MIT Open Access Articles

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

Citation	Parker, B. J. et al. "New Families of Human Regulatory RNA Structures Identified by Comparative Analysis of Vertebrate Genomes." Genome Research 21.11 (2011): 1929–1943. © 2011 by Cold Spring Harbor Laboratory Press
As Published	http://dx.doi.org/10.1101/gr.112516.110
Publisher	Cold Spring Harbor Laboratory Press
Version	Final published version
Citable link	http://hdl.handle.net/1721.1/73065
Terms of Use	Creative Commons Attribution-NonCommercial 3.0 Unported License
Detailed Terms	http://creativecommons.org/licenses/by-nc/3.0/

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

Brian J. Parker, Ida Moltke, Adam Roth, et al.

Supplemental Material	http://genome.cshlp.org/content/suppl/2011/10/05/gr.112516.110.DC1.html
Related Content	**Structured RNAs in the ENCODE selected regions of the human genome** Stefan Washietl, Jakob S. Pedersen, Jan O. Korbel, et al. Genome Res. June , 2007 17: 852-864 **Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes** Michael F. Lin, Pouya Kheradpour, Stefan Washietl, et al. Genome Res. November , 2011 21: 1916-1928 **Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa** Irwin Jungreis, Michael F. Lin, Rebecca Spokony, et al. Genome Res. December , 2011 21: 2096-2113
References	This article cites 79 articles, 36 of which can be accessed free at: **http://genome.cshlp.org/content/21/11/1929.full.html#ref-list-1** Article cited in: **http://genome.cshlp.org/content/21/11/1929.full.html#related-urls**
Creative Commons License	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Method

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

Brian J. Parker,[1,6] Ida Moltke,[1] Adam Roth,[2] Stefan Washietl,[3] Jiayu Wen,[1] Manolis Kellis,[3] Ronald Breaker,[4] and Jakob Skou Pedersen[1,5,6]

[1]The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark; [2]Department of Molecular, Cellular and Developmental Biology, Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06520-8103, USA; [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [4]Department of Molecular, Cellular and Developmental Biology, Howard Hughes Medical Institute, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8103, USA; [5]Department of Molecular Medicine, Aarhus University Hospital, Skejby, Brendstrupgaardsvej 100, DK-8200 Aarhus N, Denmark

Regulatory RNA structures are often members of families with multiple paralogous instances 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, including 48 families with known structural RNA elements. Known families identified include both noncoding RNAs, e.g., miRNAs and the recently identified MALAT1/MEN β lincRNA family; and cis-regulatory structures, e.g., iron-responsive elements. We also identify tens of new families supported by strong evolutionary evidence and other statistical evidence, such as GO term enrichments. For some of these, detailed analysis has led to the formulation of specific functional hypotheses. Examples include two hypothesized 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 predicted MAT2A structures. 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 exemplify the diversity of post-transcriptional regulation and provide a resource for further characterization of new regulatory mechanisms and families of noncoding RNAs.

[Supplemental material is available for this article and these data can also be found at http://moma.ki.au.dk/prj/mammals.]

The large and versatile role of RNA in protein-coding gene regulation is by now well supported. On the one hand, noncoding 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 the 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 (Butter et al. 2009). 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.

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), detect this signal by analyzing the substitution pattern along genomic alignments and use it to identify these conserved RNA structures (Pedersen et al. 2006).

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 genomics pipeline, EvoFam, based on the evolutionary substitution signal, to perform the first general genome-wide screen for families of human regulatory RNAs. As part of the 29 Mammals Sequencing and Analysis Consortium (Lindblad-Toh et al. 2011), we used the deep genomic vertebrate alignments generated, which include 29 mammals, 21 of which are sequenced at low coverage, and 10 additional vertebrates (Supplemental Figs. 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 10 genomes for later validation purposes. This method uncovers 220 new high-confidence families, many of which lend themselves to specific functional hypotheses, in some cases with apparent medical implications. Although 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 structural families among genome-wide structure prediction sets. Pedersen et al. (2006) identify families among EvoFold predictions limited to primary sequence similarity, whereas Will et al. (2007) identify families among RNAz (Gruber et al. 2010) predictions in *Ciona intestinalis*. Other methods for identifying families of structures are based on motif discovery of RNA structures that are shared among a selected set of unaligned sequences, rather than using 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. (2008) report a search for common structured motifs in yeast, mouse, and fly; and Khladkar et al. (2008) in human and mouse orthologs. 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. An overview of the pipeline is given below (see Fig. 1; for details, see Methods).

First, we generated a genome-wide prediction set of structural RNAs in human (Fig. 1, Step 1). For this, we applied EvoFold (Pedersen et al. 2006) to all conserved segments (spanning 5.6% of the genome) of a 31-way subset (28 placental mammals, opossum, chicken, and tetraodon) (Fig. 1B; Supplemental Fig. S1) of the vertebrate genome alignment made by the 29 Mammals Sequencing and Analysis Consortium (Lindblad-Toh et al. 2011). After eliminating predictions of either poor quality or residing in repetitive regions (for details, see Methods), this set contains 37,381 elements. Compared to the initial human EvoFold screen of an eight-way vertebrate alignment, this set has 21% fewer predictions, an elementwise overlap of 38%, and a 4% higher recall rate of 39% on a set of known, conserved structural RNAs (Pedersen et al. 2006). For the downstream analysis, we excluded predictions in protein-coding regions, since their elevated false-positive rate (Pedersen et al. 2006) combined with extensive primary sequence homology in gene families led to higher rates of apparent false family predictions. The final prediction set contains 27,012 structural RNA predictions.

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 (Fig. 1, Step 2a; Eddy and Durbin 1994; Durbin et al. 1998). The large evolutionary span of the alignments (7.1 expected substitutions per neutrally evolving position) allows the profile models to capture the general sequence and structure constraints acting on a given type of structural RNA, and the models are therefore sensitive in homology searches.

We then detected homology between the structural RNA predictions, using 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 between probabilistic models, based on the statistical significance of the similarity between sequences generated by the corresponding pSCFGs (Fig. 1, Step 3; see Supplemental Methods).

Based on this all-against-all similarity evaluation, we constructed a similarity graph and identified 1254 candidate families as densely interconnected subgraphs (Fig. 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 Supplemental Methods). 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 10 withheld vertebrates ("independent set") (see Supplemental Fig. S2). 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 (Fig. 1, Step 5).
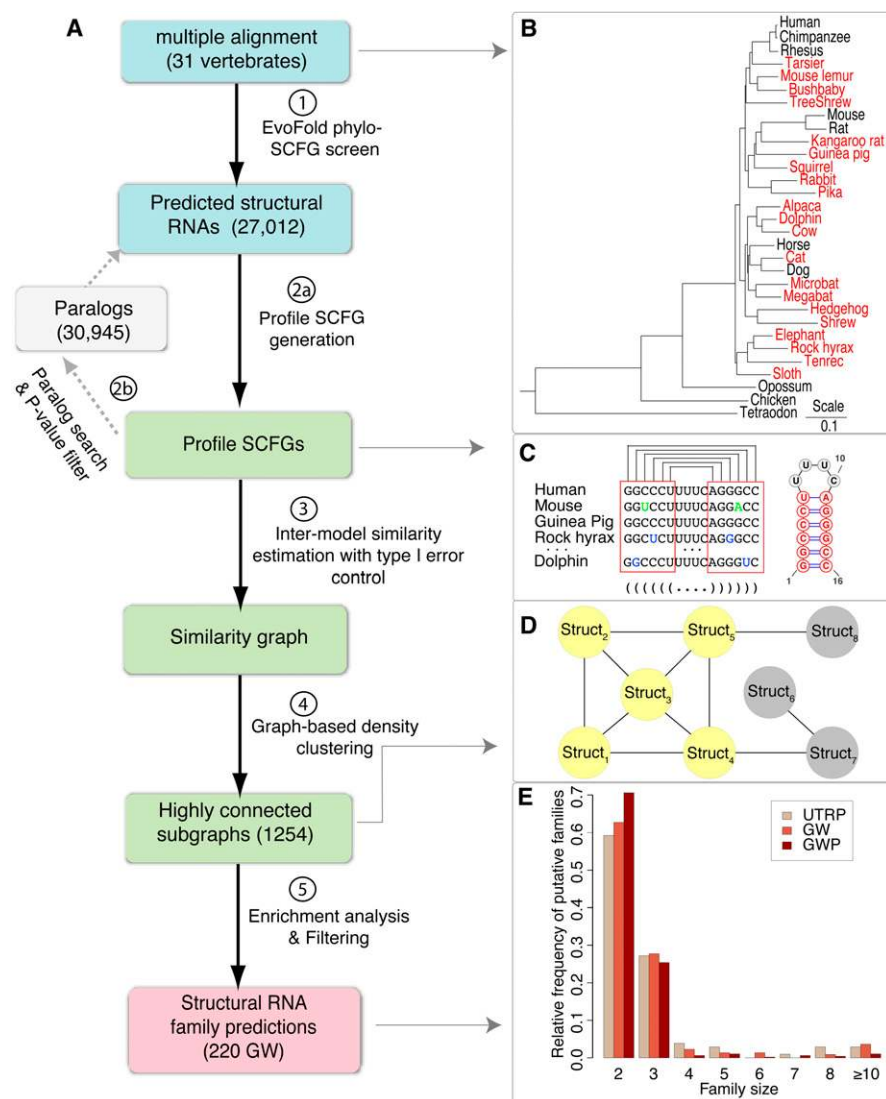
Figure 1. EvoFam family identification pipeline. (*A*) Overview of EvoFam analysis and data flow. (*B*) Phylogenetic tree relating the 31 species of the alignment screened by EvoFold. (*C*) Each structure prediction is converted into a profile SCFG model. These models describe the nucleotide (or di-nucleotide) 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.

Most families detected consist purely of hairpins (97.2%; 88.5% in EvoFold background), although some detected hairpins may represent components of larger RNA structures. The families were structurally homogeneous, with a mean fraction of matched base pairs between members of 89.0%, and mean sequence identity was 77.0% (see Supplemental Methods and Supplemental Data File S3). See Table 1, Supplemental Tables S1 and S4, and Figure 5 below for the families and structures described in the sequel. There are many families with long structures (50.5% >11 bp on average); however, this is partly caused by their explicit inclusion through the selection criteria (cf. 28% >11 bp when length is removed as a selection criterion). The larger full set of unfiltered candidate families is dominated by short structures (90.2% ≤11 bp on average), as is the background set of EvoFold predictions (77.9% ≤11 bp) (Supplemental Fig. S4). The intergenic predictions appear to be truly non-protein-coding, as the base-level overlap with ab initio protein-coding exon predictions (Siepel and Haussler 2004) was only 2.5%, which was not statistically significant (*P*-value = 1.0; binomial test) (see Supplemental Data File S2).

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 (Fig. 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 vs. 220 families), but has about the same average family size (2.7 vs. 3.3 members) (see Fig. 1E). The other set was made in the same way but was based on only UTR Evo-Fold predictions and their UTR paralogs (UTRP set). In support of our focus on *cis*-regulatory structures, this smaller focused set allows a more sensitive similarity threshold in the family classification of the UTR regions without increasing the false-positive rate and includes 103 families (see Supplemental Tables S2, S3).

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://moma.ki.au.dk/prj/mammals.

The included families were either strongly supported by double substitution evidence in the 31-way alignment or the 10 withheld species (*P*-value < 0.05; EvoP test with *P*-values combined by Fisher's method); enriched for a specific genomic region (*P*-value < 0.01; χ^2 test); had significant gene ontology term enrichment (*P*-value < 1×10^{-3}; Fisher's test); or long structures (>11 bp on average). This filter resulted in 220 high-confidence families containing 725 individual human structures (EvoFold predictions) (see Supplemental Data File S1).

The majority of families were small with two or three elements (90.5%) (see Fig. 1E; Supplemental Fig. S3). Compared to the EvoFold background, the families show strongest enrichment for UTR structures (UTR: 22.2% in families vs. 13.6% among EvoFold background; intronic: 34.8% vs. 35.6%; and intergenic: 43.0% vs. 50.8%) (*P*-value = 8×10^{-11}; χ^2 test) (Supplemental Figs. S4, S5).

Table 1. Details on families from GW, GWP, and UTRP sets described in the text

EvoFam ID	Count	Family P-value (EvoP dep./indep.)	Mean length (bp)	Known functional RNAs	Description
GW69	45	0.56/1.0	6	Histone 3′ stem–loop	Histone genes (3′-UTR)
GW210/UTRP35	5	1.0/1.0	11	IRE	*TFRC* (3′-UTR)
GWP266	2	0.1/0.1	18	linc-RNA	*MALAT1/MEN β* family (intergenic)
GW4/UTRP1	2	1.3×10^{-6}/1.7×10^{-4}	16		*MAT2A* (3′-UTR)
UTRP38	8	1.0/1.0	6	Constitutive decay element (CDE)	Immunity-enriched (e.g., *TNF, CTLA4, FOS*)
UTRP36	8	1.0/1.0	6		Immunity-enriched (e.g. *NFKBIZ, NFKBID*)
UTRP40	13	0.97/1.0	7	Stem–loop decay element (SLDE)	Immunity-enriched (e.g., *CSF3*)
GW138	2	1.0/1.0	18	RNA editing hairpins	*GRIA2, 4* (intronic)
GW129	2	0.025/1.0	14		*CACNA2* family (intronic)
GW177	3	1.0/1.0	15		*GRIA1, 3, 4* (intronic)
GW118/UTRP19	3	2.8×10^{-4}/1.0	19		*SCN* sodium channel (3′-UTR)
GW168	3	0.03/1.0	19	tRNA-glycine	*POP1* family (intronic)
UTRP5	3	1.0/1.4×10^{-1}	8	Collagen 5′-stem–loop	*COL* collagen family (5′-UTR)
UTRP37	8	0.8/1.0	8		Immunity-enriched (e.g., *CYLD, UBE2W*)
UTRP2	3	7.4×10^{-2}/3.1×10^{-4}	9		*BCL11A/B* (3′-UTR)
GW59	3	1.0/0.73	7		*ANK3, CACNB4, KCNMA1*
GW159	3	1.0/1.0	22	mir-362	*CLCN5*
GW45	2	0.72/0.49	32		Intergenic
GW103	2	0.0082/1.0	31		Intergenic
GWP786	2	1.0/1.0	16.5	linc-RNA	*XIST/MAK* (intronic)

See Supplemental Table S1 for a list of the top-10 GW families.

Recovery of known families

The EvoFam pipeline correctly identifies many of the known structural *cis*-regulatory and ncRNA families (see Supplemental Tables S4–S6). Only a few families of human *cis*-regulatory structures are described in the literature (we found less than 10) (Gardner et al. 2009; Jacobs et al. 2009). Among these, we recover the families of (a) histone 3′-UTR stem–loops (45 of 67 known); (b) hairpins regulating translation in collagen genes (3 of 3) (Stefanovic and Brenner 2003); (c) and iron-responsive elements (IRE) in the 3′-UTR of the transferrin (*TFRC*) gene (5 of 5). Some *cis*-regulatory families have low recovery 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).

Among known ncRNAs, miRNAs are recovered with good sensitivity (139 of 441 known, conserved, miRNA genes in 42 families), whereas only a few snoRNAs (two in one family) and tRNAs (two in one 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. 2006).

The precision (positive predictive value) of the approach is estimated by the high fraction of known structural RNAs (88%) in families with any known members (see Supplemental Methods). The paralog search used for the GWP set picks up many of the initially missed family members (Fig. 1, Step 2b) and increases the overall number of known structural RNAs relative to the GW set (226 vs. 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* [Wilusz et al. 2008] and *MEN β*; GWP266). In both cases, two Evo-Fold hairpin predictions, well supported by substitution evidence, are found upstream of the clover-leaf-shaped structures (Supplemental Fig. S6). The *MEN β* paralog structures and their role in 3′-end processing were recently discovered and published, during paper preparation (Sunwoo et al. 2009; Wilusz and Spector 2010).

False discovery rate 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 false discovery rate (FDR) based on randomly shuffled versions of the observed data, maintaining the original length and structural complexity distributions (see Supplemental Methods). We estimated the FDR separately for pairs of structures and for families of size three or more. The pairwise FDR was estimated at 27% by using randomly shuffled (while maintaining structure) multiple alignments; the pairwise FDR was estimated at 17% when restricted to the set of 6-bp hairpins (GW set), demonstrating that the similarity measure maintains control of the false-positive rate even for short structures. The family-wise FDR was estimated at 34% for families of size three or more (41% for size three; <1% for size greater than three) by random shuffling of the edges of the similarity graph (UTRP set).

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.

Enrichment analyses of family predictions

Comparing the median EvoFold log odds model fit scores of family members (20; GW set) versus a set of 356 known functional RNAs (22) and the background EvoFold predictions (11) shows that the families have similar enrichment to known functional RNAs.

To further substantiate these results by an alternative computational method, we applied RNAz to the EvoFam families and background set of EvoFold predictions. In contrast to EvoFold, which 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), this mutual overlap is relatively low albeit highly significant (~12-fold enrichment over random) (Supplemental Fig. S7), 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 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 ~50% roughly suggest that ~60%–80% of 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 the RNAz sensitivity on the true EvoFam predictions is similar to that of the current set of known structural RNAs. This may not be fulfilled, as EvoFam detects many families of short hairpins, which are difficult to detect by the windowing approach used by RNAz and not well represented in the benchmark set of known structural RNAs.

Furthermore, the individual structures within families are predominantly predicted on the transcribed strand when overlapping known genes (63% computationally detected by EvoFold on the transcribed strand for UTR structures of the GW set; P-value = 1×10^{-3}; binomial test), which suggests that the majority of these are indeed cis-regulatory. Note that the EvoFold strand predictions have overall low accuracy, since they mostly rely on the weak signal of substitutions involving GU base pairs (Pedersen et al. 2006): The true fraction on the transcribed strand is therefore likely higher (e.g., 80% were predicted on the transcribed strand when limited to structures showing more than three single substitutions).

In van Bakel et al. (2010) it was noted that intergenic ncRNAs detected using RNA-seq evidence were highly correlated with DNase hypersensitivity sites. Similarly, we find that 32% of the novel GW intergenic predictions overlap ENCODE UW DNase I hypersensitivity clusters (The ENCODE Project Consortium 2007), which is a 1.4× enrichment over a random conserved intergenic background (P-value < 4×10^{-3}; permutation test), 4.5× over a random intergenic background (P-value < 1×10^{-4}; permutation test), and significantly greater than the input EvoFold enrichment relative to conserved background (25%, P-value = 2×10^{-2}; Fisher's exact test) (see Supplemental Methods).

Using an Illumina ribo-depleted, non-poly(A) selected, total RNA RNA-Seq library from 16 pooled tissues, GW intergenic and intronic novel structures were significantly expressed compared with a shuffled set of random structure positions chosen from conserved intergenic and intronic regions. The novel intergenic elements showed a mean coverage (reads per base) of 1.70-fold relative to random (P-value < 1×10^{-3}; permutation test); the novel intronic elements showed a mean coverage of 1.89-fold relative to random (P-value < 1×10^{-3}; permutation test).

To estimate whether the expression of the members of the detected families show a positive correlation due to shared regulation, e.g., function as RNA regulons (Keene 2007), the mean pairwise correlation of expression within families, across 16 tissues using a large RNA-seq data set from Illumina, was compared with the distribution of randomly shuffled families. This showed a small but statistically significant mean Pearson correlation coefficient of 0.17 compared with a mean of 0.05 for a randomized set, for novel members of the GW set (P-value < 1×10^{-3}; permutation test) (see Supplemental Fig. S8).

Putative new cis-regulatory structure families

Post-transcriptional regulation of protein-coding genes is mediated by cis-regulatory elements, often structured, that bind proteins or other $trans$-acting factors. This widespread regulation takes place at various stages of the mRNA life cycle and can be categorized into (a) mRNA stability; (b) pre-mRNA processing (e.g., editing and alternative splicing); (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 (for review, see Moore 2005; Garneau et al. 2007). Below we describe several novel families of candidate cis-regulatory structures identified by EvoFam, selected as examples where available evidence allows specific functional hypotheses to be proposed within these various forms of post-transcriptional regulation.

Families hypothesized to be involved in mRNA stability

The dynamic balance of the rate of transcription and the rates 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; cmsearch) (Fig. 2A). These hairpins are characterized by a loop motif (Fig. 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 (Fig. 2C; Hofacker et al. 1994). These extended parts of the predictions are also supported by substitution evidence, although less so than the core part (Supplemental Figs. S9–S14). Some $MAT2A$ mRNAs and ESTs in GenBank show long 3'-UTRs that include all six hairpins, but most are short and include only hairpin A. Interestingly, a few alternatively spliced ESTs and RNA-seq reads show evidence of an alternative intron in the 5' part of the 3'-UTRs, with hairpins C–F located downstream from this intron, and the 3'-splice site located 6 bp before hairpin C (Fig. 2A), potentially modulating the regulatory effect of the hairpins. The ratio of 3'-UTR expression over the region of hairpins C–F compared with the region over the putative retained intron varies across tissues, broadly matching the known tissue distribution of methionine adenosyltransferase (Spearman correlation coefficient 0.94, P-value = 8×10^{-3}) (see Supplemental Fig. S15), suggesting a regulatory function of the alternative 3'-UTRs.

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,
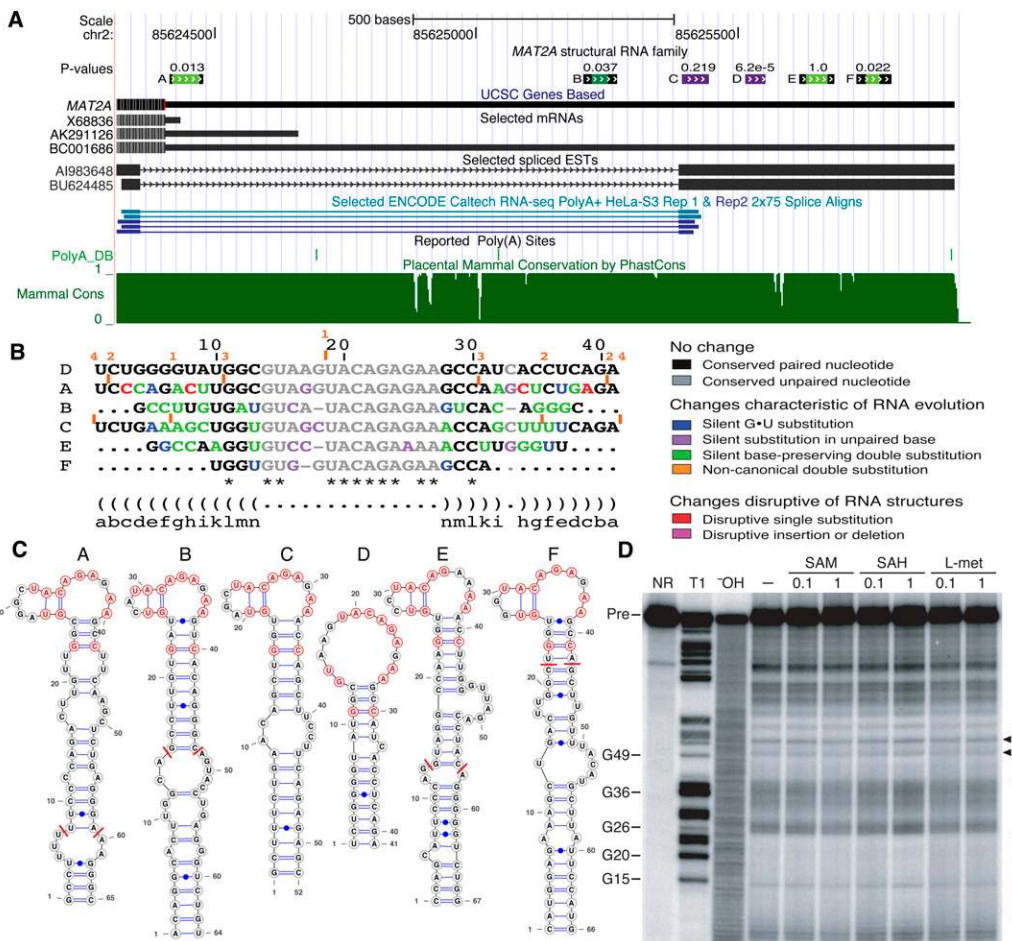
Parker et al.

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). A putative 3′-UTR alternative intron is indicated by spliced EST and RNA-seq evidence. (*B*) Color-coded alignment of the human sequence from all family members. The alignment is referenced by D. Location and length of relative insertions in the other sequences are indicated (orange bars and numbers). The loop region reveals a motif of bases that are completely conserved among all six members (*). (*C*) Structures of all six extended hairpins showing the boundary of single sequence predictions (red bars) as well as the fully conserved motif (red nucleotides). Note that hairpin D can also form the two base pairs of the loop regions seen for the other hairpins, although 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 panel *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) Putative bulged nucleotides C50 and A55.

involved in a broad range of processes including polyamine biosynthesis and gene regulation through DNA methylation. SAM is found in all kingdoms of life. In humans, the *MAT2A* mRNA half-life varies from 100 min to >3 h depending on SAM availability (Martinez-Chantar et al. 2003), and we hypothesize that the 3′-UTR hairpins are involved in this regulation.

Riboswitches 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, interestingly in UTR introns in some cases (Bocobza and Aharoni 2008).

To test the possibility that the conserved hairpins in the 3′-UTRs of vertebrate *MAT2A* homologs might mediate gene regulation as riboswitches, 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 (Supplemental Fig. S16). These RNAs, which contained from one to three conserved hairpin loops, were [32]P 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 (Fig. 2D; Supplemental Figs. 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 (Fig. 2D; Supplemental Figs. 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 (Fig. 2D; Supplemental Figs. 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 (Fig. 2D; Supplemental Figs. 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 (Fig. 2D; Supplemental Figs. 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. For hairpins A, C, and D, a nucleotide-level structural comparison could be performed, showing that the base-pairing status of 80%, 92%, and 100% of the nucleotides, respectively, is supported by the structure-probing data (Supplemental Fig. S21).

Thus, 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). Other possible mechanisms include other protein-binding RNA switch mechanisms involving conformational change in structure (Ray et al. 2009), and/or splicing enhancer/silencer functionality. These alternative 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's test) and immunity-related GO terms (e.g., leukocyte migration; P-value = 4.5×10^{-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 (Fig. 3A; Supplemental Table S7).

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]; 5×10^{-4} [GW219]; χ^2 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 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 ↔ AU). In the aligned human sequences of the family members, the stems show strong sequence conservation and weaker conservation of the loop (Fig. 3A).

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 (Fig. 3B):

1. Tumor necrosis factor α (TNF), which produces a key cytokine mediating the inflammatory response, contains a 15-nt 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 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 candidate example is the hairpin in the 3'-UTR of 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.

Families involved in pre-mRNA processing

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. The majority of human editing sites are found in inverted repeats, which are not conserved and therefore not detected here (Li et al. 2009). In contrast, the functionally
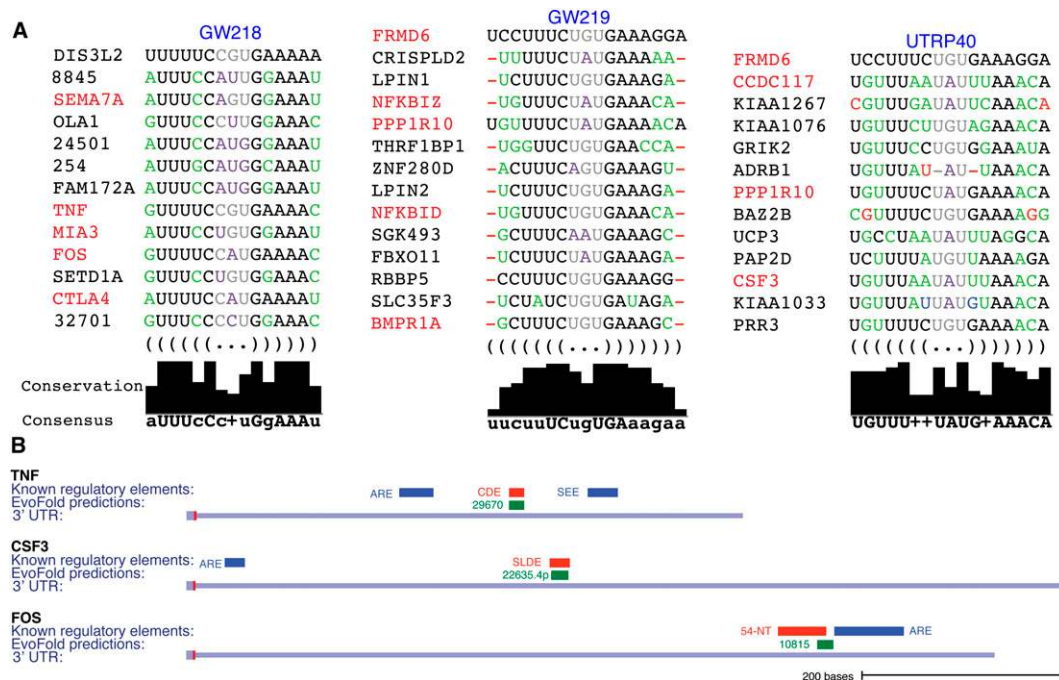
Figure 3. Immune-related families. (*A*) Alignment of human sequences of members of three immune-related families. UTRP40 includes some additional members not found in the GW families. The families are enriched for macrophage-related genes and GO immunity term association (red). Substitutions are color-coded as in Figure 2. The stems are generally more conserved (black bars) than the loops (predictions without gene symbols are labeled by EvoFold id). (*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).

characterized editing sites are found in deeply conserved hairpins and therefore potentially detectable. 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 A-to-I editing sites (GW138 containing *GRIA2* and *GRIA4*) (Lomeli et al. 1994).

All known, functional editing sites are exonic, with the exception of a site in an intron of *ADARB1*, which creates a 3′-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 intronic families of long hairpins in calcium channel genes (*CACNA2D1*, *CACNA2D2*) and glutamate receptors (*GRIA1*, *GRIA3*, *GRIA4*; GW129 and GW177, respectively) (see Supplemental Table S8). 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, these hairpins are also candidate intronic editing sites, which may change the function of splicing enhancers or other *cis*-regulatory elements. Alternatively, they may be involved in regulation of exonic editing events. Finally, we also identify a family in the 3′-UTRs of three sodium channels genes (*SCN1A*, *SCN2A*, *SCN3A*; UTRP19), which is well supported by substitution evidence (*P*-value $< 3 \times 10^{-4}$; EvoP dependent set) and may be an example of novel editing sites.

Hypothesized 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 Fig. 4). The *POP1* intronic structural element is well supported by double substitutions and thus not a tRNA pseudo-gene (*P*-value $< 1.0 \times 10^{-3}$; EvoP dependent set). We rule out that this is simply an alignment artifact caused by including tRNAs from elsewhere in the genome, as the structure shows conserved synteny with the exons of *POP1* (Fig. 4A; Kent et al. 2003). Although the *POP1* intronic structural element shows characteristics of tRNAs, including loop motifs shared with the tRNA members of the family, it is likely not a functional tRNA as it shows a shorter D stem with only a single base between it and the acceptor stem, potentially compromising the tertiary structure, and has an anticodon loop 1 base longer than the canonical tRNA structure. Furthermore, tRNAscan-SE detects it, but with a low score (Lowe and Eddy 1997). However, the structure is similar enough to a tRNA to be a predicted substrate for human ribonuclease P (RNase P) (Lundblad and Altman 2010). Interestingly, *POP1* encodes a protein subunit of RNase P, the riboprotein complex that matures tRNA molecules by cleaving the 5′-ends of precursor tRNAs; in particular, the POP1 protein component has been shown to interact with the 5′-end of tRNAs (Butter et al. 2009). Based on this, we hypothesize that this tRNA-like structure is involved in auto-regulation of the *POP1* transcripts, with the RNase P complex binding and potentially cleaving the transcript. Figure 4A shows RNA-seq evidence of such a cleavage product, showing a cytosolic ncRNA with a 5′-end precisely matching the predicted RNase P cleavage site. Also, there is anti-correlated expression of a localized region overlapping this structure compared with the upstream exon, across 11 tissues (Pohl et al. 2009) (Pearson correlation coefficient −0.63; *P*-value = 0.018; one-sided test), consistent with feedback regulation (see Supplemental Methods).
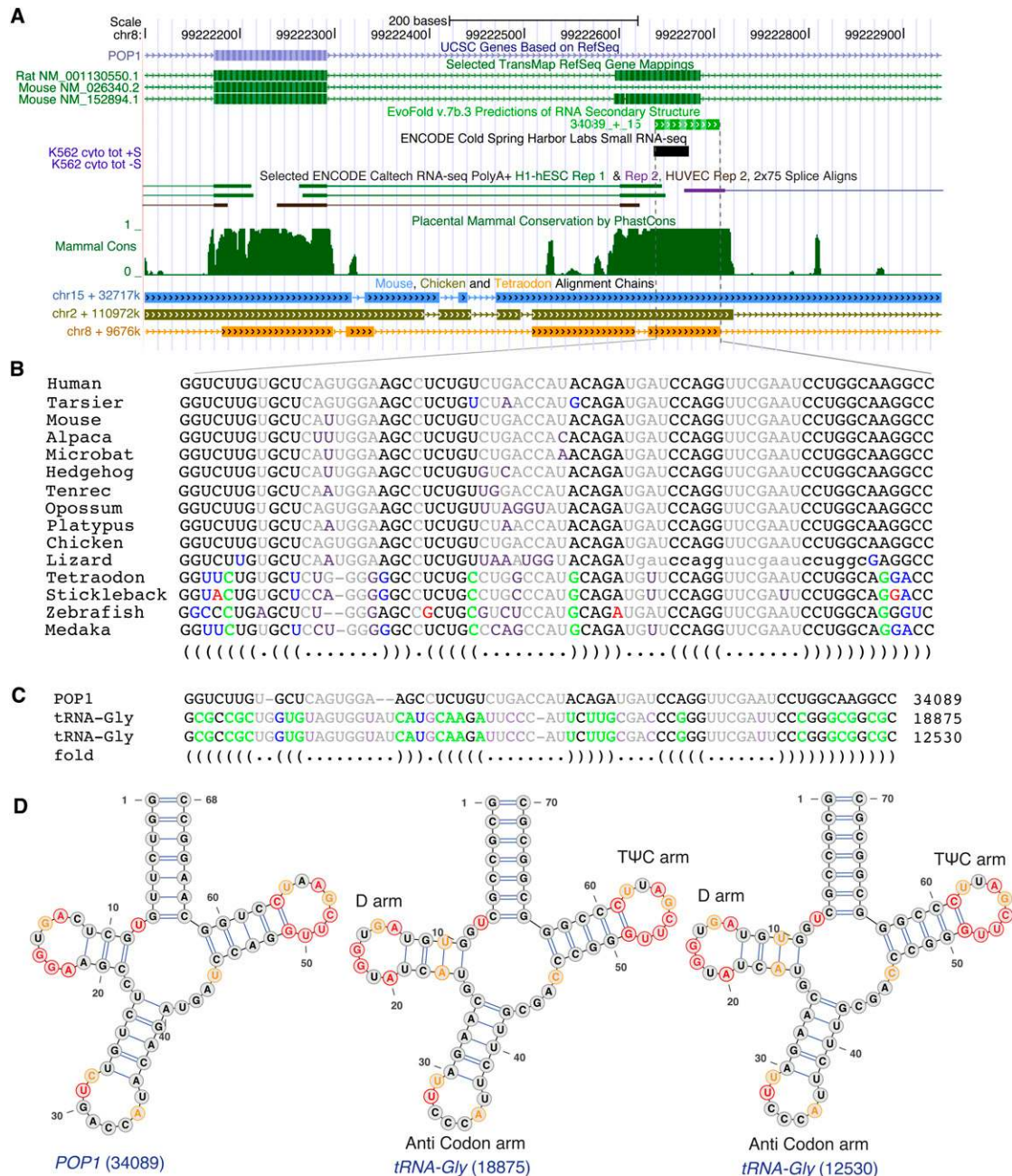
Figure 4. tRNA-like structure in intron of *POP1*. (*A*) Intronic location of the structure. The ENCODE CSHL small RNA-seq track (The ENCODE Project Consortium 2007) for cell line K562 represents three uniquely mapped cytoplasmic reads with 5′-ends aligned with the predicted RNase P cleavage site (the cloning protocol generates directional libraries that are read from the 5′-ends of the inserts, which should largely correspond to the 5′-ends of the mature RNA). Spliced reads suggest splicing activity and possible cassette exon in the region of the structure; mapped RefSeqs (TransMap) show cassette exons from mouse and rat that overlap the structure position. (*B*) Alignment with a subset of species selected to show all observed substitutions (colors as in Fig. 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]).

Families 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 hairpin downstream from the TGA recoding site of the *SELK* SECIS element (UTRP3), likely a new example of a selenocysteine codon redefinition element (SRE) (Pedersen et al. 2006; Howard et al. 2007), we recover a known family of hairpins overlapping the start codon of *COL1A1*, *COL1A2*, and *COL3A1* (GW36) known to control translation (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* (for details, see Supplemental Table S9; Supplemental Results).

Putative miRNAs, lincRNAs, and other ncRNAs

Included among the families of known miRNAs are examples of putative new miRNAs or other functional long hairpins (see Supplemental Table S10). 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 (Supplemental Fig. S22; Supplemental Table S11) (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, GW45 and GW103 (>30 bp) (see Fig. 5), which are both highly conserved, with GW103 supported by strong double substitution evidence (P-value $< 8 \times 10^{-3}$; EvoP dependent set) and both supported by compatible single substitutions ($n = 26$ and 27, respectively). Supplemental Figure S23 shows RNA-seq expression evidence for GW45. The length of the hairpins and a lack of typical RNA-seq evidence suggest that these hairpins may function other than as miRNA precursors.

From the GWP set, 119 families with 158 members (28%) overlap long intervening noncoding RNAs (lincRNAs) (Guttman et al. 2009; Khalil et al. 2009). This includes some known families, such as the previously mentioned *MALAT1* / *MEN β* family as well as 52 known miRNAs and snoRNAs. A novel and well-supported structure in *XIST*/*TSIX* was detected in the 3′ region, distinct from known 5′ structures (Maenner et al. 2010), and which overlies a region of high expression within the chromatin cellular component (217 uniquely mapped RNA-seq reads, K562 cell line, compared with a background coverage over *XIST* of 5.6 reads/base [ENCODE CSHL small RNA-seq]). Interestingly, it shows some evidence of homology with an intronic structure in male germ-cell-associated kinase (*MAK*) (see Supplemental Fig. S24; Supplemental Results).

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
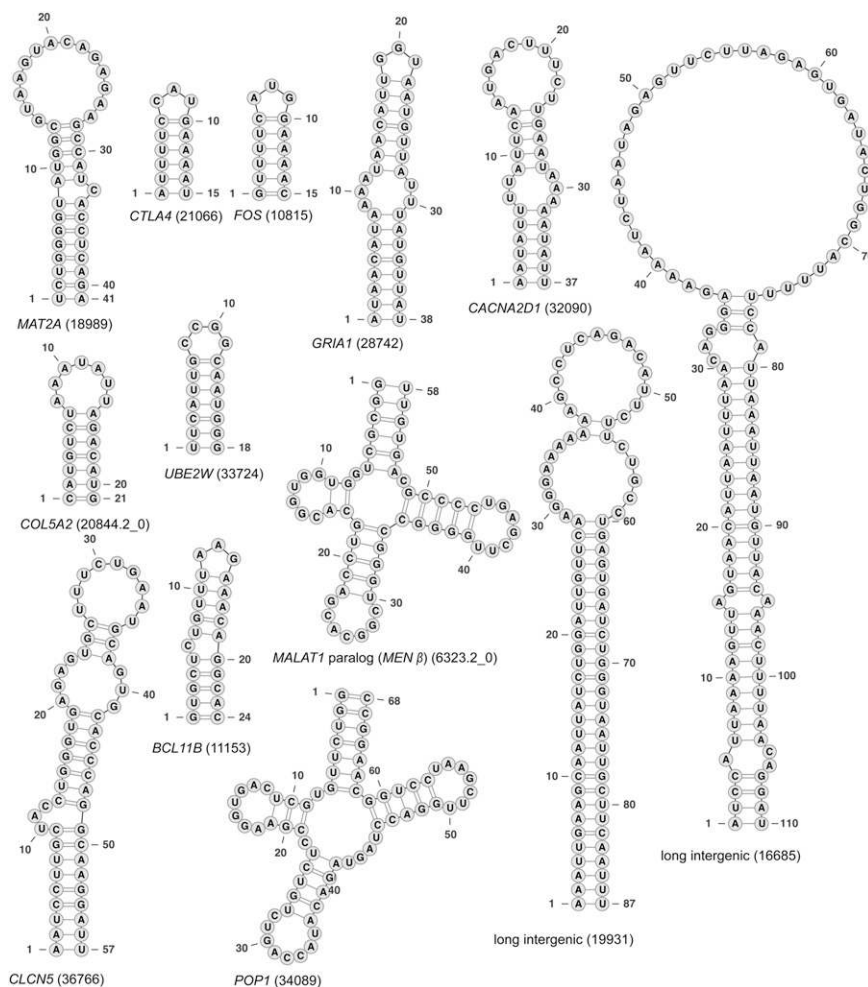

Figure 5. Examples of novel structures from families discussed in the text. Labeled by gene symbol where available (EvoFold id in brackets).

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-value $= 4.5 \times 10^{-4}$), histone methyltransferase activity (3.1×10^{-3}), and various terms related to cellular adhesion (see Supplemental Table S12). This coincides with the results of an independent 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 potentially 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* / *MEN β* lincRNA family. Furthermore, novel members were added to known families in some cases, e.g., a collagen 5′-UTR hairpin and *CLCN5* miRNA. We also found

strong evidence for a large number of completely novel families. Among our 220 high-confidence families, we found 172 novel families containing an estimated 40%–60% 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 and the overall false discovery rate (FDR) is therefore high. In the initial EvoFold screen (Pedersen et al. 2006), FDR was 62% overall (log odds score cutoff of 0) but decreasing for increasingly stringent cutoff. This study showed that the EvoFam family classification approach is an efficient means for defining a high-confidence set of predictions based on an initial inclusive set, as demonstrated by (1) the enrichment for known structures, (2) the significantly enriched overlap with DNase hypersensitivity sites, (3) the high agreement with RNAz predictions, (4) the support for specific novel families revealed by detailed analysis, and (5) 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 that is robust to noise, enables a genome-wide analysis with low overall error rates.

The availability of deep (41-way) vertebrate alignments (K Lindblad-Toh, M Garber, O Zuk, MF Lin, BJ Parker, S Washietl, P Kheradpour, J Ernst, G Jordan, E Mauceli, et al., in prep.) allowed both discovery and validation on the same data set. This was done by withholding 10 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 allows the members of a family to be studied as a whole and can allow the formulation 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 4.1× enrichment compared to the set of input conserved elements. In-depth analysis revealed several families supported by circumstantial evidence, for which we presented specific functional hypotheses based on the principles given above. These appear to provide novel examples of post-transcriptional regulation at various stages of the mRNA life cycle. Combined, the identification of these putative families potentially 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 riboswitches or via protein factors. In-line probing assays supported the predicted structures but did not support the function as riboswitches. We therefore favor the involvement of protein factors in this regulation. (2) For *POP1*, we hypothesized that the transcript is bound by RNase P and perhaps cleaved. Similar examples of auto-regulation have been reported in the literature for *ADARB1*, 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.

In this analysis, EvoFold was applied to highly conserved regions (mean number of substitutions per alignment column = 0.47) (see Supplemental Methods). EvoFold relies on an initial sequence-based multiple alignment and thus works well in such conserved regions. Methods that adjust the multiple alignment during structure modeling, e.g., CMFinder, may provide a complementary approach that is more sensitive in regions of low conservation and sequence identity (Mathews and Turner 2002; Havgaard et al. 2005; Torarinsson et al. 2008). An advantage of the

more constrained EvoFold model in conserved regions is that, with fewer degrees of freedom in the model, there is lower risk of overfitting of the structure, and it allows reliable estimation of statistical significance using the substitution signal.

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 172 new high-confidence families supported by strong evolutionary 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 29 Mammals Sequencing and Analysis Consortium, along with two out-group vertebrate species (chicken and tetraodon). An additional 10 species, primarily nonmammalian vertebrates that were not used for structure inference, were used as an independent test set (for full species list, see Supplemental Figs. S1, S2).

EvoFam pipeline (see Fig. 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. 2006). Low-confidence predictions that were short (<6 bp), harbored excessive amount of bulges, or were based on shallow or low-quality alignments or overlapped repeats or pseudogenes were eliminated from the prediction set. Finally, overlaps between predictions were resolved according to EvoFold score. For details on the screen, see the Supplemental Methods.

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 greatest overlap with. Protein-coding regions were excluded from the study to focus it on noncoding 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 because of the unequal substitution rate at different codon positions.

Profile SCFG generation

Because structure is a key feature of the family members, we used both sequence and structure information in detecting the regula-

tory RNA families. For each EvoFold prediction, we fitted a profile stochastic context-free grammar (pSCFG, aka covariance model) model using the Infernal RNA tools v1.0 (cmbuild utility) (Nawrocki et al. 2009). pSCFGs describe individual unpaired and paired positions by single states (see Fig. 1C). 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.

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-Mb 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.

Inter–pSCFG similarity estimation and type I error control

A similarity graph between structural predictions was defined based on an all-against-all similarity (homology) estimation between 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 the 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

$$\tilde{D}(M_1 \| M_2) = E(S(seq_1^{human}, M_2)) - E(S(seq_1^{human}, M_1))$$

where $S(seq_m^{human}, M_n)$ 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-Mb database. For the derivation of this measure, see the Supplemental Methods.

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 trade-off for inclusion of edges, to control FDR (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).

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:

1. We evaluated the statistical significance of the compensatory substitutions supporting each member of a family (EvoP test) (see Supplemental Methods) on the 31-way alignment and, importantly, on the independent set of 10 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 Fisher's method and used as an overall measure of evidence as well as for ranking.

2. 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 thus 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.

3. The degree of enrichment of family members for a particular genomic region (5′-UTR, 3′-UTR, intron, intergenic) was computed by χ^2 statistic relative to the background proportions of the entire EvoFold prediction set.

4. We calculated the mean structure length in terms of pairing bases for each family: Longer structures have a lower prior probability (see Supplemental Fig. 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 a 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 homologs of mouse macrophage-related genes as defined in Korb et al. (2008) was estimated by Fisher's exact 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 lincRNA 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 lincRNA-intersecting families.

To estimate the overall family false discovery rate for the candidate families, a permutation approach was used. The all-against-all similarity graph between structures was randomly shuffled to produce a null set with no genuine homologous fam-

ilies, but leaving the original pairwise similarity distribution intact. Note that size 2 clusters are invariant under such shuffling: The FDR of size 2 families was similarly estimated by comparison with a null set generated using randomly shuffled multiple alignments (see Supplemental Methods).

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′-TAATACGACTCA CTATAGGGACAGCTTCCCATGGGAAGTGCCC) and 3′A (5′-CATG TCATTGACTAGAGTGACTGCAACTGG). 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-^{32}P]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 ~40 h at 25°C in 10-μL volumes containing 50 mM Tris-HCl (pH 8.3 at 23°C), 20 mM MgCl$_2$, 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.

References

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22:** 1600–1607.

Anderson P. 2008. Post-transcriptional control of cytokine production. *Nat Immunol* **9:** 353–359.

Bocobza SE, Aharoni A. 2008. Switching the light on plant riboswitches. *Trends Plant Sci* **13:** 526–533.

Brown TA. 2007 *Genomes 3*. Garland Science, New York.

Brown CY, Lagnado CA, Goodall GJ. 1996. A cytokine mRNA-destabilizing element that is structurally and functionally distinct from A+U-rich elements. *Proc Natl Acad Sci* **93:** 13721–13725.

Butter F, Scheibe M, Morl M, Mann M. 2009. Unbiased RNA–protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci* **106:** 10626–10631.

Chen JM, Ferec C, Cooper DN. 2006. A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3′ UTR variants. *Hum Genet* **120:** 301–333.

Cochrane JC, Lipchock SV, Strobel SA. 2007. Structural investigation of the GlmS ribozyme bound to its catalytic cofactor. *Chem Biol* **14:** 97–105.

Darty K, Denise A, Ponty Y. 2009. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25:** 1974–1975.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Eddy SR, Durbin R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res* **22:** 2079–2088.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Erlitzki R, Long JC, Theil EC. 2002. Multiple, conserved iron-responsive elements in the 3′-untranslated region of transferrin receptor mRNA enhance binding of iron regulatory protein 2. *J Biol Chem* **277:** 42579–42587.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37:** D136–D140.

Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8:** 113–126.

Genome 10K Community of Scientists. 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J Hered* **100:** 659–674.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154–D158.

Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. 2010. RNAz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput* **15:** 69–79.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Hampel KJ, Tinsley MM. 2006. Evidence for preorganization of the glmS ribozyme ligand binding pocket. *Biochemistry* **45:** 7861–7871.

Han J, Pedersen JS, Kwon SC, Belair CD, Kim YK, Yeom KH, Yang WY, Haussler D, Blelloch R, Kim VN. 2009. Posttranscriptional crossregulation between Drosha and DGCR8. *Cell* **136:** 75–84.

Hartuv E, Shamir R. 2000. A clustering algorithm based on graph connectivity. *Inf Process Lett* **76:** 175–181.

Havgaard JH, Lyngso RB, Gorodkin J. 2005. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res* **33:** W650–W653.

Hel Z, Di Marco S, Radzioch D. 1998. Characterization of the RNA binding proteins forming complexes with a novel putative regulatory region in the 3′-UTR of TNF-α mRNA. *Nucleic Acids Res* **26:** 2803–2812.

Hofacker I, Fontana W, Stadler P, Bonhoeffer L, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh Chem* **125:** 167–188.

Howard MT, Moyle MW, Aggarwal G, Carlson BA, Anderson CB. 2007. A recoding element that stimulates decoding of UGA codons by Sec tRNA[Ser]Sec. *RNA* **13:** 912–920.

Jacobs GH, Chen A, Stevens SG, Stockwell PA, Black MA, Tate WP, Brown CM. 2009. Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res* **37:** D72–D76.

Jepson JE, Reenan RA. 2008. RNA editing in regulating gene expression in the brain. *Biochim Biophys Acta* **1779:** 459–470.

Keene JD. 2007. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8:** 533–543.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci* **100:** 11484–11489.

Khaladkar M, Liu J, Wen D, Wang JT, Tian B. 2008. Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment. *BMC Genomics* **9:** 189. doi: 10.1186/1471-2164-9-189.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci* **106:** 11667–11672.

Klein DJ, Ferre-D'Amare AR. 2006. Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science* **313:** 1752–1756.

Korb M, Rust AG, Thorsson V, Battail C, Li B, Hwang D, Kennedy KA, Roach JC, Rosenberger CM, Gilchrist M, et al. 2008. The Innate Immune Database (IIDB). *BMC Immunol* **9:** 7. doi: 10.1186/1471-2172-9-7.

Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34:** D158–D162.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210–1213.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. Evolutionary constraint in the human genome based on 29 eutherian mammals. *Nature* **477.** doi: 10.1038/nature10530.

Lomeli H, Mosbacher J, Melcher T, Hoger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* **266:** 1709–1713.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955–964.

Lundblad EW, Altman S. 2010. Inhibition of gene expression by RNase P. *New Biotechnol* **27:** 212–221.

Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, Dubois A, Sanglier-Cianferani S, Van Dorsselaer A, Clerc P, Avner P, et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8:** e1000276. doi: 10.1371/journal.pbio.1000276.

Malquori L, Carsetti L, Ruberti G. 2008. The 3′ UTR of the human CTLA4 mRNA can regulate mRNA stability and translational efficiency. *Biochim Biophys Acta* **1779:** 60–65.

Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. 2003. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113:** 577–586.

Martinez-Chantar ML, Latasa MU, Varela-Rey M, Lu SC, Garcia-Trevijano ER, Mato JM, Avila MA. 2003. L-Methionine availability regulates expression of the methionine adenosyltransferase 2A gene in human hepatocarcinoma cells: Role of S-adenosylmethionine. *J Biol Chem* **278:** 19885–19890.

Mathews DH, Turner DH. 2002. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* **317:** 191–203.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10:** 155–159.

Montange RK, Batey RT. 2008. Riboswitches: Emerging themes in RNA structure and function. *Annu Rev Biophys* **37:** 117–133.

Moore MJ. 2005. From birth to death: The complex lives of eukaryotic mRNAs. *Science* **309:** 1514–1518.

Namy O, Rousset JP, Napthine S, Brierley I. 2004. Reprogrammed genetic decoding in cellular gene expression. *Mol Cell* **13:** 157–168.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25:** 1335–1337.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2:** e33. doi: 10.1371/journal.pcbi.0020033.

Pohl AA, Sugnet CW, Clark TA, Smith K, Fujita PA, Cline MS. 2009. Affy exon tissues: exon levels in normal tissues in human, mouse and rat. *Bioinformatics* **25:** 2442–2443.

Putland RA, Sassinis TA, Harvey JS, Diamond P, Coles LS, Brown CY, Goodall GJ. 2002. RNA destabilization by the granulocyte colony-stimulating factor stem–loop destabilizing element involves a single stem–loop that promotes deadenylation. *Mol Cell Biol* **22:** 1664–1673.

Rabani M, Kertesz M, Segal E. 2008. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci* **105:** 14885–14890.

Ray PS, Jia J, Yao P, Majumder M, Hatzoglou M, Fox PL. 2009. A stress-responsive RNA switch regulates VEGFA expression. *Nature* **457:** 915–919.

Regulski EE, Breaker RR. 2008. In-line probing analysis of riboswitches. *Methods Mol Biol* **419:** 53–67.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311–1323.

Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2:** 8. doi: 10.1186/1471-2105-2-8.

Roth A, Breaker RR. 2009. The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem* **78:** 305–334.

Roth A, Nahvi A, Lee M, Jona I, Breaker RR. 2006. Characteristics of the glmS ribozyme suggest only structural roles for divalent metal ions. *RNA* **12:** 607–619.

Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399:** 75–80.

Siepel A, Haussler D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* **11:** 413–428.

Stefanovic B, Brenner DA. 2003. 5′ stem–loop of collagen α1(I) mRNA inhibits translation in vitro but is required for triple helical collagen synthesis in vivo. *J Biol Chem* **278:** 927–933.

Stoecklin G, Anderson P. 2006. Posttranscriptional mechanisms regulating the inflammatory response. *Adv Immunol* **89:** 1–37.

Stoecklin G, Lu M, Rattenbacher B, Moroni C. 2003. A constitutive decay element promotes tumor necrosis factor alpha mRNA degradation via an AU-rich element-independent pathway. *Mol Cell Biol* **23:** 3506–3515.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. 2009. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19:** 347–359.

Svoboda P, Di Cara A. 2006. Hairpin RNA: a secondary structure of primary importance. *Cell Mol Life Sci* **63:** 901–908.

Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J. 2008. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res* **18:** 242–251.

Tseng HH, Weinberg Z, Gore J, Breaker RR, Ruzzo WL. 2009. Finding non-coding RNAs through genome-scale clustering. *J Bioinform Comput Biol* **7:** 373–388.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8:** e1000371. doi: 10.1371/journal.pbio.1000371.

Wang JX, Breaker RR. 2008. Riboswitches that sense *S*-adenosylmethionine and *S*-adenosylhomocysteine. *Biochem Cell Biol* **86:** 157–168.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* **102:** 2454–2459.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17:** 852–864.

Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, et al. 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res* **35:** 4809–4819.

Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR. 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol* **11:** R31. doi: 10.1186/gb-2010-11-3-r31.

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3:** e65. doi: 10.1371/journal.pcbi.0030065.

Wilusz JE, Spector DL. 2010. An unexpected ending: Noncanonical 3′ end processing mechanisms. *RNA* **16:** 259–266.

Wilusz JE, Freier SM, Spector DL. 2008. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135:** 919–932.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.

Xu N, Chen CY, Shyu AB. 1997. Modulation of the fate of cytoplasmic mRNA by AU-rich elements: Key sequence features controlling mRNA deadenylation and decay. *Mol Cell Biol* **17:** 4611–4621.

Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, Tompa M, Ruzzo WL. 2007. A computational pipeline for high-throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. *PLoS Comput Biol* **3:** e126. doi: 10.1371/journal.pcbi.0030126.

Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322:** 750–756.