Article ID: WMC001600 2046-1690

RNA Structures Affected By Single Nucleotide Polymorphisms In Transcribed Regions Of The Human Genome

Corresponding Author:
Dr. Andrew D Johnson,
Research Fellow, National Heart, Lung and Blood Institute - United States of America

Submitting Author:
Dr. Andrew D Johnson,
Research Fellow, National Heart, Lung and Blood Institute - United States of America

Article ID: WMC001600

Article Type: Research articles

Submitted on:24-Feb-2011, 03:23:19 PM GMT **Published on:** 25-Feb-2011, 10:19:15 PM GMT

Article URL: http://www.webmedcentral.com/article_view/1600

Subject Categories:BIOINFORMATICS

Keywords:SNP, Polymorphism, RNA, Structure, Gene Expression, Functional, GWAS, Genome, Human, Disease

How to cite the article: Johnson A D, Trumbower H , Sadee W . RNA Structures Affected By Single Nucleotide Polymorphisms In Transcribed Regions Of The Human Genome . WebmedCentral BIOINFORMATICS 2011;2(2):WMC001600

Source(s) of Funding:

Andrew Johnson was supported by an OSU Distinguished University Fellowship, American Heart Association Predoctoral Fellowship (AHA 0515157B), and National, Heart, Lung and Blood Institute Postdoctoral IRTA Fellowship and Lenfant Biomedical Research Fellowship Award. Wolfgang Sadee is supported by U01 GM092655. Heather Trumbower was funded by the National Human Genome Research Institute.

Additional Files:

Additional File 1

RNA Structures Affected By Single Nucleotide Polymorphisms In Transcribed Regions Of The Human Genome

Author(s): Johnson A D, Trumbower H , Sadee W

Abstract

Single-stranded RNAs fold into base-paired structures sometimes critical for RNA function.We report the first genome-wide computational analysis of mRNA structures, obtaining predicted minimum free energy structures (MFE) and ensembles of top-scoring structures.Evaluation of mRNA structures from >12,450 genes indicates that thermodynamically favorable structures preferentially form within specific regions of some human mRNAs.High energy and conserved structures occur more often than expected within untranslated regions, and in coding exons in proximity to translation start sites and in-frame methionine codons.Several genetic variants within these structures are already known to contribute to human diseases.To investigate the impact of polymorphisms on RNA structures more broadly, we examined the effects of 153,397 single nucleotide polymorphisms (SNPs) in exons of the human genome on RNA structures, including splice variants.The majority of human exonic SNPs are predicted to alter RNA structure to some degree (~60% affect MFEs and >90% the ensembles). Predicted secondary structural effects are localized to SNP regions within ~50 base-pairs. We find structural (epistatic) interactions between SNP bases within some haplotypes suggesting possible covariation preserving specific structures.Further analyses of SNPs within known functional RNA structures (IREs, SECIS, miRNAs, snoRNAs) identify a number of putative functional SNPs, many within genes already associated with human phenotypes, and suggest that conformational RNA polymorphisms substantially contribute to human phenotypic variability. Comparison of SNP-RNA structures against published genome-wide association (GWAS) results indicates that some strong genetic signals from GWAS (e.g., *LPL, AGER*) may exhibit their functional effects through modified RNA structures that could impact RNA stability, alternative RNA or protein processing, or translational efficiency. Our genome-wide evaluation of RNA structures, combined with the structural effects predicted for human SNPs, yields information useful for structure-function studies on RNAs and genetic variants.

Introduction

We provide here a genome-wide analysis of the distribution of RNA transcript folding structures and predicted structural effects of single nucleotide polymorphisms (SNPs). Algorithms for predicting RNA secondary structure have seen wide application mostly with a focus on functional RNAs (tRNA and rRNAs). Recent studies support the notion that SNPs in transcribed regions of protein coding genes also affect structure and function (Illustration 1). Comparisons of secondary structures predicted from mRNAs versus shuffled sequences in a variety of genomes indicate genome-wide selection for secondary structure, particularly in eubacterial organisms and one eukaryote, Saccharomyces cerevisiae (Katz and Burge 2003). Conclusions regarding secondary structure in human mRNAs have been mixed (Clote et al. 2005; Katz and Burge 2003; Meyer and Miklos 2005; Seffens and Digby 1999; Workman and Krogh 1999).

Regardless of genome level bias for or against mRNA secondary structure there is good evidence of functional roles for structure within specific human mRNAs. The effect of structure in 5' untranslated regions (UTR) in both inhibiting and promoting translation has been documented in mammalian cells (Babendure et al. 2006; Kozak 1986; Kozak 1990; Kozak 2005). Specific structural motifs, selenocysteine insertion sequence elements (SECIS), are known to function in selenocysteine incorporation in human coding regions (Kryukov et al. 2003). Iron-responsive elements (IREs) in both 5' and 3'UTRs are known to regulate translation of specific genes (Gray and Hentze 1994). Indeed, evidence is mounting that structure in untranslated and coding exonic regions of human mRNAs and small non-coding RNAs may affect the rate of transcription, their processing (e.g., splicing, polyadenylation, editing (Athanasiadis et al. 2004)), hybridization (Kuo et al. 1997; Vickers et al. 2000), decay (Fialcowitz et al. 2005; Lopez de Silanes et al. 2005; Lopez de Silanes et al. 2004), transport, targeting and initiation, and rate of translation (Kozak

2005; Martineau et al. 2004; Russcher et al. 2005b). Throughout their life cycle, mRNAs interact with many proteins. These interactions may both limit the RNA conformations assumed in vivo (Zhang et al. 2006) and be influenced by the nascent RNA structure (Moore 2005). Pre-mRNA structure may also have functional relevance (Buratti and Baralle 2004; Meyer and Miklos 2005). Similarly, the importance of structure in eukaryotic small RNAs such as miRNAs has been realized (Bentwich et al. 2005; Bonnet et al. 2004). Functional RNA secondary structural motifs are typically small in size and fairly well-predicted in shorter sequences (Mathews et al. 1999). Ribonucleic acids fluctuate between different energetically favorable configurations due to stochastic molecular motion and other constraints, though they most likely favor the thermodynamic minimum free energy (MFE) structure over an ensemble of suboptimal structures within a free energy range. Given an RNA sequence, secondary structure prediction programs can generate both an MFE structure and an ensemble of suboptimal structures. Analysis of the MFE structure and the ensemble of suboptimal structures can provide evidence for well-determined structures likely to form in vivo (Mathews et al. 1999).

An additional dimension not often considered is how functional RNA structures are influenced by sequence variation. One method for detection of genetic variants, single-strand conformation polymorphism (SSCP), relies on differences in the structural conformation of variants in single-stranded DNA or RNA (Lenz et al. 1995; Ren 2000). A large portion of variants in exonic sequences are detectable by changes in RNA structure via RNA-SSCP, showing that most SNPs have the potential to affect RNA structure (Lenz et al. 1995; Sarkar et al. 1992). Experiments employing enzymes that specifically cleave paired and unpaired bases to create structural maps show that prokaryote and human sequences differing by a single SNP can have different mRNA secondary structures (Shen et al. 1999). On the other hand, sequence variations may often be neutral with regards to effects on structure, in many cases preserving an evolved functional state (Ancel and Fontana 2000). However, even variations that are neutral in their effect on the predicted MFE structure may alter the characteristics of suboptimal structures within the ensemble of conformations, the relative time an mRNA spends in the MFE state, or create a sequence that is likely to change the structure if additional variation occurs (Ancel and Fontana 2000). Variants that alter structure may be further compensated for by additional variants that epistatically preserve the ground state structure (Chen et al. 1999). On the other hand, any single nucleotide change can be viewed as enabling the formation of new states and functions subject to evolutionary constraints or selection (Draghi et al. 2010).

A deeper understanding of how population variants influence RNA structure may help explain inter-individual and inter-species differences in gene expression and function. Studies in humans have supported the idea that some inter-individual genetic differences alter RNA structures and affect RNA functions, in some cases contributing to disease (Illustration 1, (Florentz and Sissler 2001; Johnson et al. 2005; Wang et al. 2005; Zhang et al. 2005)). Vilmi et al. resequenced 22 tRNA genes in the mitochondrial genomes of 477 Finns, and examined 435 European tRNA sequences from the MitoKor database. They found that MFE structures predicted among the 96 polymorphic tRNA sequences showed a significantly different distribution than wild-type tRNAs, with low frequency alleles yielding the greatest predicted change in MFE (Vilmi et al. 2005). Pathological evidence also indicates that SNPs disrupting IRE structure in the 5'UTR of FTL are a genetic cause of hereditary hyperferritinemia-cataract syndrome (HHCS) (Aguilar Martinez et al. 1997; Allerson et al. 1999; Beaumont et al. 1995; Camaschella et al. 2000; Campagnoli et al. 2002; Cazzola et al. 1997; Girelli et al. 1995; Martin et al. 1998; McLeod et al. 2002; Mumford et al. 1998). Clinical cases of rigid spine syndrome have been attributed to SNPs in the structure-encoding sequence of the SECIS-containing 3'UTR of SEPN1 (Allamand et al. 2006; Moghadaszadeh et al. 2001). Thus, we hypothesize that synonymous, nonsynonymous and UTR variants can potentially act in mildly deleterious and, in some cases, pathological fashion on pre- and post-translational levels through changes in RNA structure.

Bioinformatics databases and tools have been developed to predict the potential functional effects of genetic differences, particularly for nonsynonymous SNPs that alter amino acid coding or those that fall near splicing borders and predicted protein-DNA binding sites (for reviews see (Johnson et al. 2005, Johnson 2009)). However, computational analysis of variants altering human RNA structures has typically been investigated only for single genes or variants following experimental observations. Here we present the first report of the predicted effects of known variants on RNA structure in a large portion of the human genome. We used the Vienna RNA secondary structure program (Hofacker 2003) to create a human genome SNP-structural conformation dataset for analysis. We determine how frequently variants are predicted to alter RNA structures, whether allele

frequencies associate with structural differences, and if proposed functional variants from the literature differ from the normal distribution of variance. We analyze structure results with multiple analytical approaches and report a set of SNPs predicted to potentially alter RNA processing or expression via changes in RNA structure. We also present a number of challenges and solutions specific to genome-wide surveys of SNPs and consideration of RNA sequence contexts. Aside from genetic variation, this is also one of the largest examinations of mRNA structure in the human genome (about 17,900 transcriptional units represented) as previous reports examined 12 (Meyer and Miklos 2005), 1,855 (Katz and Burge 2003), and an overlapping set of 41 (Clote et al. 2005), 46 (Workman and Krogh 1999), and 51 (Seffens and Digby 1999) human mRNAs. Results of our analysis of human mRNA structures indicate there is considerable potential for favorable structures to form within specific regions of many mRNAs.

Results

Computational analysis predicts SNPs as a common cause of conformational variation in human

RNA secondary structure

We predicted mRNA structures derived from regions surrounding 153,397 SNPs in ~17,900 Refseq genes. We filtered this set to reduce sources of ambiguity (see Methods) yielding structures and thermodynamic MFEs (ΔG) around 34,557 SNPs in ~12,450 Refseq genes. This approach allows both a general analysis of mRNA structures in a genome-wide set of RNAs, and an examination of changes in structures due to polymorphism (e.g., by analysis of changes in energies: ΔΔG). For all Figures and Tables where ΔG (thermodynamic minimum free energy in kcal/mol) and ΔΔG (change in thermodynamic minimum free energy from SNP major to minor allele) are reported, negative values correspond to lower free energies (more thermodynamically favorable). The thermodynamic distributions of RNA structures (ΔG) around SNPs across the genome are displayed in Illustration 2 (panel A). Notably wide ranges of ΔG values are observed in all SNP contexts. The distributions are slightly right-shifted toward less favorable structures, in particular in the case of UTR structures. The left tails (more favorable structures) are considerably longer than the right and at the extreme they are above a calculated normal distribution. This is in part due to the zero energy barrier, but may also indicate potential selection for a small proportion of highly favorable structures.

SNP contexts differ significantly in the distribution of Δ G values, ranking from most favorable to least favorable: synonymous, nonsynonymous, and UTR (p

To analyze the effect of SNPs on structure, we used the bracket notation for structures (e.g., Illustration 5, panel A) to compare structures between major and minor alleles for 34,557 validated, biallelic SNPs. We find that the majority of these human genetic differences alter the MFE RNA structure as well as the profile of the ensemble of top-rated suboptimal RNA structures. Some minor alleles are predicted to have identical MFE structures (34.1%), while few are predicted to have near identical structure ensembles (6.4%). The predicted changes in thermodynamic energy values for minor alleles relative to major alleles (ΔΔG) are depicted in Illustration 2 (panel B).

Analysis of SNPs predicted to change mRNA structures

We further characterized the relationship between sequence variation and RNA structure in this genomic dataset of SNPs that predicted altered structures (n=22,785). The greatest frequency of structure change is predicted for transversion SNPs involving guanine exchanges. Other transversions are the next most frequent, followed by transitions (Illustration 3, panel B). Mean results for structure predictions are displayed in Illustration 4. Synonymous major and minor allele structures have on average slightly more negative ΔG values and larger increases in ΔΔG with minor alleles (less favorable change) than nonsynonymous and UTR SNPs. In keeping with this, synonymous structure contexts generally have a greater percentage of bases involved in helical pairs, thus tending to form more favorable structures that would be more likely to be affected by a SNP. We analyzed the predicted structural behavior for each base position of each structure pair. There are four possible patterns of behavior at each base position of each structure (major>minor allele): unpaired>unpaired, unpaired>paired, paired>unpaired, paired>paired. This analysis reveals that changes from major to minor alleles favor disruption of helical pairing interactions over the creation of new helices, regardless of SNP category (Illustration 4).

Examination of the frequency of predicted structural changes at surrounding bases relative to the SNP base in the sequence reveals a strong central tendency in change despite varying flanking sequence sizes (Illustration 6), showing that alterations in secondary structure interactions typically localize to nearby regions. The relationship between SNP minor allele frequency (MAF), ΔΔG, and change in structure pairing behavior was analyzed by ANOVA at different

MAF cutoffs (similar to the tRNA allele analysis in (Vilmi et al. 2005)). There is an overall trend toward greater change in structure and larger ΔΔG among SNPs with lower MAFs (below 10% MAF vs. above 10% MAF; ΔΔG, $p < 0.01$, structure pairing change, p=0.052, Bonferroni correction).

Discrimination of putative functional mRNA variants

We used different analyses to identify putative functional mRNA variants including screening of sequences and structures with evidence for evolutionary conservation and proximity to likely functional sites such as untranslated regions. In all cases we searched for favorable structures (low ΔG) with SNPs predicted to create large changes (extremely high or low ΔΔG). All of the SNPs isolated by these approaches were visualized in their genome context via the UCSC Genome Browser to ensure accurate annotation of their position within the gene context, CpG island predictions, proximity to alternative processing events, and observation of nearby poly-repeats and inframe methionine codons.

We examined the most stable structures predicted in the single SNP human genome structure-variation dataset (Z-score >3; -52.0 kcal/mol cutoff for MFE structures). This set contains 152 variants (36 nonsynonymous, 40 synonymous, 48 5'UTR, 28 3'UTR) (Additional File 1, Tab 1). Many structures with long helices are observed, up to a maximum helical stretch of 25 bp in the most favorable structure found flanking the translation start site of NDST1 (rs3733935 G allele: -69.0 kcal/mol). A number of trends are observed among the gene regions in this set. Structures in (48) and near (38) 5'UTRs are over-represented relative to their presence within the full set. Of the synonymous and nonsynonymous structures half are located in 5' exons within 100 bp of translation start sites, and 40% are in close proximity within the mRNA to one or more downstream inframe methionine codons. At least eleven of the 152 variants are in or near methionine codons known to be sites of alternative translation. More than two-thirds of the structures (68%) are within predicted CpG islands (Gardiner-Garden and Frommer 1987). In three cases, extended exonic poly-repeats are observed near strongly predicted structures (rs367398, Leu[11] in NOTCH4; rs1799925, Pro[9] in WT1; rs3021525, Ala[16] in FOXE1). No 3'UTRs shorter than 300 bp are observed, and the average 3'UTR length is greater than 1 kb. One quarter of the predicted stable structures (38) are located in exonic regions but not in close proximity to either translation start or stop regions.

In a separate analysis, cutoffs for Boltzmann probabilities were used to identify 512 variants that have a set of well-predicted structures (Z-score >3; 0.145 cutoff). We hypothesized that structures with high Boltzmann probabilities might contain biologically functional structures because of their relatively confined and similar set of suboptimal structures (Miklos et al. 2005). We further screened these structures for those with 1) large ΔΔG due to the SNP, 2) a high degree of sequence conservation and regulatory potential based on UCSC genome annotation, and 3) low (favorable) ΔG. The resulting set of 129 candidate variants (38 nonsynonymous, 43 synonymous, 14 5'UTR, 34 3'UTR) predicted to substantially alter a favorable structure is found in Additional File 1, Tab 2. In some cases we note that the candidates coincided with EvoFold predictions indicating the structures, and not simply the sequences, are likely preserved across mammals (Pedersen et al. 2006).

Based on the success of EvoFold in predicting some functional RNA structures (Pedersen et al. 2006), we used UCSC Table Browser to create a merged set containing EvoFold predictions in mRNAs and known SNPs (n=936). There is a distinctive distribution in the location of these structures in the mRNAs to: coding exons (70.5%), 3'UTRs (27.6%), and 5'UTRs (1.8%). We find exonic EvoFold structures in this set overlap inframe methionine codons in 42% of cases, which is 1.84 fold greater than expected based on their sequence length and the calculated human RefSeq protein frequency of methionine codons (2.14%). Some of these form stem-loop structures with the loop between two methionine codons that overlap in the stem. Instances of two consecutive methionine codons occur within this structure set 3.34 fold greater than expected assuming independent assortment of codons. A strikingly large portion of the 3'UTR structures in this set are located at the extreme distal end of the transcript (48.3%). We also note that some gene categories are extensively represented within the EvoFold-SNP structure set: histones, ribosomal proteins, ribonucleoprotein-related, translation initiation factor-related, vesicular-related, ubiqutin-related, myosins, and neurotransmitter receptors. A list of selected candidates from the EvoFold-SNP analysis is presented in Additional File 1, Tab 3.

SNPs in structures near alternative mRNA processing sites

Our initial analysis of the 50 bp flanking structure set reveals that inclusion of sequences of total length less than 100 bp exaggerated differences between the UTR and coding variant sequence contexts, and in particular resulted in higher (less favorable) average Δ G and higher average Boltzmann probabilities for the

UTR category (data not shown). Correspondingly, SNPs in the UTR category have the greatest percentage of sequences with total length less than 100 bp due to their proximity to one end of the transcript (UTR – 7.3%, nonsynonymous – 2.0%, synonymous – 1.3%). These shorter sequences form less secondary structure because of their diminished length, and the smaller number of potential structural interactions yield greater mean Boltzmann probabilities.

We hypothesized that this set of shorter length sequences (n=1,203) would include many in close proximity to 5' transcription and translation initiation sites, and that those sequences predicted to form stable structures might influence the initiation or rate of transcription or translation. Additionally, we hypothesized that the set of SNPs initially removed because of multiple sequence contexts (n=634) would include many genes exhibiting alternative transcription or translation start sites, splicing, termination, or poly-adenylation. The SNP contexts for shorter length sequences and multiple sequence contexts (n=1,653 due to overlap) were visualized in UCSC Genome Browser to collect further annotation. Many of the SNPs in multiple sequence contexts (n=634) map to multiple sites in the genome or are in regions where transcripts are derived from both strands. However, a portion of these SNPs are found near sites of known or putative alternative gene processing (n=196).

The majority of SNPs (53.4%) in the shorter sequence length set are located at the 5' end of transcripts, in contrast to our estimates for the full dataset where 3'UTR SNPs (80%) far outnumber 5'UTR SNPs (20%). This is consistent with prior observations that 3'UTRs are generally longer than 5'UTRs in eukaryotes (and thus likely to accumulate relatively more variation). Filtering based on sequence length and multiple sequence contexts effectively enriches for SNPs at the 5' end of transcripts that are near transcription and translation start sites. This group of SNPs (n=196) was examined to find candidates with 1) a large $\Delta\Delta G$ predicted between alleles (>2 kcal/mol), and 2) low ΔG (one allele more favorable than -27.0 kcal/mol) and/or long helical structures. Boltzmann probability was not considered as this was observed to be influenced by sequence length. A list of candidate SNPs and structures near alternative processing sites is displayed in Additional File 1, Tab 4.

RNA structures containing interacting SNP bases

Analysis of RNA structures generated from sequences containing two SNPs identified 568 pairs where the two SNP bases are predicted to structurally interact in 1 or more haplotype, indicating potential multi-allelic interactions. Only 34 of the 568 SNP pairs were further examined since in these cases both SNPs were annotated as validated by dbSNP. Among these we find SNP pairs interacting in only one haplotype (n=21), those following the canonical GC-AU substitution pattern often observed across species (n=8), and other patterns: GU-UA (n=1), GC-CG (n=1), AU-UA (n=1) and GC>U (n=2). In some cases (CIT, TNFAIP2, BCKDHB) the allele frequencies of structurally interacting alleles complemented the substitution pattern expected to preserve RNA structures. Notably these three genes have been associated with human disease phenotypes, and the regions containing the predicted structures are more conserved across species than surrounding sequences. The structure predicted in the 3'UTR of LPL is notable since a haplotype containing two alleles in this precise location is associated with increased enzyme activity and metabolic phenotypes among a Mexican-American cohort (Goodarzi et al. 2005). Further analysis of structurally interacting alleles may be useful in finding functional structures and indicate possible instances of covariation within human populations to form or preserve RNA structure. However, estimation of the phased multi-SNP haplotypes that truly exist in human populations is inexact, making extensive computational analysis problematic. In this study, analysis was mostly done for single validated SNPs, where the two haplotypes are known and phasing is not an issue. SNP pairs displaying predicted RNA structure interactions are presented in Additional File 1, Tab 5.

Polymorphisms in known functional RNA structures

Based on previously described functional structures (see Methods for database sources) we analyzed SNPs and multi-nucleotide variants occurring in or near described IRE, SECIS, miRNA and snoRNA structures. This examination reveals that most contain no variation or variation that is predicted to be neutral in structural effect, but a few structures harbor validated variants that may affect biological function.

Sequence from Allerson et al. was used to identify the IRE structural element in the 5'UTR of FTL (L-ferritin) (Allerson et al. 1999). None of the numerous pathological variants reported in the literature within the FTL IRE were found in dbSNP build 124, consistent with a recent report that many medically relevant, CLIA-tested variants are not mapped in current databases such as dbSNP (Johnson et al., 2010). Thus, these variants were not available for analysis in our results set (Aguilar Martinez et al. 1997; Allerson et al. 1999; Beaumont et al. 1995; Camaschella et al. 2000; Campagnoli et al. 2002; Cazzola et al. 1997; Girelli et al. 1995; Martin et al. 1998; McLeod et al. 2002; Mumford et al. 1998). A

multi-nucleotide variant (rs11553230) was noted in the IRE of FTL but this variant has not been validated. A SNP reported in the FTH1 IRE was also absent from dbSNP (Kato et al. 2001). Known IRE structures in FTH1, ALAS2, ACO2, SLC40A1, TFRC, SLC11A2, CDC14A and CDC42BPA (Cmejla et al. 2006) were examined but no variants were observed. The comparative genomics program EvoFold (Pedersen et al. 2006) effectively predicted many but not all of the known IREs. BLAT search (Kent 2002) for many of the IRE encoding sequences revealed similarity in retroposed regions, pseudogenes and intronic and intergenic regions throughout the human genome, but no additional putative IREs within known protein-coding mRNAs.

The regions encoding the known functional stem-loop structures controlling selenocysteine incorporation (SECIS) were examined in 25 human genes for nearby variants (Kryukov et al. 2003). A total of 26 elements were examined since the SELP 3'UTR contains two elements. Eleven of the SECIS elements contained no known polymorphisms in dbSNP build 124. The previously described pathological variants in SEPN1 were not found in dbSNP (Allamand et al. 2006; Moghadaszadeh et al. 2001). Four SECIS elements (GPX1, C11orf31 (SELH), SELK, SELO) harbored variants in dbSNP but these were excluded from consideration because they are computationally predicted and not validated in human populations.

Potential novel variants were discovered in BLAT alignments against the human genome for one SELP SECIS and the DIO2 SECIS. A single base (A>U) difference was noted in SELP from the sequence published in Kryukov et al (Kryukov et al. 2003). Differences corresponding to the deletion of three single cytosines were noted in the alignment of the DIO2 SECIS to the genome. Ten of the SECIS elements contained variants that have been validated in human populations. Because selenocysteine structures involve considerable non-Watson-Crick base-pairing they are only partially predicted by standard settings for secondary structure algorithms like Vienna. Thus, we used SECISearch 2.19 to predict structures for all alleles (Kryukov et al. 2003).

All wild type allele sequences predict the functional SECIS structural elements as expected: Helix I – Internal loop – UGAN SECIS quartet core – Helix II – Apical Loop. Some variant alleles described below are predicted to alter SECIS structures and may affect the incorporation of selenocysteine or translational readthrough in these proteins. The putative single base change in one SELP SECIS is an A>U shift 2 bp upstream of the critical quartet and is predicted to close the internal loop via an U-A interaction. The putative cytosine indels in DIO2 are predicted to affect structure in the Helix I and apical loop areas of that element. SNPs in SEP15 (15kDA) (rs5859) and TXNRD3 (rs14682) are predicted to alter structure in the apical loop region of those SECIS elements. Notably, previous functional studies of SEP15 highlight a link between rs5859 and another SNP (rs5845) in the 3'UTR suggesting a link to cancer etiology (Hu et al. 2001; Kumaraswamy et al. 2000). A previously studied SNP in GPX4 (rs713041) is predicted to significantly alter structure in the Helix I region (Villette et al. 2002), and a validated SNP in SEPX1 (rs4987018) is found 2 bp downstream of the SECIS quartet and predicted to disrupt a portion of Helix II. A SNP in TXNRD2 (rs1044732) is predicted to weaken an interaction in Helix I. A number of variants are not predicted to appreciably affect SECIS elements: an indel in the second SELP still formed a stable Helix I (rs10569610, -/AGUA), and SNPs in GPX3 (rs4661), SELI (rs7588538) and DIO1 (rs12095080) are distanced enough that they are not predicted to affect the core SECIS structure.

Fourteen of the SECIS-containing human genes include one or more additional inframe UGA codon located in considerably distant upstream locations. In bacterium, structure immediately downstream of UGA codons is critical to selenocysteine incorporation, while in eukaryotes the SECIS elements (e.g., those above) are located, often distantly, in the 3'UTR (Berry et al. 1991). However, it has been noted that stable helical structures are located downstream of the human inframe UGA codons in SECIS-containing genes, and these structures may stabilize readthrough (Kryukov et al. 2003). We observe that EvoFold (Pedersen et al. 2006) predicts helices conserved across mammals downstream of inframe upstream UGA codons in SEPN1 (Exon 10), SELT (Exon 2), and SELK (Exon 4). Moreover, experimental mutagenesis of the helix immediately downstream of the SEPN1 upstream UGA codon shows this structure does facilitate selenocysteine readthrough (Howard et al. 2005). Thus, we further examined structures near inframe UGA codons in SECIS encoding human genes for variation. Only three variants are noted in regions near inframe UGA codons: rs11552989 (not validated, 1 bp downstream of the C11orf31 (SELH) exon 2 UGA), rs6440687 (4 bp upstream of the SELT exon 2 UGA), and rs2272853 (33 bp downstream of SELO). In our structural results database stable helices are predicted downstream of the SELO and C11orf31 UGA codons (-34.8 kcal/mol and -36.4 kcal/mol, respectively), and the SNPs are not predicted to disrupt the helices.

Of 21 variants found in human pre-microRNA (pre-miRNA) regions, twelve seem unlikely to disrupt

structure and processing of a miRNA because they preserve Watson-Crick interactions and are not located near mature miRNA sequences. The remaining 9 SNPs are distributed as follows: disruptive of a helical interaction and located in the mature miRNA (hsa-mir-520c; rs7255628, hsa-mir-125a; rs12975333), creating a novel helical interaction and located complementary to the mature miRNA (hsa-mir-146a; rs2910164), disruptive of a helical interaction within 10 bp of the mature miRNA (hsa-mir-521-2; rs13382089, hsa-mir-140; rs7205289, hsa-mir-27a; rs11671784, hsa-mir-516-3; rs10583889), disruptive of a helical interaction greater than 10 bp from the mature miRNA (hsa-mir-492; rs2289030), and a multi-nucleotide insertion in a hairpin loop (hsa-mir-516-3; rs10670323). Notably only 2 of these 9 potentially disruptive miRNA variants are validated in multiple human populations (rs2289030, rs2910164). None of the variants analyzed here overlap with a similar analysis conducted on miRNA variants in a Japanese population (Iwai and Naraba 2005).

Small nucleolar RNAs (snoRNAs) characterized with RNA structure motifs as C/D box snoRNPs and H/ACA box snoRNPs, and Cajal body-specific RNAs (scaRNAs) were examined to identify variants in their functional domains. The majority of the 42 variants analyzed in these RNAs are located greater than 10 bp away from known functional domains, or are not expected to significantly disrupt Watson-Crick pairs or RNA targeting interactions. However, three variants are potentially disruptive to functional domains: one in the D box of HBII-52-32 adjacent to a 5HT-2C complementary domain (rs12910266) (Cavaille et al. 2000), one within SNORD1B that is predicted to be involved in 2'O-ribose methylation of 28S rRNA (rs16969028), and one within SNORA44 in the stem of a pseudoknot predicted to guide the pseudouridylation of 18S rRNA (rs16837624) (Kiss et al. 2004). Of these three potentially disruptive SNPs, two are known to be valid in multiple human populations (rs16969028, rs16837624). Results of variants in known functional RNA structures are summarized in Additional File 1, Tab 6.

RNA structure-SNPs in strong linkage disequilibrium with GWAS SNPs

We queried the SNPs affecting known or putative functional RNA structures (Additional File 1, Tabs 1-6) and their strong LD proxies (r2>0.9), and compared them against a catalog of GWAS results searching for SNPs in common that have strong genetic associations (defined as pAGER in an exon which may also serve as a 5'UTR for an alternative protein isoform based on current genome annotations (Additional File 1, Tab 4). This SNP was strongly associated with lung volume traits in 2 large, distinct GWAS (Repapi et al., 2010, pLPL 3'UTR, are also found among strong GWAS results for lipid levels (e.g., Kathiresan et al., 2008, rs328, p2=0.93). The SNP rs1059611 is predicted to be found within a structurally interacting haplotype in LPL (Additional File 1, Tab 5). Notably rs328 is a nonsense variant (S447X) and may account for the mechanism explaining observed associations with cholesterol, or it may be a combination of multiple functional variants that are in high linkage disequilibrium. All of the GWAS-RNA structure SNP overlaps found with r2>0.9 are described in Additional File 1, Tab 7.

Discussion

The complex and varied interactions an mRNA engages in from genesis to processing, transport, translation, and recycling provide many regulatory setpoints. Throughout its life cycle there is potential for any RNA to form physicochemical structural interactions. Whether or not an RNA forms and maintains particular structures is influenced by the sequence of the RNA, as well as constraints including ion concentrations, temperature, timing of processing, and interactions with proteins, other nucleic acids, and further cellular substrates. An RNA also fluctuates to varying degrees among an ensemble of many possible, and often similar, structures. Critical biological roles for RNA structure have been realized in tRNAs, rRNAs, miRNAs, viral RNA genomes, and eukaryotic mRNAs (IREs, SECIS, editing sites).

There is some debate over the importance of mRNA structure in eukaryotes. This was previously addressed through analysis of MFE secondary structures for mRNA sequences from limited numbers of genes across species, including humans, and by comparison with shuffled sequences (Clote et al. 2005; Katz and Burge 2003; Meyer and Miklos 2005; Seffens and Digby 1999; Workman and Krogh 1999). However, based on experiments of Kozak and others (for a review see (Kozak 2005), (Babendure et al. 2006)) showing the importance of structure in eukaryotic 5' gene regulation, work on human functional mRNA IREs and SECIS (Kryukov et al. 2003), and a number of clinical associations with mRNA structure variants (Illustration 1), firm experimental evidence now supports human mRNA structures harboring biological functions, and that polymorphic alteration of such structure can exert biological and even clinical effects.

This study applied human genome-wide computation of mRNA secondary structures in conjunction with all

available human SNPs in transcribed regions to analyze structures in a large group of human mRNAs, and to address the effects of human genetic variation on secondary structure. We did not compare structures to those from shuffled sequences because our purpose is to describe human alleles and structures that may truly exist in vivo. This is the largest study thus far employing RNA structure prediction in combination with validated human population genetic information. This was accomplished by programmatic integration of UCSC and NCBI genome databases along with application of Perl programs to automate RNA secondary structure prediction (Vienna version 1.4 (Hofacker 2003)) and analysis. Structures were predicted for 1,059,195 unique alleles. Importantly, this larger set of sequences is inherently biased toward over-representation of highly polymorphic regions since these generate more possible haplotypes that are of high sequence similarity. In reality, given particular sets of human SNPs, all possible haplotypes commonly do not occur in populations. This initial broad approach was adopted to provide a comprehensive set of structures as an analytical starting point.

To conduct analyses on a less ambiguous dataset than available in the full dataset, annotation was employed to obtain a set of 34,577 validated, biallelic SNPs without additional sequence complexity (see Methods). Analysis of structural differences between major and minor alleles in this dataset reveals that the majority (65.9%) of human coding SNPs alter nearby MFE secondary structure, and a wider majority (93.6%) alter the profile of the ensemble of nearby secondary structures including suboptimal structures. Thus, most human SNPs alter RNA structures and should be detectable by their alteration of structure. Interestingly, single-strand conformation polymorphism (SSCP) SNP detection and genotyping approaches have high success rates (~90-95%), but some SNPs prove difficult to detect (Lenz et al. 1995; Ren 2000). This suggests that analysis of MFE structure alone in some cases is an incomplete predictor of structural change compared with analysis of the ensemble of all predicted confirmations, as the latter agrees more with observations regarding SNP detection rates by RNA-SSCP.

Many of the articles cited in Illustration 1 have relied only on MFE structures in their analysis. We applied a Z-score analysis of $\Delta\Delta G$ (MFE and ensemble) to those previously reported functional variants that appeared in our dataset. These functional variants have an average deviation from normal near 1 SD, with minor alleles biased toward thermodynamically less favorable MFE structures and ensembles, and in most cases $\Delta\Delta G$ was similar for MFE and ensemble structures (Pearson correlation; r=0.948) (Illustration 1). Correlations between MFE and ensemble ΔG values are also high in the larger datasets (Pearson correlations in both 22,785 and 34,557 SNP datasets: ΔGwt (r=0.995), ΔGsnp (r=0.995) and $\Delta\Delta G$ (r=0.94)). These results indicate that variants affecting functional structures generally change $\Delta\Delta G$ at the MFE and ensemble levels to at least a moderate extent, and that use of these measures is a potential predictor. However, reliance on thermodynamic favorability (ΔG) of MFE structures or ensembles, the free energy changes due to alleles ($\Delta\Delta G$), or Boltzmann probabilities are only partial indicators of potential functional changes, especially given the ranges of values observed in Illustration 1. We suggest that the combination of multiple analytical approaches be applied to identify variant structures as demonstrated below. In specific SNP cases, structures and ensembles should be thoroughly analyzed (e.g., (Zhang et al. 2005)), and experimentally verified by structural mapping or other methods demonstrated in articles in Illustration 1. Using such an approach we identified a region of structural integrity in the mu-opioid receptor gene OPRM1 that harbored functional variants affecting gene expression in target brain tissues and related clinical phenotypes (Zhang et al., 2005). Subsequently, Shabalina and colleagues described another variant in OPRM1 that lies in a conserved internal ribosome entry site, likely affecting a functional structure, and is also associated with differences in mRNA expression, translation efficiency and pain perception (Shabalina et al., 2009).

Notably, studies employing secondary structure prediction algorithms can potentially miss biologically important factors like long-range secondary interactions or complex tertiary motifs like pseudoknots. Nevertheless, we find that secondary structure predictions over a small sequence range capture the main portion of secondary structure change due to sequence variations (Illustration 6). We compared the behavior of major and minor allele structures at each position and find that changes in structure as a result of allele differences have a strong central tendency to alter interactions within a highly localized sequence space (Illustration 6). Sequence windows of 100 bp capture much of the predicted change due to variants, while 50 bp windows appear too narrow to allow a sufficient structural interaction space. One implication of this observation is that genetic variants must exist in close sequence space to functional RNA structures to exert a significant change, unless that change is mediated by tertiary interactions.

Comparing total sequence windows of 100 bp and 150 bp in Illustration 6 indicates that our dataset misses a relatively small portion of secondary structural change due to allele changes affecting sequence interactions beyond a 50 bp radius.

Synonymous SNP contexts form energetically more favorable structures on average than nonsynonymous and UTR contexts. This is in large part attributable to the nucleotide composition of the contexts (Illustration 3, panel D). The substantially higher frequency of guanine and uracil content among synonymous polymorphism sequence contexts results in more favorable interactions because these nucleotides may each pair in two helical interactions (G-C/G-U/A-U). This is also reflected by a higher percentage of bases helically paired on average in synonymous context structures (Illustration 4). This is consistent with the prior idea that the wobble position favors more stable structural interactions and may be adaptive (Varani & McClain, 2000; Shabalina et al. 2006; Resch et al. 2007) Structure in exon regions and within pre-mRNAs may influence alternative splicing (Buratti and Baralle 2004), and the significant difference between MFE for coding SNP contexts and UTR contexts of identical length that we observe may support some functional role for structures in the coding region.

We examined the distribution of the 12 SNP types among the functional categories and their effects on predicted structures. The four transversion types involving guanine were found to have the largest impact on structures, changing the MFE structure in close to 90% of cases, followed by other transversions (~75% change) and transitions (~60%) (Illustration 3, panel B). However, the fairly balanced distribution in Illustration 2 (panel B) indicates that change is almost as often predicted to be thermodynamically favorable as it was to be detrimental. Although we suspect most functional variants will impede the formation of favorable helical structures, cases where variants have closed functional loops have also been reported, and so variants at both sides of the $\Delta\Delta G$ distribution are potentially important (Illustration 2, panel B). Thus, the patterns of SNP type and $\Delta\Delta G$ in Illustration 3 (panel C) do not strictly match those in Illustration 3 (panel B) (e.g., a C>G transversion often forms a thermodynamically more favorable structure but in some contexts it is an unfavorable change). Both observations fit the thermodynamic underpinnings of secondary structure prediction, revealing that, in general, variants tend to follow the expected patterns dictated by Watson-Crick interactions, but the sequence context of individual variants ultimately also great influences their predicted impact upon structure. The importance of sequence context is further noted in analysis of variable length sequences and sequences differing due to alternative splicing. In particular the prediction of secondary structures of alternatively spliced mRNAs is relatively unique to this study. As expected we find that alternative splicing has significant potential to yield different mRNA structures (data not shown). One study has shown that alternative splicing of the human proinsulin gene results in a difference in 5'UTR RNA structure and altered translational efficiency (Shalev et al. 2002). This type of observation may be important given the extensive amount of splicing in vertebrate genomes (Le Texier et al. 2006).

We analyzed the minor allele frequencies (MAF) of SNPs and their predicted effect on MFE ($\Delta\Delta G$) and structure pairing behavior. We find that at MAF cutoffs varying from 1% to 40% we always observe the group of rarer alleles has larger average changes in structure and less favorable minor alleles ($\Delta\Delta G$). These differences are most significant near a MAF cutoff of 10%. Further analysis and results in Illustration 3 (panel C) indicate that the effect is in part due to a difference in the distribution of SNP types with respect to MAF. In particular, ranking the 12 SNP types by their average MAF in this set reveals that SNPs creating more often structurally favorable guanine alleles have higher MAF (A>G – 1st, U>G – 3rd) than those that abolish a guanine allele (G>A – 9th, G>U – 12th). Although this analysis indicates that variants with lower MAF are slightly more likely to create a change in RNA structure, it seems unlikely this is due to widespread population selection to specifically preserve RNA structures, since most variants are expected to be biologically neutral in their effects. Indeed, although G>A SNPs have slightly lower average MAFs, they are considerably more abundant in number in human coding regions than A>G SNPs (Illustration 3, panel A).

Given that our genome-wide analysis indicates most variants have the potential to alter RNA structure without a high degree of prediction specificity, we sought to identify alleles that alter structures most likely to convey a putative function. In these analyses we applied filters to both major and minor allele structures without preference, and considered large thermodynamic changes in either direction as potentially significant. First, selecting mRNA structures with extreme thermodynamic favorability (ΔG values between -54.0 and -69.0 kcal/mol), we observe examples of extensive helical pairing with an over-representation of sequences near translation initiation sites. This suggests that some of these structures may have functional roles in translation codon selection, initiation or ribosomal processivity.

CpG islands are highly correlated with thermodynamically favorable structures, indicating that regions susceptible to DNA methylation may form more stable RNA transcript structure than other regions.

We next selected structures with high Boltzmann probabilities (>0.145), with the hypothesis that the constrained thermodynamic ensemble may make these structures more likely to form in vivo (Miklos et al. 2005). This approach reveals a set of 46 SNP candidates altering likely RNA structures, including several in candidate genes with strong links to disease etiology. We also analyzed structures originally discarded from the single SNP dataset because of length constraints (n=1,203) or due to alternative sequence contexts (n=634). This set is significantly enriched for SNPs near the 5' end of transcripts as well as near sites of alternative splicing and translation initiation. Finally, we analyzed a merge of a database of evolutionarily conserved RNA structures identified through comparative genomics (Pedersen et al. 2006) and dbSNP (n=936). A large portion of these conserved, favorable structures are within the translated regions of mRNAs (synonymous and nonsynonymous), consistent with our findings (Illustration 3, panel A, Illustration 4). We observe a remarkably high co-localization of these exon structures with methionine codons, which indicates some of these structures likely play a role in promoting or inhibiting translation of downstream inframe start codons. A high percentage of EvoFold/SNP intersections are near the distal ends of 3'UTRs, indicating a potential conserved regulatory mechanism for mRNA structure at these sites, perhaps affecting RNA degradation rates, polyadenylation or miRNA targeting. Through alternative analyses we also identify thermodynamically favorable structures located in 5' regions and in proximity to translation start sites. Taken together these results suggest that human mRNA structures have multiple functional modalities and there is significant potential for human variants to alter favorable mRNA structures. Select putative functional variants are presented in Additional File 1.

We scanned non-coding and mRNA structures with known biological function for human genetic variation. We find support for previously investigated variants and additionally discover validated variants in functional RNA structures that have not yet been investigated. At the same time, variants known in dbSNP are notably absent, or distantly located in the bulk of functional RNAs examined, indicating that these structures could be selectively preserved due to their biological roles, or that a significant proportion of variation was not yet mapped to these RNAs at the time of our analysis. Further study of genotypes in functional RNA structures including some identified here is warranted as some variants are suggested as causative for diseases (e.g., (Allamand et al. 2006; Allerson et al. 1999)). We observe that a significant number of Mendelian disease-associated SNPs are absent from databases (consistent with Johnson et al., 2010). The proportion of non-validated, monoallelic and computationally predicted SNPs is also considerable. Thus, more effort is needed to annotate human disease-associated alleles reported in the literature but not submitted to databases. Furthermore, genome-wide projects relying on SNP annotations must generally guard against both false negatives and positives as we have done via SNPmasking and additional filtering steps before analysis.

Given our results and work of others (Illustration 1, (Pedersen et al. 2006; Washietl et al. 2005, Halvorsen et al. 2010)), we conclude there is sufficient evidence for the existence of many functional RNA structures in the human genome, and that a portion of functional human genetic variation exerts effects through the alteration of RNA conformations. Recent studies have suggested that gene expression quantitative trait loci (eQTLs) may account for a significant proportion of GWAS findings (Fransen et al. 2010; A.D. Johnson, unpublished results), and some of these eQTLs are likely to be expressed through modifications of functional RNA structures. Notably, a survey of the variants highlighted here against known GWAS findings that have survived stringent statistical thresholds, revealed several instances of overlap (Additional File 1, Tab 7, e.g., LPL, AGER, ZPBP2), suggesting that changes in functional RNA structures may represent the underlying functional mechanism that accounts for these GWAS associations. Clinical association studies have traditionally focused on amino acid changing variants and given less consideration to regulatory regions and variants that affect RNA maturation. Expanding the scope of functional genetic diversity, a mounting number of examples reveal variants of all types (synonymous, nonsynonymous, UTR, intronic) significantly altering function at a pre-translational level (Johnson et al. 2005, Sadee et al. 2011).

The impact of genetic variation on RNA structure is likely greater than our results suggest because we have analyzed only a portion of all alleles. More recent surveys of human variation are not included with this study, such as later HapMap phases or the 1000 Genomes Project data. Also, extensive haplotype analysis is likely to reveal additional variations affecting RNA structures. Our analysis of 2 SNP

haplotypes indicates that multi-allele interactions may mediate the preservation or disruption of functional structures. Although SNPs are the most common variant type, structures may also be affected by changes due to other types of variations (e.g., indels, tandem repeats, translocations). Poly-repeats associated with disease have already been observed to influence biologically important human RNA structures (Illustration 1). Furthermore, recent analyses indicate that there may be many non-coding and intergenic RNA structures left to be characterized in the human genome (Pedersen et al. 2006; Washietl et al. 2005), and these may also harbor functional genetic variation. Further work must be done to develop high-throughput methods to experimentally validate structures, understand the biologically relevant structures and the functional mechanisms that are affected, and to reveal population associations of known and putative functional RNA conformational polymorphisms.

Method

Generation of custom sequence databases via UCSC Genome Browser tools

Our sequence database is based on the May 2004 human assembly available as hg17 from genome.ucsc.edu. We used the RefSeq gene annotations as of November 2005 and dbSNP build 124. The UCSC source code includes a set of snpMask utilities: snpMaskChrom.c, snpMaskGenes.c and snpMaskFlank.c. The concept of "snpMasking" is to produce nucleotide sequence where single base substitutions are represented by IUPAC codes (Cornish-Bowden 1985). This method was independently developed by both UCSC and the WUSTL SNP Research Facility, and snpMasked sequences were compared and found to be identical (Koboldt et al. 2006). The method was more recently extended to a modified IUPAC code (Johnson 2010). snpMasking includes detection of dbSNP clustering errors, creating a blended observation of all alleles. At the current time, if the reference assembly differs from the observed SNP, this is not included as an additional variation. UCSC source code is available at http://hgdownload.cse.ucsc.edu/admin/jksrc.zip.

We used snpMaskFlank.c, which first stores all gene annotations in UCSC genePred format. This format includes a list of exon coordinates for each gene. snpMaskFlank.c then iterates through all genes, examining all exons. For each exon, snpMaskFlank.c selects all SNPs where class = 'snp', locType = 'exact' and chromEnd = chromStart + 1. It generates masked sequence for all exons using the absolute coordinates of the SNPs, and then constructs flanking sequence on two sides, connecting exons together as necessary. snpMaskFlank.c has a configurable #FLANKSIZE parameter, which was set to 25, 50 and 75 for our analysis.

These sequence sizes (50, 100, 150 bp total) were chosen to allow sufficient space for structural interactions and based on the widely held assumptions that RNA secondary structure motifs are generally small and predicted well locally, while predictions for large sequences are generally less accurate (Mathews et al. 1999; Meyer and Miklos 2005). We also tested this assumption by analyzing the frequency of structure changes at positions relative to the SNP base, and adopted a total sequence length of 100 bp for subsequent analyses (see Results, Illustration 6). Adoption of smaller sequence contexts reduces the sequence complexity due to additional genetic and alternative splicing variants (see below). Sequence retrieved was only exonic or untranslated region up to the maximum flanking length or to the end of each annotated RNA, whichever was reached first. The UCSC AltSplicing track was also incorporated to retrieve multiple flanking sequence contexts if a SNP is within a spliced region. Additional SNPs that fell within the flanking sequences were also represented by IUPAC codes and enumerated in the sequence header to allow for generation of haplotypes. An example of the filtering with IUPAC codes is given in Illustration 7 (panel A).

All IUPACs were adopted to reflect the coding strand orientation since genotypes are often reported on noncoding strands. A filter limit of 9 or less IUPAC codes per sequence was then imposed before structure prediction in order to reduce high computational load from exponential numbers of haplotypes due to highly polymorphic regions and duplicons in the genome. This resulted in the exclusion of 1,736 SNPs but decreased the size of the sequence space by more than 8.5 X 1012 alleles. A Perl script was used to generate all possible unique coding contexts/haplotypes for the remaining 153,397 unique coding region SNPs. Alternative RNA contexts (Illustration 7, panel B), SNPs with more than two alleles, and variable numbers of additional SNPs in the sequence regions (Illustration 7, panel C) contributed to generate a database of 1,059,195 unique mRNA haplotype allele sequences. These sequences were separated by chromosome and used in structure prediction.

Identification of putative or known functional RNA structures

Sets of known functional RNAs in the genome that

contained variants were analyzed separately from the large RefSeq set. Small RNAs containing variants were identified using a merge of the SNP and sno/miRNA tracks in UCSC genome Table browser (Griffiths-Jones 2004; Weber 2005), and information from two papers (Bentwich et al. 2005; Iwai and Naraba 2005). The predicted effects of variants in these RNAs were analyzed based on structures published in the miRNA Registry at the Wellcome Trust Sanger Institute (Griffiths-Jones 2004) and the snoRNA-LBME-DB at the Laboratoire de Biologie Moleculaire Eucaryote (Lestrade and Weber 2006). IREs and SECIS were determined based on previous works. BLAT searches were applied to the sequences for IREs and SECIS in an attempt to identify other similar regions in the human genome (Kent 2002). We also created a merge of dbSNP and EvoFold (Pedersen et al. 2006), a comparative genomics program that found many known and novel putative functional RNA structures in the human genome. To analyze the representation of methionine within EvoFold structures, we calculated the genome frequency of methionine codons from all RefSeq protein sequences. EvoFold structure lengths were trimmed to exclude intronic sequence, methionine codons were enumerated based on RefSeq annotation, and structures counted only once even if they contained multiple SNPs.

mRNA secondary structure prediction

The RNA sequence files were distributed over a cluster of 3, 64 bit machines, 2 with single processors and one with 4 processors, all running a UNIX O/S. A custom Perl script managed a pipeline feeding sequences to the Vienna RNA package (Hofacker 2003)(version 1.4) and processing results. The command line argument for prediction of each secondary structure followed a pattern:

RNAfold –p < InputSeq > OutputStructure

The –p flag for RNAfold specifies the calculation of a partition function and base pairing probability matrix in addition to the typical MFE structure calculation. Structures were calculated with default settings (370C, GU pairs allowed). An example and description of a text-based structure result in bracket notation is given in Illustration 5 (panel A). These textual structure results served as the basis for subsequent filtering and analysis. Additionally for every sequence, a graphical view of the MFE structure and the dot-plot was generated, archived and stored (Illustration 5, panel B).

Annotation and clustering of sequences and variants by functional type

In order to conduct analyses, annotations for sequences and variants were combined from a number of sources. A unique merge of dbSNP and UCSC annotation for more than 10 million variants was created through the use of Perl scripts. Though complete information was not available for every SNP, many contained the following information: rsID#, genome position, strand orientation, major allele, IUPAC, dbSNP validation status, average heterozygosity among genotyped populations, SD of average heterozygosity, functional classification (i.e., synonymous, UTR, intron), and number of genotyped individuals and frequencies of alleles. This annotation was critical to identifying the major and minor allele(s) in each sequence, and analyzing by allele frequency and functional categorization.

Major and minor alleles were determined by the relative allele frequencies combined from all genotyped samples for each SNP in dbSNP build 124. Untranslated region SNPs were not categorized as 5' or 3' in the human genome at the time of this method. We annotated UTR SNPs as 5' or 3' to estimate their abundance in the datasets using SNP genome position in combination with UCSC tables for gene boundaries (rnaCluster) and RefSeq genes (refGene). Merges of SNP sets and refGene were used to annotate geneIDs and estimate the number of unique genes represented in each of the datasets presented here (17,891 genes in 153,397 SNP set; 12,453 genes in 34,557 SNP set; 10,501 genes in 22,785 SNP set). Components of the annotations were also used in the filtering processes described below.

Analysis of 34,577 SNPs in a low ambiguity dataset

Ambiguity in the full sequence dataset arises from a number of sources: 1) n biallelic SNPs results in 2n possible haplotypes, some of which do not exist in human populations, 2) SNPs without validated allele frequencies, 3) SNPs with multiple functional categorizations (i.e., combinations of synonymous, nonsynonymous, UTR, intron), 4) SNPs with additional sequence contexts due to alternative splicing, 5) SNPs with flanking sequences of varying length due to transcript boundaries.

To analyze structure results in a low ambiguity context, we created a subset of 34,577 SNPs in RefSeq RNAs. All sequences in this subset contained 1) only one known SNP in the sequence, 2) SNP allele frequencies validated in one or more human population, 3) a single functional categorization for the SNP, 4) the SNP existed in only one sequence context (no alternative splicing), and 5) total sequence length of 100 bp. This subset was generated from another subset of 44,185 SNPs that contain no additional known variants in their flanking sequences. SNPs were removed from that set as follows: ambiguity in functional categorization (n=7,771), total sequence

length less than 100 bp (n=1,203), and existence in multiple sequence contexts (n=634). The final analytical subset contained 8,905 nonsynonymous, 10,702 synonymous, and 14,950 UTR region SNPs. Perl scripts were used to calculate further heuristics on the structure data for these SNPs including: change in the MFE and ensemble thermodynamic energies between major and minor allele structures, number of bases undergoing predicted structural changes from helical to base-paired and base-paired to helical between alleles, the frequency of changes in structural behavior predicted at {SNP-n…SNP…SNP+n} positions, and the change in the Boltzmann probability of the MFE structure within the ensemble.

Search for haplotypes with structurally interacting SNPs

We used a Perl script to search for human variant combinations that preserved RNA structures among instances where two SNPs were located within a 100 bp window as determined by the number of IUPACs (e.g., Illustration 7, panel C). The SNP base positions were determined by contrast of the sequences. MFE structures for each of 4 possible haplotypes were interrogated to find instances where both SNP bases helically paired. To do this a simple count up-count down ladder analysis of the helical structures in bracket notation (Illustration 5, panel A) was applied. Structures were selected as interacting instances when SNP bases were 1) both located at the bottom of the ladder, and 2) counting down only reached the base of the ladder once and exactly at the 3' SNP position.

Query of RNA-structure SNPs against GWAS results

We generated a list of RNA-structure related SNPs (all those appearing in Additional File 1, tabs 1-6). We used SNAP (Johnson et al., 2008) to query all proxies for RNA-structure related SNPs, and to expand the SNP list to include all known aliases for each SNP. The full list was then queried for direct matches to GWAS SNPs appearing in the NHGRI GWAS catalog (last accessed October 27, 2010) (Hindorff et al.).

Illustration and Additional File captions

Illustration 1. Experimental and computational investigations to date on the effects of human mRNA variation upon RNA structure, expression and disease. The table is separated as follows, 1A: a single functional SNP is investigated, 1B: multiple SNPs are known or predicted based on available information, 1C: multinucleotide and indel variants. Known functional variants in human tRNAs are not included. For review of these see (Florentz and Sissler 2001; Vilmi et al. 2005). Results from our structure-SNP database are reported for those variants for which haplotypes were unambiguous: Z-scores were calculated relative to categorical SDs (Illustration 2). Structure alteration was judged based on differences in the bracket notations for the alleles ('-' no difference, '+' difference at two or fewer positions, "++" difference at a moderate number of positions, "+++" difference at many positions). Those references not cited elsewhere in the paper are: (Carpen et al. 2005), (Steinberger et al. 2004), (Goodarzi et al. 2005), (Mas et al. 2004), (Puga et al. 2005), (Russcher et al. 2005a; Russcher et al. 2005b), (Duan et al. 2003), (Ding et al. 2005), (Michlewski and Krzyzosiak 2004), (Myers et al. 2004), (Sobczak et al. 2003), (Popowski et al. 2003), (Ly et al. 2003), (Barrette et al. 2001), (de Leon et al. 2000), (Maffei et al. 1997)

Illustration 2. A: Histogram distributions of the ΔG (kcal/mol) values for the MFE structures predicted for 69,114 human mRNA sequences. The structures include those predicted from sequences surrounding both major and minor alleles for 34,557 validated human SNPs. Histograms are separated by three SNP categories (nonsynonymous, synonymous and UTR). Means ± 1 SD are: nonsynonymous (-25.69 kcal/mol ± 8.59), synonymous (-26.40 kcal/mol ± 8.21) and UTR (-23.24 kcal/mol ± 9.42). B: Histogram distributions of the $\Delta\Delta$G (kcal/mol) values for the MFE structures predicted for 22,785 validated human SNPs. Not represented are 11,772 SNPs for which both alleles predicted identical MFE structures and $\Delta\Delta$G values of 0 kcal/mol. Their exclusion prevents kurtotic spikes for all three SNP contexts. Means ± 1 SD are: nonsynonymous (0.27 kcal/mol ± 2.06), synonymous (0.32 kcal/mol ± 1.90) and UTR (0.20 kcal/mol ± 1.95). Calculated normal curves with identical means and standard deviations are superimposed (SPSS, version 13). Histogram distributions for ensemble ΔG and $\Delta\Delta$G values showed similar patterns (data not shown).

Illustration 3. Relative frequencies in the dataset of 12 SNP types among functional categories for 34,557 validated human SNPs for which mRNA structures were analyzed. Panels A, B, and C are organized from the most common to least common SNP types (left to right) in human mRNAs. A: Relative frequencies of 12 SNP types among functional categories for 34,557 validated human SNPs. B: Percentage of SNPs of each type and functional category for which the major and minor alleles predicted identical MFE structures. C: Boxplots indicating the median (horizontal bars), interquartile ranges (boxes) and outliers for $\Delta\Delta$G of

MFE structures for 12 SNP types by sequence context. D: Percent sequence composition of the four nucleotides by SNP context type among 69,114 human mRNA sequences for which structures were analyzed.

Illustration 4. Comparative description and analysis of structures from 22,785 validated human SNPs whose major and minor alleles predict non-identical MFE structures. All reported values are means. Statistical analysis of differences by category was done by ANOVA with Bonferroni correction in SPSS (version 13).

Illustration 5. A: Example mRNA secondary structure results for a single SNP (rs10778) with two alleles. The SNP base is in bold. ID and sequences are indicated on lines 1, 2, 6 and 7 in identical format to Illustration 7. Lines 3 and 8 indicate the predicted MFE structures for the two alleles with the MFE (∆G in kcal/mol) indicated in parentheses. Structure is in the bracket notation adopted in Vienna: '.' unpaired base, '(' base paired with 3' partner, ')' base paired with 5' partner. Lines 4 and 9 indicate the consensus structure for the ensemble including all suboptimal structures: '.' unpaired base in > 2/3 of structures, '(' and ')' paired in > 2/3 of structures, ',' '{' '}' are weaker versions of these preferences, and '|' indicates that a base is paired in 2/3 of structures without a clear preference for a 5' or 3' partner. The free energy for the thermodynamic ensemble (∆G in kcal/mol) is indicated in braces. Lines 5 and 10 indicate the Boltzmann probabilities; assuming thermodynamic equilibration this is a representation of the amount of time spent in the MFE structure relative to all structures in the ensemble. B: Graphical depiction of the MFE structure and a dotplot representation (C) of the structure ensemble for rs10778 (G allele). MFE structures can be reconstituted from bracket notations.

Illustration 6. Frequencies of changes in structural RNA folding between major and minor alleles among 34,557 validated human SNPs. The x-axisindicates base positions 5' to 3' (-75 to +75; -50 to +50; -25 to +25) relative to the centrally located SNPs. Analysis was done with a Perl program that compared the structural pairing behavior of individual base pairs between alleles using the bracket notation for structure (e.g., Illustration 5, panel B). In overlap regions the 50 bp and 75 bp flanking results are not significantly different from each other, but both are different from the 25 bp flanking results (p

Illustration 7. A: Generalized view of IUPAC and mRNA haplotype databases. Each sequence header line indicates the following: a – dbSNP reference id for SNP, b – unique sequence context number for given SNP (0…n) (e.g., alternative splicing context), c – haplotype allele sequence number (0,1…n), d – total number of SNPs located in sequence. Major allele and coding strand are also indicated. A second line gives the sequence with IUPAC codes or the unique haplotype sequence. B: Example of a biallelic SNP (rs2205447) that has no additional SNPs in the flanking sequence but exists in two unique sequence contexts due to alternative splicing. Four mRNA haplotypes are generated for this SNP. C: Example of a biallelic SNP (rs4113746) that has one additional SNP in the flanking sequence region. Four RNA haplotypes are also generated but the resulting numbering scheme differs from the example in B.

Additional File 1, Tab 1. Structures and information on 152 thermodynamically most favorable RNA structures that contain SNPs (one allele with ∆G more favorable than -50 kcal/mol). SNP frequencies of 0.00 indicate a reported frequency below 1%. Validation codes: 'C' – by cluster, 'F' – by frequency, '2' – by 2-hit-2-allele, 'G' – by genotype, 'P' – by population.

Additional File 1, Tab 2. Structures and information for 46 SNPs that have one or more allele with a Boltzmann value >0.145, and displaying a significant change in energy between alleles. SNP frequencies of 0.00 indicate a reported frequency below 1%. Validation codes: 'C' – by cluster, 'F' – by frequency, '2' – by 2-hit-2-allele, 'G' – by genotype, 'P' – by population.

Additional File 1, Tab 3. Information on 936 SNPs intersecting with EvoFold annotations in the human genome. This set has not been filtered and contains non-validated SNPs.

Additional File 1, Tab 4. Structures and information on 97 SNPs selected from a set of structures enriched for 5'UTR boundaries and sites of alternative RNA processing and translation. SNP frequencies of 0.00 indicate a reported frequency below 1%. Validation codes: 'C' – by cluster, 'F' – by frequency, '2' – by 2-hit-2-allele, 'G' – by genotype, 'P' – by population.

Additional File 1, Tab 5. Structures and information for 34 2-SNP haplotypes where SNP bases interact in RNA structures in at least 1 of the 4 predicted haplotypes. For each pair there are two sets of 4 haplotypes, depending on which SNP is centered within the flanking sequence. The haplotypes where SNP bases interact are indicated and correspond to haplotypes numbers indicated in Columns K-V. The position of each SNP within the haplotype sequences are indicated in Columns I and J (0…n numbering). For those pairs with multiple haplotypes displaying interactions, substitution patterns corresponding to the structural observations are given in Column H.

Additional File 1, Tab 6. Summary of results from a survey of known functional structures in the human

genome. Included in the table are only those SNPs that have been validated; additional non-validated but potentially functional SNPs are discussed in the text.

Additional File 1, Tab 7. RNA-structure SNPs or their strong LD proxies (r2>0.9) which coincide with strong GWAS results from the NHGRI Catalog (last accessed October 27, 2010).

References

1. Aguilar Martinez, P., Biron, C., Blanc, F., Masmejean, C., Jeanjean, P., Michel, H., and Schved, J.F. 1997. Compound heterozygotes for hemochromatosis gene mutations: may they help to understand the pathophysiology of the disease? Blood Cells Mol Dis 23: 269-276.

2. Allamand, V., Richard, P., Lescure, A., Ledeuil, C., Desjardin, D., Petit, N., Gartioux, C., Ferreiro, A., Krol, A., Pellegrini, N. et al. 2006. A single homozygous point mutation in a 3'untranslated region motif of selenoprotein N mRNA causes SEPN1-related myopathy. EMBO Rep 7: 450-454.

3. Allerson, C.R., Cazzola, M., and Rouault, T.A. 1999. Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. J Biol Chem 274: 26439-26447.

4. Ancel, L.W. and Fontana, W. 2000. Plasticity, evolvability, and modularity in RNA. J Exp Zool 288: 242-283.

5. Athanasiadis, A., Rich, A., and Maas, S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol 2: e391.

6. Babendure, J.R., Babendure, J.L., Ding, J.H., and Tsien, R.Y. 2006. Control of mammalian translation by mRNA structure near caps. Rna.

7. Barrette, I., Poisson, G., Gendron, P., and Major, F. 2001. Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. Nucleic Acids Res 29: 753-758.

8. Beaumont, C., Leneuve, P., Devaux, I., Scoazec, J.Y., Berthier, M., Loiseau, M.N., Grandchamp, B., and Bonneau, D. 1995. Mutation in the iron responsive element of the L ferritin mRNA in a family with dominant hyperferritinaemia and cataract. Nat Genet 11: 444-446.

9. Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E. et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37: 766-770.

10. Berry, M.J., Banu, L., Chen, Y.Y., Mandel, S.J., Kieffer, J.D., Harney, J.W., and Larsen, P.R. 1991. Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. Nature 353: 273-276.

11. Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20: 2911-2917.

12. Buratti, E. and Baralle, F.E. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 24: 10505-10514.

13. Camaschella, C., Zecchina, G., Lockitch, G., Roetto, A., Campanella, A., Arosio, P., and Levi, S. 2000. A new mutation (G51C) in the iron-responsive element (IRE) of L-ferritin associated with hyperferritinaemia-cataract syndrome decreases the binding affinity of the mutated IRE for iron-regulatory proteins. Br J Haematol 108: 480-482.

14. Campagnoli, M.F., Pimazzoni, R., Bosio, S., Zecchina, G., DeGobbi, M., Bosso, P., Oldani, B., and Ramenghi, U. 2002. Onset of cataract in early infancy associated with a 32G-->C transition in the iron responsive element of L-ferritin. Eur J Pediatr 161: 499-502.

15. Carpen, J.D., Archer, S.N., Skene, D.J., Smits, M., and von Schantz, M. 2005. A single-nucleotide polymorphism in the 5'-untranslated region of the hPER2 gene is associated with diurnal preference. J Sleep Res 14: 293-297.

16. Cavaille, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C.I., Horsthemke, B., Bachellerie, J.P., Brosius, J., and Huttenhofer, A. 2000. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A 97: 14311-14316.

17. Cazzola, M., Bergamaschi, G., Tonon, L., Arbustini, E., Grasso, M., Vercesi, E., Barosi, G., Bianchi, P.E., Cairo, G., and Arosio, P. 1997. Hereditary hyperferritinemia-cataract syndrome: relationship between phenotypes and specific mutations in the iron-responsive element of ferritin light-chain mRNA. Blood 90: 814-821.

18. Chen, Y., Carlini, D.B., Baines, J.F., Parsch, J., Braverman, J.M., Tanda, S., and Stephan, W. 1999. RNA secondary structure and compensatory evolution. Genes Genet Syst 74: 271-286.

19. Clote, P., Ferre, F., Kranakis, E., and Krizanc, D. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. Rna 11: 578-591.

20. Cmejla, R., Petrak, J., and Cmejlova, J. 2006. A novel iron responsive element in the 3'UTR of human

MRCKalpha. Biochem Biophys Res Commun 341: 158-166.
21. Cornish-Bowden, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res 13: 3021-3030.
22. de Leon, J.H., Vatsis, K.P., and Weber, W.W. 2000. Characterization of naturally occurring and recombinant human N-acetyltransferase variants encoded by NAT1. Mol Pharmacol 58: 288-299.
23. Ding, D., Xu, L., Menon, M., Reddy, G.P., and Barrack, E.R. 2005. Effect of GGC (glycine) repeat length polymorphism in the human androgen receptor on androgen action. Prostate 62: 133-139.
24. Draghi, J.A., Parsons, T.L., Wagner, G.P., Plotkin, J.B. 2010. Mutational robustness can facilitate adaptation. Nature 463: 353-355.
25. Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., and Gejman, P.V. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet 12: 205-216.
26. Fialcowitz, E.J., Brewer, B.Y., Keenan, B.P., and Wilson, G.M. 2005. A hairpin-like structure within an AU-rich mRNA-destabilizing element regulates trans-factor binding selectivity and mRNA decay kinetics. J Biol Chem 280: 22406-22417.
27. Florentz, C. and Sissler, M. 2001. Disease-related versus polymorphic mutations in human mitochondrial tRNAs. Where is the difference? EMBO Rep 2: 481-486.
28. Fransen, K., Visschedijk, M.C., van Sommeren, S., Fu, J.Y., Franke, L., Festen, E.A., Stokkers, P.C., van Bodegraven, A.A., Crusius, J.B., Hommes, D.W. et al. 2010. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk gene for Crohn's disease. Hum Mol Genet 19: 3482-3488.
29. Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. J Mol Biol 196: 261-282.
30. Girelli, D., Corrocher, R., Bisceglia, L., Olivieri, O., De Franceschi, L., Zelante, L., and Gasparini, P. 1995. Molecular basis for the recently described hereditary hyperferritinemia-cataract syndrome: a mutation in the iron-responsive element of ferritin L-subunit gene (the "Verona mutation"). Blood 86: 4050-4053.
31. Goodarzi, M.O., Wong, H., Quinones, M.J., Taylor, K.D., Guo, X., Castellani, L.W., Antoine, H.J., Yang, H., Hsueh, W.A., and Rotter, J.I. 2005. The 3' untranslated region of the lipoprotein lipase gene: haplotype structure and association with post-heparin plasma lipase activity. J Clin Endocrinol Metab 90: 4816-4823.
32. Gray, N.K. and Hentze, M.W. 1994. Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs. Embo J 13: 3882-3891.
33. Griffiths-Jones, S. 2004. The microRNA Registry. Nucleic Acids Res 32: D109-111.
34. Halvorsen, M., Martin, J.S., Broadaway, S., Laederach, A. 2010. Disease-associated mutations that alter the RNA structural ensemble. PLoS Genet 6: e1001074.
35. Hancock, D.B., Eijgelsheim, M., Wilk, J.B., Gharib, S.A., Loehr, L.R., Marciante, K.D., Franceschini, N., Durme, Y.M.T.A., Chen T., Barr, R.G. 2009. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Gen 42: 45-52.
36. Hindorff, L.A., Junkins, H.A., Hall, P.N., Mehta, J.P., Manolio, T.A. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed [October 27, 2010]
37. Hofacker, I.L. 2003. Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429-3431.
38. Howard, M.T., Aggarwal, G., Anderson, C.B., Khatri, S., Flanigan, K.M., and Atkins, J.F. 2005. Recoding elements located adjacent to a subset of eukaryal selenocysteine-specifying UGA codons. Embo J 24: 1596-1607.
39. Hu, Y.J., Korotkov, K.V., Mehta, R., Hatfield, D.L., Rotimi, C.N., Luke, A., Prewitt, T.E., Cooper, R.S., Stock, W., Vokes, E.E. et al. 2001. Distribution and functional consequences of nucleotide polymorphisms in the 3'-untranslated region of the human Sep15 gene. Cancer Res 61: 2307-2310.
40. Iwai, N. and Naraba, H. 2005. Polymorphisms in human pre-miRNAs. Biochem Biophys Res Commun 331: 1439-1444.
41. Johnson, A.D., Wang, D., and Sadee, W. 2005. Polymorphisms affecting gene regulation and mRNA processing: broad implications for pharmacogenetics. Pharmacol Ther 106: 19-38.
42. Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., de Bakker, P.I.W. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938-2939.
43. Johnson, A.D. 2009. SNP bioinformatics: a comprehensive review of resources. Circ. Cardio. Gen. 2: 530-536.
44. Johnson, A.D., Bhimavarapu, A., Benjamin, E.J., Fox, C., Levy, D., Jarvik, G.P., O'Donnell, C.J. 2010. CLIA-tested genetic variants on commercial SNP arrays: potential for incidental findings in genome-wide association studies. Genet. Med. 12: 355-363.

45. Johnson, A.D. 2010. An extended IUPAC nomenclature code for polymorphic nucleic acids. Bioinformatics 26: 1386-1389.
46. Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., et al. 2008. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or tryiglycerides in humans. Nat Genet 40: 189-197.
47. Kato, J., Fujikawa, K., Kanda, M., Fukuda, N., Sasaki, K., Takayama, T., Kobune, M., Takada, K., Takimoto, R., Hamada, H. et al. 2001. A mutation, in the iron-responsive element of H ferritin mRNA, causing autosomal dominant iron overload. Am J Hum Genet 69: 191-197.
48. Katz, L. and Burge, C.B. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 13: 2042-2051.
49. Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.
50. Kiss, A.M., Jady, B.E., Bertrand, E., and Kiss, T. 2004. Human box H/ACA pseudouridylation guide RNA machinery. Mol Cell Biol 24: 5797-5807.
51. Koboldt, D.C., Miller, R.D., and Kwok, P.Y. 2006. Distribution of human SNPs and its effect on high-throughput genotyping. Hum Mutat 27: 249-254.
52. Kozak, M. 1986. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci U S A 83: 2850-2854.
53. Kozak, M. 1990. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc Natl Acad Sci U S A 87: 8301-8305.
54. Kozak, M. 2005. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.
55. Kryukov, G.V., Castellano, S., Novoselov, S.V., Lobanov, A.V., Zehtab, O., Guigo, R., and Gladyshev, V.N. 2003. Characterization of mammalian selenoproteomes. Science 300: 1439-1443.
56. Kumaraswamy, E., Malykh, A., Korotkov, K.V., Kozyavkin, S., Hu, Y., Kwon, S.Y., Moustafa, M.E., Carlson, B.A., Berry, M.J., Lee, B.J. et al. 2000. Structure-expression relationships of the 15-kDa selenoprotein gene. Possible role of the protein in cancer etiology. J Biol Chem 275: 35540-35547.
57. Kuo, K.W., Leung, M.F., and Leung, W.C. 1997. Intrinsic secondary structure of human TNFR-I mRNA influences the determination of gene expression by RT-PCR. Mol Cell Biochem 177: 1-6.
58. Le Texier, V., Riethoven, J.J., Kumanduri, V., Gopalakrishnan, C., Lopez, F., Gautheret, D., and Thanaraj, T.A. 2006. AltTrans: Transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. BMC Bioinformatics 7: 169.
59. Lenz, H.J., Danenberg, K.D., Schnieders, B., Banerjee, D., Bertino, J.R., Leichman, L., and Danenberg, P.V. 1995. Identification of mutations by RNA conformational polymorphism "bar code" analysis. Genomics 30: 120-122.
60. Lestrade, L. and Weber, M.J. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res 34: D158-162.
61. Lopez de Silanes, I., Galban, S., Martindale, J.L., Yang, X., Mazan-Mamczarz, K., Indig, F.E., Falco, G., Zhan, M., and Gorospe, M. 2005. Identification and functional outcome of mRNAs associated with RNA-binding protein TIA-1. Mol Cell Biol 25: 9520-9531.
62. Lopez de Silanes, I., Zhan, M., Lal, A., Yang, X., and Gorospe, M. 2004. Identification of a target RNA motif for RNA-binding protein HuR. Proc Natl Acad Sci U S A 101: 2987-2992.
63. Ly, H., Blackburn, E.H., and Parslow, T.G. 2003. Comprehensive structure-function analysis of the core domain of human telomerase RNA. Mol Cell Biol 23: 6849-6856.
64. Maffei, A., Pozzo, G.D., Prisco, A., Ciullo, M., Harris, P.E., Reed, E.F., and Guardiola, J. 1997. Polymorphism in the 5' terminal region of the mRNA of HLA-DQA1 gene: identification of four groups of transcripts and their association with polymorphism in the alpha 1 domain. Hum Immunol 53: 167-173.
65. Martin, M.E., Fargion, S., Brissot, P., Pellat, B., and Beaumont, C. 1998. A point mutation in the bulge of the iron-responsive element of the L ferritin gene in two families with the hereditary hyperferritinemia-cataract syndrome. Blood 91: 319-323.
66. Martineau, Y., Le Bec, C., Monbrun, L., Allo, V., Chiu, I.M., Danos, O., Moine, H., Prats, H., and Prats, A.C. 2004. Internal ribosome entry site structural motifs conserved among mammalian fibroblast growth factor 1 alternatively spliced mRNAs. Mol Cell Biol 24: 7622-7635.
67. Mas, C., Taske, N., Deutsch, S., Guipponi, M., Thomas, P., Covanis, A., Friis, M., Kjeldsen, M.J., Pizzolato, G.P., Villemure, J.G. et al. 2004. Association of the connexin36 gene with juvenile myoclonic epilepsy. J Med Genet 41: e93.
68. Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911-940.
69. McLeod, J.L., Craig, J., Gumley, S., Roberts, S., and Kirkland, M.A. 2002. Mutation spectrum in Australian pedigrees with hereditary

hyperferritinaemia-cataract syndrome reveals novel and de novo mutations. Br J Haematol 118: 1179-1182.
70. Meyer, I.M. and Miklos, I. 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acids Res 33: 6338-6348.
71. Michlewski, G. and Krzyzosiak, W.J. 2004. Molecular architecture of CAG repeats in human disease related transcripts. J Mol Biol 340: 665-679.
72. Miklos, I., Meyer, I.M., and Nagy, B. 2005. Moments of the Boltzmann distribution for RNA secondary structures. Bull Math Biol 67: 1031-1047.
73. Moghadaszadeh, B., Petit, N., Jaillard, C., Brockington, M., Roy, S.Q., Merlini, L., Romero, N., Estournet, B., Desguerre, I., Chaigne, D. et al. 2001. Mutations in SEPN1 cause congenital muscular dystrophy with spinal rigidity and restrictive respiratory syndrome. Nat Genet 29: 17-18.
74. Moore, M.J. 2005. From birth to death: the complex lives of eukaryotic mRNAs. Science 309: 1514-1518.
75. Mumford, A.D., Vulliamy, T., Lindsay, J., and Watson, A. 1998. Hereditary hyperferritinemia-cataract syndrome: two novel mutations in the L-ferritin iron-responsive element. Blood 91: 367-368.
76. Myers, S.J., Huang, Y., Genetta, T., and Dingledine, R. 2004. Inhibition of glutamate receptor 2 translation by a polymorphic repeat sequence in the 5'-untranslated leaders. J Neurosci 24: 3489-3499.
77. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., Cox, N.J. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6: e1000888.
78. Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. PLoS Comput Biol 2: e33.
79. Popowski, K., Sperker, B., Kroemer, H.K., John, U., Laule, M., Stangl, K., and Cascorbi, I. 2003. Functional significance of a hereditary adenine insertion variant in the 5'-UTR of the endothelin-1 gene. Pharmacogenetics 13: 445-451.
80. Puga, I., Lainez, B., Fernandez-Real, J.M., Buxade, M., Broch, M., Vendrell, J., and Espel, E. 2005. A polymorphism in the 3' untranslated region of the gene for tumor necrosis factor receptor 2 modulates reporter gene expression. Endocrinology 146: 2210-2220.
81. Ren, J. 2000. High-throughput single-strand conformation polymorphism analysis by capillary electrophoresis. J Chromatogr B Biomed Sci Appl 741: 115-128.
82. Repapi, E., Sayers, I., Wain, L.V., Burton, P.R., Johnson, T., Obeidat, M., Zhao, J.H., Ramasamy, A., Zhai, G., Vitart, V. 2009. Genome-wide association study identifies five loci associated with lung function. Nat Gen 42: 36-45.
83. Resch, A.M., Carmel, L., Marino-Ramirez, L., Ogurtsov, A.Y., Shabalina, S.A., Rogozin, I.B., Koonin, E.V. 2007. Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol 24: 1821-1831.
84. Russcher, H., Smit, P., van den Akker, E.L., van Rossum, E.F., Brinkmann, A.O., de Jong, F.H., Lamberts, S.W., and Koper, J.W. 2005a. Two polymorphisms in the glucocorticoid receptor gene directly affect glucocorticoid-regulated gene expression. J Clin Endocrinol Metab 90: 5804-5810.
85. Russcher, H., van Rossum, E.F., de Jong, F.H., Brinkmann, A.O., Lamberts, S.W., and Koper, J.W. 2005b. Increased expression of the glucocorticoid receptor-A translational isoform as a result of the ER22/23EK polymorphism. Mol Endocrinol 19: 1687-1696.
86. Sadee, W., Wang, D., Papp, A.C., Pinsonneault, J.K., Smith, R.M., Moyer, R.A., Johnson, A.D. 2011. Pharmacogenomics of the RNA world: structural RNA polymorphisms in drug therapy. Clin Pharamcol Ther 89: 355-365.
87. Sarkar, G., Yoon, H.S., and Sommer, S.S. 1992. Screening for mutations by RNA single-strand conformation polymorphism (rSSCP): comparison with DNA-SSCP. Nucleic Acids Res 20: 871-878.
88. Seffens, W. and Digby, D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res 27: 1578-1584.
89. Shabalina, S.A., Ogurtsov, A.Y., Spiridonov, N.A. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res 34: 2428-2438.
90. Shabalina, S.A., Zaykin, D.V., Gris, P., Gauthier, J., Shibata, K., Tchivileva, I.E., Belfer, I., Mishra, B., Kiselycznyk, C., Wallace, M.R. et al. 2009. Expansion of the human mu-opioid receptor gene artchitecture: novel functional variants. Hum Mol Genet. 18: 1037-1051.
91. Shalev, A., Blair, P.J., Hoffmann, S.C., Hirshberg, B., Peculis, B.A., and Harlan, D.M. 2002. A proinsulin gene splice variant with increased translation efficiency is expressed in human pancreatic islets. Endocrinology 143: 2541-2547.
92. Shen, L.X., Basilion, J.P., and Stanton, V.P., Jr. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci U S A 96: 7871-7876.

93. Sobczak, K., de Mezer, M., Michlewski, G., Krol, J., and Krzyzosiak, W.J. 2003. RNA structure of trinucleotide repeats associated with human neurological diseases. Nucleic Acids Res 31: 5469-5482.
94. Steinberger, D., Blau, N., Goriuonov, D., Bitsch, J., Zuker, M., Hummel, S., and Muller, U. 2004. Heterozygous mutation in 5'-untranslated region of sepiapterin reductase gene (SPR) in a patient with dopa-responsive dystonia. Neurogenetics 5: 187-190.
95. Varani, G., McClain, W.H. 2000. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. EMBO Rep. 1: 18-23.
96. Vickers, T.A., Wyatt, J.R., and Freier, S.M. 2000. Effects of RNA secondary structure on cellular antisense activity. Nucleic Acids Res 28: 1340-1347.
97. Villette, S., Kyle, J.A., Brown, K.M., Pickard, K., Milne, J.S., Nicol, F., Arthur, J.R., and Hesketh, J.E. 2002. A novel single nucleotide polymorphism in the 3' untranslated region of human glutathione peroxidase 4 influences lipoxygenase metabolism. Blood Cells Mol Dis 29: 174-178.
98. Vilmi, T., Moilanen, J.S., Finnila, S., and Majamaa, K. 2005. Sequence variation in the tRNA genes of human mitochondrial DNA. J Mol Evol 60: 587-597.
99. Wang, D., Johnson, A.D., Papp, A.C., Kroetz, D.L., and Sadee, W. 2005. Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. Pharmacogenet Genomics 15: 693-704.
100. Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., and Stadler, P.F. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23: 1383-1390.
101. Weber, M.J. 2005. New human and mouse microRNA genes found by homology search. Febs J 272: 59-73.
102. Workman, C. and Krogh, A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res 27: 4816-4822.
103. Zhang, Q., Sun, X., Watt, E.D., and Al-Hashimi, H.M. 2006. Resolving the motional modes that code for RNA adaptation. Science 311: 653-656.
104. Zhang, Y., Wang, D., Johnson, A.D., Papp, A.C., and Sadee, W. 2005. Allelic expression imbalance of human mu opioid receptor (OPRM1) caused by variant A118G. J Biol Chem 280: 32618-32624.

Acknowledgments

The authors thank Dan Janies for kind use of his cluster for computation and Jakob Pedersen and Kirk Mykytyn for reading and critical comments on the manuscript.

Illustrations

Illustration 1

Experimental and computational investigations to date on the effects of human mRNA variation upon RNA structure, expression and disease.

Illustration 1A: Single SNPs				ΔG		ΔΔG				Structure alteration		Boltzmann probability	
Article	Human Gene(s)	dbSNPid	Region	MFE wt	Z	MFE	Z	Ens	Z	MFE	Ens	Major	Minor
Allamand (2006)	*SEPN1*	-	3'UTR	Structural differences are observed									
Carpen (2005)	*PER2*	rs2304672	5'UTR	-26.80	0.38	-0.90	-0.56	-0.78	-0.56	+++	+++	<0.01	<0.01
Steinberger (2004)	*SPR*	-	5'UTR	Structural differences are observed									
Vilette (2002)	*GPX4*	rs713041	3'UTR	-37.00	1.46	1.70	0.77	1.74	0.92	+	+	0.42	0.45
Hu (2001)	*SEP15*	rs5845	3'UTR	-18.30	-0.52	1.45	0.64	1.93	1.03	++	+++	0.01	0.02
	SEP15	rs5859	3'UTR	-26.00	0.29	2.00	0.92	1.67	0.88	+	++	0.20	0.12
Shen (1999)	*AARS*	rs2070203	Syn	-33.40	0.85	-0.70	-0.46	1.41	0.67	+++	+++	<0.01	<0.01
	RPA1	rs2230931	Syn	-23.40	-0.37	0.70	0.26	0.59	0.18	-	+	<0.01	<0.01
Illustration 1B: Haplotypes													
Goodarzi (2005)	*LPL*	2 haplotypes	3'UTR	Structural differences are observed									
Mas (2005)	*GJA9*	rs3743123	Syn	-22.61	-0.46	-1.42	-0.92	-0.63	-0.56	+	++	<0.01	0.01
Puga (2005)	*TNFRSF1B*	5 haplotypes	3'UTR	Structural differences are observed									
Russcher (2005)	*NR3C1*	rs6189	Syn, NS	-24.60	-0.22	2.20	0.99	2.54	1.36	++	++	0.07	0.12
Wang (2005)	*ABCB1*	rs1045642	Syn	-28.40	0.24	-0.10	-0.22	0.70	0.24	++	++	<0.01	0.02
Zhang (2005)	*OPRM1*	rs1799971	NS	-37.70	1.40	-4.90	-2.51	-4.18	-2.43	+	++	0.01	0.04
Duan (2003)	*DRD2*	rs6277	Syn	Structural differences are observed									
Illustration 1C: Insertions and multinucleotide variants													
Ding (2005)	*AR*	-	G,Q repeats										
Michlewski (2004)	*ATXN3, CACNA1A, ATN1*	-	Q repeats										
Myers (2004)	*GRIA2*	-	5'UTR										
Sobczak (2003)	-	*In vitro*	Repeats										
Popowski (2003)	*EDN1*	rs10478694	5'UTR										
Ly (2003)	*TERC*	Multi-allelic	Pseudo-knots										
Shalev (2002)	*INS*	-	5'UTR										
Barette (2001)	*PRNP*	-	24bp repeats										
de Leon (2000)	*NAT1*	-	3'UTR										
Maffei (1997)	*HLA-DQA1*	-	5'UTR										

Illustration 1. Experimental and computational investigations to date on the effects of human mRNA

Illustration 2

Histogram distributions of the deltaG (kcal/mol) values for the MFE structures predicted for 69,114 human mRNA sequences.

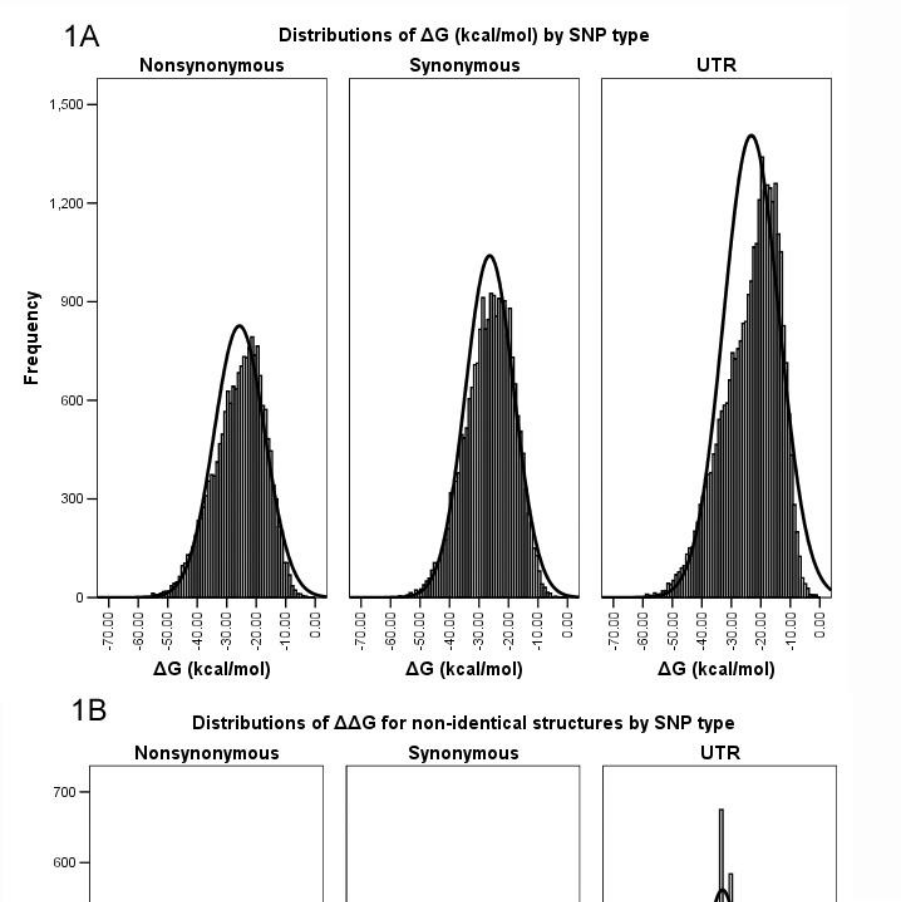


Illustration 2. A: Histogram distributions of the ΔG (kcal/mol) values for the MFE structures predicted for 69,114 human mRNA sequences. The structures include those predicted from sequences surrounding both major and minor alleles for 34,557 validated human SNPs. Histograms are separated by three SNP categories (nonsynonymous, synonymous and UTR). Means ± 1 SD are: nonsynonymous (-25.69 kcal/mol ± 8.59), synonymous (-26.40 kcal/mol ± 8.21) and UTR (-23.24 kcal/mol ± 9.42). **B:** Histogram distributions of the ΔΔG (kcal/mol) values for the MFE structures predicted for 22,785 validated human SNPs. Not represented are 11,772 SNPs for which both alleles predicted identical MFE structures and ΔΔG values of 0 kcal/mol. Their exclusion prevents kurtotic spikes for all three SNP contexts. Means ± 1 SD are: nonsynonymous (0.27 kcal/mol ± 2.06), synonymous (0.32 kcal/mol ± 1.90) and UTR (0.20 kcal/mol ± 1.95). Calculated normal curves with identical means and standard deviations are superimposed (SPSS, version 13). Histogram distributions for *ensemble* ΔG and ΔΔG values showed similar patterns (data not shown).

Illustration 3

Frequencies of changes in structural mRNA folding between major and minor alleles among 34,557 validated human SNPs under varying window sizes.

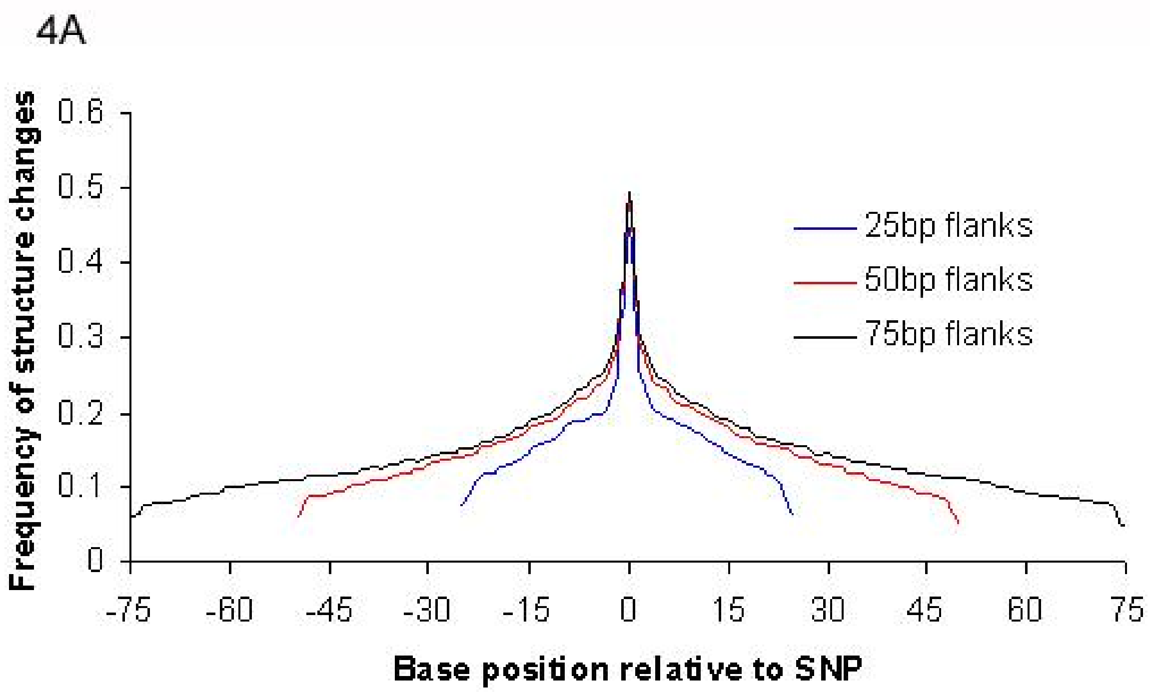


Illustration 6. Frequencies of changes in structural RNA folding between major and minor alleles among 34,557 validated human SNPs. The *x-axis* indicates base positions 5' to 3' (-75 to +75; -50 to +50; -25 to +25) relative to the centrally located SNPs. Analysis was done with a Perl program that compared the structural pairing behavior of individual base pairs between alleles using the bracket notation for

Illustration 4

Comparative description and analysis of structures from 22,785 validated human SNPs whose major and minor alleles predict non-identical MFE structures.

Illustration 4	n	Nonsynonymous (n=6041)	Synonymous (n=6886)	UTR (n=9858)	ANOVA (Bonferroni correction) NS/Syn	Syn/UTR	UTR/NS
ΔG major allele (kcal/mol)	MFE	-25.83	-26.31	-23.38	0.004	<0.001	<0.001
ΔG minor allele (kcal/mol)	MFE	-25.56	-26.00	-23.18	0.011	<0.001	<0.001
	ΔΔG (kcal/mol)	0.27	0.31	0.20	0.549	0.001	0.109
ΔG major allele (kcal/mol)	Ensemble	-28.91	-29.45	-26.44	0.001	<0.001	<0.001
ΔG minor allele (kcal/mol)	Ensemble	-28.65	-29.15	-26.25	0.003	<0.001	<0.001
	ΔΔG (kcal/mol)	0.26	0.30	0.19	0.561	<0.001	0.029
Average Boltzmann probability	Major allele	0.027	0.025	0.025	0.098	1.000	0.015
	Minor allele	0.026	0.024	0.024	0.002	1.000	0.003
Average % of bases helically paired	Major allele	52.73%	53.21%	52.61%			
	Minor allele	52.37%	52.83%	52.47%			
	Average MAF	0.163	0.181	0.188			
Helices disrupted: Helices formed	MFE	1.037	1.039	1.014			
	Ensemble	1.048	1.054	1.037			

Illustration 4. Comparative description and analysis of structures from 22,785 validated human SNPs whose major and minor alleles predict non-identical MFE structures. All reported values are means. Statistical analysis of differences by category was done by ANOVA with Bonferroni correction in SPSS (version 13).

Illustration 5

Relative frequencies in the dataset of 12 SNP types among functional categories for 34,557 validated human SNPs for which mRNA structures were analyzed.

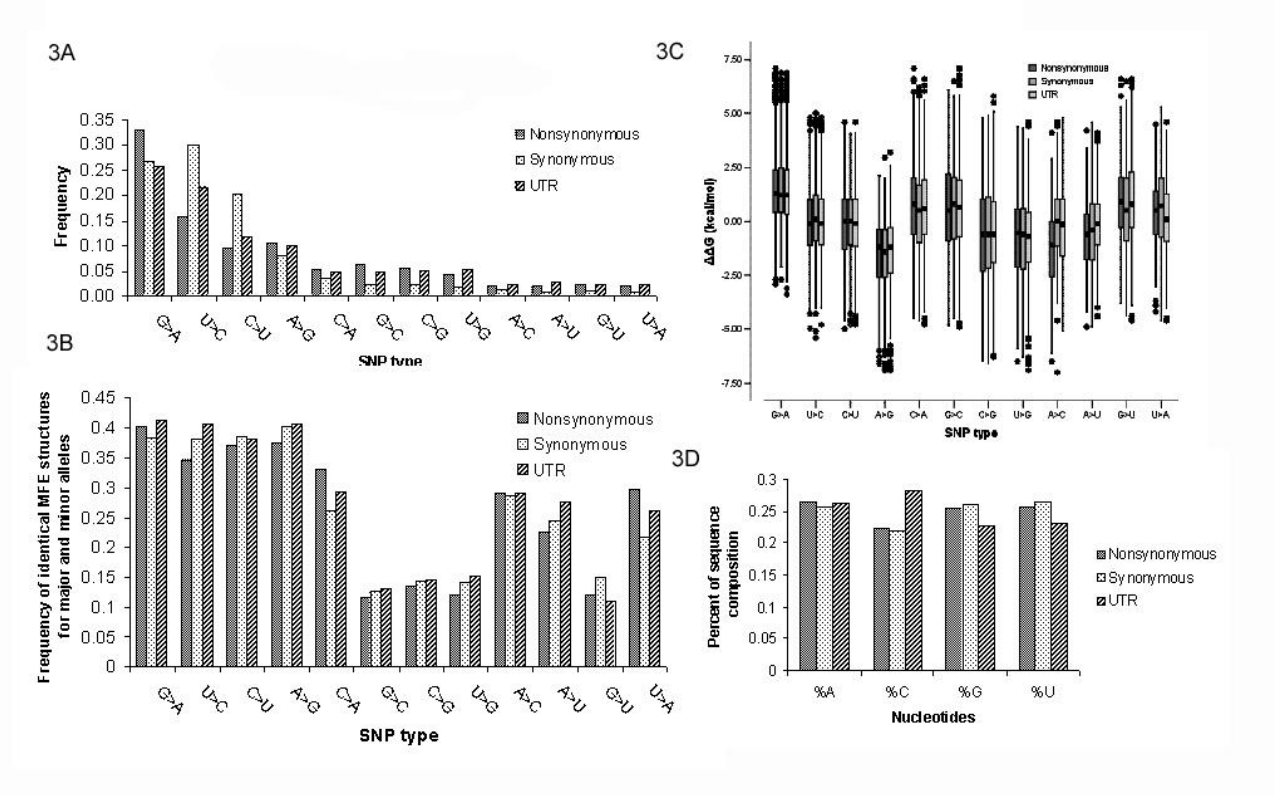
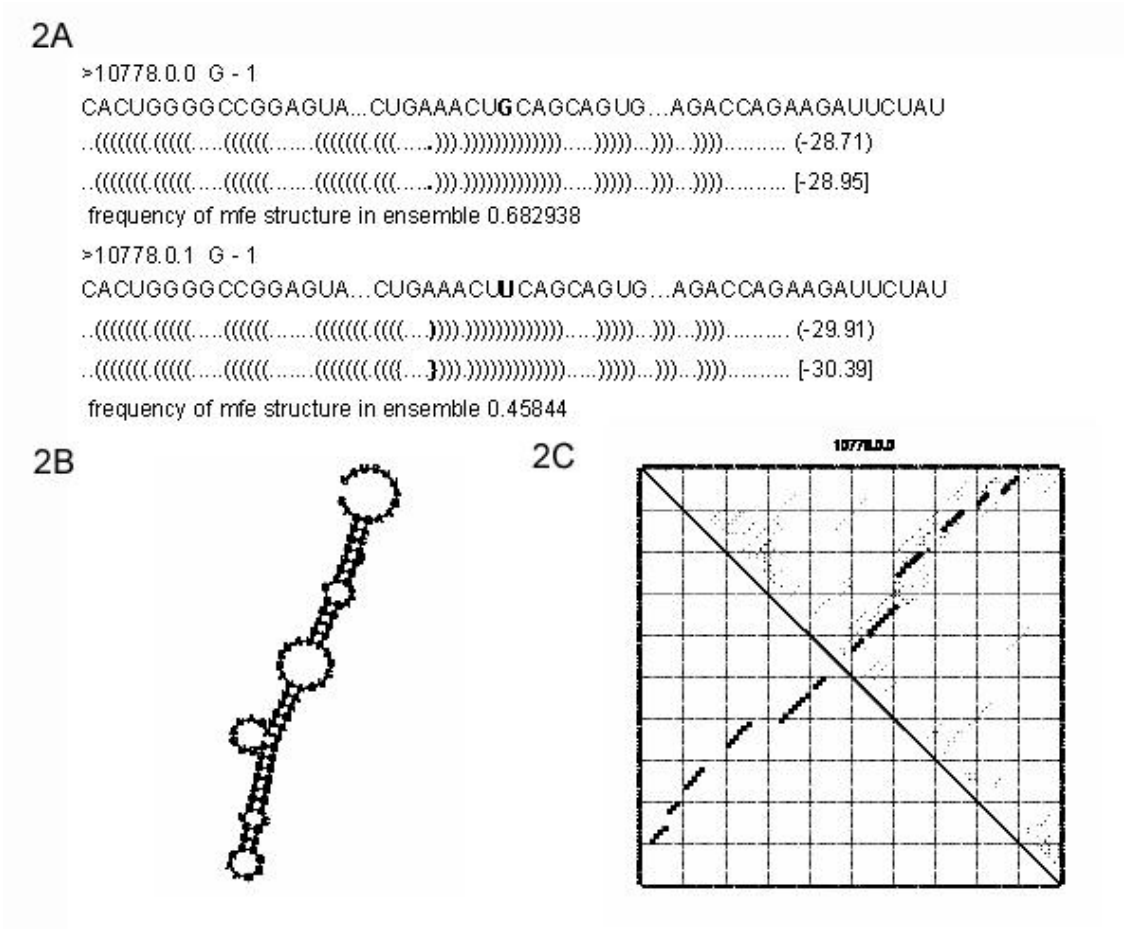


Illustration 3 - Relative frequencies in the dataset of 12 SNP types among functional categories for 34,557 validated human SNPs for which mRNA structures were analyzed. Panels A, B, and C are organized from the most common to least common SNP types (left to right) in human mRNAs. **A:**

Illustration 6

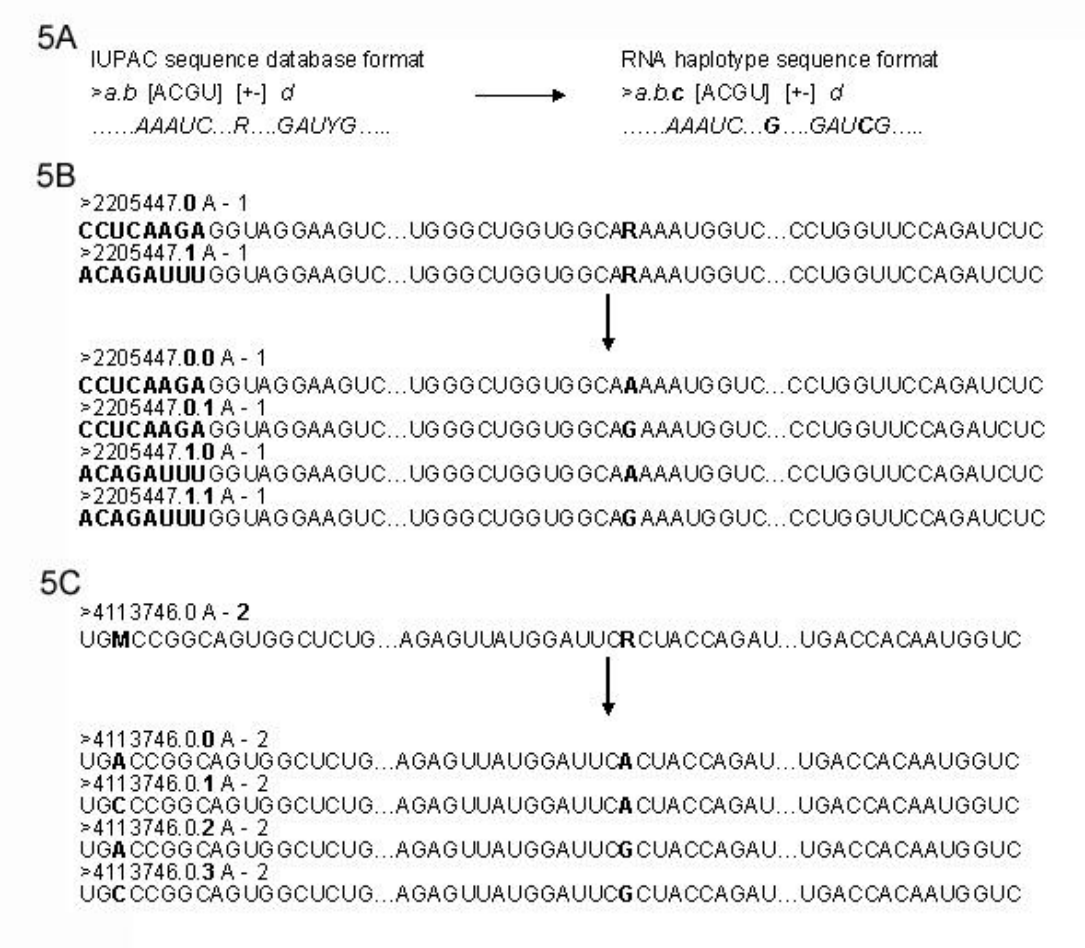
Example mRNA secondary structure results for a single SNP (rs10778) with two alleles.



depiction of the MFE structure and a dotplot representation (**C**) of the structure ensemble for rs10778 (G allele). MFE structures can be reconstituted from bracket notations.

Illustration 7

Generalized view of IUPAC and mRNA haplotype databases.



Disclaimer