Systematic identification and characterization of chicken (Gallus gallus) ncRNAs

Yong Zhang¹, Jie Wang², Shoujun Huang¹, Xiaopeng Zhu², Jun Liu³, Ning Yang⁴, Dandan Song⁵, Rimao Wu¹, Wei Deng², Geir Skogerbo², Xiu-Jie Wang³, Runsheng Chen²,* and Dahai Zhu¹,*

¹National Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100005, ²Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, ³State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Beijing 100101, ⁴Department of Animal Genetics and Breeding, College of Animal Science and Technology, China Agricultural University, Beijing 100094 and ⁵Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Received December 2, 2008; Revised July 25, 2009; Accepted August 10, 2009

ABSTRACT

Recent studies have demonstrated that non-coding RNAs (ncRNAs) play important roles during development and evolution. Chicken, the first genome-sequenced non-mammalian amniote, possesses unique features for developmental and evolutionary studies. However, apart from microRNAs, information on chicken ncRNAs has mainly been obtained from computational predictions without experimental validation. In the present study, we performed a systematic identification of intermediate size ncRNAs (50–500 nt) by ncRNA library construction and identified 125 chicken ncRNAs. Importantly, through the bioinformatics and expression analysis, we found the chicken ncRNAs has several novel features: (i) comparative genomic analysis against 18 sequenced vertebrate genomes revealed that the majority of the newly identified ncRNA candidates is not conserved and most are potentially bird/chicken specific, suggesting that ncRNAs play roles in lineage/species specification during evolution. (ii) The expression pattern analysis of intronic snoRNAs and their host genes suggested the coordinated expression between snoRNAs and their host genes. (iii) Several spatio-temporal specific expression patterns suggest involvement of ncRNAs in tissue development. Together, these findings provide new clues for future functional study of ncRNAs during development and evolution.

INTRODUCTION

The completion of an increasing number of genome sequencing projects have revealed that sequences corresponding to protein-coding genes only comprise around 1.5–2% of vertebrate genomes (1–4). Accumulating evidence suggests that considerable proportions of eukaryotic genomes are transcribed as non-coding RNAs (ncRNAs) (5–8). A number of different types of ncRNAs have been identified (9–14), ranging in size from around 20 nt, e.g. miRNAs and piRNAs, to over 10 000 nt. Importantly, a growing body of evidence has shown that ncRNAs can act as key regulators of development and other biological processes, supporting the notion that ncRNAs are indispensable players in the control of eukaryotic life/biology (15–20).

The chicken (Gallus gallus) possesses unique features for the study of developmental processes of organogenesis and has for several decades been used as a model organism for vertebrate developmental biology. The domesticated chicken originates from the red jungle fowl (G. gallus), and shares a common ancestor with mammals ~310 million years ago (Mya). The chicken was the first non-mammalian amniote to have its genome sequenced; the genome size is about 40% of the average mammal...
genome and contains 20,000–23,000 predicted protein-coding genes. As a representative of the birds, chicken occupies an important evolutionary position which bridges the evolutionary gap between mammals and other vertebrates (4), and some genomic syntenic regions are more conserved between human and chicken than between human and mouse (21). Therefore, comparative genomic analyses of the chicken genome could provide important information for functional and evolutionary studies of vertebrate protein coding and non-protein-coding genes.

Genome-wide identification of chicken ncRNAs has mainly relied on computational predictions by searching for sequences with homology to ncRNAs identified in other species. The first draft of the chicken genome sequencing project only predicted 571 RNA genes, mainly including highly conserved tRNAs, rRNAs, snRNAs, RNase P, Y RNA, SRP, 7SK and a few conserved miRNAs and snoRNAs (4). The updated data from ENSEMBL (Release 52) annotated 1026 chicken ncRNA genes due to an increased number of predicted snoRNAs and miRNAs based on searching for homologs of human and/or mouse RNAs (560 miRNAs and 148 snoRNAs) (22). Very recently, Qu's group reported 155 chicken snoRNAs predicted by snoSeeker (23). Thus far, 5732 human and 3287 mouse RNA genes, respectively, have been annotated by the ENSEMBL Release 52, and based on these numbers it is reasonable to expect that there exist at least 3000–4000 ncRNA genes in the chicken genome that have not yet been identified. The identification of ncRNAs escaping bioinformatical/computational prediction will rely on experimental strategies and systematic screening. Screenings for chicken microRNAs have been performed by several groups (24–26), but genome-wide identification of other types of chicken ncRNAs by experimental RNomics has yet not been reported and this prompted us to perform a profiling of the chicken non-miRNA RNomics.

Identification and classification of chicken ncRNAs

Identification and classification of chicken ncRNAs

Total RNA was isolated from chicken tissues at embryonic Day 14, post-hatch 1 day and post-hatch 4 weeks. At each time-point, five randomly selected male individuals were sacrificed for tissue sampling. Full-length ncRNA specific libraries of both capped and uncapped transcripts were generated according to a previously described method (30) with modifications. Total RNA was fractionated on Qiagen-tips with a 0.6–1.0 M NaCl gradient elution of QRW2 buffer (Qiagen RNA/DNA handbook). Highly abundant rRNAs (5.8S rRNAs and 5S rRNAs) and snRNAs (U1 snRNA, U2 snRNA, U4 snRNA and U5 snRNA) were removed from the small RNA fraction (50–500 nt) with
the Ambion MicroArray kit. The remaining RNAs were dephosphorylated with calf intestine alkaline phosphatase (Fermentas), and ligated to the 3' adaptor with T4 RNA ligase (Fermentas). After removal of excessive 3' adaptor, the ligation products were split into two aliquots, of which one was treated with PolyNucleotide Kinase (PNK, Fermentas) to phosphorylate non-capped RNA, and the other was treated with Tobacco Acid Pyrophosphatase (TAP, Epicentre) to remove 5'-end methyl-guanosine caps from capped RNA. Thereafter, both samples were ligated to the 5' adaptor and reverse transcribed with Thermoscript reverse transcriptase (Invitrogen) using oligo 3RT as the RT primer. The cDNA was amplified by PCR for 13 cycles using Platinum Taq (Invitrogen) with 3RT and 5CD primers, cloned into pGEM-T vector (Promega) and sequenced. All oligonucleotide sequences used in this study are shown in Supplementary Data file 4.

After removing redundant sequences, the obtained unique ncRNA sequences were first used as queries in a BlastN search against the NCBI nucleotide database and the Ensemble annotated chicken genome sequences to remove matches to rRNAs, tRNAs and mRNAs. The remaining sequences were used for ncRNA prediction and classification using the INFERNAL (42) and SnoReport software (43). The genomic loci of snoRNAs were obtained with the BLAT program (44), and comparisons of cloned snoRNAs with snoRNAs recorded in NCBI, Ensemble and Rfam databases was performed according to the genomic loci of snoRNAs. Known snoRNAs from other species were downloaded from the Rfam database. Transcripts that could not be annotated by any former databases or tools were regarded as unclassified ncRNAs.

For the experimentally identified snoRNA candidates, their potential targets were computationally predicted based on their base complementary information by using pairwise alignment. We searched against several known chicken non-coding transcripts including four rRNAs (18S, 28S, 5.8S, 5S) and eight snRNAs (U1, U2, U4a, U4b, U4b1, U4c, U4x and U5a).

Northern blot hybridization
Total RNA extracted from eight chicken tissues, including heart, liver, brain, lung, kidney, intestine, spleen and skeletal muscle at different development stages, were separated by 8% PAGE (7M urea) and transferred to Nylon membrane (N+, Amersham). Antisense RNAs against the ncRNAs were labeled with Digoxigenin-11-UTP by in vitro transcription with T7 and SP6 RNA polymerase. The RNA blots were hybridized in ULTRAhyb (Ambion) at 68°C overnight, washed with 2× SSC/0.1% SDS washing buffer at 68°C for 2×5 min, following stringent washing with 0.1× SSC/0.1% SDS washing buffer at 68°C for 2×30 min. Thereafter, the RNA blots were blocked with blocking buffer for 30–60 min at room temperature and incubated for 30 min with anti-DIG-AP (1:10000 diluted in blocking buffer). Hybridization signals were detected by CDP-star (Roche). Chemiluminescent signals were exposed to X-ray film. For detection of snoRNA expression in different tissues, RNA samples from chicken at 2 weeks post hatching were used.

Analysis of snoRNA host genes
Intronic ncRNAs were identified according to the ENSEMBL release 52 Gallus gallus genome annotation version WASHUC2. The Gene Ontology analysis of host genes of intronic snoRNAs was carried out using the GOEAST software (45). The microarray data used in this study covered expression profiles of chicken skeletal muscle tissues at post-hatch 1 day, 2, 4, 6 and 8 weeks. Three biological replicates were performed for each time point, and the average signal for each gene over the three biological replicates were used for further analysis. The expression changes shown in Figure 4B were calculated as the expression of a gene at a time point minus its expression at the immediate preceding time point.

Conservation analysis of ncRNAs
The genomic sequences of chicken and 18 vertebrate genomes were downloaded from the UCSC genome browser. The conservation of chicken ncRNAs in the 18 vertebrate genomes was determined by using chicken ncRNAs as queries in BLAT searches of other genomes (44). Conservation scores were calculated based on the maximal alignment length and the identity of the alignment of the BLAT hits on each vertebrate genome. The conservation status of all cloned ncRNAs was clustered hierarchically according to their conservation scores in the 18 genomes using the cluster package in R.

Preparation and transfection of chicken fetal myoblasts
Chicken fetal myoblasts (CFMs) were prepared from pectoralis muscle tissue from 11th day White Leghorn chicken embryos according to a modified protocol described by Yang et al. (46). Briefly, the pectoralis muscle tissue was collected and washed with phosphate buffered saline (PBS). The tissue was minced and digested with 0.1% collagenase type I in PBS buffer at 37°C for 30 min with constantly shaking. The digested tissue was collected by centrifuging, resuspended in PBS buffer by repeated pipetting and filtrated to remove aggregated cells and myotubes. The collected cells were resuspended in z-MEM medium, seeded into a 100-mm dish, and cultured for 1 h at 37°C. Rapidly adhering cells were mainly composed of fibroblasts and were discarded. Myoblasts remaining in suspension were collected, transferred to a six-well gelatin-coated plate at 1×10⁵ cells/cm² with z-MEM containing 10% Horse Serum, 100 IU/ml penicillin and 100 μg/ml streptomycin (Invitrogen), and cultured for 24 h at 37°C in humidified 95% air–5% CO₂. The myoblasts were transfected with the Lipofectamine 2000 reagent (Invitrogen) according to the manufacturer’s instructions. The transfected cells were lysed after 48 h of transfection, and total RNA was extracted using the TRIzol reagent (Invitrogen). The primers for pfkm minigene generation are shown in Supplementary Data file 4.
RESULTS

Identification and Classification of Chicken ncRNAs

Full-length intermediate size ncRNAs (50–500 nt) enriched libraries were constructed according to a previously described method (30) with some modifications. This procedure ensured that the library contained a substantial fraction of full-length ncRNA clones with defined 5' and 3' termini. The total RNA used for library construction was extracted from a mix of 10 chicken tissues (heart, liver, brain, lung, kidney, intestine, glandular stomach, gizzard, spleen, skeletal muscle) harvested at three developmental stages (14 days old embryo, 1 day and 4 weeks chicken after hatching). Altogether, 3468 clones from two full-length cDNA libraries were sequenced. After removing matches to tRNAs and rRNAs, the remaining 201 sequences were taken as ncRNA candidates. Among these, 33 clones with a match to EST or RefSeq RNA sequences were excluded in this report because they can not be detected the same size band as obtained from library sequencing, suggesting they could be mRNA degradation products. The 43 other ncRNA candidates were also excluded because they did not yield detectable expression signals by northern blot. The remaining 125 ncRNAs candidates validated by northern blot included 90 ncRNA loci predicted by the ENSEMBL or Rfam (22,47) and 35 novel candidates (Figure 1, Supplementary Data files 1 and 2). The previously predicted 90 ncRNAs included 77 snoRNAs, 10 snRNA, two Y RNAs and one SRP RNA. The 35 novel ncRNA candidates were comprised of 25 transcripts with possible snoRNA characteristics, one transcript predicted to be a snRNA, and nine transcripts that could not be assigned to any known class of ncRNAs (Table 1). Among the 25 novel snoRNA candidates, three C/D box snoRNA candidates could target chicken rRNAs, whereas 10 of the novel snoRNA candidates might be orphan snoRNAs with no apparent target in chicken rRNAs or snRNAs. The other 12 novel snoRNA candidates were also independently predicted by Qu's group (23) (Table 1 and Supplementary Data file 3). The expression of the 35 novel ncRNA candidates was shown in Figure 2. They are ubiquitously expressed in the eight tested chicken tissues, the single exception being Ggn39, which is specifically expressed in skeletal muscle and heart tissue. The expression patterns of the other ncRNAs were shown in Supplementary Figure S1.

Potentially bird-specific ncRNAs

To analyze the evolutionary conservation of the ncRNAs, we searched for homologs of all the 125 chicken ncRNAs in the genomes of the lizard and 6 fish and 11 mammalian species, and hierarchically clustered the ncRNAs according to their conservation scores. The 125 chicken ncRNAs could be grouped into seven clusters according to their sequence conservation scores in each of the 18 queried genomes (Figure 3A), and one ncRNA from each of the seven clusters was selected as a representative and shown in Figure 3B. As shown in Figure 3A, compared to the highly conserved snRNAs, previously predicted snoRNAs are less conserved, whereas the newly identified snoRNAs and unclassified ncRNAs are poorly or not conserved. Ten snRNAs, two Y RNA and SRP RNA are highly conserved and grouped into clusters 1 and 2 (e.g. Ggn130_U4 and Ggn85_Y RNA in Figure 3B), and only the predicted snRNA (Ggn103) shows relatively low conservation (cluster 7). Unlike snoRNAs, snoRNAs are generally less conserved. Of the 77 previously predicted snoRNAs, none are grouped into cluster 1 and only two are grouped into cluster 2 (Ggn112_SCARNA8 and Ggn6_SNORD18), whereas 36 (46.7%) belong to cluster 6, which is composed of transcripts with very limited sequence conservation (e.g. Ggn75_SNORD36 in Figure 3B). Interestingly, 16 snoRNAs, fall into clusters 3–5 (e.g. Ggn46, Ggn9 and Ggn120 in Figure 3B) and are only conserved in the lizard and/or platypus genomes but not in other mammals, suggesting the existence of an ancient group of snoRNAs that have been preserved in the reptile–bird lineage.

The 35 newly identified snoRNAs and unclassified ncRNA candidates are all not well conserved. In total, 22 of 25 novel snoRNA candidates (88%) are grouped into cluster 7 which show no conservation in any other examined genome/organism (Figure 3A). Six of the nine unclassified ncRNA candidates (67%) are also unconserved and are grouped into cluster 7 (e.g. Ggn39 in Figure 3B). Consistent with their low conservation score in mammalian genomes, only one ncRNA out of 35 novel ncRNAs can be detected in human and mouse tissues by northern blot (Figure 2). Interestingly, three snoRNAs (Ggn5, Ggn51 and Ggn112) can be detected in human but not in mouse tissues (Figure 3C and Supplementary Figure 2). In support of this observation, genomic sequence analysis showed these three snoRNAs have higher sequence similarity between chicken and human than between chicken and mouse, indicating that these chicken snoRNAs are shared by chicken and a limited number of mammalian lineages.

Figure 1. Classification of cloned chicken ncRNAs. (A) Classification of unique ncRNA sequences. (B) Distribution of cloned snoRNAs.
We further searched for homologs of 35 novel ncRNA candidates in the recently released zebra finch (*Taeniopygia guttata*) genome. Twenty one of the novel transcripts have homologs in the zebra finch with sequence identity above 85%, whereas the remaining 14 ncRNAs had no hits in the current version of the zebra finch genome (Table 1). Although this may be due to the incomplete genomic sequence available for the zebra finch it is also possible that some of these ncRNAs are specific to chicken.

### Host gene analysis of intron-encoded snoRNAs

Of the 102 snoRNAs, 89 were encoded by intronic regions of protein-coding genes and 11 were located in intergenic regions (Table 2). For two of the snoRNA transcripts no genomic loci could be found in the current version of the chicken genome sequence. The 11 non-intronic snoRNAs appear all to be unique transcriptional units. For the intronic snoRNAs, we found 12 protein-coding genes each hosting two intronic snoRNAs, two genes hosting three snoRNAs and two genes hosting four snoRNAs (Supplementary Data file 2). Only in two cases, a single intron encodes the two ncRNA loci, and in both cases the two loci within the same intron encoded identical ncRNA transcripts (Supplementary Figure S3).

We performed functional analysis of the snoRNA host genes based on Gene Ontology annotations. Host genes of both C/D box and H/ACA box snoRNAs were enriched with functional annotation relating to ribosome biogenesis, translation and cellular protein metabolic processes. We also found that host genes of both C/D box and H/ACA box snoRNAs were enriched for some other functions, including initiation of DNA duplication, protein transportation, regulation of translation initiation, regulation of cell cycle and other processes (Figure 4A and Supplementary Figure S4).

To further explore the possible coordinated expression pattern between intronic snoRNAs and their host genes, we examined the expression of snoRNA host genes using our recently published Affymetrix microarray data (48). RNA samples from chicken skeletal muscle tissues at 1 day, or 2, 4, 6 and 8 weeks after hatching were used for the array analysis. The results show that in general, the expression of snoRNA host genes did not vary much.

### Table 1. General information for 35 novel ncRNA candidates identified in this study

<table>
<thead>
<tr>
<th>GGN_ID</th>
<th>GenBank_ID</th>
<th>Reads_no</th>
<th>Length</th>
<th>ncRNA_type</th>
<th>Target</th>
<th>Genomic_location in_chick</th>
<th>chr_location in_zebra finch</th>
<th>Seq_identity with_zebra finch</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGN11</td>
<td>EU240230</td>
<td>2</td>
<td>85</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN20</td>
<td>EU240238</td>
<td>1</td>
<td>61</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN37</td>
<td>EU240254</td>
<td>1</td>
<td>70</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN86</td>
<td>EU240302</td>
<td>9</td>
<td>96</td>
<td>CD</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN120</td>
<td>EU240333</td>
<td>1</td>
<td>71</td>
<td>CD</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN138</td>
<td>EU240346</td>
<td>1</td>
<td>58</td>
<td>CD</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN148</td>
<td>EU240352</td>
<td>2</td>
<td>66</td>
<td>CD</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN1100</td>
<td>EU240315</td>
<td>3</td>
<td>78</td>
<td>CD</td>
<td>18S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN345</td>
<td>EU240252</td>
<td>1</td>
<td>77</td>
<td>CD</td>
<td>18S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN1105</td>
<td>EU240321</td>
<td>6</td>
<td>68</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN520</td>
<td>EU240268</td>
<td>1</td>
<td>66</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN345</td>
<td>EU240252</td>
<td>1</td>
<td>77</td>
<td>CD</td>
<td>18S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN1106</td>
<td>EU240322</td>
<td>2</td>
<td>99</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN808</td>
<td>EU240296</td>
<td>4</td>
<td>82</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN828</td>
<td>EU240298</td>
<td>2</td>
<td>67</td>
<td>CD</td>
<td>18S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN176</td>
<td>EU240236</td>
<td>1</td>
<td>70</td>
<td>CD</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN79</td>
<td>EU240295</td>
<td>3</td>
<td>133</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN1126</td>
<td>EU240339</td>
<td>1</td>
<td>123</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN72</td>
<td>EU240288</td>
<td>5</td>
<td>129</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN87</td>
<td>EU240303</td>
<td>3</td>
<td>130</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN58</td>
<td>EU240274</td>
<td>11</td>
<td>134</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN56</td>
<td>EU240272</td>
<td>2</td>
<td>135</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN320</td>
<td>EU240250</td>
<td>1</td>
<td>136</td>
<td>HACA</td>
<td>Orphan</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN1230</td>
<td>EU240336</td>
<td>5</td>
<td>130</td>
<td>HACA</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN740</td>
<td>EU240290</td>
<td>2</td>
<td>137</td>
<td>HACA</td>
<td>28S rRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN103</td>
<td>EU240318</td>
<td>4</td>
<td>122</td>
<td>snRNA</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN141</td>
<td>EU240348</td>
<td>1</td>
<td>123</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN767</td>
<td>EU240237</td>
<td>2</td>
<td>69</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN136</td>
<td>EU240344</td>
<td>1</td>
<td>115</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN105</td>
<td>EU240320</td>
<td>3</td>
<td>62</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN68</td>
<td>EU240284</td>
<td>1</td>
<td>227</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN46</td>
<td>EU240262</td>
<td>1</td>
<td>76</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN46</td>
<td>EU240262</td>
<td>1</td>
<td>76</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN147</td>
<td>EU240351</td>
<td>1</td>
<td>96</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN16</td>
<td>EU240235</td>
<td>1</td>
<td>56</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
<tr>
<td>GGN39</td>
<td>EU240256</td>
<td>1</td>
<td>67</td>
<td>Unclassified</td>
<td>Intergenic</td>
<td>No hit</td>
<td>No hit</td>
<td>No hit</td>
</tr>
</tbody>
</table>

*ncRNAs independently predicted by Qu’s group (23).*
across the examined developmental stages, and that genes with ribosome related functions had least variation in expression (Figure 4B). Host genes with functions related to cell cycle and apoptosis showed the most variable expression across developmental stages (Figure 4B). Correspondingly, snoRNAs encoded in intronic regions of ribosome related genes also exhibited constant expression across different developmental stages, whereas snoRNAs hosted by non-ribosome related genes showed variable expression levels during postnatal development (Supplementary Figure S5). It is worth noting that snoRNA host genes with ribosome and translation related functions had much higher expression levels than the average coding gene expression, whereas the expression levels of host genes with other known functions were relatively low (Figure 4C). Similarly, such expression discrepancy also observed in the expression pattern of snoRNAs during chicken muscle development (Supplementary Figure S5).

**Tissue and developmental stage-specific expression of chicken ncRNAs**

The ncRNA expression patterns may yield important clues to their functions. Thus, we investigated the expression of ncRNAs in various chicken tissues, including heart, liver, lung, spleen, kidney, intestine, skeletal muscles and brain. All examined ncRNAs were detected by northern blot analysis (Supplementary Figure S1). Unlike snRNAs and Y-RNAs, which showed constant expression levels across all examined tissues, some of the snoRNAs tended to have lower expression in brain, muscle and heart than in other tissues (Figure 5A). Interestingly, we found that one of the unclassified ncRNA (Ggn39) was only expressed in the chicken heart and skeletal muscle (Figure 5A). To look at the possible functional roles of the ncRNAs during skeletal muscle development, we examined the expression pattern of these ncRNAs in chicken skeletal muscle tissues at different developmental stages by northern blot analysis. Our results showed that several of the snoRNAs had developmentally regulated expression patterns (Figure 5B and Supplementary Figure S5). For example, snoRNAs Ggn74, Ggn3 and Ggn7 were highly expressed during embryonic development, whereas snoRNA Ggn5 was more abundantly expressed after hatch (Figure 5B). Ggn39, which is specifically expressed in skeletal muscle and heart, was predominantly expressed at the post-hatching stages (Figure 5B), suggesting that this unclassified ncRNA might have specific functions in myogenesis during postnatal development.

To further investigate the possible involvement of ncRNAs in chicken myogenesis, we analyzed the expression pattern of the 105 ncRNAs in the skeletal muscle tissues of two chicken lines with divergent

![Figure 2. Tissue expression pattern of 35 novel ncRNA candidates. The expression pattern of 35 novel ncRNA candidates was examined by northern blot analysis of total RNA from chicken heart, liver, lung, spleen, kidney, intestine, skeletal muscles and brain. Total RNA from human and mouse skeletal muscle were included in each blot in order to test for possible expression of the ncRNA candidates in different species. The ncRNA candidates are labeled as GgnX_class_size on the left of each northern blot. (For example: Ggn37_CD_70 nt means Ggn37 is a predicted C/D box snoRNA with length 70 nt. ‘Un’ denotes ‘unclassified’. Asterisks indicate that ncRNAs were independently predicted by Qu’s group (23).) The expression patterns of 35 novel ncRNA candidates are shown in the left panel. The expression patterns of 35 novel ncRNA candidates are shown in the right panel.](image)
muscle growth rates (fast growing ‘broiler’ chicken for meat production and slow growing ‘layer’ chicken for egg production). We found that several ncRNAs were differentially expressed in skeletal muscle tissues between broilers and layers. The expression levels of Ggn68, Ggn91 and Ggn89 were significantly higher in layers than in broilers at post-hatching 2 and 4 weeks (Figure 5C), whereas Ggn44, Ggn4 and Ggn29 showed higher expression in broilers than in layers (Figure 5D). The differential expression patterns of these ncRNAs suggest that they may be important regulators for the divergent muscle growth phenotype of broilers and layers.

Figure 3. Comparative genomics analysis of 125 chicken ncRNAs. (A) Heatmap of the conservations scores of 125 chicken ncRNAs in 18 vertebrate genomes. The conservation degree decreases as color changes from green to red. The ncRNAs were hierarchically clustered based on the conservation scores in the 18 genomes. (B) Representatives of ncRNAs from each of the seven clusters. (C) Representatives of conserved ncRNAs. The identity of each ncRNA is shown on the left of each blot. The tested chicken tissues are indicated on the upper of the figure. Human and mouse total RNA samples used in the experiments are as described in Figure 2. (D) Representatives of less conserved ncRNAs. The U5 snRNA was included as a loading control.
Ggn39 is generated from the pfkm gene in a splicing-dependent manner

Mapping the novel transcripts to the chicken genome showed that Ggn39 is located in the second intron of the phosphofructokinase M (pfkm) gene (Figure 6A). The size of this intron is 68 nt, whereas the cloned Ggn39 transcript is 67 nt long corresponding to 1–67 bp of intron sequence. Northern blot analysis showed that Ggn39 and its host gene pfkm were specifically detected in skeletal and heart muscle tissues, and are most abundantly expressed during postnatal myogenesis (Figure 6B). The similarity in spatio-temporal expression-patterns of the ncRNA and host gene loci suggests that Ggn39 may be generated by intron splicing from the pfkm pre-mRNA. We tested this possibility by transfecting the chicken myoblasts with a wild-type pfkm minigene (pcDNA3-pfkm) comprising exons 1, 2, 3 and introns 1, 2 and the first 37 bp of intron 3, and a mutant pfkm minigene containing a mutation in the 5’ splicing site (GT to CT) of the second intron (pcDNA3-pfkm-mutant) (Figure 6C). Northern blot analysis demonstrated that the Ggn39 ncRNA was only detected in the myoblasts transfected with the wild-type minigene but not in myoblasts transfected with the mutant construct (Figure 6D), suggesting that Ggn39 is generated in a splicing-dependent manner.

**DISCUSSION**

Increasing numbers of ncRNAs are being identified from various model organisms by experimental RNomics combined with bioinformatics analyses (49,50). Herein, we have performed the first systematic profiling of the chicken intermediate size ncRNA transcriptome by experimental RNomics, and cloned 125 chicken ncRNAs. Comparative genomic analysis with blast against 18 sequenced vertebrate genomes showed that only a few (5 out of 35) of the novel ncRNAs have homologs in non-avian vertebrates, and the remaining novel ncRNAs are likely to be specific to the bird or chicken lineages. Our findings are in line with recent studies on other species. Computer analysis of genomic tiling array data revealed 300 putative candidates for primate specific ncRNAs (51). Unclassified novel ncRNAs often show little sequence conservation and low cellular concentrations (30). Less
Figure 4. Analysis of snoRNA host genes. (A) Functional analysis of snoRNA host genes according to the Biological Process category of Gene Ontology. GO terms enriched among host genes of C/D box snoRNAs and H/ACA box snoRNAs are shown as red and green boxes, respectively, and GO terms enriched among host genes of both C/D box and H/ACA box snoRNAs are shown as yellow boxes. The P-values for the enrichment of each GO term are shown within the boxes. Non-enriched GO terms in the hierarchical tree of GO terms are shown as dots to reduce the figure size. (B) Expressional changes of snoRNA host gene in chicken skeletal muscle tissue across different developmental stages. The log₂ value of the expression difference between post-hatch time points 1 day, 2, 4, 6 and 8 weeks are shown. Genes with expression variance less than 1.5-fold between any two continues time points are shown as green lines, genes with expression variance larger than 1.5-fold but less than 2-fold between any two continues time points are shown as blue lines, and genes with expression variance larger than 2-fold between any two continues time points are shown as red lines. (C) Expression of snoRNA host genes in chicken skeletal muscle tissues. The log₂ values of the average expression level of each snoRNA host gene across all examined time points are shown on the X-axis. The proportions of snoRNA host genes within each expression level range are shown on the Y-axis. The expression of all genes detected on the microarray is used as the reference gene set.
conserved ncRNAs might be important factors in animal development (8), and a remaining challenge for a better understanding of ncRNA function is therefore the detection of less conserved, low abundance ncRNAs in different organisms. Together with previous reports, the identification of chicken specific ncRNAs indicated that there may exist numerous types of ncRNAs that are specific to a lineage or even a single species.

It has been reported that a majority of the identified vertebrate snoRNAs are encoded in introns of protein-coding genes whose functions are mostly related to synthesis, structure or function of the translational apparatus (36). This mode of gene organization and expression provides a molecular basis for functional prediction for recently discovered ncRNAs. Similar analyses have been proved invaluable for predicting the functional roles of miRNAs host genes (52). In this study, we have sought to combine GO annotation on host gene function and cellular localizations with experimentally validated expression profiles of snoRNA host genes to predict possible functions of the intronic snoRNAs. We have found that snoRNA host genes encode proteins with a broad spectrum of biological roles ranging from DNA duplication initiation, protein transportation, translation initiation, regulation of cell cycle and other developmental processes. Our results indicate that the host genes of H/ACA have a broader spectrum of cellular locations than do host genes of C/D box snoRNAs, suggesting that H/ACA and C/D box snoRNAs may function in different cellular processes. This observation was further supported by the microarray data on host gene expression during skeletal muscle development. snoRNA host genes with functions related to ribosome biogenesis and translation were much more abundantly expressed during myogenesis than host genes with other functions (such as cell cycle and apoptosis), which showed lower and more variable expression across the developmental stages. Our findings could provide information for further studies of the functions of snoRNAs from host genes with different biological functions, as well as for the study of the functional and regulatory coordination between snoRNAs and their host genes during development.

The spatio-temporally regulated expression patterns of the novel ncRNAs provide further functional information about the role of these ncRNAs during development. It is well accepted that some ncRNAs, especially some microRNAs and piRNAs, have specific spatio-temporal expression patterns and play functional roles in cell differentiation and organogenesis during development. For instance, miRNAs with muscle-specific expression are involved in development and disease of the skeletal and heart muscle (53,54). In general, the spatio-temporally regulated expression of other classes of ncRNAs, particularly the unclassified ncRNAs, has been less well explored. We investigated this question by northern blot...
analysis of a substantial number of the ncRNAs identified in the present work. Our data showed that the majority of the novel ncRNAs exhibited ubiquitous tissue distribution; some were abundantly expressed only in a few tissues. Most interestingly, we found that one of the unclassified ncRNAs (Ggn39) was generated from the second intron of \textit{pfkm} gene in a splicing-dependent manner, and that Ggn39 ncRNA, like its host gene \textit{pfkm}, showed specific expression in the chicken heart and skeletal muscle cells, especially during postnatal development. As a key enzyme in the glucose metabolism, PFKM plays important roles in the metabolism and development of skeletal muscle (55). The coordinated expression pattern of Ggn39 and its host gene \textit{pfkm} suggests a functional correlation between Ggn39 and PFKM which may provides useful information for further studies of the metabolic and developmental functions of Ggn39 and its host gene \textit{pfkm}.

Recently, brain-specific snoRNAs have been implicated in brain development and neurological disease (41,56,57). In chicken we identified an H/ACA-box snoRNA (Ggn146) encoded in the intron of the chicken ortholog of the Williams Beuren syndrome chromosome region 22 gene. Since Williams Beuren syndrome is caused by microdeletion within the syntenic region containing this gene in human (58), it is an intriguing possibility that an ortholog of Ggn146 may also be involved in the pathogenesis of the Williams Beuren syndrome in human. The broiler and layer chickens are an excellent model systems for the study of molecular mechanisms involved in skeletal muscle development because of their significant muscle growth differences during myogenesis. Up to now, the molecular genetics underlying the divergent muscle growth rate between broilers and layers have remained unclear. Similar to muscle-specific miRNAs (e.g. miRNA-1), ncRNAs that are differentially expressed in skeletal muscle tissues between broiler and layer chickens might function as muscle growth regulators. Identification of these ncRNAs could facilitate ncRNA functional studies and provide essential

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure6.png}
\caption{Ggn39 is generated in a splicing-dependent manner. (A) The genomic organization of Ggn39. Boxes and lines represent exons (E1–E9) and introns (1–8), respectively, of the \textit{pfkm} gene. (B) The spatio-temporal expression pattern of Ggn39 and its host gene \textit{pfkm}. The expression patterns of Ggn39 and \textit{pfkm} were examined by northern blot analysis of total RNA from chicken tissues. The corresponding antisense RNA labeled with DIG-11-UTP were used as probes. U6 snRNA was used as loading control. (C) Schematic structure of the \textit{pfkm} minigene and the second intron 5’ splicing site mutant. (D) The biogenesis of Ggn39 is dependent on splicing of the second intron of the \textit{pfkm} gene. Chicken primary myoblasts were transfected with wild-type (WT) and mutant (MT) \textit{pfkm} minigenes. GFP denotes pcDNA3-EGFP that served as transfection control. Northern blot analysis was performed on total RNAs prepared from the transfected cells. As the Ggn39 is highly expressed in skeletal muscle, chicken skeletal muscle RNA was used as positive control for the northern blot experiments.}
\end{figure}
information for a better understanding of the functional roles of ncRNAs during development.

ACCESSION NUMBERS

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

FUNDING

Conflict of interest statement. None declared.

REFERENCES