncoding RNAs in cancer: from function to translation. Trends in cancer, 2015: 1(2); 93-109.
21. Kim T.-K., Hemberg M., Gray J.M., Costa A.M., Bear D.M., Wu J., et al., Widespread transcription at neuronal activity-regulated enhancers. Nature, 2010: 465(7295); 182-187.
22. Marques A.C., Hughes J., Graham B., Kowalczyk M.S., Higgs D.R.,Ponting C.P., Chromatin

signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biology, 2013: 14(11); R131.

23. De Santa F., Barozzi I., Mietton F., Ghisletti S., Polletti S., Tusi B.K., et al., A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol, 2010: 8(5); e1000384.

24. Orom U.A., Derrien T., Beringer M., Gumireddy K., Gardini A., Bussotti G., et al., Long noncoding RNAs with enhancer-like function in human cells. Cell, 2010: 143(1); 46-58.

25. Lai F., Orom U.A., Cesaroni M., Beringer M., Taatjes D.J., Blobel G.A., et al., Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature, 2013: 494(7438); 497-501.

26. Feng J., Bi C., Clark B.S., Mady R., Shah P.,Kohtz J.D., The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev, 2006: 20(11); 1470-84.

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

28. Dimitrova N., Zamudio J.R., Jong R.M., Soukup D., Resnick R., Sarma K., et al., LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol Cell, 2014: 54(5); 777-90.

29. Gonzalez I., Munita R., Agirre E., Dittmer T.A., Gysling K., Misteli T., et al., A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature. Nat Struct Mol Biol, 2015: 22(5); 370-6.

30. Mercer T.R., Dinger M.E., Sunkin S.M., Mehler M.F.,Mattick J.S., Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A, 2008: 105(2); 716-21.

31. Derrien T., Johnson R., Bussotti G., Tanzer A., Djebali S., Tilgner H., et al., The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res, 2012: 22(9); 1775-89.

32. Cabili M.N., Trapnell C., Goff L., Koziol M., Tazon-Vega B., Regev A., et al., Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011: 25(18); 1915-27.

33. Quinn J.J.,Chang H.Y., Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet, 2016: 17(1); 47-62.

34. Djebali S., Davis C.A., Merkel A., Dobin A., Lassmann T., Mortazavi A., et al., Landscape of transcription in human cells. Nature, 2012: 489(7414); 101-8.

35. Ulitsky I., Shkumatava A., Jan C.H., Sive H.,Bartel D.P., Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell, 2011: 147(7); 1537-50.

36. Li L.,Chang H.Y., Physiological roles of long noncoding RNAs: insight from knockout mice. Trends in Cell Biology, 2014: 24(10); 594-602.

37. Ulitsky I., Shkumatava A., Jan Calvin H., Sive H.,Bartel David P., Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell, 2011: 147(7); 1537-1550.

38. Nam J.W.,Bartel D.P., Long noncoding RNAs in C. elegans. Genome Res, 2012: 22(12); 2529-40.

39. Pauli A., Valen E., Lin M.F., Garber M., Vastenhouw N.L., Levin J.Z., et al., Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res, 2012: 22(3); 577-91.

40. Iyer M.K., Niknafs Y.S., Malik R., Singhal U., Sahu A., Hosono Y., et al., The landscape of long noncoding RNAs in the human transcriptome. Nat Genet, 2015: 47(3); 199-208.

41. Dinger M.E., Amaral P.P., Mercer T.R., Pang K.C., Bruce S.J., Gardiner B.B., et al., Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res, 2008: 18(9); 1433-45.

42. Faghihi M.A., Modarresi F., Khalil A.M., Wood D.E., Sahagan B.G., Morgan T.E., et al., Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med, 2008: 14(7); 723-30.

43. Leucci E., Vendramin R., Spinazzi M., Laurette P., Fiers M., Wouters J., et al., Melanoma addiction to the long non-coding RNA SAMMSON. Nature, 2016: 531(7595); 518-522.

44. Ghosh S., Sati S., Sengupta S.,Scaria V., Distinct patterns of epigenetic marks and transcription factor binding sites across promoters of sense-intronic long noncoding RNAs. Journal of genetics, 2015: 94(1); 17-25.
45. Guttman M., Amit I., Garber M., French C., Lin M.F., Feldser D., et al., Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009: 458(7235); 223-227.
46. Pang K.C., Dinger M.E., Mercer T.R., Malquori L., Grimmond S.M., Chen W., et al., Genome-wide identification of long noncoding RNAs in CD8+ T cells. J Immunol, 2009: 182(12); 7738-48.
47. Liao Q., Liu C., Yuan X., Kang S., Miao R., Xiao H., et al., Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res, 2011: 39(9); 3864-78.
48. Lv Y., Liang Z., Ge M., Qi W., Zhang T., Lin F., et al., Genome-wide identification and functional prediction of nitrogen-responsive intergenic and intronic long non-coding RNAs in maize (Zea mays L.). BMC Genomics, 2016: 17(1); 350.
49. Wang Y., Xue S., Liu X., Liu H., Hu T., Qiu X., et al., Analyses of long non-coding RNA and mRNA profiling using RNA sequencing during the pre-implantation phases in pig endometrium. Scientific reports, 2016: 6.
50. Cogill S.B.,Wang L., Co-expression Network Analysis of Human lncRNAs and Cancer Genes. Cancer Inform, 2014: 13(Suppl 5); 49-59.
51. Weikard R., Hadlich F.,Kuehn C., Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. BMC Genomics, 2013: 14(1); 789.
52. Billerey C., Boussaha M., Esquerré D., Rebours E., Djari A., Meersseman C., et al., Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. BMC Genomics, 2014: 15(1); 499.
53. Li T., Wang S., Wu R., Zhou X., Zhu D.,Zhang Y., Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. Genomics, 2012: 99(5); 292-298.
54. Ren H., Wang G., Chen L., Jiang J., Liu L., Li N., et al., Genome-wide analysis of long non-coding RNAs at early stage of skin pigmentation in goats (Capra hircus). BMC Genomics, 2016: 17(1); 67.
55. Al-Tobasei R., Paneru B.,Salem M., Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. PLOS ONE, 2016: 11(2); e0148940.
56. Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013: 29(1); 15-21.
57. Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J., et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010: 28(5); 511-5.
58. You B.H., Yoon S.H.,Nam J.W., High-confidence coding and noncoding transcriptome maps. Genome Res, 2017: 27(6); 1050-1062.
59. Li B.,Dewey C.N., RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics, 2011: 12(1); 323.
60. Greco S., Gorospe M.,Martelli F., Noncoding RNA in age-related cardiovascular diseases. J Mol Cell Cardiol, 2015: 83; 142-55.
61. Cech T.R.,Steitz J.A., The noncoding RNA revolution-trashing old rules to forge new ones. Cell, 2014: 157(1); 77-94.
62. Felsenstein J., PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics, 1989: 5; 164-166.
63. Krueger F.,Andrews S.R., Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 2011: 27(11); 1571-1572.
64. Huang da W., Sherman B.T.,Lempicki R.A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009: 4(1); 44-57.
65. Wu L., Murat P., Matak-Vinkovic D., Murrell A.,Balasubramanian S., Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins. Biochemistry, 2013: 52(52); 9519-9527.
66. Zhao J., Sun B.K., Erwin J.A., Song J.J.,Lee J.T., Polycomb proteins targeted by a short

repeat RNA to the mouse X chromosome. Science, 2008: 322(5902); 750-6.
67. Chalei V., Sansom S.N., Kong L., Lee S., Montiel J.F., Vance K.W., et al., The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. eLife, 2014: 3; e04530.
68. Di Ruscio A., Ebralidze A.K., Benoukraf T., Amabile G., Goff L.A., Terragni J., et al., DNMT1-interacting RNAs block gene-specific DNA methylation. Nature, 2013: 503(7476); 371-376.
69. Machanick P.,Bailey T.L., MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics, 2011: 27(12); 1696-1697.
70. Gupta S., Stamatoyannopoulos J.A., Bailey T.L.,Noble W.S., Quantifying similarity between motifs. Genome Biology, 2007: 8(2); R24.
71. Wingender E., Kel A.E., Kel O.V., Karas H., Heinemeyer T., Dietze P., et al., TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. Nucleic Acids Res, 1997: 25(1); 265-8.
72. Mi H., Dong Q., Muruganujan A., Gaudet P., Lewis S.,Thomas P.D., PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Research, 2010: 38(suppl_1); D204-D210.
73. Thomas P.D., Campbell M.J., Kejariwal A., Mi H., Karlak B., Daverman R., et al., PANTHER: a library of protein families and subfamilies indexed by function. Genome research, 2003: 13(9); 2129-2141.
74. Li W., Notani D.,Rosenfeld M.G., Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet, 2016: 17(4); 207-223.
75. Seki R., Li C., Fang Q., Hayashi S., Egawa S., Hu J., et al., Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features. Nat Commun, 2017: 8; 14229.
76. Kim T.-K., Hemberg M., Gray J.M., Costa A.M., Bear D.M., Wu J., et al., Widespread transcription at neuronal activity-regulated enhancers. Nature, 2010: 465(7295); 182-187.
77. Wang D., Garcia-Bassets I., Benner C., Li W., Su X., Zhou Y., et al., Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature, 2011: 474(7351); 390-394.
78. Ito S., Wakamatsu K.,Ozeki H., Chemical analysis of melanins and its application to the study of the regulation of melanogenesis. Pigment Cell Res, 2000: 13 Suppl 8; 103-9.
79. Ito S.,Wakamatsu K., Chemistry of mixed melanogenesis--pivotal roles of dopaquinone. Photochem Photobiol, 2008: 84(3); 582-92.
80. Dunn K.J., Williams B.O., Li Y.,Pavan W.J., Neural crest-directed gene transfer demonstrates Wnt1 role in melanocyte expansion and differentiation during mouse development. Proceedings of the National Academy of Sciences, 2000: 97(18); 10050-10055.
81. Guo H., Xing Y., Liu Y., Luo Y., Deng F., Yang T., et al., Wnt/β-catenin signaling pathway activates melanocyte stem cells in vitro and in vivo. Journal of Dermatological Science, 2016: 83(1); 45-51.
82. Wu T.D.,Watanabe C.K., GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics, 2005: 21(9); 1859-1875.
83. Andrews S., FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
84. Zhao Y., Li H., Fang S., Kang Y., wu W., Hao Y., et al., NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Research, 2016: 44(D1); D203-D208.
85. Kong L., Zhang Y., Ye Z.-Q., Liu X.-Q., Zhao S.-Q., Wei L., et al., CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research, 2007: 35(suppl_2); W345-W349.
86. Yeo G.,Burge C.B., Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol, 2004: 11(2-3); 377-94.
87. Anders S.,Huber W., Differential expression analysis for sequence count data. Genome Biol, 2010: 11(10); R106.

Figure 1

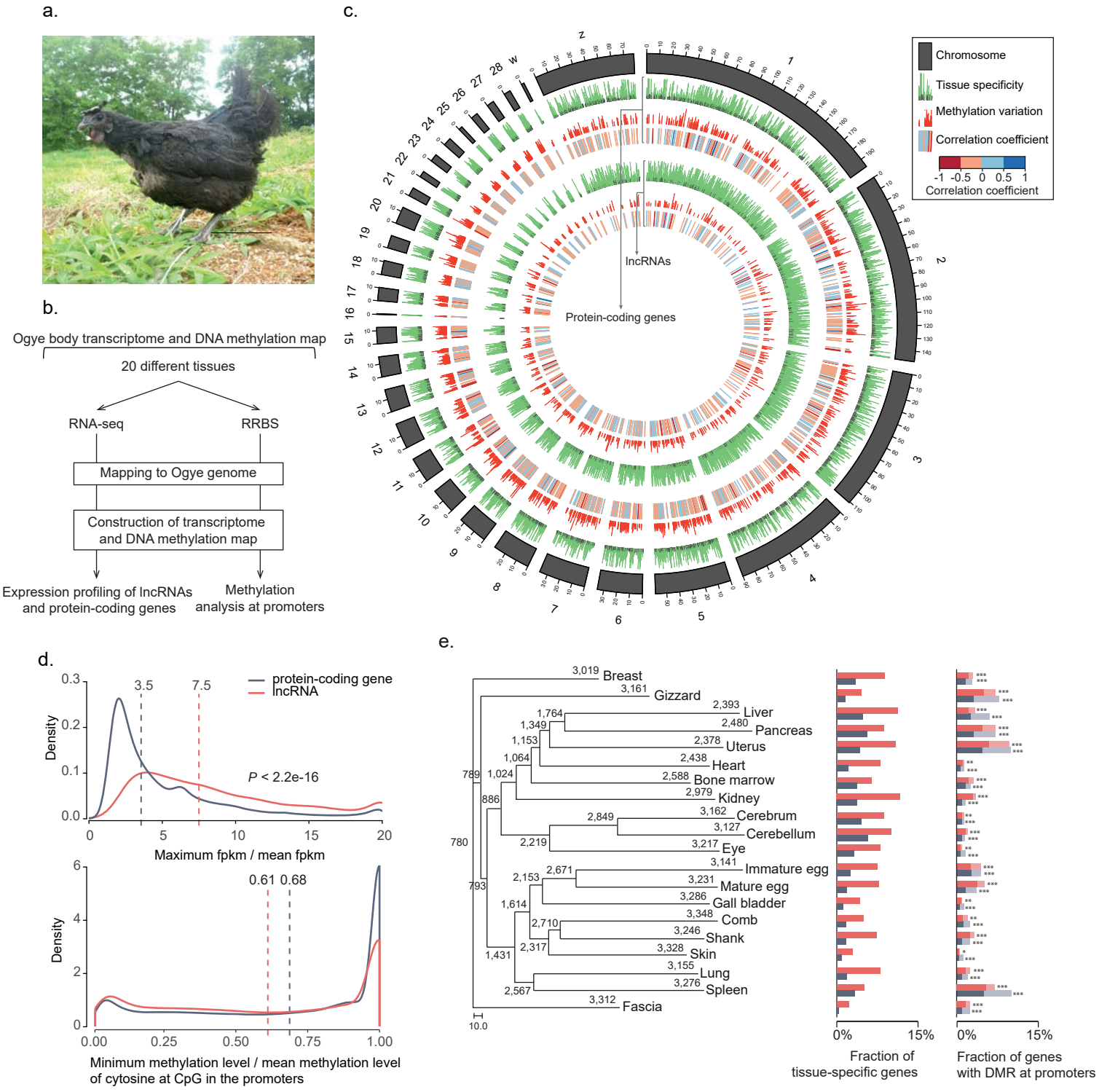

a.
b.
Ogye body transcriptome and DNA methylation map
20 different tissues
RNA-seq
RRBS
Mapping to Ogye genome
Construction of transcriptome and DNA methylation map
Expression profiling of lncRNAs and protein-coding genes
Methylation analysis at promoters
c.
Chromosome
Tissue specificity
Methylation variation
Correlation coefficient
-1 -0.5 0 0.5 1
Correlation coefficient
lncRNAs
Protein-coding genes
d.
protein-coding gene
lncRNA
3.5
7.5
P < 2.2e-16
Density
Maximum fpkm / mean fpkm
0.61 0.68
Minimum methylation level / mean methylation level of cytosine at CpG in the promoters
e.
Breast
Gizzard
Liver
Pancreas
Uterus
Heart
Bone marrow
Kidney
Cerebrum
Cerebellum
Eye
Immature egg
Mature egg
Gall bladder
Comb
Shank
Skin
Lung
Spleen
Fascia
10.0
0%
15%
Fraction of tissue-specific genes
Fraction of genes with DMR at promoters

Figure 2

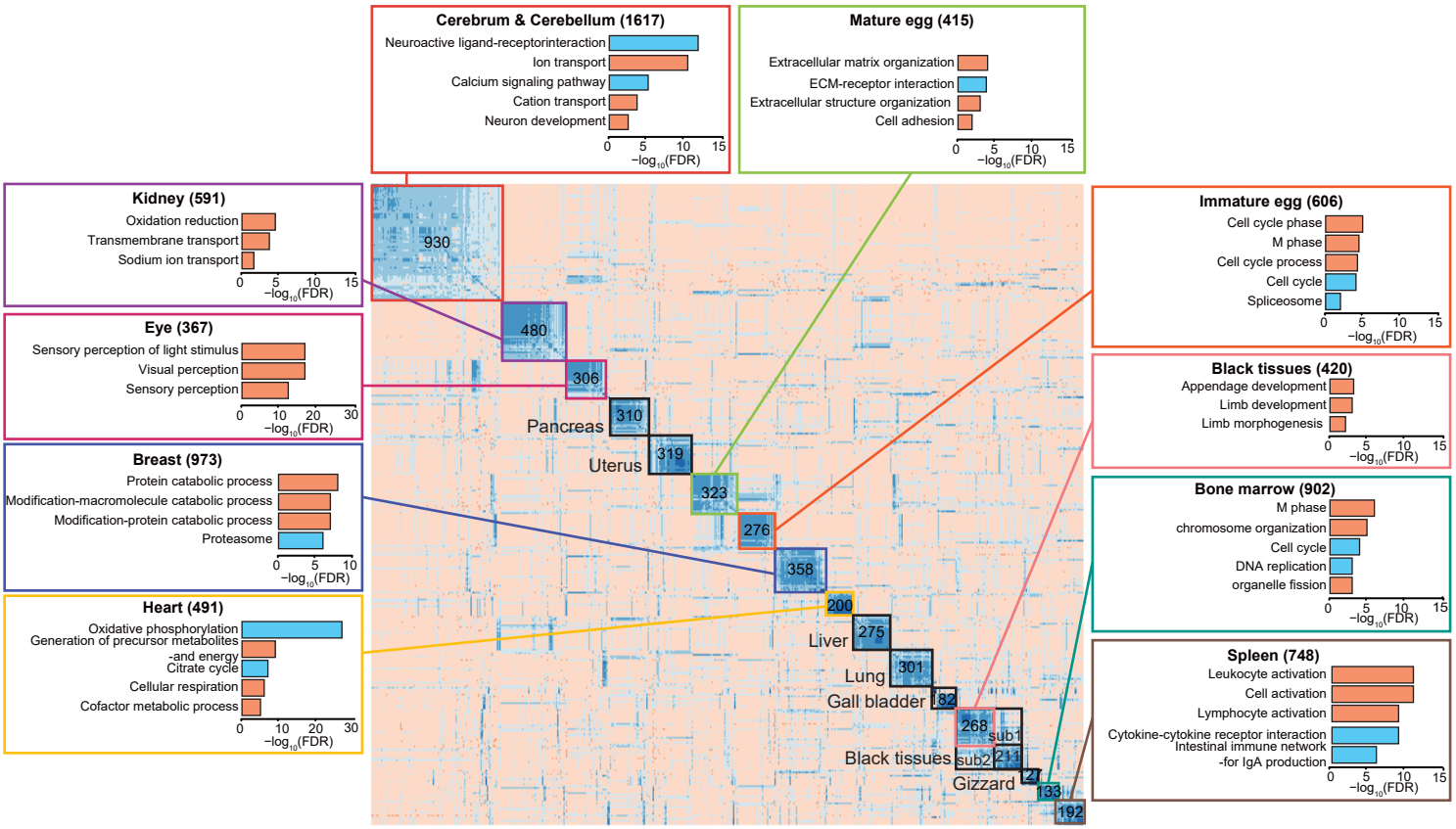

Cerebrum & Cerebellum (1617)
Neuroactive ligand-receptorinteraction
Ion transport
Calcium signaling pathway
Cation transport
Neuron development
0 5 10 15
−log10(FDR)
Mature egg (415)
Extracellular matrix organization
ECM-receptor interaction
Extracellular structure organization
Cell adhesion
0 5 10 15
−log10(FDR)
Kidney (591)
Oxidation reduction
Transmembrane transport
Sodium ion transport
0 5 10 15
−log10(FDR)
Eye (367)
Sensory perception of light stimulus
Visual perception
Sensory perception
0 10 20 30
−log10(FDR)
Breast (973)
Protein catabolic process
Modification-macromolecule catabolic process
Modification-protein catabolic process
Proteasome
0 5 10
−log10(FDR)
Heart (491)
Oxidative phosphorylation
Generation of precursor metabolites
-and energy
Citrate cycle
Cellular respiration
Cofactor metabolic process
0 10 20 30
−log10(FDR)
930
480
306
310
Pancreas
319
Uterus
323
276
358
200
275
Liver
301
Lung
Gall bladder
182
268
sub1
211
Black tissues
sub2
127
Gizzard
133
192
Immature egg (606)
Cell cycle phase
M phase
Cell cycle process
Cell cycle
Spliceosome
0 5 10 15
−log10(FDR)
Black tissues (420)
Appendage development
Limb development
Limb morphogenesis
0 5 10 15
−log10(FDR)
Bone marrow (902)
M phase
chromosome organization
Cell cycle
DNA replication
organelle fission
0 5 10 15
−log10(FDR)
Spleen (748)
Leukocyte activation
Cell activation
Lymphocyte activation
Cytokine-cytokine receptor interaction
Intestinal immune network
-for IgA production
0 5 10 15
−log10(FDR)

Figure 3

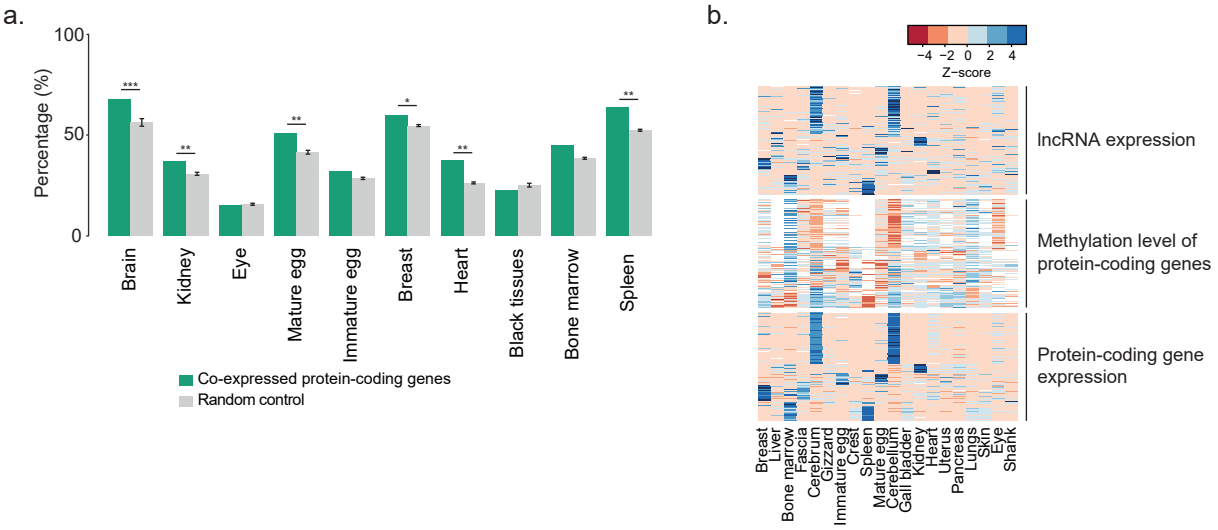
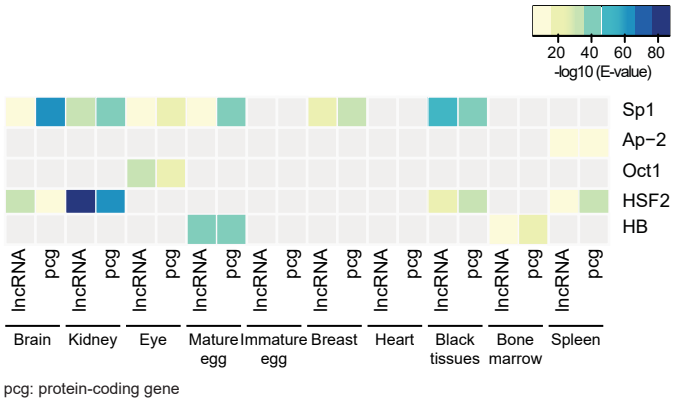
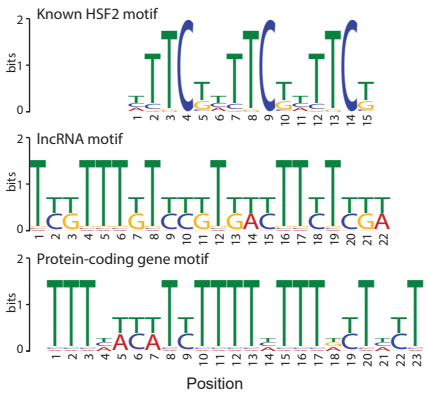

a.
100
50
0
Percentage (%)
***
**
**
*
**
**
Brain
Kidney
Eye
Mature egg
Immature egg
Breast
Heart
Black tissues
Bone marrow
Spleen
Co-expressed protein-coding genes
Random control
b.
−4 −2 0 2 4
Z−score
lncRNA expression
Methylation level of protein-coding genes
Protein-coding gene expression
Breast
Liver
Bone marrow
Fascia
Cerebrum
Gizzard
Immature egg
Crest
Spleen
Mature egg
Cerebellum
Gall bladder
Kidney
Heart
Uterus
Pancreas
Lungs
Skin
Eye
Shank

Figure 4

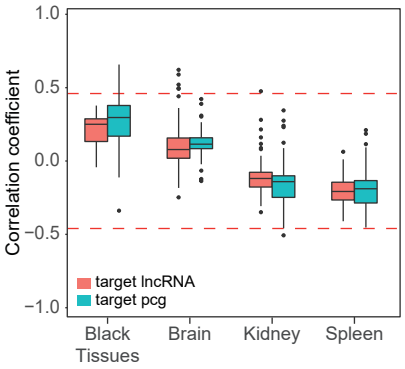
a.



b.



c.



d.

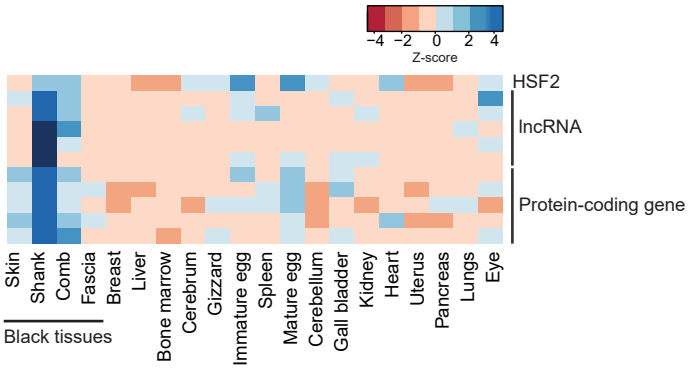


Figure 5

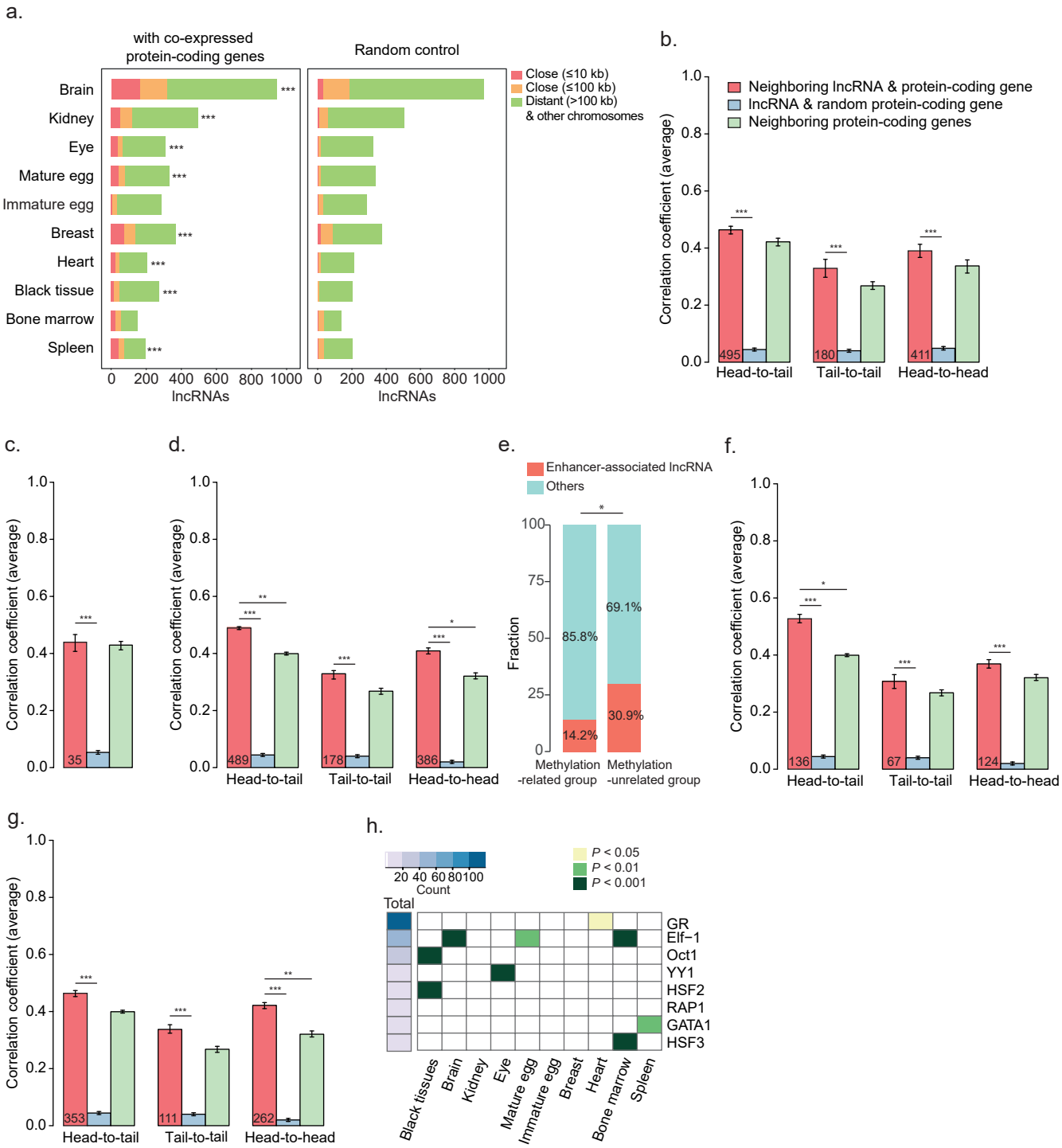

a.
with co-expressed protein-coding genes
Random control
Brain
Kidney
Eye
Mature egg
Immature egg
Breast
Heart
Black tissue
Bone marrow
Spleen
Close (≤10 kb)
Close (≤100 kb)
Distant (>100 kb) & other chromosomes
lncRNAs
b.
Neighboring lncRNA & protein-coding gene
lncRNA & random protein-coding gene
Neighboring protein-coding genes
Correlation coefficient (average)
Head-to-tail
Tail-to-tail
Head-to-head
c.
d.
e.
Enhancer-associated lncRNA
Others
Fraction
85.8%
14.2%
69.1%
30.9%
Methylation -related group
Methylation -unrelated group
f.
g.
h.
Count
Total
P < 0.05
P < 0.01
P < 0.001
GR
Elf−1
Oct1
YY1
HSF2
RAP1
GATA1
HSF3
Black tissues
Brain
Kidney
Eye
Mature egg
Immature egg
Breast
Heart
Bone marrow
Spleen

Figure 6

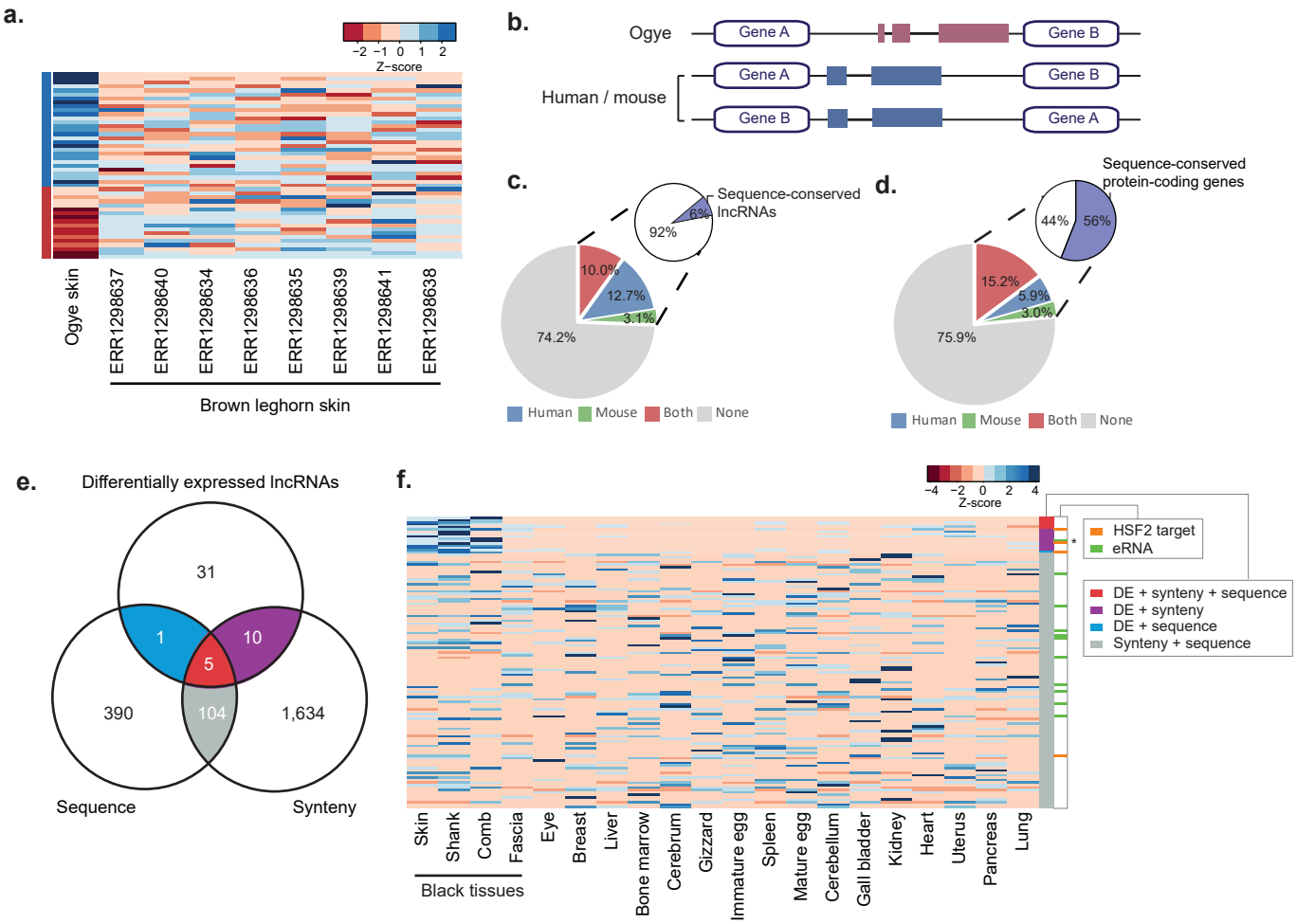
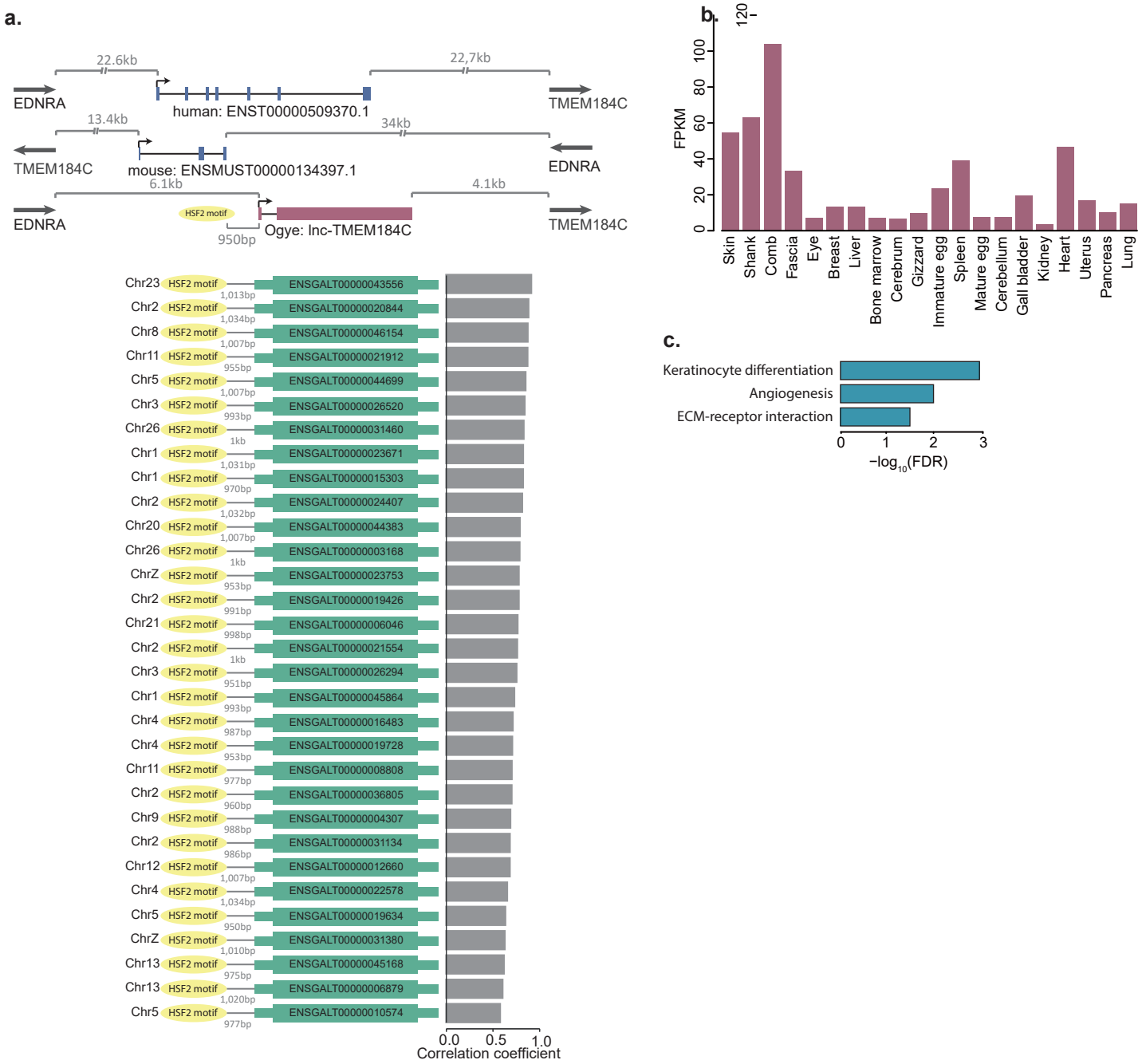

a.
−2 −1 0 1 2
Z−score
Ogye skin
ERR1298637
ERR1298640
ERR1298634
ERR1298636
ERR1298635
ERR1298639
ERR1298641
ERR1298638
Brown leghorn skin
b.
Ogye
Gene A
Gene B
Human / mouse
Gene A
Gene B
Gene B
Gene A
c.
Sequence-conserved lncRNAs
6%
92%
10.0%
12.7%
3.1%
74.2%
Human
Mouse
Both
None
d.
Sequence-conserved protein-coding genes
44%
56%
15.2%
5.9%
3.0%
75.9%
Human
Mouse
Both
None
e.
Differentially expressed lncRNAs
31
1
10
5
390
104
1,634
Sequence
Synteny
f.
−4 −2 0 2 4
Z-score
HSF2 target
eRNA
DE + synteny + sequence
DE + synteny
DE + sequence
Synteny + sequence
Skin
Shank
Comb
Fascia
Eye
Breast
Liver
Bone marrow
Cerebrum
Gizzard
Immature egg
Spleen
Mature egg
Cerebellum
Gall bladder
Kidney
Heart
Uterus
Pancreas
Lung
Black tissues

Figure 7


a.
22.6kb
22,7kb
EDNRA
human: ENST00000509370.1
TMEM184C
13.4kb
34kb
TMEM184C
mouse: ENSMUST00000134397.1
EDNRA
6.1kb
4.1kb
EDNRA
HSF2 motif
950bp
Ogye: lnc-TMEM184C
TMEM184C
Chr23 HSF2 motif 1,013bp ENSGALT00000043556
Chr2 HSF2 motif 1,034bp ENSGALT00000020844
Chr8 HSF2 motif 1,007bp ENSGALT00000046154
Chr11 HSF2 motif 955bp ENSGALT00000021912
Chr5 HSF2 motif 1,007bp ENSGALT00000044699
Chr3 HSF2 motif 993bp ENSGALT00000026520
Chr26 HSF2 motif 1kb ENSGALT00000031460
Chr1 HSF2 motif 1,031bp ENSGALT00000023671
Chr1 HSF2 motif 970bp ENSGALT00000015303
Chr2 HSF2 motif 1,032bp ENSGALT00000024407
Chr20 HSF2 motif 1,007bp ENSGALT00000044383
Chr26 HSF2 motif 1kb ENSGALT00000003168
ChrZ HSF2 motif 953bp ENSGALT00000023753
Chr2 HSF2 motif 991bp ENSGALT00000019426
Chr21 HSF2 motif 998bp ENSGALT00000006046
Chr2 HSF2 motif 1kb ENSGALT00000021554
Chr3 HSF2 motif 951bp ENSGALT00000026294
Chr1 HSF2 motif 993bp ENSGALT00000045864
Chr4 HSF2 motif 987bp ENSGALT00000016483
Chr4 HSF2 motif 953bp ENSGALT00000019728
Chr11 HSF2 motif 977bp ENSGALT00000008808
Chr2 HSF2 motif 960bp ENSGALT00000036805
Chr9 HSF2 motif 988bp ENSGALT00000004307
Chr2 HSF2 motif 986bp ENSGALT00000031134
Chr12 HSF2 motif 1,007bp ENSGALT00000012660
Chr4 HSF2 motif 1,034bp ENSGALT00000022578
Chr5 HSF2 motif 950bp ENSGALT00000019634
ChrZ HSF2 motif 1,010bp ENSGALT00000031380
Chr13 HSF2 motif 975bp ENSGALT00000045168
Chr13 HSF2 motif 1,020bp ENSGALT00000006879
Chr5 HSF2 motif 977bp ENSGALT00000010574
0.0 0.5 1.0
Correlation coefficient
b.
FPKM
0 20 40 60 80 100 120
Skin
Shank
Comb
Fascia
Eye
Breast
Liver
Bone marrow
Cerebrum
Gizzard
Immature egg
Spleen
Mature egg
Cerebellum
Gall bladder
Kidney
Heart
Uterus
Pancreas
Lung
c.
Keratinocyte differentiation
Angiogenesis
ECM-receptor interaction
0 1 2 3
−log10(FDR)