# Non-coding RNAs of bird genomes

Paul P. Gardner*[1,2] , Jana Hertel[5] , Sarah W. Burge[3] , Maria Ninova[4] , Stephanie Kehr[5] , Mario Fasold[5] , Tammy E. Steeves[1] , Sam Griffiths-Jones[4] and Peter Stadler*[5]

[1] School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. [2] Biomolecular Interaction Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. [3] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK. [4] Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom. [5] Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany

Email: Paul P. Gardner*- paul.gardner@canterbury.ac.nz; Jana Hertel*- jana@bioinf.uni-leipzig.de; sb30@sanger.ac.uk; Maria.Ninova@postgrad.manchester.ac.uk; steffi@bierdepot.bioinf.uni-leipzig.de; mario@bierdepot.bioinf.uni-leipzig.de; tammy.steeves@canterbury.ac.nz; sam.griffiths-jones@manchester.ac.uk; Peter Stadler*- studla@bioinf.uni-leipzig.de;

*To whom correspondence should be addressed

## Abstract

Here we present the results of a large-scale bioinformatic annotation of non-coding RNAs in 48 avian genomes. Our approach uses probabilistic models of hand-curated families from the Rfam database to infer conserved RNA families within each avian genome. We supplement these annotations with predictions from the tRNA annotation tool, tRNAscan-SE and microRNAs from miRBase. Although we show extensive conservation of classical ncRNAs (e.g., tRNAs and rRNAs) and more recently discovered ncRNAs (e.g., snoRNAs and miRNAs) in birds, we also demonstrate apparent "losses" in several RNA families. These include the divergence of some classical ncRNAs and the loss of several snoRNAs and microRNAs. In contrast, we show a significant number of lncRNAs are surprisingly well conserved between birds and mammals including several intriguing cases where the reported mammalian lncRNA function is not conserved in birds. These combined results illustrate the utility of applying homology based methods for annotating vertebrate genomes and illustrate many complex evolutionary patterns within the avian ncRNA cohort.

## Introduction

Non-coding RNAs (ncRNAs) are an important class of genes. However, they pose serious research challenges, particularly for the field of genomics. They lack the strong statistical signals associated with protein coding genes, e.g. open reading frames, G+C content and codon-usage biases [1].

Non-coding RNAs are classified into hierarchical groupings by the Rfam database. The basic units are "families" which are groups of homologous, alignable sequences; "clans" which are groups of un-alignable (or functionally distinct), homologous families; and "classes" which are groups of clans and families with related biological functions e.g. spliceosomal RNAs, miRNAs and snoRNAs [2–6]. The major RNA families include the classical, highly conserved RNAs, sometimes called "molecular fossils", such as the transfer RNAs, ribosomal RNAs, RNA components of RNase P and the signal recognition particle [7]. Other classes appear to have have evolved more recently, e.g. the small nucleolar RNAs (snoRNAs), microRNAs (miRNAs) and the long non-coding RNAs (lncRNAs) [8]. The publication of avian genomes provides exciting opportunities to to explore ncRNA conservation in unprecedented detail.

Despite promising techniques such as RNA-seq [9], homology based methods namely covariance models (CMs) are state of the art for ncRNA analyses [10–12]. The CM based approach for annotating ncRNAs in genomes requires reliable alignments and consensus secondary structures of representative sequences of RNA families. These are used to train probabilistic models for each family. These models can be used to generate sequences with similar properties, score the likelihood that a sequence is was generated by the same evolutionary processes as the training sequences and to build alignments based upon sequence and structural information [10–12]. The tRNAscan-SE software package uses CMs to accurately predict transfer RNAs [13, 14]. The Rfam database contains thousands of curated alignments and consensus structures for diverse classes of ncRNAs [2–6]. Independent benchmarks of bioinformatic annotation tools have shown that the CM approaches dramatically out-perform alternative methods [15].

The CM based approach works well for almost all classes of ncRNA, but the long non-coding RNAs (lncRNAs) are a particular challenge [16]. Recent technological advances have led to dramatic speed and memory-usage enhancements for CM analyses [12, 17–19]. However, CMs cannot model the exon-intron structures of spliced lncRNAs, nor can they deal simply with the repeats that many lncRNAs host. Consequently in the latest release of Rfam the lncRNA families that were added were composed of local conserved and structured elements within lncRNAs, analogous to the "domains" housed within protein sequences [6]. The functions determined to date for lncRNAs range from regulating chromatin status to chromosomal inactivation [20, 21]. Yet functional characterisation of these genes is a lengthy and expensive

process [16].

## Results

In the following we explore the conservation patterns of the major classes of ncRNA in further detail. The collection of ncRNA sequences is generally biased towards model organisms [8, 22]. Consequently the bulk of the families discussed below are either conserved across the vertebrates or have been apparently lost or diversified in the avian lineage(s).

### Classic RNAs: LUCA and LECA

Many RNA families constitute the most evolutionarily conserved genes across all life on this planet [7]. Examples of RNAs derived from the Last Universal Common Ancestor (LUCA) include, the transfer RNAs (tRNA), ribosomal RNAs (rRNA), RNA components of RNase P (RNase P RNA), RNase MRP (RNase MRP RNA) and the signal recognition particle (SRP RNA). Other classes of RNA are likely to have been components of the Last Eukaryotic Common Ancestor (LECA) include the telomerase RNA, major spliceosomal RNAs (U1, U2, U4, U5, and U6) and the minor spliceosomal RNAs (U11, U12, U4atac, and U6atac) [8].

Unsurprisingly, the bulk of these classes of RNAs are well represented across the bird genomes (See Figure 1). However, there appear to have been "losses" of a few of these RNAs in certain bird species. Some of these may be due to sequence divergence, of which there are several notable examples e.g. [23–27]. Other losses may be due to incomplete genome coverage.

A number of the classic RNAs are incorporated into RNA-protein complexes (RNPs) involved in core cellular processes. An example of this are the spliceosomal RNAs. Based upon the presence/absence patterns of the major spliceosomal RNAs they are all well represented in these genome sequences. Except for the U4 RNA in cormorant and the U5 RNA in the bee eater. Again these two genomes are low coverage suggesting these genes weren't captured in the current assembly. The minor spliceosomal RNAs are more interesting, the U4atac and U11 snRNAs show widespread patterns of loss, even in some of the high coverage genomes. These RNAs are frequently missed in bioinformatic screens. Indicating either frequent loss [28] or sequences that have diverged beyond the ability of detection by covariance models [29].

The telomerase RNA is also largely missing from the avian annotations. This RNA acts as a template for the telomerase enzyme that extends the telomeres found on chromosome ends. It is only found in the chicken, bald eagle, kea, budgerigar, crow and zebrafinch. Homology searches searches with the telomerase

reverse transcriptase (TERT) protein show that the protein component of the telomerase RNP is conserved across all the bird genomes (data not shown). This pattern of presumably divergent telomerase RNA and conserved telomerase protein has been noted previously, most notably in the fungi [23, 24].

The the RNA components of RNase P and RNase MRP also appear to have undergone dramatic losses within the bird lineage. RNase P is required for the maturation of tRNA, the paralogous enzyme, RNase MRP is required for the maturation of rRNA. Each RNP cleaves smaller RNAs from larger transcripts [30]. It is unlikely that the these genes have been lost in any of the birds. Homology searches with the RNase associated protein coding genes (POP1, POP4, POP5, POP7, RPP1, RPP14, RPP25, RPP38, RPP40 and RPR2), identified viable homologs of each in all of the bird genomes [31] (data not shown). This suggests that the bird RNase P and MRP RNAs may have diverged slightly from the canonical models.

The 5.8S component of the ribosome in the turtle, turkey bustard, hoatzin, flamingo, tropicbird, seriema, owl, cuckoo roller, trogon, bee eater and falcon appears to have been lost (See Figure 1). However, the genomes for these species are also low-coverage. Therefore gaps in these families are more likely to be due to incomplete genome assemblies.

### Small nucleolar RNAs

Small nucleolar RNAs (snoRNAs) are important ncRNAs that participate in the maturation of other functional RNAs [22]. The bulk of the characterised snoRNAs guide either methylation or pseudouridylation modifications, primarily of rRNAs but also spliceosomal RNAs. The two types of modifications are guided by two different types of RNA, the C/D and the H/ACA box snoRNAs respectively, each with a characteristic cohort of motifs and secondary structures [32].

There are 66 ribosomal modification sites, guided by 59 snoRNA families, that are preserved between *H. sapiens* and *S. cerevisiae* [33]. Of these, 45 snoRNA families are conserved in the bird dataset.

Over a third of the apparent losses of the yeast-human conserved snoRNA families appear to cluster on 2 locii of the ancestral vertebrate genome. We investigated these losses further.

The first cluster is found at chr11:62620797-62622484 on the human genome (hg19) and contains SNORD27, SNORD29 and SNORD31 of the human-yeast conserved snoRNAs. These snoRNAs are located in the inside-out gene SNHG1 which hosts a total of eight C/D box snoRNAs: SNORD25, SNORD26, SNORD27, SNORD28, SNORD29, SNORD22, SNORD30 and SNORD31 [34]. Each of which are also found in the alligator and turtle genomes within a 3-4 KB locus, yet these have largely been lost in the birds. Notable exceptions are five of the eight are located in the tinamou genome, these are located on

the same scaffold and are within 2 KB of each other. Implying that SNHG1 is conserved in the tinamou. Three of the eight are located in the ostrich and zebrafinch genomes, again within 2 KB of each other on the same scaffolds. This complex pattern of loss could be attributed to many different models, e.g. multiple losses in birds, poor homology modelling or incomplete genome sequences.

The second cluster is located at chr19:49993222-49994231 on the human genome (hg19) and contains two copies of SNORD33 and one SNORD34 all within a 1 KB genomic region. The turtle and alligator genomes retain the two copies of SNORD33 yet don't have an obvious SNORD34 gene on the same scaffold. Within the bird genomes, the crow and rifleman each retain a single SNORD33 and SNORD34 gene on the same scaffold. While the ground-finch and zebrafinch retain a single SNORD33 and the bald eagle and seriema retain a single SNORD34 (See Figure 2). In human these snoRNAs are intronic to the host gene, ribosomal protein L13a (RPL13A). Based on BLASTP (version 2.2.18) homology searches for the RPL13A gene, the protein is conserved in the human and turtle genomes and in the bald eagle, crow, rifleman and zebrafinch avian genomes (data not shown). Therefore the conservation patterns of the RPL13A host gene and corresponding intronic snoRNAs largely mimic each other, each independently supports a pattern of loss of the RPL13A gene and the intronic snoRNAs that it hosts in the bird genomes.

### MicroRNAs

MicroRNAs are an important class of non-coding RNA. They have been found in the genomes of Chromalveolata [35, 36], Metazoa [37–39], Mycetozoa [40, 41], Viridiplantae [42–45] and Viruses [46–49]. The miRNAs have been shown to regulate the expression of large numbers of messenger RNAs [50]. The mature miRNA product is generally 22 nucleotides long which is usually processed from a larger RNA that is characterised by a stable hairpin-shaped secondary structure.

Chicken and zebrafinch are the only birds with previously annotated microRNAs. We searched for homologs of these and other vertebrate microRNAs in the genomes of the 48 birds, American alligator and green turtle. Overall, we annotate a total of 16617 putative microRNA loci, homologous to 543 known microRNA genes, of which 487 are annotated in chicken and/or zebra finch, while 56 have been so far known only in non-avian vertebrates. The numbers of annotated loci in the individual species are approximately equal - 300-400 per species, except for the turkey (Meleagris gallopavo) where we identified 543 sequences homologous to known microRNAs.

Nearly all the microRNAs that are broadly conserved in fish, amphibians and mammals are also conserved in the birds. Nevertheless, there are obvious instances of microRNAs lost in all birds. For example,

mammalian, amphibian and fish genomes contain three loci of clustered microRNAs from the mir-17 and mir-92 families. One of these clusters (mir-93 25) was not found in turtles, crocodiles and birds. In addition, we can confidently identify a further 3 microRNA families that are present in mammals, and turtle and/or crocodile, but not in any avian genome (mir-150, mir-208, mir-590). This suggests that these sequences were lost in the last common ancestor of archosaurs or birds. There are also a number of microRNAs that are predicted to be present in turtles and/or crocodiles, and only a small number of bird genomes. Indeed, there are many missing annotations, species-specific and otherwise, that are not consistent with the consensus phylogeny, and could be due to either incomplete genomes or widespread microRNA loss.

The turkey genome contains a high number (190) of microRNAs so far found only in chicken, which account for the higher number of annotated sequences in this genome compared with other birds. This is consistent with its phylogenetic position as the closest chicken relative among the examined birds. However, 101 chicken microRNAs have no homolog in the turkey or other bird genomes, suggesting that these genes are chicken-specific. This is consistent with previous reports of large number of speciesspecific microRNAs in all animals, and supports the view of fast microRNA turnover during animal evolution.

### Long non-coding RNAs

**MAY BE MOVED INTO THE lncRNA PAPER!**

As mentioned in the introduction, lncRNAs are a diverse group of RNAs that have been implicated in a multitude of functional processes [16, 20, 21]. These RNAs have mainly been characterised in mammalian species, particularly human and mouse.

In general Rfam cannot include the entire length of any large, spliced RNAs. This is a limitation of the covariance-models used for the homology-searches Rfam runs [12]. Consequently, only short, well-conserved regions with evolutionarily conserved secondary structures are included in Rfam, we will refer to these as RNA-domains [6].

When analysing the RNA-domain annotations it is striking that many of the lncRNAs with multiple RNA-domains appear to be consistently preserved in the birds. The annotations of these domains are all in the same genomic region supporting a high degree of evolutionary conservation for the entire lncRNA. In particular the HOXA11-AS1, PCA3, RMST, Six3os1, SOX2OT and ST7-OT3 lncRNAs have multiple well conserved RNA-domains (See Figure 3).

The HOX cluster lncRNAs, HOTAIRM1, HOTTIP and HOXA11-AS1, are remarkably well conserved. In

the human genome the corresponding RNA-domains are located at chr7:27135743-27245922 (hg19) and are modelled by 15 RNA-domains in total, the lncRNAs are ordered HOTAIRM1 (5 RNA-domains) → HOXA11-AS1 (6 RNA-domains) → HOTTIP (4 RNA-domains). In the alligator and turtle genomes five and six of the HOX RNA-domains are conserved, each supporting an ancestral gene order of HOTAIRM1 → HOXA11-AS1 → HOTTIP. In the birds, where two or more of the HOX cluster lncRNA RNA-domains were predicted on the same scaffold, this gene order was preserved.

Many of the lncRNAs have been associated with cancer, sparking a minor review industry [51, 52]. Three examples of these that are also conserved in the birds are described below.

The RMST (Rhabdomyosarcoma 2 associated transcript) RNA-domains 6, 7, 8, and 9 are conserved across the birds. In each bird the gene order was also consistent with the human ordering. In the alligator and turtle an additional RNA-domain was predicted in each, these were RNA-domains 2 and 4 respectively, again the ordering of the domains was consistent with human. This suggests that the RMAST lncRNA is highly conserved. However, little is known about the function of this RNA. It was originally identified in a screen for differentially expressed genes in two Rhabdomyosarcoma tumor types [53].

In addition, the lncRNA DLEU2, is well conserved conserved across the vertebrates, it is a host gene for two miRNA genes, miR-15 and miR-16, both of which are also well conserved across the vertebrates (See Figure 2). DLEU2 is thought to be a tumor-suppressor gene as it is frequently deleted in malignant tumours [54, 55].

The NBR2 lncRNA and BRCA1 gene share a bidirectional promotor [56]. Both are expressed in a broad ranch of tissues. Extensive research on BRCA1 has shown that it is involved in DNA repair []. The function of NBR2 remains unknown, yet its conservation across the vertebrates certainly implies a function (See Figure 3).

### Cis-regulatory elements

The cis-regulatory RNAs are a group of RNA structures encoded on mRNAs. Generally they are involved in regulating the expression of the mRNA they are encoded within. Others may recode the translated protein product into an alternate sequence.

This group includes the iron response element (IRE) [57] and the histone 3′ UTR (histone3) [58]. These are structured motifs bound by regulatory proteins. The selenocysteine insertion sequence (SECIS) is a structured motif that recodes UGA stop codons to selenocysteines [59] and the GABRA3 stem-loop is a structure recognised by the ADAR enzyme family. This enzyme edits adenine nucleotides to inosine, in this

case recoding an isoleucine codon to methionine in exon 9 of the GABRA3 gene [60].

These regulatory elements and others, including an internal ribosome entry site (IRES), potassium channel RNA editing signal (K chan RES), Antizyme RNA frameshifting stimulation element (Antizyme FSE), vimentin 3′ UTR protein-binding region (Vimentin3) and a connective tissue growth factor (CTGF) 3′ UTR element (CAESAR) are conserved across a diverse group of vertebrates, including the bird lineages explored here (See Figure 1).

### Exceptional RNAs

A number of other ncRNAs can also be found in Eukaryotic genomes. These do not fit into the main classes of RNA but still perform vital roles in the function and evolution of Eukaryotes. Their functions are diverse and many have not yet been characterised.

An example of an uncharacterised RNA is the ultraconserved element, uc.338 (also known as TUC338) [61–63]. The uc.338 element is derived from a short interspersed element (SINE) called the lobe-finned fishes SINE (LF-SINE) as it is conserved between the coelacanth and mammals [62]. Analysis of the expression of uc.338 implies that it plays a role in the progression of hepatocellular carcinoma, possibly by influencing cell growth [63]. This RNA is conserved in the birds and appears to have been duplicated in several lineages.

The Y RNA is an enigmatic ncRNA where we know very little about the function. It was discovered in the 1980s in ribonucleoprotein complexes [64]. The function of the Y RNAs remain unknown, but evidence is emerging that they may be associated with DNA replication [65]. There are 4 functional Y RNAs encoded in the human genome, Y1, Y3, Y4 and Y5. However, there are hundreds of pseudogenised copies of the Y RNA scattered throughout the human genome [66]. In the birds and other lizards, we identify between two and seven Y RNA paralogs (See Figure 1).

The Vault RNA forms a major component of the vault ribonucleoprotein complex, this is one of the largest particles found in the vertebrate cell; In fact, it is larger than the ribosome [67]. As yet not much is known about the function of Vault. The Vault RNA has been shown to be broadly conserved in metazoans [68]. However, in the bird lineages

### Pseudogenes

The human genome is littered with decaying, pseudogenised copies of ncRNAs. The Dfam database of human repeats estimates that approximately 54% of the human genome is composed of repeat

elements [69]. Of these annotations XXX% are derived from ncRNAs, including the large Alu family of repeats, which is derived from a retransposed SRP RNA pseudogene [].

tRNA, RSV RNA, [70]

*Endogenous retroviruses:* RSV RNA, Alfamo CPB

### Contamination

Bacterial families can be used to identify problematic sequences that are likely to be the result of contamination from non-avian sources.

## Conclusions

## Methods

Bird genomes were searched using the cmsearch program from INFERNAL 1.1 and the covariance models from the Rfam database v11.0 [2–6]. All matches above the curated GA threshold were included. Subsequently, all hits with an E value greater than 0.0005 were discarded, so only matches which passed the model-specific GA threshold and had an E-value smaller than 0.0005 were retained.

tRNA-scan version 1.3.1, ... [13, 14]

Rfam matches and the tRNA-scan results for families belonging to the same clan were then competed so that only the best match was retained. Matches were further sorted into taxonomic groups **ONLY MATCHES IN 5 OR MORE ...**

999 microRNA sequence families, previously annotated in at least one vertebrate, were retrieved from miRBase (v19). Individual sequences or multiple sequence alignments were used to build covariance models with INFERNAL (v1.1rc3), and these models were searched against the 48 bird genomes, and the genomes of the american alligator and the green turtle as outgroups. Hits with e-value ¡10 realigned with the query sequences and the resultant multiple sequence alignments manually inspected and edited using RALEE.

## Acknowledgements

Erich Jarvis, Guojie Zhang and Tom Gilbert. Magnus Alm Rosenblad.

## References

1. Rivas E, Eddy SR: **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics* 2000, **16**(7):583–605.
2. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**:439–41.

3. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33**(Database issue):D121–4.

4. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37**(Database issue):D136–40.

5. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A: **Rfam: Wikipedia, clans and the "decimal" release.** *Nucleic Acids Res* 2011, **39**(Database issue):D141–5.

6. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A: **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Res* 2013, **41**(Database issue):D226–32.

7. Jeffares DC, Poole AM, Penny D: **Relics from the RNA world.** *J Mol Evol* 1998, **46**:18–36.

8. Hoeppner MP, Gardner PP, Poole AM: **Comparative analysis of RNA families reveals distinct repertoires for each domain of life.** *PLoS Comput Biol* 2012, **8**(11):e1002752.

9. Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using RNA-seq.** *Curr Opin Microbiol* 2010, **13**(5):619–24.

10. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, Underwood RC, Haussler D: **Stochastic context-free grammars for tRNA modeling.** *Nucleic Acids Res* 1994, **22**(23):5112–20.

11. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**(11):2079–88.

12. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–7.

13. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**(5):955–64.

14. Chan PP, Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence.** *Nucleic Acids Res* 2009, **37**(Database issue):D93–7.

15. Freyhult EK, Bollback JP, Gardner PP: **Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA**. *Genome Res* 2007, **17**:117–125.

16. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458**(7235):223–7.

17. Eddy SR: **A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.** *BMC Bioinformatics* 2002, **3**:18.

18. Nawrocki EP, Eddy SR: **Query-dependent banding (QDB) for faster RNA similarity searches.** *PLoS Comput Biol* 2007, **3**(3):e56.

19. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol* 2011, **7**(10):e1002195.

20. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.** *Cell* 2007, **129**(7):1311–23.

21. Chow JC, Yen Z, Ziesche SM, Brown CJ: **Silencing of the mammalian X chromosome.** *Annu Rev Genomics Hum Genet* 2005, **6**:69–92.

22. Gardner PP, Bateman A, Poole AM: **SnoPatrol: how many snoRNA genes are there?** *J Biol* 2010, **9**:4.

23. Leonardi J, Box JA, Bunch JT, Baumann P: **TER1, the RNA subunit of fission yeast telomerase.** *Nat Struct Mol Biol* 2008, **15**:26–33.

24. Webb CJ, Zakian VA: **Identification and characterization of the Schizosaccharomyces pombe TER1 telomerase RNA.** *Nat Struct Mol Biol* 2008, **15**:34–42.

25. Mao C, Bhardwaj K, Sharkady SM, Fish RI, Driscoll T, Wower J, Zwieb C, Sobral BW, Williams KP: **Variations on the tmRNA gene.** *RNA Biol* 2009, **6**(4):355–61.

26. Lai LB, Chan PP, Cozen AE, Bernick DL, Brown JW, Gopalan V, Lowe TM: **Discovery of a minimal form of RNase P in Pyrobaculum.** *Proc Natl Acad Sci U S A* 2010, **107**(52):22493–8.

27. Chan PP, Cozen AE, Lowe TM: **Discovery of permuted and recently split transfer RNAs in Archaea.** *Genome Biol* 2011, **12**(4):R38.

28. Dávila López M, Rosenblad MA, Samuelsson T: **Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components.** *Nucleic Acids Res* 2008, **36**(9):3001–10.

29. Marz M, Kirsten T, Stadler PF: **Evolution of spliceosomal snRNA genes in metazoan animals.** *J Mol Evol* 2008, **67**(6):594–607.

30. López MD, Rosenblad MA, Samuelsson T: **Conserved and variable domains of RNase MRP RNA.** *RNA Biol* 2009, **6**(3).

31. Rosenblad MA, López MD, Piccinelli P, Samuelsson T: **Inventory and analysis of the protein subunits of the ribonucleases P and MRP provides further evidence of homology between the yeast and human enzymes.** *Nucleic Acids Res* 2006, **34**(18):5145–56.

32. Marz M, Gruber AR, Höner Zu Siederdissen C, Amman F, Badelt S, Bartschat S, Bernhart SH, Beyer W, Kehr S, Lorenz R, Tanzer A, Yusuf D, Tafer H, Hofacker IL, Stadler PF: **Animal snoRNAs and scaRNAs with exceptional structures.** *RNA Biol* 2011, **8**(6).

33. Lestrade L, Weber MJ: **snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.** *Nucleic Acids Res* 2006, **34**(Database issue):D158–62.

34. Tycowski KT, Shu MD, Steitz JA: **A mammalian gene with introns instead of exons generating stable RNA products.** *Nature* 1996, **379**(6564):464–6.

35. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury JM, Badger JH, Beszteri B, Billiau K, Bonnet E, Bothwell JH, Bowler C, Boyen C, Brownlee C, Carrano CJ, Charrier B, Cho GY, Coelho SM, Collén J, Corre E, Da Silva C, Delage L, Delaroque N, Dittami SM, Doulbeau S, Elias M, Farnham G, Gachon CM, Gschloessl B, Heesch S, Jabbari K, Jubin C, Kawai H, Kimura K, Kloareg B, Küpper FC, Lang D, Le Bail A, Leblanc C, Lerouge P, Lohr M, Lopez PJ, Martens C, Maumus F, Michel G, Miranda-Saavedra D, Morales J, Moreau H, Motomura T, Nagasato C, Napoli CA, Nelson DR, Nyvall-Collén P, Peters AF, Pommier C, Potin P, Poulain J, Quesneville H, Read B, Rensing SA, Ritter A, Rousvoal S, Samanta M, Samson G, Schroeder DC, Ségurens B, Strittmatter M, Tonon T, Tregear JW, Valentin K, von Dassow P, Yamagishi T, Van de Peer Y, Wincker P: **The Ectocarpus genome and the independent evolution of multicellularity in brown algae.** *Nature* 2010, **465**(7298):617–21.

36. Huang A, He L, Wang G: **Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis.** *BMC Genomics* 2011, **12**:337.

37. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75**(5):843–54.

38. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.** *Science* 2001, **294**(5543):858–62.

39. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, of Bioinformatics Computer Labs 2004 S, 2005 C: **The expansion of the metazoan microRNA repertoire.** *BMC Genomics* 2006, **7**:25.

40. Hinas A, Reimegård J, Wagner EG, Nellen W, Ambros VR, Söderbom F: **The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway.** *Nucleic Acids Res* 2007, **35**(20):6714–26.

41. Avesson L, Reimegård J, Wagner EG, Söderbom F: **MicroRNAs in Amoebozoa: deep sequencing of the small RNA population in the social amoeba Dictyostelium discoideum reveals developmentally regulated microRNAs.** *RNA* 2012, **18**(10):1771–82.

42. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP: **MicroRNAs in plants.** *Genes Dev* 2002, **16**(13):1616–26.

43. Fattash I, Voss B, Reski R, Hess WR, Frank W: **Evidence for the rapid expansion of microRNA-mediated regulation in early land plant evolution.** *BMC Plant Biol* 2007, **7**:13.

44. Axtell MJ, Snyder JA, Bartel DP: **Common functions for diverse small RNAs of land plants.** *Plant Cell* 2007, **19**(6):1750–69.

45. Molnár A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC: **miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii.** *Nature* 2007, **447**(7148):1126–9.

46. Pfeffer S, Zavolan M, Grässer FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T: **Identification of virus-encoded microRNAs.** *Science* 2004, **304**(5671):734–6.

47. Ouellet DL, Plante I, Landry P, Barat C, Janelle ME, Flamand L, Tremblay MJ, Provost P: **Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element.** *Nucleic Acids Res* 2008, **36**(7):2353–65.

48. Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grässer FA, van Dyk LF, Ho CK, Shuman S, Chien M, Russo JJ, Ju J, Randall G, Lindenbach BD, Rice CM, Simon V, Ho DD, Zavolan M, Tuschl T: **Identification of microRNAs of the herpesvirus family.** *Nat Methods* 2005, **2**(4):269–76.

49. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, FoÃ  R, Schliwka J, Fuchs U, Novosel A, MÃ$\frac{1}{4}$ller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, De Vita G, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Di Lauro R, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T: **A mammalian microRNA expression atlas based on small RNA library sequencing.** *Cell* 2007, **129**(7):1401–14.

50. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.** *Nature* 2005, **433**(7027):769–73.

51. Prensner JR, Chinnaiyan AM: **The emergence of lncRNAs in cancer biology.** *Cancer Discov* 2011, **1**(5):391–407.

52. Spizzo R, Almeida MI, Colombatti A, Calin GA: **Long non-coding RNAs and cancer: a new frontier of translational research?** *Oncogene* 2012, **31**(43):4577–87.

53. Chan AS, Thorner PS, Squire JA, Zielenska M: **Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between rhabdomyosarcoma subtypes.** *Oncogene* 2002, **21**(19):3029–37.

54. Lerner M, Harada M, LovÃ©n J, Castro J, Davis Z, Oscier D, Henriksson M, Sangfelt O, GrandÃ©r D, Corcoran MM: **DLEU2, frequently deleted in malignancy, functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1.** *Exp Cell Res* 2009, **315**(17):2941–52.

55. Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, Ambesi-Impiombato A, Califano A, Migliazza A, Bhagat G, Dalla-Favera R: **The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia.** *Cancer Cell* 2010, **17**:28–40.

56. Xu CF, Brown MA, Nicolai H, Chambers JA, Griffiths BL, Solomon E: **Isolation and characterisation of the NBR2 gene which lies head to head with the human BRCA1 gene.** *Hum Mol Genet* 1997, **6**(7):1057–62.

57. Stevens SG, Gardner PP, Brown C: **Two covariance models for iron-responsive elements.** *RNA Biol***8**(5):792–801.

58. López D, Samuelsson T: **Early evolution of histone mRNA 3' end processing.** *RNA* 2008, **14**:1–10.

59. Lambert A, Lescure A, Gautheret D: **A survey of metazoan selenocysteine insertion sequences.** *Biochimie* 2002, **84**(9):953–9.

60. Ohlson J, Pedersen JS, Haussler D, Ohman M: **Editing modifies the GABA(A) receptor subunit alpha3.** *RNA* 2007, **13**(5):698–703.

61. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**(5675):1321–5.

62. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultraconserved exon are derived from a novel retroposon.** *Nature* 2006, **441**(7089):87–90.

63. Braconi C, Valeri N, Kogure T, Gasparini P, Huang N, Nuovo GJ, Terracciano L, Croce CM, Patel T: **Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma.** *Proc Natl Acad Sci U S A* 2011, **108**(2):786–91.

64. Lerner MR, Boyle JA, Hardin JA, Steitz JA: **Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus.** *Science* 1981, **211**(4480):400–2.

65. Christov CP, Gardiner TJ, Szüts D, Krude T: **Functional requirement of noncoding Y RNAs for human chromosomal DNA replication.** *Mol Cell Biol* 2006, **26**(18):6993–7004.

66. Mosig A, Guofeng M, Stadler BM, Stadler PF: **Evolution of the vertebrate Y RNA cluster.** *Theory Biosci* 2007, **126**:9–14.

67. Kong LB, Siva AC, Rome LH, Stewart PL: **Structure of the vault, a ubiquitous celular component.** *Structure* 1999, **7**(4):371–9.

68. Stadler PF, Chen JJ, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, Kretzschmar AK, Mosig A, Prohaska SJ, Qi X, Schutt K, Ullmann K: **Evolution of vault RNAs.** *Mol Biol Evol* 2009, **26**(9):1975–91.

69. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD: **Dfam: a database of repetitive DNA based on profile hidden Markov models.** *Nucleic Acids Res* 2013, **41**(Database issue):D70–82.

70. International Chicken Genome Sequencing Consortium C: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**(7018):695–716.

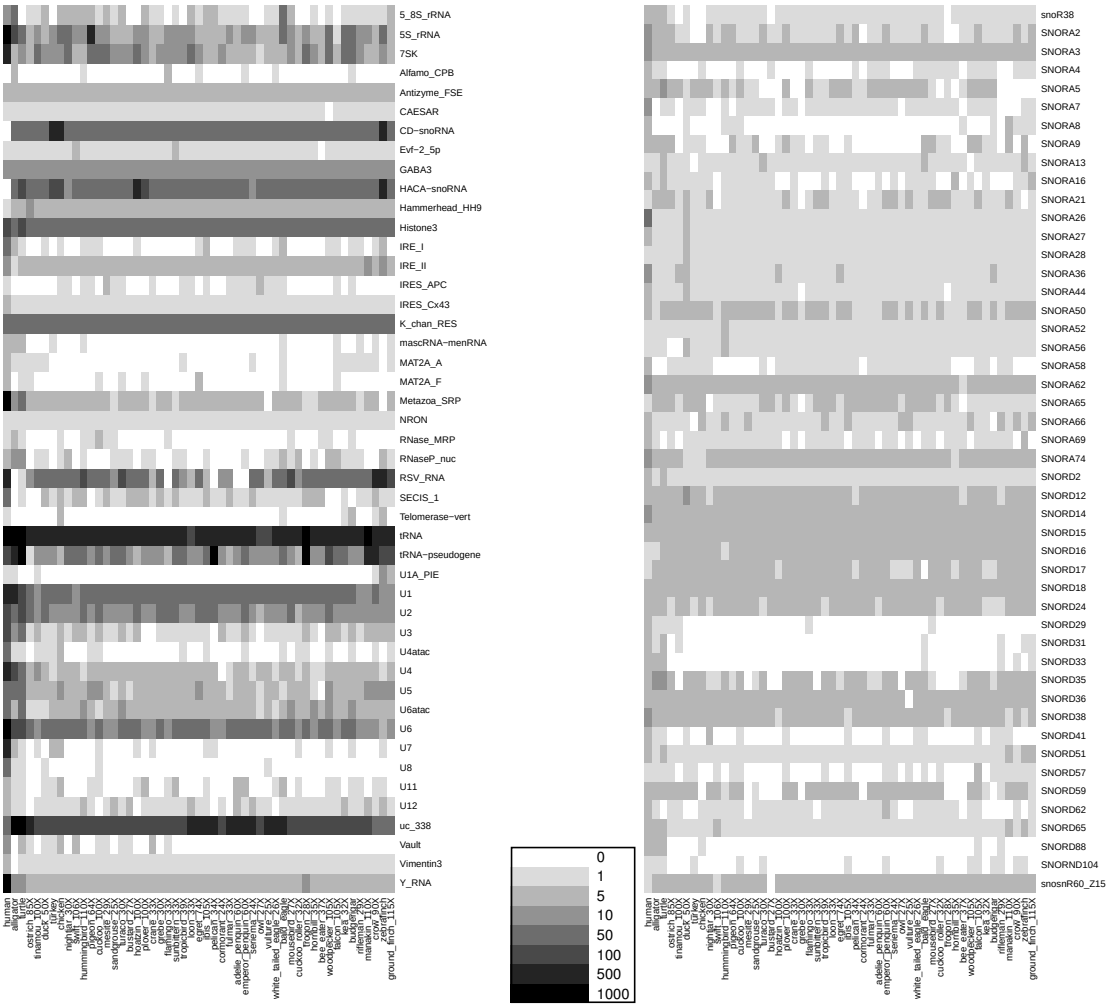


Figure 1: This figure illustrates the number of copies of Rfam and tRNAscan annotations for each RNA family in each avian species and outgroups. The families are ordered alphabetically vertically and the species are in phylogenetic order horizontally. Darker shades indicate high copy numbers, lighter shades indicate fewer, white indicates no predictions were made.

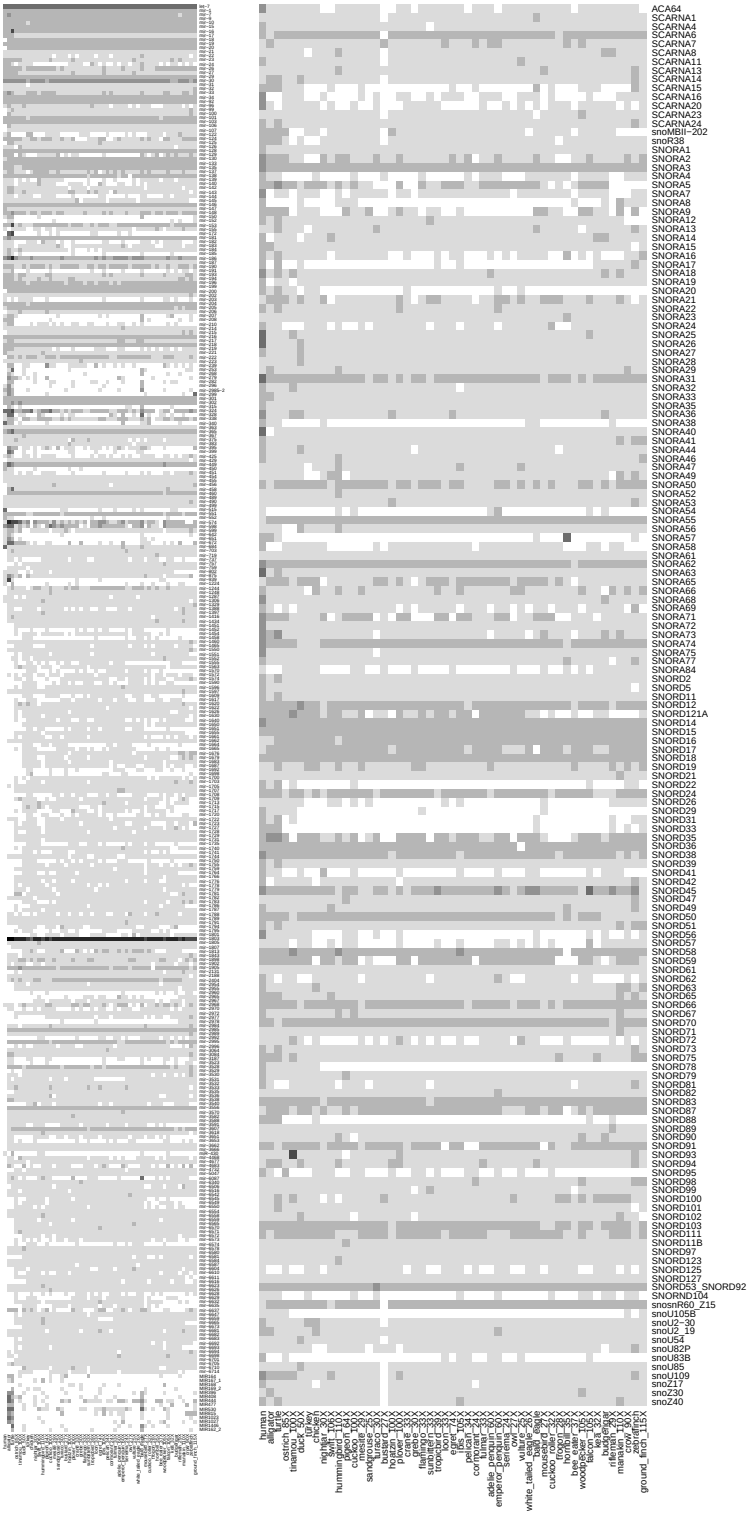
ACA64
SCARNA1
SCARNA4
SCARNA6
SCARNA7
SCARNA8
SCARNA11
SCARNA13
SCARNA14
SCARNA15
SCARNA16
SCARNA20
SCARNA23
SCARNA24
snoMBII−202
snoR38
SNORA1
SNORA2
SNORA3
SNORA4
SNORA5
SNORA7
SNORA8
SNORA9
SNORA12
SNORA13
SNORA14
SNORA15
SNORA16
SNORA17
SNORA18
SNORA19
SNORA20
SNORA21
SNORA22
SNORA23
SNORA24
SNORA25
SNORA26
SNORA27
SNORA28
SNORA29
SNORA31
SNORA32
SNORA33
SNORA35
SNORA36
SNORA38
SNORA40
SNORA41
SNORA44
SNORA46
SNORA47
SNORA49
SNORA50
SNORA52
SNORA53
SNORA54
SNORA55
SNORA56
SNORA57
SNORA58
SNORA61
SNORA62
SNORA63
SNORA65
SNORA66
SNORA68
SNORA69
SNORA71
SNORA72
SNORA73
SNORA74
SNORA75
SNORA77
SNORA84
SNORD2
SNORD5
SNORD11
SNORD12
SNORD121A
SNORD14
SNORD15
SNORD16
SNORD17
SNORD18
SNORD19
SNORD21
SNORD22
SNORD24
SNORD26
SNORD29
SNORD31
SNORD33
SNORD35
SNORD36
SNORD38
SNORD39
SNORD41
SNORD42
SNORD45
SNORD47
SNORD49
SNORD50
SNORD51
SNORD56
SNORD57
SNORD58
SNORD59
SNORD61
SNORD62
SNORD63
SNORD65
SNORD66
SNORD67
SNORD70
SNORD71
SNORD72
SNORD73
SNORD75
SNORD78
SNORD79
SNORD81
SNORD82
SNORD83
SNORD87
SNORD88
SNORD89
SNORD90
SNORD91
SNORD93
SNORD94
SNORD95
SNORD98
SNORD99
SNORD100
SNORD101
SNORD102
SNORD103
SNORD111
SNORD11B
SNORD97
SNORD123
SNORD125
SNORD127
SNORD53_SNORD92
SNORND104
snosnR60_Z15
snoU105B
snoU2−30
snoU2_19
snoU54
snoU82P
snoU83B
snoU85
snoU109
snoZ17
snoZ30
snoZ40
human
elephant
turtle
frog
ostrich
tinamou
duck
turkey
hoatzin
nightjar
owl
wild_turkey
hummingbird
cuckoo
osprey
mesite
sandgrouse
turaco
hoatzin
parrot
crane
limpkin
flamingo
sunbittern
tropicbird
loon
grebe
tropicbird
flamingo
penguin
adelie_penguin
emperor_penguin
sunbittern
owl
grebe
white_tailed_eagle
turaco
mousebird
cuckoo
hoatzin
bee_eater
woodpecker
falcon
kea
hummingbird
manakin
crow
ground_finch

Figure 2: .

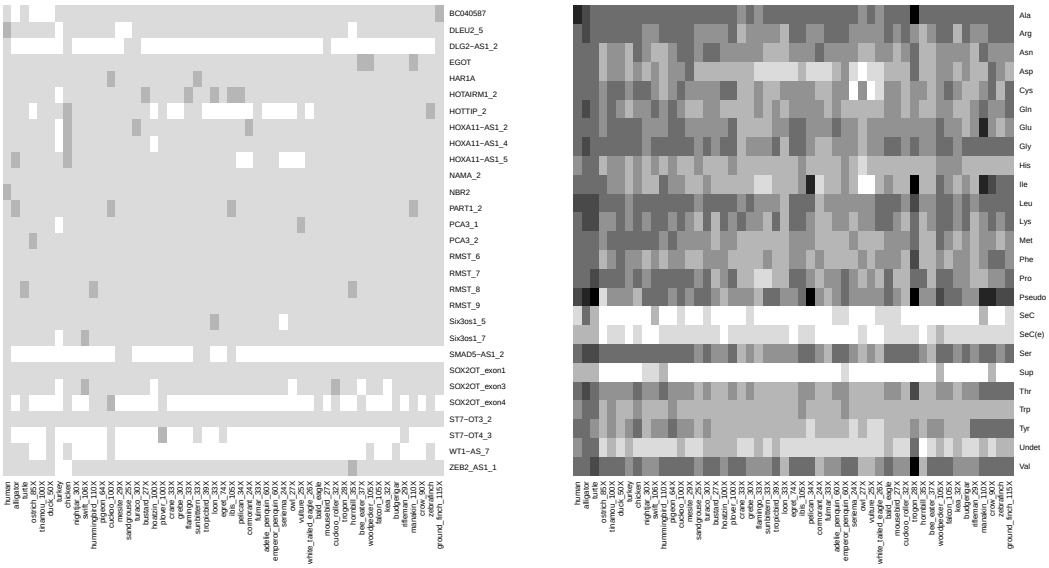
BC040587
DLEU2_5
DLG2−AS1_2
EGOT
HAR1A
HOTAIRM1_2
HOTTIP_2
HOXA11−AS1_2
HOXA11−AS1_4
HOXA11−AS1_5
NAMA_2
NBR2
PART1_2
PCA3_1
PCA3_2
RMST_6
RMST_7
RMST_8
RMST_9
Six3os1_5
Six3os1_7
SMAD5−AS1_2
SOX2OT_exon1
SOX2OT_exon3
SOX2OT_exon4
ST7−OT3_2
ST7−OT4_3
WT1−AS_7
ZEB2_AS1_1
human
alligator
turtle
X_tropicalis
osprich_85X
tinamou_100X
duck_90X
turkey
chicken
nightjar_30X
swift_100X
hummingbird_110X
mouse_lemur_64X
cuckoo_100X
crane_29X
sandgrouse_29X
turaco_30X
hoatzin_27X
prairie_chicken_100X
crane_30X
grebe_30X
flamingo_30X
sunbittern_33X
tropicbird_38X
loon_33X
egret_74X
ibis_105X
pelican_34X
cormorant_24X
tuna_23X
adelie_penguin_60X
emperor_penguin_60X
seriema_24X
ow_27X
vulture_30X
white_rabbit_eagle_26X
bald_eagle
mousebird_27X
cuckoo_roller_32X
rockpen_30X
hornbill_38X
bee_eater_27X
woodpecker_105X
falcon_105X
kea_32X
budgerigar
hummingbird_105X
dikkop_23X
mouse_110X
crow_90X
wallaby_90X
ground_finch_115X
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Pseudo
SeC
SeC(e)
Ser
Sup
Thr
Trp
Tyr
Undet
Val

Figure 3: .