New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes

Brian J. Parker, Ida Moltke, Adam Roth, Stefan Washietl, Jean Wen, Manolis Kellis, Ronald Breaker, and Jakob Skou Pedersen.

ABSTRACT

Regulatory RNA structures are often members of families with multiple paralogous instances spread across the genome. Family members share functional and structural properties, which allow them to be studied as a whole, facilitating both bioinformatic and experimental characterization. We have developed a comparative method, EvoFam, for genome-wide identification of families of regulatory RNA structures, based on primary sequence and secondary structure similarity.

We apply EvoFam to a 41-way genomic vertebrate alignment. Genome-wide we identify 220 human, high-confidence families outside protein-coding regions, comprising 725 individual structures. Among these are 47 known families, including both noncoding RNAs, e.g., miRNAs and the recently identified MALAT1 / MEN1 lincRNA family, and cis-regulatory structures, e.g., iron responsive elements and histone 3’UTR stem-loops. We also identify tens of new families supported by multiple lines of evidence. For some of these, detailed analysis has led to the specification of specific functional hypotheses. Examples include two auto-regulatory feedback mechanisms: one involving six long hairpins in the 3’UTR of MAT2A, a key metabolic gene that produces the primary human methyl donor S-adenosylmethionine; the other involving a tRNA-like structure in the intron of the tRNA maturation gene POP1. We experimentally validate the MAT2A structures and show that they are unlikely to act as riboswitches. Finally, we identify potential new regulatory networks, including large families of short hairpins enriched in immunity-related genes, e.g., TNF, FOS and CTLA4, which include known transcript destabilizing elements. Our findings underscore the diversity of post-transcriptional regulation and provide a resource for further characterization of new regulatory mechanisms, regulatory networks, and families of noncoding RNAs.

INTRODUCTION

The large and versatile role of RNA in protein-coding gene regulation is by now well supported. On the one hand, non-coding RNAs (ncRNAs) are known to regulate gene expression at virtually every possible stage, ranging from chromatin packaging to mRNA translation (Rinn et al. 2007; Zhao et al. 2008; Mercer et al. 2009). On the other hand, cis-regulatory elements within messenger RNAs (mRNAs) mediate post-transcriptional gene regulation to determine aspects of mRNA life-cycle such as stability, localization, and translational efficiency (Namy et al. 2004; Garneau et al. 2007).

Despite the progress in identifying novel ncRNAs, as well as protein-coding genes under post-transcriptional regulation, the functional characterization of specific instances and the elucidation of the involved regulatory mechanisms remain challenging and largely unsolved. A central problem is the identification of the specific functional regions responsible for the regulatory function within long
ncRNAs or mRNAs, and, in particular, common families of such regions. Their identification will allow the hypothesis-driven experiments necessary for functional characterization and mechanistic understanding, such as site-directed mutagenesis and screens for trans-acting protein factors. We developed a comparative method, EvoFam, to identify such families of human regulatory RNAs genome-wide by focusing on the large subset of these regions that function through well-defined RNA structures.

These regulatory RNA structures are often highly conserved in evolution because of their functional importance. They therefore evolve with a characteristic substitution pattern that often preserves base-pair interactions over primary sequence, resulting in compensatory double substitutions (e.g., AU ↔ GC) and compatible single substitutions (e.g., AU↔GU). EvoFold, and related programs (Rivas and Eddy 2001; Washietl et al. 2005), detects this signal by analyzing the substitution pattern along genomic alignments and uses it to identify these conserved RNA structures (Pedersen et al. 2006a).

Both ncRNAs and cis-regulatory elements are often members of families with multiple paralogous (evolutionarily related) instances spread across the genome. Due to their shared ancestry, members of a family normally share functional and structural properties. Identification of families among the individual structural RNA candidates identified in a genomic screen eases their study in several ways:
1) Finding that a predicted structural RNA is part of a larger family raises confidence in both the individual prediction and the family as a whole. For instance, high confidence members with significant substitution evidence will lend credence to the whole family and thereby to members with insignificant evidence. 2) Substitutions between family members may be observed for even slowly evolving structural RNAs due to the long divergence times, which can raise confidence in the predicted structure and its functional importance. 3) Functionally, families can be studied as a whole, which benefits both bioinformatic and experimental studies. Existing functional annotations and experimental results can be compared between family members.

We developed the comparative genomic pipeline, EvoFam, based on the evolutionary substitution signal, to perform the first general genome-wide screen for human families of regulatory RNAs. As part of the 2x Mammals Sequencing and Analysis Consortium (2X Mammals Sequencing and Analysis Consortium 2010), we employed the deep genomic vertebrate alignments generated, which include 29 mammals, 21 of which are sequenced at low coverage, and ten additional vertebrates (figure S1 & S2). This is an unprecedented resource for comparative analysis in general and structural RNA identification in particular. The extent of the full 41-way alignment allows us to withhold ten genomes for later validation purposes. This uncovers 220 new high-confidence families, many of which lend themselves to specific functional hypotheses, in some cases with apparent medical implications. Though this approach identifies families in both mRNAs and ncRNAs, our main focus is on the cis-regulatory elements, which cannot currently be discovered using high-throughput methods.

To our knowledge, only a small number of previous studies have identified families among genome-wide structure predictions sets. Pedersen et al. identify families among EvoFold predictions limited to primary sequence similarity (Pedersen et al. 2006a), whereas Will et al. identify families among RNAz predictions based on a combined sequence and structure alignment metric in Ciona intestinalis (Will et al. 2007). Other methods for identifying families of structures are based on motif-discovery of RNA
structures that are shared amongst a selected set of unaligned sequences, rather than utilizing genomic alignments. For example, CMFinder (Weinberg et al. 2007) has been applied across sets of orthologous prokaryotic genes (Yao et al. 2007; Tseng et al. 2009); Rabani et al. reports a search for common structured motifs in yeast, mouse and fly (Rabani et al. 2008); and Khaladkar et al. in human and mouse orthologues (Khaladkar et al. 2008). The motif-finding approaches are complementary to, and can be distinguished from, our approach by being based on a predefined set of input regions, which may be identified based on a functional hypothesis or experimental results.

The combination of de novo structure identification in deep genomic alignments with unbiased genome-wide all-against-all family detection, as described here, allows identification of completely novel and unanticipated families of structures from across the genome.

**RESULTS**

**Family identification**

Our structural RNA family identification pipeline, EvoFam, was used to screen the human genome for families of regulatory structures (outlined in figure 1; see Methods for details). The individual steps of this analysis are detailed below.

First, we generated a genome-wide prediction set of structural RNAs in human (figure 1, step 1). For this, we applied EvoFold (Pedersen et al. 2006b) to all conserved segments (spanning 5.6% of the genome) of a 31-way subset (28 placental mammals, opossum, chicken, and tetraodon; figure 1(b) & figure S1) of the vertebrate genome alignment made by the 2x Mammals Sequencing and Analysis Consortium (2X Mammals Sequencing and Analysis Consortium 2010). We excluded predictions in protein-coding regions from the downstream analysis, since their elevated false positive rate (Pedersen et al. 2006a) combined with extensive primary sequence homology in gene families led to higher rates of apparent false family predictions. The final prediction set contains 27,014 structural RNA predictions, after eliminating predictions of either poor quality or residing in repetitive regions (37,381 including protein-coding regions) (See Methods for details).

Next, we created a probabilistic model for each structural RNA prediction using profile stochastic context free grammars (pSCFGs), which capture the observed sequence variation at each position through the 31-way alignment as well as the predicted base pair interactions (figure 1, step 2a) (Eddy and Durbin 1994). The large evolutionary span of the alignments (7.1 expected substitutions per neutrally evolving position) allows the profiles to capture the general sequence and structure constraints acting on a given type of structural RNA, which are therefore sensitive in homology searches.

We then detected homology between the structural RNA predictions utilizing both sequence and structural similarities. As structural RNAs can vary dramatically in size and complexity, a key issue in such a genome-wide analysis is controlling for model-dependent false positive rates, to avoid false positive homology matches derived from low complexity structures (e.g. short hairpins) from generating spurious families. This was achieved by the use of a new similarity measure based on the
statistical significance of the similarity between sequences generated by the corresponding pSCFGs (see supplementary methods) (figure 1, step 3).

Based on this all-against-all similarity evaluation, we constructed a similarity graph and identified 1254 candidate families as densely inter-connected subgraphs (figure 1, step 4) (12.2% of the EvoFold predictions were included in families).

To evaluate the significance of the observed substitution evidence for a predicted structure, we developed a Monte Carlo-based statistical-significance measure (EvoP test), which evaluates how surprising the number of observed double substitutions is given the total number of observed substitutions, the phylogenetic tree relating the genomes, and the given secondary structure (see supplementary methods section). For each predicted structure, we calculated this p-value for the double substitution support in both the 31-way alignment (“dependent set”) and on branches leading to ten withheld vertebrates (“independent set”). Importantly, the substitution evidence in the withheld species is independent of the evidence used for the structure prediction.

Finally, we defined a set of high-confidence families based on their evolutionary, structural, and biological support (figure 1, step 5). The included families were either strongly supported by double substitution evidence in the 31-way alignment or the ten withheld species (p-value < 0.05; EvoP test combined by Fisher method); enriched for a specific genomic location (p-value < 0.01; chi-squared test); had significant gene ontology term enrichment (p-value < 1E-3; Fisher test); or long structures (> 11 base pairs in average). This resulted in 220 high-confidence families containing 725 individual structures.

The majority of families were small with two or three elements (90.5%; see figure 1-D). Compared to the EvoFold background, the families show strongest enrichment for UTR structures (UTR: 22.2% in families versus 13.6% among EvoFold background; intronic: 34.8% vs. 35.6%; and intergenic: 43.0% vs. 50.8%) (p-value=8E-11; chi-squared test) (figures S4 and S5). Most families detected consist purely of hairpins (97%), although some detected hairpins may represent components of larger RNA structures. See table 1 and figure 5 for families and structures described in the sequel. There are many families with long structures (50.5% > 11 base pairs on average), however, this is partly caused by their explicit inclusion through the selection criteria (cf. 28% > 11 base pairs when length is removed as a selection criteria). The larger full set of unfiltered candidate families is dominated by short structures (90.2% <= 11 base pairs on average), as is the background set of EvoFold predictions (77.9% <= 11 base pairs) (figure S4).

In addition to the set of families identified by the genome-wide analysis presented above (GW set), we also defined two other sets of families. One was based on an extended set of genome-wide structure predictions, which includes initially missed paralogs of the EvoFold predictions (GWP set). We identified EvoFold paralogs by searching all the conserved regions of the human genome with the profile models defined above (Figure 1, step 2b). Only significant hits (E-value<0.1) that showed strong double substitution evidence (p-value<0.05; EvoP test) in the alignment were included (see methods) (n=30,945). The GWP set is much larger than the GW set (949 versus 220 families), but has about the same average family size (2.7 versus 3.3 members; see figure 1-D). The other set was made
in the same way, but was based on only UTR EvoFold predictions and their UTR paralogs (UTRP set). This smaller focused set allows a more sensitive similarity threshold in the family classification without increasing the false positive rate and includes 103 families.

The following analysis is based on the GW set unless otherwise stated. We note that the majority of families discussed are found in all three sets. The full set of results, including raw data files, annotated families, and links to the UCSC Genome Browser is available at: http://athena.binf.ku.dk/~jsp/mammals.

**Recovery of known families**

The EvoFam pipeline correctly identifies many of the known structural cis-regulatory and ncRNA families. Only few families of human cis-regulatory structures are described in the literature (we found less than ten) (Gardner et al. 2009; Jacobs et al. 2009). Among these, we recover the families of: (a) histone 3' UTR stem-loops (49 of 67 known); (b) hairpins regulating translation in collagen genes (3 of 3) (Stefanovic and Brenner 2003); (c) and iron responsive elements in the 3'UTR of the transferrin (TFRC) gene (5 of 5). Some cis-regulatory families are missed since EvoFold only predicts parts of the individual member structures, likely due to structural evolution. This is for instance the case for the family of selenocysteine insertion sequences (SECIS).

Amongst known ncRNAs, miRNAs are recovered with good sensitivity (139 of 441 known, conserved, miRNA genes, in 42 families) whereas only few snoRNAs (2 in 1 family) and tRNAs (2 in 1 family) are recovered. This is due to the low numbers of snoRNAs (n=40) and tRNAs (n=13) in the initial EvoFold prediction set. snoRNAs are likely missed due to alignment problems with pseudo-genes and most tRNAs are annotated as repeats and masked out (Pedersen et al. 2006b). The specificity of the approach is indicated by the high fraction of known structural RNAs (88%) in families with any known members.

The paralog search used for the GWP set picks up many of the initially missed family members (Figure 1, step 2b) and increases the overall number of known structural RNAs relative to the GW set (226 versus 191). The power of EvoFam’s comparative approach is also exemplified by the identification of a two-member family of clover-leaf-shaped structures in two long intergenic ncRNAs (MALAT1 and MEN1; GWP266). In both cases, two EvoFold hairpin predictions, well supported by substitution evidence, are found upstream of the clover-leaf-shaped structures (figure S13). These structures and their role in 3'-end processing were recently discovered in three independent dedicated studies, two of which were published during paper preparation (Wilusz et al. 2008; Sunwoo et al. 2009; Wilusz and Spector 2010). Had these not been discovered already, all six structures and their family relationship would have been a clear-cut result of our genome-wide de-novo family identification.

**Specificity of family predictions**

The specificity of the family predictions cannot be measured directly, since we do not know a priori how many of the novel predictions are true. Instead we estimated the family false discovery rate (FDR) based on randomly shuffled data sets that preserved the original observed distribution of structure similarities (see Methods). For computational reasons, this analysis was run on the UTRP set
only, and was applied to larger families only (size > 3) (as size-two families are invariant under the shuffling approach used - see Methods). The FDR was estimated at 0.8% for such families.

Enrichment of high confidence structures in families

Comparing the median EvoFold log odds scores of family members versus known functional RNAs and the background EvoFold predictions shows that the families have similar enrichment to known functional RNAs (median log odds: EvoFold background 11; known functional RNA 22; GW set 20).

To further substantiate these results by an alternative computational method, we applied RNAz (Gruber et al. 2010) to the EvoFam families and background set of EvoFold predictions. In contrast to EvoFold that is based on a purely probabilistic model, RNAz uses a thermodynamic model augmented by an ad hoc covariance score to predict functional RNA structures. RNAz classifies 9.7% of the initial EvoFold prediction set as functional RNAs. Although, consistent with previous studies (Washietl et al. 2007), the overlap is relatively low albeit highly significant (~12 fold enrichment over random, Supplementary Fig. S6), the fraction of RNAz predictions drastically increases in the clustered families. In the full, unfiltered GW set 22.9% of structures are predicted as functional RNAs by RNAz (~23 fold enrichment over random); for the high-confidence GW set we even observe 40.2% (50.2 fold enrichment). The latter is close to the fraction of positive RNAz predictions in a set of 356 known structural RNAs (50.2%). These results show that the EvoFam approach effectively enriches for high confidence RNA structures. 30-40% of positive RNAz predictions and a sensitivity of about 50% suggest that about 60%-80% structures in the predicted families correspond to true functional structures (40-60% if miRNAs and other known structures are excluded). A premise for this comparison is that EvoFam and RNAz detects similar types of true structures. This may not be fulfilled, as EvoFam detects many families of short hairpins, which will be difficult to detect by the windowing approach used by RNAz. This suggests that the above-calculated fractions of true predictions may be conservative. Furthermore, the individual structures within families are predominately predicted on the transcribed strand when overlapping known genes (61% on transcribed strand, p-value:5.3e-06; binomial test) which suggests that the majority of these are indeed cis-regulatory.

To estimate the contribution of structure versus sequence in the detection of the novel families, we reran the UTRP analysis with profile models stripped of structural information. This purely sequence-based comparison resulted in only 42 final families compared with 103 in the full sequence and structure-based comparison, and did not detect several of the novel 3'UTR-based families described below, including the MAT2A family (family identifier UTRP1) and the immunity related families (UTRP36; UTRP38; UTRP40). Overall, the sequence-only analysis identified only 37% of the structures in the UTRP set, demonstrating that shared structure is an important aspect of many of these families.

New cis-regulatory structure families

Post-transcriptional regulation of protein-coding genes is mediated by cis-regulatory elements, often structured, which bind proteins or other trans-acting factors. This wide-spread regulation takes place at various stages of the mRNA life cycle and can be categorized into: a) mRNA stability; b) pre-mRNA processing; c) nuclear export and subcellular localization; and d) translational regulation. The structured cis-regulatory elements are most commonly located in the 3' UTR, but can be located
across the mRNA (reviewed in (Moore 2005; Garneau et al. 2007). Below we describe several novel families of candidate cis-regulatory structures identified by EvoFam that we hypothesize belong to these various forms of post-transcriptional regulation, apart from (c).

Families hypothesized to be involved in mRNA stability
The dynamic balance of the rate of transcription and the rate of degradation and sequestration determines the final level of mRNA abundance. We hypothesize that several of the predicted families are involved in regulating mRNA transcript stability.

Family of hairpins in 3’UTR of MAT2A
The highest-ranked UTRP family by independent double-substitution evidence consists of a cluster of three long (12-18 bp) hairpins in the 3’UTR of the key metabolic gene methionine adenosyltransferase II, alpha (MAT2A). A more sensitive directed homology search revealed six matching 3’ UTR hairpins in total (E-value cutoff < 1.0) (figure 2a). These hairpins are characterized by a loop motif (figure 2b) with strong evolutionary conservation, indicative of a critical biological role. The hairpins initially identified by EvoFold can be extended, as seen by folding only the human sequence using an energy minimization method (Hofacker et al. 1994) (figure 2c). These extended parts of the predictions are also supported by substitution evidence, although less so than the core part (figures S7-S12). Most MAT2A mRNAs and ESTs in GenBank show short 3’UTRs, which only include hairpin A. Interestingly, a few alternatively spliced ESTs have 3’UTRs that instead include C-F, with the 3’ splice site located six base pairs before hairpin C (figure 2a), potentially modulating the regulatory effect of the hairpins.

The MAT2A gene product, methionine adenosyltransferase, is responsible for the synthesis of S-adenosylmethionine (SAM aka AdoMet), which is the primary methyl donor in human cells, involved in a broad range of processes including polyamine biosynthesis and gene regulation through DNA methylation. SAM is found in all five kingdoms of life. In humans, the MAT2A mRNA half-life varies from 100 min to more than three hours depending on SAM availability (Martinez-Chantar et al. 2003), and we hypothesize that the 3’ UTR hairpins are involved in this regulation.

Riboswitches, which occur typically in UTRs of mRNAs, are conserved RNA structures that interact directly with metabolites to control gene expression (Roth and Breaker 2009). Nearly 20 different metabolites are targeted by unique riboswitch classes, including four distinct classes that selectively recognize SAM (Wang and Breaker 2008; Weinberg et al. 2010). Generally, riboswitches are widely distributed in eubacteria, but among eukaryotes just one such class of RNA has been identified, occurring only in plants, fungi, and algae (Bocobza and Aharoni 2008).

To test the possibility that the conserved hairpins in the 3’ UTRs of vertebrate MAT2A homologs might mediate gene regulation through direct interactions with SAM or related metabolites, and to validate the predicted RNA structures, we performed binding assays using four RNA constructs corresponding to different regions of the human MAT2A 3’ UTR (Supplementary Figure S16). These RNAs, which contained from one to three conserved hairpin loops, were 32P 5’-end-labeled and subjected to in-line probing, a technique that can reveal ligand-induced structural changes (Regulski and Breaker 2008). Overall, each of the constructs appears to be largely unstructured, as evidenced by the relatively high rates of strand cleavage over the entire lengths of the RNAs (Figure 2d and Supplementary figures...
S17-S20). Among the four constructs examined, no significant differences were observed in the patterns of spontaneous cleavage products resulting from separate incubations with SAM, S-adenosylhomocysteine, and L-methionine, indicating that no large structural changes are induced by these compounds (Figure 2d and figures S17-S20). Although not all metabolite-binding RNAs experience major structural changes upon docking of the cognate ligand (Hampel and Tinsley 2006; Klein and Ferre-D’Amare 2006; Cochrane et al. 2007; Montange and Batey 2008), the elevated cleavage levels corresponding to the highly conserved sequence within the predicted loop (Figure 2d and figures S17-S20) suggest a significant degree of flexibility in this region, and are not consistent with a highly preorganized binding site. The in-line probing data therefore suggest that there are no direct interactions between any of the test compounds and the conserved RNA hairpins, at least as they occur in the context of these segments of the 3’ UTR.

Nonetheless, the secondary structures predicted for individual hairpin elements are strongly supported by the results of the in-line probing analyses. In local zones corresponding to the conserved hairpins, the sequences that are predicted to be base-paired experience reduced cleavage levels (Figure 2d and figures S17-S20), which is consistent with these regions forming double-stranded RNA structure. Interestingly, there are isolated sites within predicted stems that experience elevated rates of strand scission (Figure 2d and figures S17-S20). These sites often correspond to nucleotides predicted to reside in bulges or internal loops, which are generally expected to be more susceptible to spontaneous cleavage. Taken together with the apparent high degree of flexibility in the putative terminal loop sequences, these observations lend strong experimental support to the secondary structure model proposed for this RNA motif.

An alternative hypothesis is that the hairpins may bind a protein-complex involving SAM, which determines transcript stability. An analogous system is known from the transferrin receptor gene TFRC, which harbors a cluster of five IREs (hairpins) in the 3’UTR. In this case, the transcript undergoes endonucleolytic cleavage, mediated by IRE binding proteins (IRP1 and 2), when environmental iron levels drop (Erlitzki et al. 2002). Interestingly, the IRE structures contain a conserved loop motif, 5’-CAGWGH-3’ which is critical for IRP recognition (Chen et al. 2006; Serganov and Patel 2008) and matches part of the MAT2A motif. Another possible mechanism is a protein-binding RNA switch mechanism (Ray et al. 2009). These hypotheses are currently being investigated experimentally.

**Large families extend known post-transcriptional regulation in the immune system**

Three structurally similar families in the UTRP and GW sets (UTRP38/GW218, UTRP36/GW219 and UTRP40) show statistically significant enrichment for macrophage-related immunity genes (25%, 25%, 31% of members; p-values: 0.039, 0.039, 0.0014 respectively, Fisher test) and immunity related GO terms (e.g. leukocyte migration; p-value: 4.5e-5 for GW218) (see methods). All three families consist of short hairpins (6-7 bp) found in the 3’UTR of many key inflammatory and immunity genes including TNF, CSF3, FOS, and CTLA4. The three families are very similar, having a 3-nt loop and an AU-rich stem with the upstream strand being A-rich and the downstream strand being U-rich (figure 3a).

The families have multiple lines of supporting evidence: in addition to the enrichment evidence for immunity and inflammation, they show strong 3’UTR enrichment compared to other genomic regions in the genome-wide set (p-values: 0.012 (GW218); 5e-04 (GW219); chi-squared test), consistent with cis-regulatory structures. The individual members are highly conserved at the primary sequence level.
and therefore few show compensatory substitutions, an exception is the hairpin in \textit{TNF}, which shows a compensatory substitution in opossum, which notably was not used for structure inference. On the other hand, 40% of them are supported by compatible single substitutions (e.g., GU\rightarrowAU). In the aligned human sequences of the family members, the stems show strong sequence conservation, and weaker conservation of the loop (figure 3a). See supplementary table S2 for full gene list of these families.

Several examples of short hairpins involved in mRNA stability control are known for inflammatory and early response genes (Stoecklin and Anderson 2006; Anderson 2008). These elements are distinct from, though often adjacent to, the well-characterized sequence-based AU-rich elements (ARE). Two members from the above-defined families correspond to such known elements (figure 3b):

1) Tumor necrosis factor \textalpha \,(\textit{TNF}), which produces a key cytokine mediating the inflammatory response, contains a 15 long sequence element that has been found to be a degradation point (Stoecklin et al. 2003), termed a constitutive decay element (CDE). A member of the UTRP38 / GW218 family precisely matches the CDE, which has been suggested to form a hairpin previously (Chen et al. 2006).

2) Granulocyte-colony stimulating factor (CSF3), which produces a pro-inflammatory cytokine, contains a hairpin in the 3’UTR, termed a stem-loop destabilizing element (SLDE) (Putland et al. 2002). It has been found to enhance mRNA decay independently of the nearby ARE (Brown et al. 1996). A member of UTRP40 corresponds exactly to the SLDE.

In addition, the proto-oncogene \textit{FOS}, which is a transcription factor that regulates cell proliferation and differentiation and is induced during activation of T lymphocytes, contains a 54-nt AU-rich region that has been found to enhance the degradation response of downstream AREs (Xu et al. 1997); a member of the UTRP38 / GW218 family overlaps the 3’ end of this region.

These characterized and functional members of the families strongly suggest that the other members of these families also have regulatory roles. One example supporting this is the hairpin in the 3’ UTR of \textit{CTLA4}, which is a key receptor expressed by T lymphocytes that suppresses the adaptive immune system and is known to be post-transcriptionally regulated by currently uncharacterized elements in the 3’UTR (Malquori et al. 2008).

Overall, these results suggest that the identified families extend and generalize previously identified post-transcriptional regulatory mechanisms mediated by short hairpins in immune-related genes, possibly in combination with ARE-mediated decay.

\textbf{Families involved in pre-mRNA processing}

\textit{Families of A-to-I RNA editing hairpins}

RNA editing is a post-transcriptional, pre-mRNA modification of bases, which may alter the encoded amino acids or the function of regulatory signals. In mammals the most common form of RNA editing is adenosine-to-inosine (A-to-I), catalyzed by ADARs (adenosine deaminases acting on RNAs), which normally target stems of long hairpins. We recover a family of long hairpins partly overlapping the
coding regions of paralogous glutamate receptors, which, when extended, overlap well-studied amino acid altering editing sites (GW138 containing GRIA1 & GRIA4) (Lomeli et al. 1994).

All known, functional editing sites are exonic, with the exception of a site in an intron of ADAR2, which creates a 5’ splice site and thereby regulates splicing (Rueter et al. 1999). Intronic sites have been challenging to discover, since editing sites are found by observing discrepancies between the genomic sequence and sequenced, mature transcripts (Li et al. 2009).

We identify several families of long hairpins in calcium channel genes and glutamate receptors (GW129, GW177). Since most known editing sites are found in ion channel genes and neurotransmitters (Jepson and Reenan 2008) and since one of these families is in glutamate receptor genes, already known to harbor several exonic sites, we speculate that these hairpins likely contain additional intronic editing sites. Finally we also identify a family in the 3’UTRs of three sodium channels genes, which is well supported by substitution evidence (p-value: 3e-4; EvoP dependent set) and may be an example of novel editing sites (UTRP19) (see Supp. Table S1).

**Auto-regulation of tRNA biogenesis gene**

A family of three cloverleaf structures (GW168) consists of two intergenic glycine tRNAs, and a tRNA-like structure in the intron of processing of precursor 1 (POP1) (see figure 4). The POP1 structure is well supported by double substitutions (p-value 1.0E-3; EvoP dependent set) and shows conserved synteny with the exons of POP1, thus ruling out that it is a tRNA pseudogene. Although the POP1 structure shows characteristics of tRNAs, including loop motifs shared with the tRNA members of the family, it shows a shorter D arm and an anticodon loop one base longer than the canonical tRNA structure. Furthermore, tRNAscan-SE detects it, but with a low score (Lowe and Eddy 1997). Interestingly, POP1 encodes a protein subunit of ribonuclease P (RNase P), the riboprotein complex that matures tRNA molecules by cleaving the 5’-ends of precursor tRNAs. Based on this, we hypothesize that this tRNA-like structure is involved in auto-regulation of the POP1 transcripts, with the RNaseP complex binding and potentially cleaving the transcript.

**Families hypothesized to be involved in translational regulation**

In some cases cis-regulatory structures regulate the translational process directly, such as in alternative translational initiation sites, regulated frameshifts, selenocysteine insertions, etc. (Namy et al. 2004). In addition to a known SECIS element (UTRP3), we recover a known family of hairpins overlapping the start codon of COL1A1, COL1A2 and COL3A1 (GW36) (Stefanovic and Brenner 2003). This family is expanded by a previously undescribed member overlapping the start codon of COL5A2 in the UTRP set (UTRP5).

**UTR families lacking specific functional hypotheses**

Many other UTR families have strong evidence of functionality, for which no definite hypothesis has been formulated. Among these are a family enriched for genes in the ubiquitin pathway (such as, BAP1, CYLD, UBE2W, and MID1); a family of 9-10 bp long hairpins in the 3’UTR of the lymphoid development genes BCL11A and B; and a family of short (6-8 bp) hairpins in the ion-channel-related genes CACNB4, KCNMA1, and ANK3 (see Table S4 and supplementary results for details).
miRNAs, lincRNAs, and other ncRNAs

Included amongst the miRNA families are examples of putative new miRNAs or other functional long hairpins (see Table S3). For example, an intron of CLCN5 harbors a family (GW159) of two long hairpins, of which one is a known miRNA (MIR362) and the other is an apparent novel miRNA, which shows evidence of miR and miR* expression in RNA-seq for multiple tissues (Figure S15) (an entry for this element has since appeared in miRBase release 15 while this paper was in preparation).

Two families contain extremely long intergenic hairpins (>30 base pairs, see figure 5) (GW45 and GW103), which are both highly conserved, but also supported by double substitution evidence (pvalue: 0.72 and 8e-3; EvoP dependent set, respectively) and compatible single substitutions (n=26 and 27 respectively). Given their length and a lack of RNA-seq evidence, these hairpins may not be miRNA precursors, but instead be a novel, uncharacterized type of ncRNA.

The GWP set includes 119 families with 158 members (28%) that overlap lincRNAs (Guttman et al. 2009; Khalil et al. 2009). This includes some known families, such as the previously mentioned MALAT1 / MEN1 family as well as 52 known miRNAs and snoRNAs. Interestingly, a novel and well-supported structure in XIST / TSIX forms a family with an intronic structure in male germ cell-associated kinase (MAK) (S14 and supp. results).

Apart from these cases, the mean length of lincRNAs that are members of families is 13.4 bp, thus showing that there are likely not many more families of local, complex structures in lincRNAs. Some of these families (64%) also contain members from UTRs and introns of protein-coding genes, which may be shared cis-regulatory structures defining regulatory networks between lincRNAs and protein-coding genes. Based on this, we analyzed the GO enrichment for the protein-coding genes containing such shared members in the full GWP set (the GW subset was too small for this type of analysis) and found them to be enriched for regulation of T-cell differentiation (P = 4.5e-04), histone methyltransferase activity (3.1e-03) and various terms related to cellular adhesion (see Table S5). This coincides with the results of a completely unrelated expression analysis, which found that lincRNAs are involved in immunity/inflammation as well as chromatin modification (Guttman et al. 2009; Khalil et al. 2009).

DISCUSSION

In this study, we developed and used a comparative approach (EvoFam) to identify families of regulatory structural RNAs in the human genome based on deep vertebrate genomic alignments. We found that this approach could successfully identify a wide range of known families de novo, including both cis-regulatory families, e.g., the iron responsive elements in TFRC, and ncRNAs, e.g., the MALAT1 / MEN1 lincRNA family. Furthermore, novel members were added to known families in some cases, e.g., the collagen 5’UTR hairpins and the CLCN5 miRNA. We also found strong evidence for a large number of completely novel families. Among our 220 high-confidence families, we found 174 novel families containing an estimated 40-60% lower bound of true, functional regulatory RNA structures. A detailed analysis of these revealed many strongly supported novel families, which in several cases
allowed the formulation of specific functional hypotheses. Several of these are currently being evaluated experimentally.

A strength of comparative RNA structure identification is that it relies on few assumptions. It therefore has the potential to identify structures of any function, shape, size, and genomic location. Although this allows novel types of structural RNAs to be identified, it also provides less confidence in individual predictions compared to dedicated searches tailored for specific types of structures. Even with the deep alignments used here, true structures will often only accumulate few supporting substitutions through evolution. Consequently, most predictions are initially not of high confidence. This study showed that the EvoFam family classification approach is an efficient means for defining a high confidence set of predictions, as demonstrated by 1) the enrichment for known structures, 2) the high agreement with RNAz predictions, 3) the support for specific novel families revealed by detailed analysis, 4) and in the case of the MAT2A family, experimental support for the predicted structures.

A key aspect of the EvoFam pipeline that enables such a genome-wide analysis for structures of varying size and complexity, from short hairpins to large clover-leaf-shaped structures, is stringent control of structure-dependent false positive (type I) error rates. Without such stringent type I error control across structure comparisons, pilot studies showed that the false positive homology matches from low complexity structures would dominate the results. We controlled for this by using a new similarity measure between probabilistic models which corrects for expected false positive rates of the compared structures. This, combined with a graph-theoretic family definition (see Methods) that is robust to noise, enables a genome-wide analysis with low over-all error rates.

The availability of deep (41-way) vertebrate alignments (2X Mammals Sequencing and Analysis Consortium 2010) allowed both discovery and validation on the same data set. This was done by withholding ten species from the initial structure discovery step and using these to later validate predictions based on their independent substitution evidence. For this, we developed a general-purpose significance test (EvoP) for evaluating the significance of double-substitutions supporting specific structure predictions. This approach was also successfully used to evaluate the substitution support for paralog hits, where only the human sequence was initially searched. The power of this validation strategy will increase as more species are sequenced.

The ultimate goal when new genomic functional elements are discovered is to connect them with known biology and to characterize them functionally and mechanistically. Family identification greatly facilitates this process. It not only allows the members of a family to be studied as a whole, it can also allow the specification of functional or mechanistic experimentally verifiable hypotheses. We found this to be most pronounced for the cis-regulatory structures, where existing knowledge on the common function or regulation of the protein-coding genes harboring the family members often allowed specific hypotheses. In addition, members shared between protein-coding genes also suggested co-regulation and the presence of regulatory networks. Finally, the inclusion of a family member with known function shed light on the common function of an entire family in several cases.

Overall, we found strong evidence for many novel families of cis-regulatory structures. UTRs were especially enriched for structure families, with a 4.1x enrichment compared to the set of input
conserved elements. In-depth analysis revealed several families strongly supported by multiple lines of evidence, for which we presented specific functional hypothesis based on the principles given above. These appear to provide novel examples of post-transcriptional regulation at various stages of the mRNA life-cycle. One example is the families of short hairpins in immunity related genes, which may present a network of destabilizing elements. Combined, the discovery of these families largely expands the set of known human UTR structure families and is an added indication of the complexity and abundance of post-transcriptional regulation.

We hypothesized that at least two of our families are involved in auto-regulation of protein-coding genes, where cis-regulatory structures regulate transcript stability and ultimately protein abundance: (1) For MAT2A, we hypothesized that the 3′UTR hairpins mediate transcript stability in response to metabolite concentration, either directly as ribo-switches or via protein factors. In-line probing assays supported the predicted structures, but did not support the function as ribo-switches. We therefore favor the involvement of protein-factors in this regulation. (2) For POP1, we hypothesized that the primary, un-spliced transcript is bound by RNaseP and perhaps cleaved. Similar examples of auto-regulation have been reported in the literature for ADAR2, which edits its own transcript and creates an alternatively spliced isoform (Rueter et al. 1999), and for DGCR8, which together with Drosha binds and cleaves a miRNA-precursor-like hairpin in its 5′UTR (Han et al. 2009). In such cases, auto-regulation provides a simple and direct regulatory mechanism. We therefore propose that auto-regulation has evolved relatively often and that it is more common than currently realized.

Some types of cis-regulatory structures may be shared between mRNAs and lincRNAs. This notion is supported by our functional analysis of mixed families containing members in both mRNAs and lincRNAs, which found an enrichment among the protein-coding genes for broadly the same functions previously reported for lincRNAs using different types of analyses (Guttman et al. 2009; Khalil et al. 2009). Such shared post-transcriptional regulation could be expected, given that lincRNAs appear to be processed similarly to mRNAs (Guttman et al. 2009).

Most of the identified families consist of short hairpins. Given that hairpins are common among known regulatory RNA structures (Svoboda and Di Cara 2006), this may represent the true distribution. However, the distribution is likely affected by the identification approach used here. For instance, the input set of structure predictions is likely biased toward hairpins: EvoFold most efficiently predicts consensus structures found in all sequences of the alignment, it may sometimes only detect core hairpins of more complex structures. Similarly, local alignment or sequencing errors will likely break up more complex structures into hairpins.

The EvoFam approach to RNA structure family identification complements the previously published comparative searches for primary sequence families (Xie et al. 2005). It benefits from the added functional information in RNA structures and can thereby identify structured instances with higher confidence than purely sequence-based approaches. Knowledge of RNA structure also allows more specific functional or mechanistic hypotheses than would otherwise be possible.

We here presented the first genome-wide comparative screen for human families of regulatory RNA structures using both sequence and structure homology, which revealed 174 new high confidence
families supported by multiple lines of evidence. We expect that this resource and the accompanying functional hypotheses will facilitate experimental characterization of new post-transcriptional regulatory mechanisms, regulatory networks, and families of ncRNAs. The planned sequencing of 10,000 vertebrate genomes will soon provide a rich data set for performing this type of screen with even higher accuracy (Genome 10K Community of Scientists 2009).

**METHODS**

**Genomic alignments**

A 41 species subset of the human-referenced (hg18) genome-wide 44-way vertebrate multiz alignment from the UCSC Genome Browser was used. For the structure prediction and profile-model training we used a 31-way subset, consisting of 29 mammalian species, mostly sequenced by the 2x Mammals Sequencing and Analysis Consortium, along with two out-group vertebrate species (chicken and tetraodorn). An additional 10 species were used as an independent test set (see figures S1 and S2 for full species list).

**Methods**

**EvoFam pipeline (see Figure 1):**

1) **EvoFold phylo-SCFG screen:** A genome-wide input set of structural RNA predictions were made by screening both strands of the conserved segments of the 31-way genomic alignment using EvoFold (v.2.0) (Pedersen et al. 2006b). Low-confidence predictions that are short (< 6 base-pairs); harbor excessive amount of bulges; based on shallow or low quality alignments; or overlap repeats or pseudogenes were eliminated from the prediction set. Finally, overlap between predictions were resolved according to EvoFold score. See supplementary methods for details on the screen.

We used the UCSC Genes set (as of May 25, 2009) to define genomic regions. Each prediction was assigned to the genomic region it had the largest overlap with. Protein-coding regions were excluded from the study to focus it on non-coding regions; also protein-coding regions show a higher than average false positive rate due to the many large families of protein-coding genes, and due to the assumptions of the double-substitution p-value (EvoP) measure not being fulfilled due to the unequal substitution rate at different codon positions.

2a) **Profile SCFG generation:** As structure is a key feature of the family members, we utilized both sequence and structure information in detecting the regulatory RNA families. For each EvoFold prediction, we fitted a profile stochastic context free grammar (pSCFG) model using the Infernal RNA tools v 1.0 (Nawrocki et al. 2009). pSCFGs describe individual unpaired and paired positions by single states (see figure 1(c)). Default sequence and entropy-weighting options for priors were enabled: this provides the models with a prior preference for canonical Watson-Crick pairings as well as including GU-wobble pairs, as updated by the actual training data across the 31 species.

2b) **Paralog search:** Paralogous matches to the EvoFold predictions were detected by searching the conserved regions of the human genome with the corresponding pSCFG (using cmsearch with global
search option). For the UTRP set, only the UTR regions were searched. The paralogous hits were filtered by requiring E-value < 0.1 (relative to a 1 Megabase database) and good double substitution evidence (p-value < 0.05; EvoP test applied to all species excluding human (< 0.2 for the smaller UTRP set)). Repeat regions and known pseudogene matches were removed (as above). Overlapping paralog hits were resolved to the hit with the lowest E-value. This set of putative paralogs was then optionally combined with the original EvoFold set and analyzed by the subsequent family identification stages.

3) **Inter-pSCFG similarity estimation and type I error control:** A similarity graph between structural predictions was defined based on an all-against-all similarity estimation of the pSCFG models. The large all-against-all comparison introduces a multiple testing issue and, as structures vary substantially in their length and complexity, stringent control of false positive (type I) error rate is essential. The false positive rate was controlled, unbiased by model size and complexity, by basing the similarity measure on an estimate of the statistical significance of the similarity between sequences generated by pairs of pSCFG models. We define the dissimilarity $D$ between profile models $M_1$ and $M_2$ as $D(M_1, M_2) = \max(\tilde{D}(M_1, M_2), \tilde{D}(M_2, M_1))$ where the (asymmetric) divergence $\tilde{D}$ between $M_1$ and $M_2$ is $\tilde{D}(M_1 \parallel M_2) = E(S(seq_{human}^1, M_2)) - E(S(seq_{human}^1, M_1))$, and where $S(seq_{human}^1, M_n)$ is the score of the alignment of the human sequence used to train model $m$ against model $n$, and $E$ is the E-value of the score computed relative to a constant 1 Mbase database. See the supplementary methods for the derivation of this measure.

4) **Graph-based density clustering:** A similarity graph $G(V, E)$ was defined with vertex set $V$ corresponding to pSCFG models of RNA structures and with edges connecting pairs of models with a dissimilarity $D(M_1, M_2)$ below a threshold $T$. The threshold $T$ was specified to vary the sensitivity/specificity tradeoff for inclusion of edges (set to 0.25, 0.25, 1.0 for GW, GWP and UTRP sets respectively). Families were defined as highly connected subgraphs $S \subset G$, where a highly connected subgraph (HCS) is defined as a subgraph of $n$ vertices with edge connectivity $k(S) > n/2$. Edge connectivity $k(S)$ is defined as the minimum number of edges whose removal disconnects $S$. These families were computed using the iterated HCS algorithm of (Hartuv and Shamir 2000).

5) **Enrichment analysis and filtering:** After initial definition of the candidate families through cluster analysis, we further evaluated the statistical significance and biological evidence for the candidate sets. The disjunction of a series of enrichment tests was used to produce the final high confidence filtered sets:

(i) We evaluated the statistical significance of the compensatory substitutions supporting each member of a family (EvoP test, see supplementary methods) on the 31-way alignment and, importantly, on the independent set of ten held-out species not used for structure inference. Considering each member as an independent test for the overall significance of a family, the p-values of all family members were combined multiplicatively using the Fisher method and used as an overall measure of evidence as well as for ranking.

(ii) For predictions within known protein-coding genes (i.e. UTR and intronic genomic regions), gene ontology (GO) enrichment statistics were computed for each cluster with three or more members, using the topGO library (Alexa et al. 2006)). We additionally required that an enriched GO term had
evidential support in two or more family members to prevent a single unusual gene flagging the entire family. The GO analysis was conducted against a background set of the original EvoFold structure predictions, and so estimated the additional enrichment of families beyond the possible enrichments or biases of the original EvoFold set. The GO analysis included inferred-by-electronic annotation (IEA) annotations. Families were filtered based on the most significant p-value in each ontology.

(iii) The degree of enrichment of family members for a particular genomic region (5' UTR, 3'UTR, intron, intergenic) was computed by chi-squared statistic relative to the background proportions of the entire EvoFold prediction set.

(iv) We calculated the mean structure length in terms of pairing bases for each family: longer structures have a lower prior probability (see figure S3) and thus higher confidence.

Using these individual family significance measures, we defined a final set of high-confidence predictions as the disjunction of the families deemed biologically significant via any of these significance estimates: those for which any of these measures had p-value smaller than a defined threshold (0.05 for double substitution p-values; < 0.005 for region enrichment; < 0.01 for maximal GO enrichment p-values); or mean base-pair length > 11. Combining these statistical measures of confidence, the original full set of candidate families was filtered to a smaller set of high-confidence families.

In addition, other enrichments were computed for annotation, but were not used in family selection. Enrichment relative to an immunity-related gene set consisting of the human homologues of mouse macrophage-related genes as defined in (Korb et al. 2008) was estimated by Fisher test.

Known structural RNA annotations were defined from human RFAM Seed (v. 9.0) entries mapped to hg18 (Gardner et al. 2009); the subset of histone 3’UTR stem-loops from RFAM Full (v. 9.0) that overlap histone-associated genes; miRBase (v. 13) (Griffiths-Jones et al. 2008); snoRNA-LBME-db (Lestrade and Weber 2006); and the Genomic tRNA Database (entries with score > 55 bits) (Lowe and Eddy 1997). After removing redundancies, this resulted in a total of 2047 known structural RNAs.

The linc-RNA sets defined in mouse (Guttman et al. 2009) and human (Khalil et al. 2009) were extracted and used to annotate family members and for GO enrichment analysis of linc-RNA-intersecting families.

To estimate the overall family false discovery rate (FDR) for the high confidence sets of candidate families, a permutation approach was used. The all-against-all similarity graph between structures was randomly shuffled (maintaining symmetry and identity properties) to produce a null set with genuine biological families shuffled, but leaving the original pair-wise similarity distribution intact. This was repeated 50 times and the numbers of families > 3 in size was compared with this randomized distribution. (Note that size 2 clusters are invariant under such shuffling, and size 2 and 3 clusters are increased due to fragmentation of larger clusters when shuffled, therefore we restricted these tests to clusters > size 3).
**RNA preparation:** The DNA template corresponding to the 186-long MAT2A RNA construct was PCR-amplified from human genomic DNA (Promega) using the oligodeoxynucleotide primers 5′A (5′-TAATACGACTCACTATAGGGACAGTTCCCATGGGAAGTGCCC) and 3′A (5′-CATGTTCATTGACTAGGTGACTGCAACTGG). As the PCR product contained an embedded T7 promoter sequence, RNA constructs were prepared by transcription in vitro using T7 RNA polymerase and gel-purified as described (Roth et al. 2006).

**In-line probing analysis:** Enzymatically synthesized RNAs were dephosphorylated with rAPid alkaline phosphatase (Roche) and radiolabeled using [γ-32P]ATP and T4 polynucleotide kinase (NEB) according to the manufacturers’ instructions. The resulting labeled RNAs were then gel-purified and subjected to in-line probing analysis essentially as described (Mandal et al. 2003). Precursor RNA at a concentration of 15 nM was allowed to undergo spontaneous transesterification for approximately 40 hours at 25 °C in 10 μL volumes containing 50 mM Tris-HCl (pH 8.3 at 23 °C), 20 mM MgCl2, and 100 mM KCl in the presence or absence of test compounds. Resulting RNA fragments were resolved using denaturing 10% PAGE and imaged with a Molecular Dynamics PhosphorImager and ImageQuaNT software.

**Acknowledgements**

BJP was funded by the Novo Nordisk Foundation and a Statistics Network Fellowship, Statistics Department, University of Copenhagen. JSP was funded by The Danish Council for Independent Research | Medical Sciences as well as the Lundbeck Foundation.

**Figure Legends**

**Figure 1.** a) Overview of EvoFam analysis and data flow. b) Phylogenetic tree relating the 31 species of the alignment screened by EvoFold. c) Each structure prediction is converted into a profile SCFG model. These models describe the nucleotide (or di-nucleotide) distributions at every base and base-pair in the structure. d) Small example of similarity graph between profile models. Maximal highly connected subgraphs (yellow) are extracted as putative families. e) Distribution of family sizes in the three final prediction sets.

**Figure 2:** Family of hairpins in 3′UTR of MAT2A. a) Location of the six hairpins (named A-F) of the MAT2A 3′UTR family. The initially predicted UTRP family consists of C, D (EvoFold predictions, purple) and B (paralog search hit, dark green). Hairpins A, E, and F were found by a dedicated, more lenient search for paralogs (light green). The well-conserved core part of the hairpins can be extended in some cases (black flanks). b) Color-coded alignment of the human sequence from all family members. Location and lengths of deletions relative to D are indicated (orange bars and numbers). The loop region reveals a motif of bases that are completely conserved between all six members (**‘**s). c) Structure of all six extended hairpins showing boundary of single sequence predictions (red bars) as well as the fully conserved motif (red bases). Note that hairpin D can also form the two base pairs of the loop regions seen for the other hairpins, though not predicted by EvoFold. d) In-line probing analysis of the 186-long MAT2A construct (including hairpin A). RNA cleavage products resulting from spontaneous transesterification during incubations in the absence (-) of any candidate ligand or in the presence of SAM, S-adenosylhomocysteine (SAH), and L-methionine (L-met), each tested at concentrations of 0.1 mM and 1 mM, were resolved by denaturing 10% PAGE. NR, no reaction; T1,
partial digest with RNase T1; 'OH, partial alkaline digest; Pre, precursor RNA. Selected bands in the T1 lane are labeled with the positions of the respective 3' terminal guanosyl residues, according to the numbering used for hairpin A in (c). Filled bars correspond to positions within hairpin A that are predicted to be largely base-paired, while the open bar corresponds to positions within the putative loop sequence. Arrowheads correspond to putative bulged nucleotides C50 and A55.

**Figure 3: Immune-related families.** a) Alignment of human sequences of members of three immune-related families. UTRP40 includes some additional members not found in the GW families. The families are enriched for macrophage-related genes and GO immunity term-associated (red). Substitutions are color-coded as in figure 2. The stems are generally more conserved (black bars) than the loops. b) Family members (green) overlap with known stabilization/destabilization elements (red). All three genes also have known AREs (blue) (including the ARE-like stability and efficiency element, SEE, (Hel et al. 1998)).

**Figure 4. tRNA-like structure in intron of POP1.** a) Intronic location of the structure and (b) alignment with a subset of species selected to show all observed substitutions (colors as in Figure 2.) c) Alignment of human sequences of family. d) Structures of family members with tRNA invariant (red) and semi-invariant (R or Y) (orange) nucleotides (Brown 2007 ). (RNA structure images generated with VARNA (Darty et al. 2009).)

**Figure 5. Examples of novel structures from families discussed in the text.** Labelled by gene symbol where available (EvoFold id in brackets).

**Table 1**

Details on families from GW, GWP and UTRP sets described in the text (see table S1 for list of top-10 GW families).

**References**


Figure 1: EvoFam family identification pipeline.
Figure 2: Family of hairpins in 3'UTR of MAT2A.
Figure 3: Immune-related families.
Figure 4. tRNA-like structure in intron of POP1.
Figure 5. Examples of novel structures from families discussed in the text.
<table>
<thead>
<tr>
<th>ID</th>
<th>Count</th>
<th>p-values (dep./indep.)</th>
<th>Mean length (bp)</th>
<th>Known functional RNA</th>
<th>Description</th>
<th>Most signif. GO term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GW4/UTRP1</td>
<td>2</td>
<td>1.3E-6/1.7E-4</td>
<td>16</td>
<td>MAT2A (3’UTR)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW177</td>
<td>3</td>
<td>1.0/1.0</td>
<td>15</td>
<td>GRIA1, 3, 4 (intrinsic)</td>
<td>glutamate signaling pathway (1.1e-04)</td>
<td></td>
</tr>
<tr>
<td>GW138</td>
<td>2</td>
<td>1.0/1.0</td>
<td>18</td>
<td>GRIA2, 4 (intrinsic)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW129</td>
<td>2</td>
<td>0.025/1.0</td>
<td>14</td>
<td>CACNA2 family (intrinsic)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW118/UTRP1</td>
<td>3</td>
<td>2.8E-4/1.0</td>
<td>19</td>
<td>SCN sodium channel</td>
<td>sodium ion transport (9.2e-07)</td>
<td></td>
</tr>
<tr>
<td>GW168</td>
<td>3</td>
<td>0.03/1.0</td>
<td>19</td>
<td>tRNA-Gly POP1 (intrinsic)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW159</td>
<td>3</td>
<td>1.0/1.0</td>
<td>22</td>
<td>mir-362 CLCN5</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW45</td>
<td>2</td>
<td>0.72/0.49</td>
<td>32</td>
<td>intergenic</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW103</td>
<td>2</td>
<td>0.0082/1.0</td>
<td>31</td>
<td>intergenic</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW266</td>
<td>2</td>
<td>0.1/0.1</td>
<td>18</td>
<td>linc-RNA MALAT1</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GW69</td>
<td>55</td>
<td>0.56/1.0</td>
<td>6</td>
<td>histone 3’SL HIST</td>
<td>nucleosome assembly (6.5e-98)</td>
<td></td>
</tr>
<tr>
<td>GW210/UTRP3</td>
<td>5</td>
<td>1.0/1.0</td>
<td>11</td>
<td>IRE TFRC (3’UTR)</td>
<td>response to iron ion (1.5e-12)</td>
<td></td>
</tr>
<tr>
<td>UTRP5</td>
<td>3</td>
<td>1.0/1.4E-1</td>
<td>8</td>
<td>translational COL collagen family (UTR)</td>
<td>skin development (1.2e-08)</td>
<td></td>
</tr>
<tr>
<td>UTRP38</td>
<td>8</td>
<td>1.0/1.0</td>
<td>6</td>
<td>CDE immunity (TNF, CYLD, UBE2W)</td>
<td>response to glucocorticoid stimulus (3.7e-04)</td>
<td></td>
</tr>
<tr>
<td>UTRP36</td>
<td>8</td>
<td>1.0/1.0</td>
<td>6</td>
<td>immunity (NFKBIZ, NFKBIP)</td>
<td>lung development (4.0e-03)</td>
<td></td>
</tr>
<tr>
<td>UTRP40</td>
<td>13</td>
<td>0.97/1.0</td>
<td>7</td>
<td>SLDE immunity-enriched (CSF3)</td>
<td>fear response (2.8e-04);</td>
<td></td>
</tr>
<tr>
<td>UTRP37</td>
<td>8</td>
<td>0.8/1.0</td>
<td>8</td>
<td>immunity (CYLD, UBE2W)</td>
<td>proteolysis (9.2e-04)</td>
<td></td>
</tr>
<tr>
<td>UTRP2</td>
<td>3</td>
<td>7.4E-2/3.1E-4</td>
<td>9</td>
<td>BCL11A/B (3’UTR)</td>
<td>T cell differentiation (1.7e-04)</td>
<td></td>
</tr>
<tr>
<td>GW59</td>
<td>3</td>
<td>1.0/0.73</td>
<td>7</td>
<td>(ANK3, CACNB4, CACNB4)</td>
<td>adult walking behavior (1.1e-04)</td>
<td></td>
</tr>
</tbody>
</table>