Fly away Peter, come back Paul: the conservation and losses of avian non-coding RNAs

Paul P. Gardner*[1,2] , Mario Fasold[5] , 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

Email: Paul P. Gardner*- paul.gardner@canterbury.ac.nz; mario@bierdepot.bioinf.uni-leipzig.de; swb@ebi.ac.uk; Maria.Ninova@postgrad.manchester.ac.uk; Jana Hertel*- jana@bioinf.uni-leipzig.de; steffi@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. 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. 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 have discovered apparent "losses" in several RNA families, these illustrate the complexity of bird genomes. In particular, issues with assembling microchromosomes using next-generation sequencing methods. 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, 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].

needs work: New sequencing technologies have dramatically expanded the rate at which ncRNAs are discovered and their functions are determined []. However, in order to determine the full range of ncRNAs across multiple species we would 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. Consequently, for this work we have concentrated on bioinformatic approaches. Primarily, homology based methods, namely covariance models (CMs). These remain state of the art for ncRNA analyses [4–6] and have well established sensitivity and specificity rates [7]. 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 generated by the same evolutionary processes as the training sequences and to build alignments based upon sequence and structural information [4–6]. The tRNAscan-SE software package uses CMs to accurately predict transfer RNAs [8, 9]. The Rfam database contains thousands of curated alignments and consensus structures for diverse classes of ncRNAs [10–14]. Independent benchmarks of bioinformatic annotation tools have shown that the CM approaches dramatically out-perform alternative methods [7], although its sensitivity is limited for the most rapidly evolving families such as vault RNAs or telomerase RNA [15].

REDUCE? too much talking about what it can't to compared to what it can do... 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 [6, 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 possibly

structured elements) within lncRNAs, analogous to the "domains" housed within protein sequences [14]. 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]. The publication of 48 avian genomes, including the previously published chicken [22], zebra finch [23] and turkey [24] with the recently published 45 avian genomes [25–31], provides exciting opportunities to explore ncRNA conservation in unprecedented detail.

In the following we explore the conservation patterns of the major classes of avian ncRNA in further detail. The collection of ncRNA sequences is generally biased towards model organisms [2, 32]. **Using accurate homology search tools and evolutionary constraints, we produce an accurate set of genome annotations for 48 bird species for ncRNAs that are conserved across the avian species. This conservative set of annotations is expected to contain the core avian ncRNAs.** We focus our report on the unusual results within the avian lineages. These are either unexpectedly well-conserved RNAs or unexpectedly poorly-conserved RNAs. The former are RNAs we would not have expected to be conserved between the birds and the organisms these genes were initially identified in; Usually, this hypothesis is based upon the function of the RNA which is not conserved in avian species. The latter are apparent losses of RNA genes that were expected to be conserved; usually, this hypothesis is based upon the conservation of these RNAs in other vertebrate species.

Here, we consider three categories of apparent loss: First, genuine gene losses in the avian lineage. Second, "divergence" where RNA genes 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" genes 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 a few larger "macrochromosomes" and many smaller "microchromosomes" [33]. This 'so many, so small' pattern presents significant assembly challenges [34]. Indeed, of the 48 published avian genomes, 20 of which are high-coverage ($> 50X$), only two are chromosomally assembled (chicken and zebra finch; [25]).

Results

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 [16, 20, 21]. 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 mammalia. Notable examples include Xist [35] and H19 [36]. There is emerging evidence for the conservation of "mammalian" lncRNAs in other vertebrates [37, 38]), however, like most lncRNAs, the function of these lncRNAs remains largely unknown. Here, we show the conservation of several well-characterised lncRNAs of known function in humans.

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 [6]. Consequently, only short, well-conserved regions with evolutionarily conserved secondary structures are included in Rfam. By analogy to protein-domains, we refer to these as RNA-domains [14]. **SOME OF THESE MAY BE DNA ELEMENTS RATHER THAN LNCRNA ELEMENTS!!!!**

When analysing the RNA-domain annotations it is striking that 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).

The conservation of these "human" lncRNAs among birds suggests they may also be functional in birds but what these functions 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 [39–41]. Birds lack a prostate but both males and females express the androgen receptor (AR or NR3C4) in gonadal and non-gonadal tissue [42–45]. 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 remarkably well conserved. 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 [46], 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 [47]. 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.

Many of the lncRNAs have been associated with cancer, sparking a minor review industry [48, 49]. 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 [50].

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 [51, 52].

The NBR2 lncRNA and BRCA1 gene share a bidirectional promotor [53]. Both are expressed in a broad ranch of tissues. Extensive research on BRCA1 has shown that it is involved in DNA repair [54]. The function of NBR2 remains unknown, yet its conservation across the vertebrates certainly implies a function (See Figure).

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 which has 92 copies in the tinamou genome, whereas it only has 1-2 copies in all the other vertebrate genomes.

RNA losses, divergence or missing data?

Much of the number of apparent losses and reduction in genomic sequence has been extensively discussed elsewhere [55]. Unsurprisingly, this reduction is reflected in the copy-number of RNA 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 copies numbers of these drop to ~ 280 and ~ 100 copies in the avian genomes.

The absence of seven well-conserved ncRNA families from many or even most bird genomes is unlikely to represent true gene losses. This concerns in particular the telomerase RNA, the RNA components of RNase P and MRP, the minor spliceosomal snRNAs U4atac and U11, the selenocystein tRNA (tRNA-Sec) as well as the vault RNAs (See Figure).

Microchromosome background...

We propose three possible models of "loss" to explain the data: Firstly, these could be genuine cases of an ancestral gene-loss along the avian lineage. Secondly, this could be a case of "divergence" where a RNA gene has undergone significant sequence and structural alterations, so much so, that homology search tools no longer detect a relationship between vertebrate exemplars and the avian varieties. Thirdly, we consider the possibility that the available genome assemblies have independently failed to capture these genes.

The third model of loss is the most likely explanation. These families range in conservation level, from being ubiquitous to cellular-life (RNase P RNA and tRNA-Sec), to present in the majority of eukaryotes (RNase MRP, U4atac and U11) to present in most Bilateria (Vault). Therefore, the loss or even diversification of these families in the avian lineage is unlikely.

One defining feature of the avian genomes is the presence of "microchromosomes". As mentioned above, these are often difficult to assemble. The most complete available avian genomes are the chicken [22] and the zebrafinch [23]. The "lost ncRNAs" (loncRNAs) were found to more prevalent in the more complete genomes 6/7 in chicken and 3/7 in the zebrafinch. We examined the distribution of these loncRNAs and discovered that only the chicken RNase MRP gene is found on a macrochromosome. A Fisher's exact test showed that the loncRNAs are significantly enriched within microchromosomes ($P < 10^{16}$) in both the chicken and the zebrafinch genomes (combined $P < 2 \times 10^{31}$).

CUT?

In order to get an idea to what extent the absence of these RNAs from the `infernal-based` annotation is caused by sequence divergence beyond the thresholds of the Rfam CMs and/or missing or incomplete data, we complemented our analysis by dedicated searches for a few of these RNA groups.

The simplest case are the selenocystein tRNAs. Here, tRNAscan is tuned for specificity and thus misses several occurrances that are easily found by `blastn` with $E \leq 10^{-30}$. In some cases the sequences appear degraded at the ends, which may be explained e.g. by low sequence quality at the very ends of contigs or scaffolds. A `blastn` search also readily retrieves additional RNase P and RNAse MRP RNAs, capturing only the best conserved regions. In many cases these additional candidates are incomplete or contain undetermined sequence, explaining why they are missed by the CMs. Overall, we identify tRNA-Sec in

most and RNAse P and MRP RNAs in the majority of the genomes. An additional candidate could also be retrieved for telomerase RNA. Telomerase is well known to exhibit very poor sequence conservation and rapid variations in size that make it notoriously hard to identify by homology search [56]. The poor return thus does not come as a surprise. Since `blastn` searches remained unsuccessfull we constructed a sauropsid-specific CM for the vault RNA. In addition to the hits identified by the Rfam model we obtained three additional homologs. Vault RNAs, with a size of about 100 nt, exhibit conserved sequence patterns only at their ends, with essentially unconstrained sequence in the central part. Their identification is one of the well-know difficult problems for homology search [57].

Our ability to find additional homologs for several RNA families that fill gaps in the abundance matrices (Figure) 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 true losses.

Vertebrate Y RNAs typically form a cluster comprising four well-defined paralog groups Y1, Y3, Y4, and Y5. In line with [58] 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 bird, we find an the conserved Y4-Y3-Y1 cluster. Apparently, broken-up clusters are in most cases consistent with breaks 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.

Nearly all 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 and ambhibian genomes contain three loci of clustered microRNAs from the mir-17 and mir-92 families [59]. 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.

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 [60]). 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 for details.

Pseudogenes

Briefly mention the reduction in number of pseudogenes. Pick a few key human ones and compare with the birds.

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 [61]. We have identified previously unrecognised conservation of ncRNAs in avian genomes as well as some surprising "losses" of otherwise well conserved ncRNAs. We suspect many of these losses are due to a combination of limitations in the homology search tools that we use for annotation and the ability of ncRNAs to tolerate large amounts of sequence variation while remaining functional, rather than *bona fide* gene loss. In some cases these losses could be due to missing data from the genome assemblies, but this unlikely to be the case for multiple independant assemblies.

These results indicate we are still 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 are also found in bird genomes.

Methods

Bird genomes were searched using the cmsearch program from INFERNAL 1.1 and the covariance models from the Rfam database v11.0 [13, 14]. 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 hierarchial 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 [10–14]; these categories have been used to classify our results.

In order to obtain good annotations of tRNA genes we also 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 [8, 9]. 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 [13]. To further increase the specificity of our annotations we filtered out families that were indentified 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 additonal snoRNA homology search was performed with snoStrip [62]. As initial queries we used deutorostomian snoRNA families from human [63], platypus [64], and chicken [65].

Our diverse sets of genome annotations were combined. We collapsed the overlapping annotations into a single annotation. We also generated heatmaps for different groups of ncRNA genes (see Figure and Sup. Figs. 1-3). All the scripts and annotations presented here are available from Github [66].

Chicken snoRNA and miRNA annotations were validated using small RNA-seq data comprising 27 samples from 14 diferent tissues. Since not all RNAs are expressed at any given time we expect to verify only a fraction of annotated RNAs using the expression data. Nevertheless 242 of 691 (35%) miRNAs as well as 328 of 376 (87%) snoRNAs of our homology-based annotations were found to be supported by the RNA-seq data.

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) for useful discussions.

We thank Fiona McCarthy (Mississippi State University) and Carl Schmidt (University of Delaware) for providing the RNA-seq data.

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

We thank Matt Schwatz (Harvard) and Igor Ulitsky (Weizmann Institute of Science) for access to the RNA-seq data.

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. 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.
5. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**(11):2079–88.
6. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–7.
7. 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.
8. 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.
9. Chan PP, Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence.** *Nucleic Acids Res* 2009, **37**(Database issue):D93–7.
10. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**:439–41.
11. 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.
12. 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.
13. 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.
14. 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.
15. Menzel P, Gorodkin J, Stadler PF: **The Tedious Task of Finding Homologous Non-coding RNA Genes**. *RNA* 2009, **15**:2075–2082.
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. 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.

23. 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.

24. 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).

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

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

27. 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.

28. 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.

29. 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.

30. 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.

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

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

33. 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.

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

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

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

47. 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.

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

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

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

55. 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.

56. Xie M, Mosig A, Qi X, Li Y, Stadler PF, Chen JJL: **Size Variation and Structural Conservation of Vertebrate Telomerase RNA**. *J. Biol. Chem.* 2008, **283**:2049–2059.

57. 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.

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

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

60. 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.

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

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

63. 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.

64. 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.

65. 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.

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

Figures

Figure 1 - Heatmaps

Heatmaps showing the prescence/abscence and approximate genomic copy number of "unusually, well conserved RNAs" (particularly the lncRNAs) on the left and families that have been identified as surprising 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.

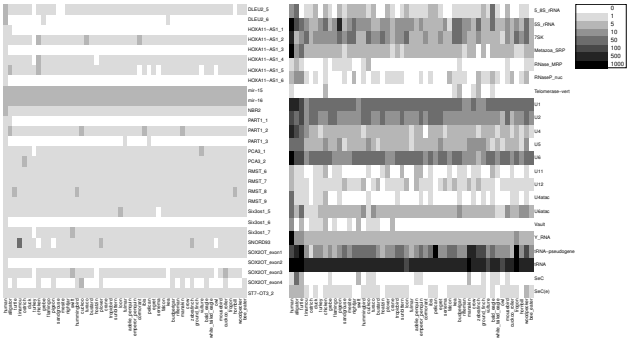

Figure 2 - Identifying missing RNAs

Additional homologs of some sparsely represented RNA families were discovered using dedicated search strategies combined with highly sensitive settings, synteny information, lineage-specific CMs and subsequent manual inspection.

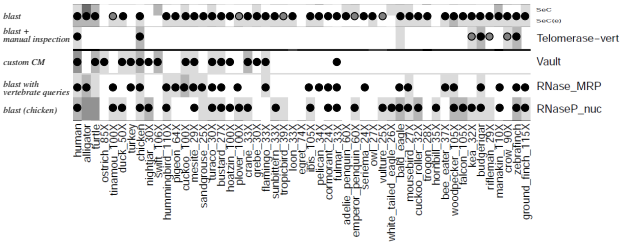

Table 1 - A summary of ncRNA genes in human, chicken and all bird genomes

This table contains the total number of annotated ncRNAs from different RNA types in human, chicken and the median number for each of the 48 birds. The number of chicken ncRNA that show evidence for expression is also indicated (the percentage is given in parentheses).

False positive rate less the 10%:

ncRNA genes in human, chicken and all bird genomes				
			Chicken ncRNAs confirmed with RNA-seq	
Number in human	median(48 birds)	Number in chicken	max(RNA_i)> 13.0	RNA type
62	25.0	34	12 (35.3%)	Long non-coding RNA
356	499.5	427	272 (63.7%)	microRNA
281	120.0	106	88 (83.0%)	C/D box snoRNA
336	85.5	68	46 (67.6%)	H/ACA box snoRNA
34	13.0	12	11 (91.7%)	Small cajal body RNA
1754	48.5	71	29 (40.8%)	Major spliceosomal RNA
58	3.0	6	2 (33.3%)	Minor spliceosomal RNA
525	82.0	122	79 (64.8%)	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	256 (85.3%)	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	817 (68.4%)	Total