Conservation and losses non-coding RNA associated loci in avian genomes

Paul P. Gardner*[1,2] , Mario Fasold[5,6] , Sarah W. Burge[3] , Maria Ninova[4] , Jana Hertel[5] , Stephanie Kehr[5] , Tammy E. Steeves[1] , Sam Griffiths-Jones[4] and Peter F. 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. [6] ecSeq Bioinformatics, Brandvorwerkstr.43, D-04275 Leipzig, Germany.

Email: Paul P. Gardner*- paul.gardner@canterburmotifs/manuscript/y.ac.nz; mario@bioinf.uni-leipzig.de; swb@ebi.ac.uk; Maria.Ninova@postgrad.manchester.ac.uk; Jana Hertel*- jana@bioinf.uni-leipzig.de; steffi@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 RNA loci 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. We identify 34 lncRNA-associated loci that are conserved between birds and mammals and validate 12 of these in chicken. These include several intriguing cases where the reported mammalian lncRNA function is not conserved in birds. We also demonstrate extensive conservation of classical ncRNAs (e.g., tRNAs) and more recently discovered ncRNAs (e.g., snoRNAs and miRNAs) in birds. Furthermore, we describe numerous "losses" of several RNA families, and attribute these to genuine loss, divergence or missing data. In particular, we show that many of these losses are due to the challenges associated with assembling Avian microchromosomes. These combined results illustrate the utility of applying homology-based methods for annotating novel vertebrate genomes.

Introduction

Non-coding RNAs (ncRNAs) are an important class of genes, responsible for the regulation of many key cellular functions. 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 [1]. 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) [2].

The ncRNAs pose serious research challenges, particularly for the field of genomics. For example, they lack the strong statistical signals associated with protein coding genes, e.g. open reading frames, G+C content and codon-usage biases [3].

New sequencing technologies have dramatically expanded the rate at which ncRNAs are discovered and their functions are determined [4]. However, in order to determine the full range of ncRNAs across multiple species we require multiple RNA fractions (e.g. long and short), in multiple species, in multiple developmental stages and tissues types. The costs of this approach are still prohibitive in terms of researcher-time and finances. Consequently, in this study we concentrate on bioinformatic approaches, primarily we use homology-based methods (i.e. covariance models (CMs)). However, we do validate the majority of these predictions using RNA-seq. The bioinformatic approaches we use remain state of the art for ncRNA bioinformatic analyses [5–7] and have well established sensitivity and specificity rates [8]. For example, the CM based approach for annotating ncRNAs in genomes requires reliable alignments and consensus secondary structures of representative sequences of RNA families, many of which can be found at Rfam [9–13]. These are used to train probabilistic models that score the likelihood that a database sequence is generated by the same evolutionary processes as the training sequences based upon both sequence and structural information [5–7]. The tRNAscan-SE software package uses CMs to accurately predict transfer RNAs [14, 15].

Independent benchmarks of bioinformatic annotation tools have shown that the CM approaches out-perform alternative methods [8], although their sensitivity can be limited for rapidly evolving families such as vault RNAs or telomerase RNA [16].

The publication of 48 avian genomes, including the previously published chicken [17], zebra finch [18] and turkey [19] with the recently published 45 avian genomes [20–26], provides an exciting opportunity to explore conservation of genomic loci that have been associated with ncRNAs in unprecedented detail.

In the following we explore the conservation patterns of the major classes of avian ncRNA loci in further detail. Using homology search tools and evolutionary constraints, we have produced a set of genome

annotations for 48 predominantly non-model bird species for ncRNAs that are conserved across the avian species. This conservative set of annotations is expected to contain the core avian ncRNA loci. We focus our report on the unusual results within the avian lineages. These are either unexpectedly well-conserved ncRNAs or unexpectedly poorly-conserved ncRNAs. The former are ncRNA loci that were not expected to be conserved between the birds and the other vertebrates, particularly those ncRNAs whose function is not conserved in birds. The latter are apparent losses of ncRNA loci expected to be conserved; Here, we consider three categories of such "loss": First, genuine gene losses in the avian lineage where ncRNAs well conserved in other vertebrates are completely absent in birds. Second, "divergence" where ncRNAs have undergone such significant sequence and structural alternations that homology search tools can no longer detect a relationship between other vertebrate exemplars and avian varieties. Third, "missing" ncRNAs that failed to be captured in the available, largely fragmented, avian genomes. We postulate that the latter category is likely to be prevalent in comparative avian genome studies given the distinctive organisation of the avian genome. Namely, the avian karyotype is characterised by a large number of chromosomes (average $2n \approx 80$) generally consisting of approximately 5 larger "macrochromosomes" and many smaller "microchromosomes" [27]. This 'so many, so small' pattern presents significant assembly challenges [28]. Indeed, of the 48 published avian genomes, 20 of which are high-coverage ($> 50X$), only two are chromosomally assembled (chicken and zebra finch; [18, 20]).

Results

There is substantial gain and loss of lncRNAs and other ncRNA associated loci over evolutionary time [2, 29, 30]. It is difficult to assess how many of these "gains" and "losses" are due to limited bioinformatic sequence alignment tools (these generally fail align correctly below 60-50% sequence identity [31]) or due to genuine gains and losses or data missing from the current genome assemblies. Nevertheless, sequence conservation, generally speaking, provides useful evidence for gene and function conservation.

We have identified 66,879 loci in 48 avian genomes that share sequence similarity with previously characterised ncRNAs and are conserved in $> 10\%$ of these avian genomes. These loci have been classified into 626 different families, the majority of which correspond to miRNAs and snoRNAs (summarised in Table 1). Out of necessity we have selected a modest number of families for further discussion. These include the lncRNAs that appear to be conserved between Mammals and Aves and the cases of apparent loss of genes that conserved in most other Vertebrates.

Unusually well conserved RNAs

The bulk of the "unusually well conserved RNAs" belong to the long non-coding RNA (lncRNA) group. The lncRNAs are a diverse group of RNAs that have been implicated in a multitude of functional processes [32–35]. These RNAs have largely been characterised in mammalian species, particularly human and mouse. Consequently, we generally do not expect these to be conserved outside of Mammals. Notable examples include Xist [36] and H19 [37]. There is emerging evidence for the conservation of "mammalian" lncRNAs in Vertebrates [38,39]), however, like most lncRNAs, the function of these lncRNAs remains largely unknown. Here, we show the conservation of several lncRNAs that have been well-characterised in humans.

The CM based approach is appropriate for most classes of ncRNA, but the lncRNAs are a particular challenge [34]. CMs cannot model the exon-intron structures of spliced lncRNAs, nor do they deal elegantly 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 possibly structured elements) within lncRNAs, analogous to the "domains" housed within protein sequences [13]. Whilst some these regions may not reflect functional RNA elements but instead regulatory regions, enhancers or insulators, their syntenic conservation still provides an indication of lncRNA conservation [40].

When analysing the RNA-domain annotations it is striking that the order (synteny) of many of the lncRNAs with multiple RNA-domains are consistently preserved in the birds. The annotations of these domains lie in the same genomic region, in the same order as in the mammalian homologs. Thus they support a high degree of evolutionary conservation for the entire lncRNA. In particular the HOXA11-AS1, PART1, PCA3, RMST, Six3os1, SOX2OT and ST7-OT3 lncRNAs have multiple, well conserved RNA-domains (See Figure 1). The syntenic ordering of these seven lncRNAs and the flanking genes are also preserved between the human and chicken genomes (data not shown). We illustrate this in detail for the HOTAIRM1 lncRNA (See Supplemental Figures 13&14).

The conservation of these "human" lncRNAs among birds suggests they may also be functional in birds. But what these functions may be is not immediately obvious. For example, PART1 and PCA3 are both described as prostate-specific lncRNAs that play a role in the human androgen-receptor pathway [41–43]. Birds lack a prostate but both males and females express the androgen receptor (AR or NR3C4) in gonadal and non- gonadal tissue [44–47]. Thus, we postulate that PART1 and PCA3 also play a role in the androgen-receptor pathway in birds but whether the expression of these lncRNAs are tissue specific is unknown at present.

The HOX cluster lncRNAs HOTAIRM1 (5 RNA-domains), HOXA11-AS1 (6 RNA-domains), and HOTTIP (4 RNA domains) are conserved across the Mammalian and Avian lineages. In the human genome they are located in the HOXA cluster (hg coordinates chr7:27135743-27245922), one of the most highly conserved regions in vertebrate genomes [48], in antisense orientation between HoxA1 and HoxA2, between HoxA11 and HoxA13, and upstream of HoxA13, respectively. Conservation and expression of HOTAIRM1 and HOXA11-AS1 within the HOXA cluster has been studied in some detail in marsupials [49]. Of the 15 RNA-domains five and six representing all three lncRNAs were recovered in the alligator and turtle genomes. All of them appear in the correct order at the expected, syntenically conserved positions within the HOXA cluster. In the birds, where two or more of the HOX cluster lncRNA RNA-domains were predicted on the same scaffold, this gene order and location within HOX was also preserved.

The majority ($> 80\%$) of genome-wide association studies of cancer identify loci outside of protein-coding genes [50]. Many of these are loci are now known to be transcribed into lncRNAs. Furthermore, many lncRNAs are differentially expressed in tumorous tissues [51], suggesting further mechanistic links with the aberrant gene expression associated cancer progression. In our work we have identified 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 RMST 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 [51].

In addition, the lncRNA DLEU2 is well 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 Supplemental Figure 2). DLEU2 is thought to be a tumor-suppressor gene as it is frequently deleted in malignant tumours [52, 53].

The NBR2 lncRNA and BRCA1 gene share a bidirectional promotor [54]. Both are expressed in a broad range of tissues. Extensive research on BRCA1 has shown that it is involved in DNA repair [55]. The function of NBR2 remains unknown, yet its conservation across the vertebrates certainly implies a function (See Figure 1). We note that the function for this locus may be at the DNA level, however, function at the RNA level cannot be ruled out at this stage.

Of the other classes of RNAs, none showed an unexpected degree of conservation or expansion within the

avian lineage. The only exception being the snoRNA, SNORD93. SNORD93 has 92 copies in the tinamou genome, whereas it only has 1-2 copies in all the other vertebrate genomes.

Unexpectedly poorly conserved ncRNAs: genuine loss, divergence or missing data?

Genuine loss

The overall reduction in avian genomic size has been extensively discussed elsewhere [56]. Unsurprisingly, this reduction is reflected in the copy-number of ncRNA genes. Some of the most dramatic examples are the transfer RNAs and pseudogenes which average ~ 900 and ~ 580 copies in the human, turtle and alligator genomes, the average copy-numbers of these drop to ~ 280 and ~ 100 copies in the avian genomes. In addition to reduction in copy-number, the absence of several, otherwise ubiquitous vertebrate ncRNAs, in the avian lineage are suggestive of genuine gene loss.

Namely, mammalian and amphibian genomes contain three loci of clustered microRNAs from the mir-17 and mir-92 families [57]. One of these clusters (cluster II, with families mir-106b, mir-93 and mir-25) was not found in turtles, crocodiles and birds (see Supplemental Figure 6). In addition, the microRNA family let-7 is the most diverse microRNA family with 14 paralogs in human. These genes also localize in 7 genomic clusters, together with mir-100 and mir-125 miRNA families (see previous study on the evolution of the let-7 miRNA cluster in [58]). In Sauropsids we observed that cluster A - which is strongly conserved in vertebrates has been completely lost in the avian lineage. Another obvious loss in birds is cluster F, containing two let-7 microRNA paralogs. Cluster H, on the other hand has been retained in all oviparous animals and completely lost later, after the split of Theria (see Supplemental Figure 7).

Divergence

In order to determine to what extent the absence of some ncRNAs from the infernal-based annotation is caused by sequence divergence beyond the thresholds of the Rfam CMs, we complemented our analysis by dedicated searches for a few of these RNA groups. Our ability to find additional homologs for several RNA families that fill gaps in the abundance matrices (Figure 1) strongly suggests that conspicuous absences, in particular of LUCA and LECA RNAs, are caused by incomplete data in the current assemblies and sequence divergence rather then genuine losses.

Vertebrate Y RNAs typically form a cluster comprising four well-defined paralog groups Y1, Y3, Y4, and Y5. In line with [59] we find that the Y5 paralog family is absent from all bird genomes, while it is still present in both alligator and turtle, see Supplemental Figure 4. Within the avian lineage, we find a

conserved Y4-Y3-Y1 cluster. Apparently, broken-up clusters are in most cases consistent with breaks (e.g. ends of contigs) in the available sequence assemblies. In several genomes we observe one or a few additional Y RNA homologs unlinked to the canonical Y RNA cluster. These sequences can be identified unambiguously as derived members of one of the three ancestral paralog groups, they almost always fit less well to the consensus (as measured by the CM bit score of paralog group specific covariance models) than the paralog linked to cluster, and there is no indication that any of these additional copies is evolutionarily conserved over longer time scales. We therefore suggest that most or all of these interspersed copies are in fact pseudogenes (see below).

Missing data

Seven families of ncRNAs were found in some avian genomes but not others (Figure 1). These families range in conservation level from being ubiquitous to cellular-life (RNase P and tRNA-sec), present in most Bilateria (vault), present in the majority of eukaryotes (RNase MRP, U4atac and U11) and present in all vertebrates (telomerase) [2]. Therefore, the genuine loss or even diversification of these ncRNA families in the avian lineage is unlikely. Rather, this lack of phylogenetic signal, combined with the fragmented nature of the vast majority of these genomes described above (i.e., of the 48 avian genomes, only the chicken and zebra finch are chromosomally assembled [18, 20]), suggests the most likely explanation is that these ncRNA families are indicative of missing data. Indeed, of the seven missing ncRNA families, six where found in the chicken genome and three were found in the zebra finch genome. Furthermore, only one of these (RNase MRP) is found on a macrochromosome, and all remaining missing ncRNAs are found on microchromosomes (see Supplemental Table 1). A Fisher's exact test showed that there is significantly more missing ncRNAs on microchromosomes than macrochromosomes ($P < 10^{16}$ for both the chicken and zebra finch). Thus, we suggest that many of these ncRNAs families are missing because: (1) they are predominantly found on microchromosomes [this study] and (2) the vast majority of avian microchromosomes remain unassembled [20, 28].

To wit, we performed dedicated searches for a selection of these missing ncRNA families. Here, tRNAscan is tuned for specificity and thus misses several occurrences of tRNA-sec that are easily found in the majority of genomes by `blastn` with $E \leq 10^{-30}$. In some cases the sequences appear degraded at the ends, which is likely due to low sequence quality at the very ends of contigs or scaffolds. A `blastn` search also readily retrieves additional RNase P and RNAse MRP RNAs in the majority of genomes, albeit only the best conserved regions are captured. In many cases these additional candidates are incomplete or contain

undetermined sequence, which explains why they are missed by the CMs [60, 61].

Pseudogenes

Non-coding RNA derived pseudogenes are a major problem for many ncRNA annotation projects. The human genome, for example, contains > 1 million Alu repeats, which are derived from the SRP RNA [62]. The existing Rfam annotation of the human genome, in particular, contains a number of problematic families that appear to have been excessively pseudogenised. The U6 snRNA, SRP RNA and Y RNA families have 1,371, 941 and 892 annotations in the human genome. These are a heterogenous mix of pseudogenised, paralogous, diverged or functional copies of these families. Unfortunately, a generalised model of RNA pseudogenes has not been incorporated into the main covariance model package, Infernal. An approach used by tRNAscan [14], is, in theory, generalisable to other RNA families but this remains a work in progress.

It is possible that the avian annotations also contains excessive pseudogenes. However, it has previously been noted that avian genomes are significantly smaller than other vertebrate species [17]. We have also noted a corresponding reduction in the number of paralogs and presumed ncRNA-derived pseudogenes in the avian genomes (See Supplemental Figure 12). The problematic human families, U6 snRNA, SRP RNA and Y RNA have, for example, just 26, 4 and 3 annotations respectively in the chicken genome and 13, 3 and 3 annotations respectively, on average, in the 48 avian genomes used here. Therefore, we conclude that the majority of our annotations are in fact functional orthologs.

Experimentally confirmed ncRNAs

The ncRNAs presented here have been identified using homology models and are evolutionarily conserved in multiple avian species. In order to further validate these predictions we have used strand-specific total RNA-seq and small RNA-seq of multiple chicken tissues. After mapping the RNA-seq data to the chicken genome (see Methods for details), we identified a threshold for calling a gene as expressed by limiting our estimated false-positive rate to approximately 10%. This FDR was estimated using a negative control of randomly selected, un-annotated regions of the genome. Since some regions may be genuinely expressed, the true FDR is potentially lower than 10%. Overall, the number of ncRNAs we have identified in this work that are expressed above background levels is 865 (72.4%) (see Table 1). This shows that 7.0 times more of our ncRNAs are expressed than expected by chance (Fisher's exact test: $P < 10^{16}$). This number is an underestimate of the fraction of our annotations that are genuinely expressed, as only a fraction of

the developmental stages and tissues of chicken have been characterized with RNA-seq. Furthermore, some ncRNAs are expressed in highly specific conditions [63, 64].

The classes of RNAs where the majority of our annotations were experimentally confirmed includes microRNAs, snoRNAs, cis-regulatory elements, tRNAs, SRP RNA and RNase P/MRP RNA. The RNA-seq data could not provide evidence for a telomerase RNA transcript, which are only generally only expressed in embryonic, stem or cancerous tissues. Only a small fraction of the 7SK RNA, the minor spliceosomal RNAs and the lncRNAs could be confirmed with the 10% FDR threshold. There are a number of possible explanations for this: the multiple copies of the 7SK RNA may be functionally redundant and can therefore compensate for one another; The minor spliceosome is, as the name suggests, a rarely used alternative spliceosome; and the lncRNAs are generally expressed at low levels under specific conditions [63, 65].

Conclusions

In this work we have provided a comprehensive annotation of non-coding RNAs in genome sequences using homology-based methods. The homology-based tools have distinct advantages over experimental-based approaches as not all RNAs are expressed in any particular tissue-type or developmental-stage, in fact some RNAs have extremely specific expression profiles, e.g. the lsy-6 microRNA [66]. We have identified previously unrecognised conservation of ncRNAs in avian genomes and some surprising "losses" of otherwise well conserved ncRNAs. We have shown that most of these losses are due to difficulties assembling avian microchromosomes rather than *bona fide* gene loss. A large fraction of our annotations have been confirmed using RNA-seq data, which also showed a 7-fold enrichment of expression within our annotations relative to unannotated regions.

The collection of ncRNA sequences is generally biased towards model organisms [2, 67]. However, we have shown that using data from well studied lineages such as mammals can also result in quality annotations of sister taxa such as Aves.

In summary, these results indicate we are in the very early phases of determining the functions of many RNA families. This is illustrated by the fact that the reported functions of some ncRNAs are mammal-specific, yet these are also found in bird genomes.

Methods

The 48 bird genome sequences used for the following analyses are available from the phylogenomics analysis of birds website [68].

Bird genomes were searched using the cmsearch program from INFERNAL 1.1 and the covariance models from the Rfam database v11.0 [12, 13]. 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. The Rfam database classifies non-coding RNAs into hierarchical groupings. 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 [9–13]; these categories have been used to classify our results.

In order to obtain good annotations of tRNA genes we ran the specialist tRNA-scan version 1.3.1 annotation tool. This method also uses covariance models to identify tRNAs. However it also uses some heuristics to increase the search-speed, annotates the Isoacceptor Type of each prediction and uses sequence analysis to infer if predictions are likely to be functional or tRNA-derived pseudogenes [14, 15]. 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 for any genomic region [12]. To further increase the specificity of our annotations we filtered out families that were identified in four or fewer of the 51 vertebrate species we have analysed in this work. These filtered families largely corresponded to bacterial contamination within the genomic sequences.

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.

An additional snoRNA homology search was performed with snoStrip [69]. As initial queries we used deutorostomian snoRNA families from human [70], platypus [71], and chicken [72].

The diverse sets of genome annotations were combined and filtered, ensuring conservation in 10% or more of the avian genomes. We collapsed the remaining overlapping annotations into a single annotation. We also generated heatmaps for different groups of ncRNA genes (see Figure 1 and Supplemental Figure 1-3). All the scripts and annotations presented here are available from Github [73].

Chicken ncRNA predictions were validated using two separate RNA-seq data sets. The first data set

(Bioproject PRJNA204941) contains 971 million reads and comprises 27 samples from 14 different chicken tissues sequenced on Illumina HiSeq2000 using a small RNA-seq protocol. The second data set (SRA accession SRP041863) contains 1,46 billion Illumina HiSeq reads sequenced from whole chicken embryo RNA from 7 stages using a strand-specific dUTP protocol. Raw reads were checked for quality and adapters clipped if required by the protocol. Preprocessed reads were mapped to the galGal4 reference genome using SEGEMEHL short read aligner [74] and then overlapped with the ncRNA annotations.

Acknowledgements

Erich Jarvis (Duke University), Guojie Zhang (BGI-Shenzhen & University of Copenhagen) and Tom Gilbert (University of Copenhagen) for access to data and for invaluable feedback on the manuscript. Magnus Alm Rosenblad (Univ. of Gothenburg) and Eric Nawrocki (HHMI Janelia Farm) for useful discussions. Matthew Walters for assistance with figures.

We thank Fiona McCarthy (University of Arizona) and Carl Schmidt (University of Delaware) as well as Matt Schwartz (Harvard) and Igor Ulitsky (Weizmann Institute of Science) for providing the RNA-seq data as part of the Avian RNAseq consortium.

Thanks to @ewanbirney for the following timely tweet: "So ... missing orthologs to chicken often mean 'gene might be on the microchromosome".

We thank the anonymous referees for providing invaluable suggestions that improved this work.

References

1. Jeffares DC, Poole AM, Penny D: **Relics from the RNA world.** *J Mol Evol* 1998, **46**:18–36.
2. 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.
3. Rivas E, Eddy SR: **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics* 2000, **16**(7):583–605.
4. Cech TR, Steitz JA: **The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones**. *Cell* 2014, **157**:77–94.
5. 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.
6. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**(11):2079–88.
7. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–7.
8. 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.
9. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**:439–41.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

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

16. Menzel P, Gorodkin J, Stadler PF: **The Tedious Task of Finding Homologous Non-coding RNA Genes**. *RNA* 2009, **15**:2075–2082.

17. 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.

18. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TA, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin YC, George J, Sweedler J, Southey B, Gunaratne P, Watson M, Nam K, Backström N, Smeds L, Nabholz B, Itoh Y, Whitney O, Pfenning AR, Howard J, Völker M, Skinner BM, Griffin DK, Ye L, McLaren WM, Flicek P, Quesada V, Velasco G, Lopez-Otin C, Puente XS, Olender T, Lancet D, Smit AF, Hubley R, Konkel MK, Walker JA, Batzer MA, Gu W, Pollock DD, Chen L, Cheng Z, Eichler EE, Stapley J, Slate J, Ekblom R, Birkhead T, Burke T, Burt D, Scharff C, Adam I, Richard H, Sultan M, Soldatov A, Lehrach H, Edwards SV, Yang SP, Li X, Graves T, Fulton L, Nelson J, Chinwalla A, Hou S, Mardis ER, Wilson RK: **The genome of a songbird.** *Nature* 2010, **464**(7289):757–62.

19. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg LA, Bouffard P, Burt DW, Crasta O, Crooijmans RP, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MA, Harkins TT, Herrero J, Hoffmann S, Megens HJ, Jiang A, de Jong P, Kaiser P, Kim H, Kim KW, Kim S, Langenberger D, Lee MK, Lee T, Mane S, Marcais G, Marz M, McElroy AP, Modise T, Nefedov M, Notredame C, Paton IR, Payne WS, Pertea G, Prickett D, Puiu D, Qioa D, Raineri E, Ruffier M, Salzberg SL, Schatz MC, Scheuring C, Schmidt CJ, Schroeder S, Searle SM, Smith EJ, Smith J, Sonstegard TS, Stadler PF, Tafer H, Tu ZJ, Van Tassell CP, Vilella AJ, Williams KP, Yorke JA, Zhang L, Zhang HB, Zhang X, Zhang Y, Reed KM: **Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis.** *PLoS Biol* 2010, **8**(9).

20. Avian Genome Project Consortium: **Genome evolution and biodiversity of the avian class**. *In press* 2014.

21. Avian Phylogenomics Consortium: **Using whole genomes to resolve the tree of life of modern birds**. *In press* 2014.

22. Huang Y, Li Y, Burt DW, Chen H, Zhang Y, Qian W, Kim H, Gan S, Zhao Y, Li J, Yi K, Feng H, Zhu P, Li B, Liu Q, Fairley S, Magor KE, Du Z, Hu X, Goodman L, Tafer H, Vignal A, Lee T, Kim KW, Sheng Z, An Y, Searle S, Herrero J, Groenen MA, Crooijmans RP, Faraut T, Cai Q, Webster RG, Aldridge JR, Warren WC, Bartschat S, Kehr S, Marz M, Stadler PF, Smith J, Kraus RH, Zhao Y, Ren L, Fei J, Morisson M, Kaiser P, Griffin DK, Rao M, Pitel F, Wang J, Li N: **The duck genome and transcriptome provide insight into an avian influenza virus reservoir species.** *Nat Genet* 2013, **45**(7):776–83.

23. Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H, Chen Y, Xia J, Luo Q, Xu P, Chen Y, Liao S, Cao C, Gao S, Wang Z, Yue Z, Li G, Yin Y, Fox NC, Wang J, Bruford MW: **Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle.** *Nat Genet* 2013, **45**(5):563–6.

24. Shapiro MD, Kronenberg Z, Li C, Domyan ET, Pan H, Campbell M, Tan H, Huff CD, Hu H, Vickrey AI, Nielsen SC, Stringham SA, Hu H, Willerslev E, Gilbert MT, Yandell M, Zhang G, Wang J: **Genomic diversity and evolution of the head crest in the rock pigeon.** *Science* 2013, **339**(6123):1063–7.

25. Howard J, Koren S, Phillippy A, Zhou S, Schwartz D, Schatz M, Aboukhalil R, Ward J, Li J, Li B, Fedrigo O, Bukovnik L, Wang T, Wray G, Rasolonjatovo I, Winer R, Knight J, Warren W, Zhang G, Jarvis E: **De novo high-coverage sequencing and annotated assemblies of the budgerigar genome**. *GigaScience Database* 2013.

26. Li J, *et al*: **The genomes of two Antarctic penguins reveal adaptations to the cold aquatic environment** 2014. [Submitted].

27. Griffin DK, Robertson LB, Tempest HG, Skinner BM: **The evolution of the avian genome as revealed by comparative molecular cytogenetics.** *Cytogenet Genome Res* 2007, **117**(1-4):64–77.

28. Ellegren H: **The avian genome uncovered.** *Trends Ecol Evol* 2005, **20**(4):180–6.

29. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.** *Genes Dev* 2011, **25**(18):1915–27.

30. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC: **Rapid turnover of long noncoding RNAs and the evolution of gene expression.** *PLoS Genet* 2012, **8**(7):e1002841.

31. Gardner PP, Wilm A, Washietl S: **A benchmark of multiple sequence alignment programs upon structural RNAs**. *Nucleic Acids Res* 2005, **33**(8):2433–2439, [http://www.hubmed.org/display.cgi?uids=15860779].

32. 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.

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

34. 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.

35. Ulitsky I, Bartel DP: **lincRNAs: genomics, evolution, and mechanisms.** *Cell* 2013, **154**:26–46.

36. Duret L, Chureau C, Samain S, Weissenbach J, Avner P: **The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene.** *Science* 2006, **312**(5780):1653–5.

37. Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D, Matthews L, Rogers J, Pask AJ, Shaw G, VandeBerg JL, McCarrey JR, SAVOIR Consortium C, Renfree MB, Reik W, Dunham I: **Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians.** *Nat Genet* 2008, **40**(8):971–6.

38. Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnár Z, Ponting CP: **Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes.** *Genome Biol* 2010, **11**(7):R72.

39. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP: **Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.** *Cell* 2011, **147**(7):1537–50.

40. Diederichs S: **The four dimensions of noncoding RNA conservation**. *Trends in Genetics* 2014.

41. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N, Isaacs WB: **DD3: a new prostate-specific gene, highly overexpressed in prostate cancer.** *Cancer Res* 1999, **59**(23):5975–9.

42. Lin B, White JT, Ferguson C, Bumgarner R, Friedman C, Trask B, Ellis W, Lange P, Hood L, Nelson PS: **PART-1: a novel human prostate-specific, androgen-regulated gene that maps to chromosome 5q12.** *Cancer Res* 2000, **60**(4):858–63.

43. Ferreira LB, Palumbo A, de Mello KD, Sternberg C, Caetano MS, de Oliveira FL, Neves AF, Nasciutti LE, Goulart LR, Gimba ER: **PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling.** *BMC Cancer* 2012, **12**:507.

44. Yoshimura Y, Chang C, Okamoto T, Tamura T: **Immunolocalization of androgen receptor in the small, preovulatory, and postovulatory follicles of laying hens.** *Gen Comp Endocrinol* 1993, **91**:81–9.

45. Veney SL, Wade J: **Steroid receptors in the adult zebra finch syrinx: a sex difference in androgen receptor mRNA, minimal expression of estrogen receptor alpha and aromatase.** *Gen Comp Endocrinol* 2004, **136**(2):192–9.

46. Fuxjager MJ, Schultz JD, Barske J, Feng NY, Fusani L, Mirzatoni A, Day LB, Hau M, Schlinger BA: **Spinal motor and sensory neurons are androgen targets in an acrobatic bird.** *Endocrinology* 2012, **153**(8):3780–91.

47. Leska A, Kiezun J, Kaminska B, Dusza L: **Seasonal changes in the expression of the androgen receptor in the testes of the domestic goose (Anser anser f. domestica).** *Gen Comp Endocrinol* 2012, **179**:63–70.

48. Pascual-Anaya J, D'Aniello S, Kuratani S, Garcia-Fernàndez J: **Evolution of *Hox* gene clusters in deuterostomes**. *BMC Developmental Biology* 2013, **13**:26.

49. Yu H, Lindsay J, Feng ZP, Frankenberg S, Hu Y, Carone D, Shaw G, Pask AJ, O'Neill R, Papenfuss AT, Renfree MB: **Evolution of coding and non-coding genes in HOX clusters of a marsupial**. *BMC Genomics* 2012, **13**:251.

50. Cheetham SW, Gruhl F, Mattick JS, Dinger ME: **Long noncoding RNAs and the genetics of cancer.** *Br J Cancer* 2013, **108**(12):2419–25.

51. 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.

52. Lerner M, Harada M, Lovén J, Castro J, Davis Z, Oscier D, Henriksson M, Sangfelt O, Grandér 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.

53. 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.

54. 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.

55. Moynahan ME, Chiu JW, Koller BH, Jasin M: **Brca1 controls homology-directed DNA repair.** *Mol Cell* 1999, **4**(4):511–8.

56. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV: **Origin of avian genome size and structure in non-avian dinosaurs.** *Nature* 2007, **446**(7132):180–4.

57. Tanzer A, Stadler P: **Molecular evolution of a microRNA cluster.** *J Mol Biol.* 2004, **339**(2):327–35.

58. Hertel J, Bartschat S, Wintsche A, C O, The Students of the Bioinformatics Computer Lab 2011, Stadler PF: **Evolution of the let-7 microRNA Family**. *"RNA Biol"* 2012. in press.

59. Mosig A, Guofeng M, Stadler B, Stadler P: **Evolution of the vertebrate Y RNA cluster**. *Theory in Biosciences* 2007, **126**:9–14.

60. Stadler PF, Chen JJL, 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**:1975–1991.

61. Kolbe DL, Eddy SR: **Local RNA structure alignment with incomplete sequence.** *Bioinformatics* 2009, **25**(10):1236–43.

62. 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.

63. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain**. *Proceedings of the National Academy of Sciences* 2008, **105**(2):716–721.

64. Johnston RJ, Hobert O: **A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans**. *Nature* 2003, **426**(6968):845–849.

65. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL: **Targeted RNA sequencing reveals the deep complexity of the human transcriptome**. *Nature biotechnology* 2012, **30**:99–104.

66. Johnston RJ, Hobert O: **A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans.** *Nature* 2003, **426**(6968):845–9.

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

68. **The phylogenomics analysis of birds website**[http://phybirds.genomics.org.cn/index.jsp].

69. Bartschat S, Kehr S, Tafer H, Stadler PF, Hertel J: **snoStrip: A snoRNA annotation pipeline** 2014. [Preprint].

70. 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.

71. Schmitz J, Zemann A, Churakov G, Kuhl H, Grützner F, Reinhardt R, Brosius J: **Retroposed SNOfall–a mammalian-wide comparison of platypus snoRNAs.** *Genome Res* 2008, **18**(6):1005–10.

72. Shao P, Yang JH, Zhou H, Guan DG, Qu LH: **Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs.** *BMC Genomics* 2009, **10**:86.

73. **Non-coding RNA annotations of bird genomes** 2014, [https://github.com/ppgardne/bird-genomes].

74. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J: **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. *PLoS computational biology* 2009, **5**(9):e1000502.

Figures

Figure 1 - Heatmaps

Heatmaps showing the presence/absence and approximate genomic copy-number of "unusually, well conserved RNAs" (particularly the lncRNAs) on the left and families that have been identified as RNA losses, divergence or missing data. In several cases functionally related families have also been included, e.g. the RNA components of the major and minor spliceosomes: U1, U2, U4, U5 and U6; and U11, U12, U4atac, U5 and U6atac, respectively.

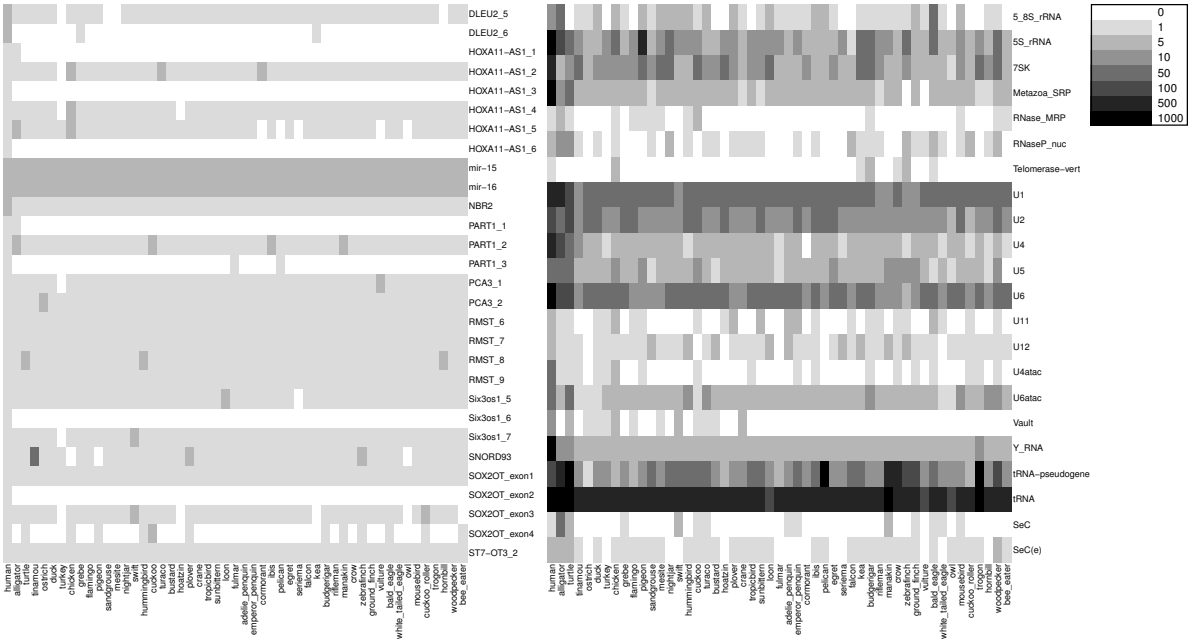

Figure 1:

ncRNA genes in human, chicken and all bird genomes				
Number in human	median(48 birds)	Number in chicken	Chicken ncRNAs confirmed with RNA-seq	RNA type
62	25.0	34	12 (35.3%)	Long non-coding RNA
356	499.5	427	280 (65.6%)	microRNA
281	120.0	106	90 (84.9%)	C/D box snoRNA
336	85.5	68	48 (70.6%)	H/ACA box snoRNA
34	13.0	12	12 (100.0%)	Small cajal body RNA
1754	48.5	71	32 (45.1%)	Major spliceosomal RNA
58	3.0	6	3 (50.0%)	Minor spliceosomal RNA
525	82.0	122	88 (72.1%)	Cis-regulatory element
316	6.5	9	3 (33.3%)	7SK RNA
1	0.0	2	0 (0.0%)	Telomerase RNA
9	0.0	2	1 (50.0%)	Vault RNA
892	3.0	3	2 (66.7%)	Y RNA
1084	173.5	300	278 (92.7%)	Transfer RNA
80	9.5	4	2 (50.0%)	Transfer RNA pseudogene
941	3.0	4	2 (50.0%)	SRP RNA
607	7.0	22	10 (45.5%)	Ribosomal RNA
4	1.0	2	2 (100.0%)	RNase P/MRP RNA
7340	1080.0	1194	865 (72.4%)	Total

Table 1:

Table 1 - A summary of ncRNA genes in human, chicken and all bird genomes

This table contains the total number of annotated ncRNAs from different RNA types in human, the median number for each of the 48 birds and chicken. The number of chicken ncRNA that show evidence for expression is also indicated (the percentage is given in parentheses). The threshold for determining expression was selected based upon a false positive rate of less than 10%.