LETTER

doi:10.1038/nature12946

Landscape and variation of RNA secondary structure across the human transcriptome

Yue Wan[1,2]*, Kun Qu[1]*, Qiangfeng Cliff Zhang[1], Ryan A. Flynn[1], Ohad Manor[3], Zhengqing Ouyang[1]†, Jiajing Zhang[1], Robert C. Spitale[1], Michael P. Snyder[4], Eran Segal[3] & Howard Y. Chang[1]

In parallel to the genetic code for protein synthesis, a second layer of information is embedded in all RNA transcripts in the form of RNA structure. RNA structure influences practically every step in the gene expression program[1]. However, the nature of most RNA structures or effects of sequence variation on structure are not known. Here we report the initial landscape and variation of RNA secondary structures (RSSs) in a human family trio (mother, father and their child). This provides a comprehensive RSS map of human coding and non-coding RNAs. We identify unique RSS signatures that demarcate open reading frames and splicing junctions, and define authentic microRNA-binding sites. Comparison of native deproteinized RNA isolated from cells versus refolded purified RNA suggests that the majority of the RSS information is encoded within RNA sequence. Over 1,900 transcribed single nucleotide variants (approximately 15% of all transcribed single nucleotide variants) alter local RNA structure. We discover simple sequence and spacing rules that determine the ability of point mutations to impact RSSs. Selective depletion of 'riboSNitches' versus structurally synonymous variants at precise locations suggests selection for specific RNA shapes at thousands of sites, including 3′ untranslated regions, binding sites of microRNAs and RNA-binding proteins genome-wide. These results highlight the potentially broad contribution of RNA structure and its variation to gene regulation.

We performed parallel analysis of RNA structure[2] (PARS) on RNA isolated from lymphoblastoid cells of a family trio (Fig. 1a). Deep sequencing of RNA fragments generated by RNase V1 or S1 nuclease (Extended Data Fig. 1a) determined the double or single-stranded regions, respectively, across the human transcriptome. We obtained over 160-million mapped reads for each individual. Transcript abundance and structure profiles are highly correlated among the individuals (Extended Data Fig. 2a, b). Summation of PARS data from the trio produced structural information for >20,000 transcripts with at least 1 read per base (load ≥ 1, Fig. 1b), and accurately identified known RSSs in RNAs (Fig. 1c and Extended Data Fig. 1b, c). We also developed methods for RNA extraction, deproteinization, and PARS under native conditions (native deproteinized samples) that accurately captured structures with known RSS, and revealed RSS for 6,524 transcripts (Extended Data Fig. 3a–d).

PARS data for thousands of transcripts afforded a genome-wide view of the structural landscape of human messenger RNAs. Metagene analysis shows that, on average, the coding region (CDS) is demarcated by focally accessible regions near the translational start site and stop codon. Contrary to yeast, human CDS is slightly more single-stranded than the untranslated regions (UTRs) (Fig. 1d), similar to previous trends in other metazoans[3]. A three-nucleotide structure periodicity is present in the CDS and absent in UTRs, consistent with prior computational prediction[4]. Both renatured and native mRNAs showed similar RSS features, suggesting that RNA sequence is a strong determinant of RSS. However, RNA structures also deviate from sequence content. In particular, human 3′ UTR has low GC content but is highly structured (Fig. 1d). We also identified 583 (5.7%) consistently different regions between native deproteinized and renatured structure profiles, providing candidate sites for regulation of RNA structure *in vivo* (Supplementary Table 1). Highly structured RNAs have fewer structure differences as compared to mRNAs (Extended Data Fig. 3e), suggesting stronger evolutionary selection for functional conformations. We note that 3.7% of bases (residing in 9.7% of transcripts) have both strong V1 and S1 reads, indicating the existence of multiple mRNA conformations.

We detected unique signatures of RSSs at sites of post-transcriptional regulation. RNA structure is believed to be important in regulating distinct splicing signals on exons and introns of pre-messenger RNAs[5]. We observed a unique asymmetric RSS signature at the exon–exon junction in both renatured and native deproteinized transcripts that is not simply explained by GC content. The terminal AG dinucleotide at the end of the 5′ exon tends to be more accessible, whereas the first nucleotides of the 3′ exon are more structured (Fig. 2a and Extended Data Fig. 3f). Hence, a specific RSS signature may contribute to RNA splicing.

Regulation of mRNAs by microRNAs (miRNAs) is an important post-transcriptional process that causes translation repression and/or mRNA degradation[6]. However the extent to which structural accessibility drives productive miRNA targeting is still unclear. Analysis of RSS from renatured RNA around predicted miRNA targets revealed that true Argonaute (AGO)-bound target sites[7] show strong structural accessibility from −1 to 3 nucleotides upstream of the miRNA-target site compared to predicted targets not bound by AGO ($P < 10^{-10}$, Wilcoxon rank-sum test; Fig. 2b, orange window, and Extended Data Fig. 4a). AGO-bound sites are also more accessible at bases 4 to 6 of the miRNA-target site ($P = 0.004$, Wilcoxon rank-sum test), agreeing with prior computational predictions[8]. To test whether our identified 5′ accessibility neighbourhood (−1 to 3 nucleotides) is truly important for AGO binding, we performed AGO individual nucleotide-resolution crosslinking and immunoprecipitation (iCLIP) on each member of the trio. Separating the predicted target sites according to average 5′ structural accessibility showed that single-stranded targets are more likely to be AGO-bound than double-stranded targets (Fig. 2c and Extended Data Fig. 4b). The most significant difference in AGO binding occurs close to our identified accessible region ($P = 0.01$, Fig. 2d). Separating predicted targets into five accessibility quantiles also demonstrated that the most accessible 20% of predicted targets are most AGO bound ($P < 10^{-19}$, Fig. 2e). Furthermore, ectopic expression of miR142 or miR148 in HeLa cells[9] resulted in greater repression of mRNAs with the 100 most accessible sites as compared to mRNAs with the 100 least accessible sites ($P < 0.005$, Wilcoxon rank-sum test; Fig. 2f and Extended Data Fig. 4c, d). This indicates that mRNAs with accessible miRNA sites are more likely to be true targets, and upstream accessibility is important for miRNA targeting.

[1]Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Stem Cell and Development, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672. [3]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovet 76100, Israel. [4]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. †Present address: The Jackson Laboratory for Genomic Medicine, 263 Farmington Avenue, ASB Call Box 901 Farmington, Connecticut 06030, USA. *These authors contributed equally to this work.

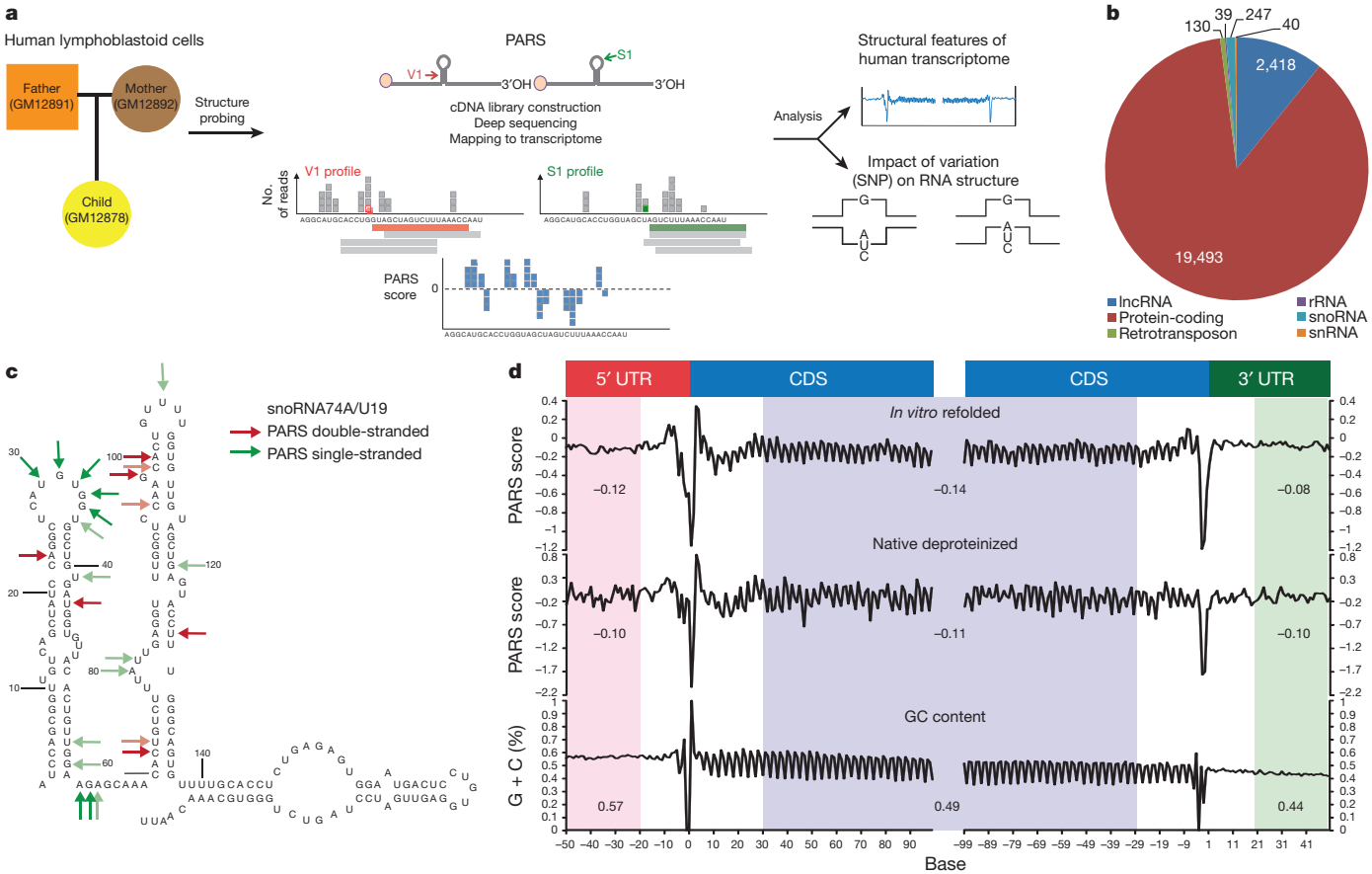

Figure 1 | PARS reveals the landscape of human RNA structure.
a, Experimental overview. Circles represent females, squares represent males. **b**, Pie chart showing the distribution of structure-probed RNAs with a coverage of at least one read per base. **c**, High (red arrows) and low (green arrows) PARS scores were mapped onto the secondary structure of small nucleolar RNA snoRNA74A. Red (positive PARS score), double-stranded regions by PARS score; green (negative PARS score), single-stranded regions by PARS score. The colour intensity reflects the magnitude of the PARS scores. Darker red and darker green, reflect more positive and more negative PARS scores (double- and single-stranded regions), respectively. **d**, PARS score (top, renatured transcripts; middle, native deproteinized transcripts) and GC content (bottom) across the 5′ UTR, the coding region, and the 3′ UTR, averaged across all transcripts, aligned by translational start and stop sites. Averaged regions are shaded in pink, blue and green for 5′ UTR, CDS and 3′ UTR, respectively.

Comparison of RNA structural landscapes between individuals revealed the impact of diverse sequence variants on RNA structure. As a class, local PARS score differences at single nucleotide variants (SNVs) were significantly greater than biological replicates of an invariant doped in RNA ($P < 0.001$ Kolmogorov–Smirnov test; Extended Data Fig. 5a). SNVs that alter RNA structure, known as 'riboSNitches', also exhibit threefold greater local structure change than replicates of the same sequence in different individuals (Extended Data Fig. 5b). At a gene level, transcripts with SNVs are significantly more disrupted, calculated using the experimental structure disruption coefficient (eSDC)[10], than transcripts without SNVs ($P = 1.3 \times 10^{-4}$, Kolmogorov–Smirnov Test; Extended Data Fig. 5c, d). Furthermore, 78.2% of all structure changing bases lie in transcripts that contain either SNVs or indels, suggesting that sequence variation is important in shaping RSS variation in the human transcriptome (Extended Data Fig. 5e). The list of the top 2,000 disrupted transcripts is shown in Supplementary Table 2.

To pinpoint riboSNitches[11], we calculated structure changes between each pair of individuals (Fig. 3a) and selected SNVs that had large PARS score differences, low false discovery rate (FDR), significant P value, and high local read coverage (Methods). Permutation analysis across genotypes and along transcripts confirmed that riboSNitches are significantly detected over random noise (Methods). We experimentally validated nine riboSNitches using independent structure probing methods such as nucleases, selective 2′ hydroxyl acylation and primer extension (SHAPE) or dimethyl sulphate (DMS), and confirmed the ability of PARS to discover riboSNitches (Extended Data Figs 6–9). The SeqFold program is used to visualize structure changes caused by riboSNitches[12] (Fig. 3b, c and Extended Data Fig. 7g, h).

We found that 1,907 out of 12,233 (15%) SNVs switched RNA structure in the trio (Fig. 3d, Extended Data Fig. 5e and Supplementary Table 3). As riboSNitches are expected to cause RSS changes in a heritable and allele-specific fashion, we performed allele-specific PARS in the cell line derived from the child by mapping uniquely across each of the two alleles for SNVs that are homozygous and different in the parents (for example, father AA and mother GG, with child AG when he or she inherits one copy from each parent) (Methods and Extended Data Fig. 6e). Out of 172 parental homozygous riboSNitches, 117 (68%) were validated by allele-specific mapping in the child. As only reads upstream of the riboSNitch can be uniquely mapped and detected, this is likely to be an underestimate. We also observed a validation rate of 61% in native deproteinized samples of the child, indicating that the structural changes are biologically relevant *in vivo* (Extended Data Fig. 9b).

The large numbers of riboSNitches identified raised the possibility that riboSNitches may have greater influence on gene regulation and human diseases than previously appreciated. Intersection with expression quantitative trait loci (eQTL) identified 211 riboSNitches that are associated with changes in gene expression (Supplementary Table 4). Overlapping riboSNitches with the NHGRI catalogue of genome-wide association studies identified 22 unique riboSNitches that are associated with diverse human diseases and phenotypes, including multiple

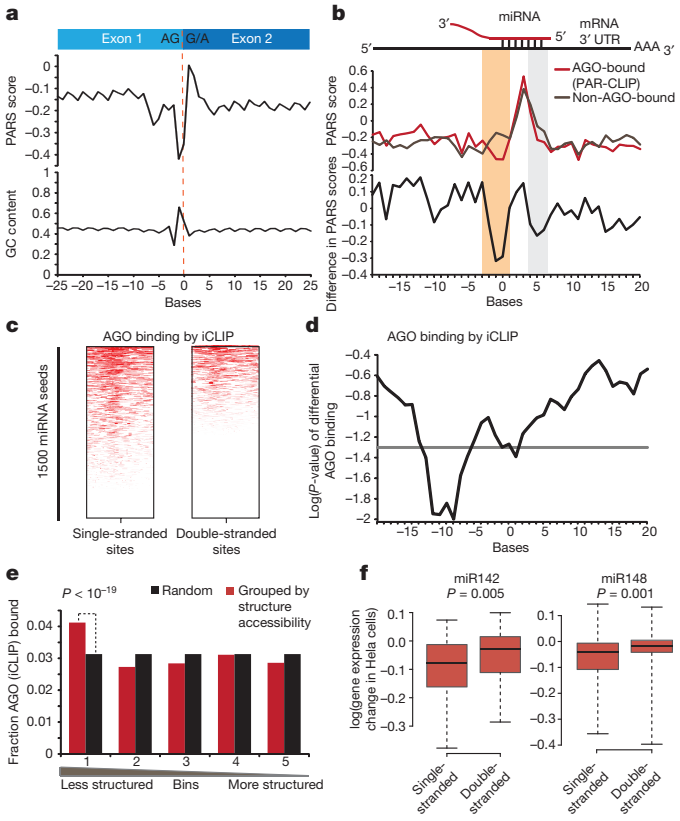

Figure 2 | RSS signatures of post-transcriptional regulation. **a**, Average PARS score and GC content across transcript exon–exon junctions. **b**, Average PARS score (top) and PARS score difference (bottom) across miRNA sites for AGO-bound (red) versus non-AGO-bound sites (grey). Structurally different regions are in orange and light grey. **c**, AGO-iCLIP binding for single- versus double-stranded miRNA target sites. **d**, *P* value for differential AGO-iCLIP binding (*t*-test, $P = 0.05$ in grey). **e**, Observed versus expected AGO binding (*P* value, chi-squared test). **f**, Expression changes of mRNAs with accessible and inaccessible miR142 (left) or miR148 (right) sites, upon miRNA overexpression (Wilcoxon rank-sum test).

sclerosis, asthma and Parkinson's disease (Supplementary Table 5). Hence, many non-coding changes in the transcriptome may alter gene function by altering RNA structure.

We also observed sequence and context rules in riboSNitches. First, riboSNitches that lie in double- or single-stranded regions tend to become more single- or double-stranded, respectively, after nucleotide change (Fig. 3e). Second, the nucleotide content of the riboSNitch is instructive of the direction of RSS change. Bases that undergo G/C to A/T changes tend to become more single-stranded, whereas bases that change from A/T to G/C tend to become more paired (Fig. 3f). This effect is stronger for homozygous riboSNitches than heterozygous riboSNitches, and typically disrupts 10 bases centred on the mutation. Third, the structural context flanking SNVs influence their transition to become more single- or double-stranded (Extended Data Fig. 10a–c). Fourth, riboSNitches have fewer SNVs around them as compared to non-structure changing SNVs, suggesting that co-variation of some SNVs may help to maintain functional RNA structures (Extended Data Fig. 10d).

The distribution of extant riboSNitches provides insights into regions of the transcriptome that require specific RNA shape. If an RSS is functionally important, a riboSNitch that disrupts the structure will be evolutionarily selected against, whereas a non-structure-changing SNV will not (Fig. 4a)[13]. We tested whether such selection occurs in the human transcriptome, and found that riboSNitches are significantly depleted at 3′ UTRs compared to control SNVs ($P < 10^{-20}$, chi-squared test; Fig. 4b). This depletion is even stronger for larger disruptions which would be expected to be less tolerated (Extended Data Fig. 10e). Additional genomic features associated with riboSNitches are also found (Extended Data Fig. 10f, Supplementary Table 6). RiboSNitches are also significantly depleted around predicted miRNA target sites ($P < 10^{-5}$, chi-squared test; Fig. 4c) and RNA binding protein (RBP) binding sites ($P = 0.004$, chi-squared test). However, depletion of riboSNitches varies for each individual RBP (Fig. 4d), suggesting that different RBPs may have different RSS requirements for binding. RiboSNitches may also influence gene regulation through splicing. Indeed, riboSNitches near splice junctions are associated with greater alternative splicing changes (defined as percentage spliced in (PSI)[14,15]; Fig. 4e), suggesting that RNA structures could regulate splicing.

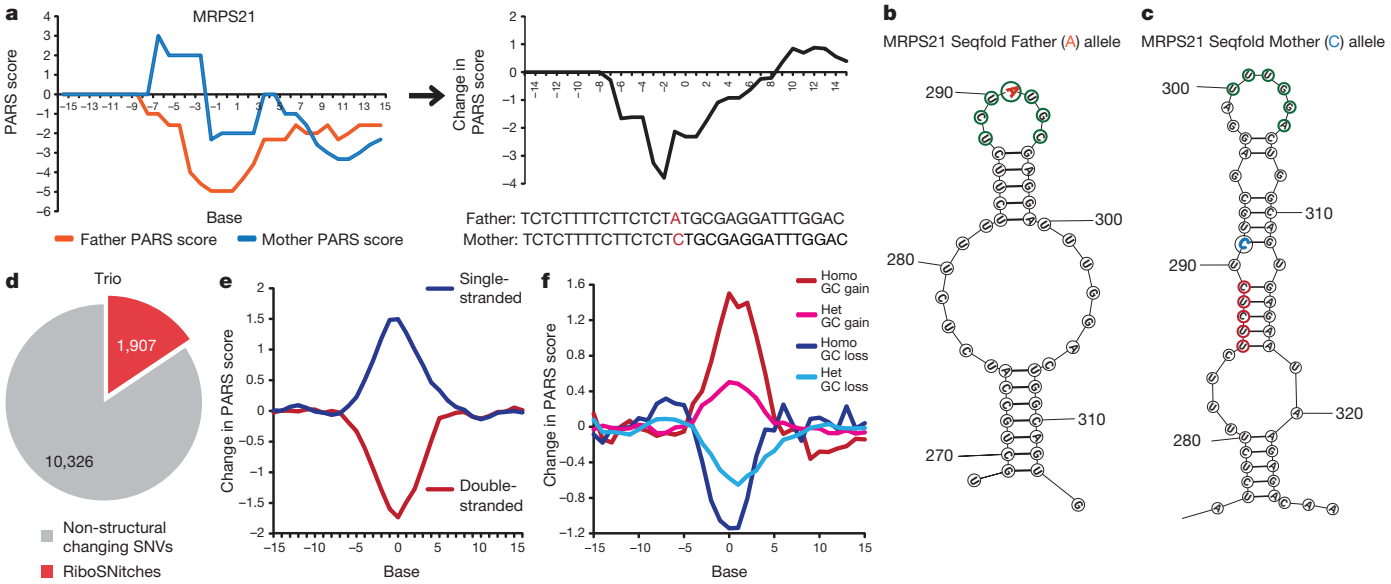

Figure 3 | PARS identifies riboSNitches genome-wide. **a**, PARS score (left) and PARS-score difference (right) of *MRPS21* father's and mother's alleles. **b**, **c**, SeqFold models of *MRPS21* A and C alleles (single-and double-stranded bases circled in green and red, respectively). **d**, Number of SNVs identified as riboSNitches in the trio. **e**, **f**, Average PARS score changes of riboSNitches that (**e**) originally reside in double-stranded (red) or single-stranded regions (blue); or (**f**) undergo nucleotide changes from A/T to G/C (red, pink) or from G/C to A/T (dark and light blue). 0 indicates the position of SNV on the *x* axis.

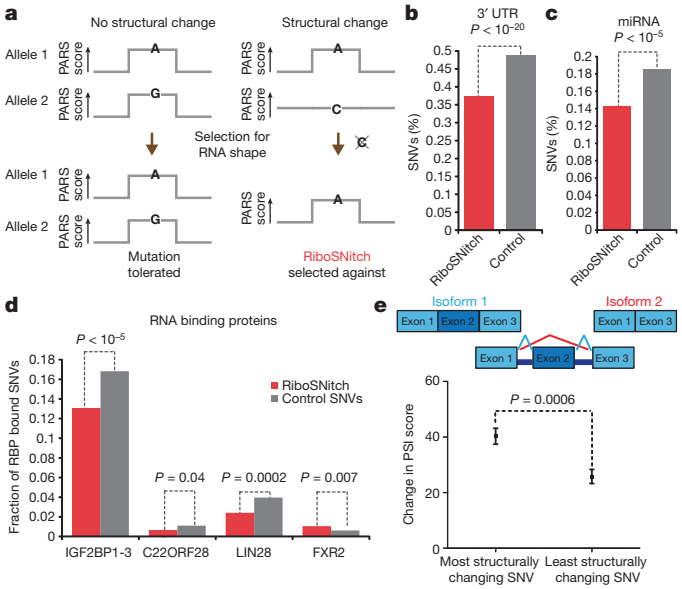

Figure 4 | Genetic evidence for functional RSS elements in the transcriptome. **a**, Schematic of RSS selection test: mutations that do not change the shape of an important RNA structure may be tolerated and accumulates (left), but a riboSNitch that changes RNA shape will be evolutionarily selected against and removed. Brown arrows, alleles that were present before and after selection for RNA shape. **b–d**, Selective depletion of riboSNiches versus structurally synonymous SNVs at 3′ UTRs (**b**); predicted miRNA target sites (**c**); specific RBP binding sites (**d**). *P* value is calculated using chi-square test. **e**, RiboSNitches impact splicing. PSI score is calculated to be the ratio of alternatively spliced isoform versus total isoforms (Methods), $P = 0.0006$, Student's *t*-test. Error bars show mean ± s.e.m.

In summary, the landscape and variation of RSS across human transcriptomes suggest important roles of RNA structure in many aspects of gene regulation. We provide the experimental and analytical frameworks to evaluate SNVs that change RSSs, and demonstrate potentially much broader roles for riboSNitches in multiple steps of post-transcriptional regulation. In the future, use of high resolution, *in vivo* probes of RSSs[16] and studies of many individuals of diverse genetic backgrounds may allow systematic determination of functional RSSs across the transcriptome.

METHODS SUMMARY

Sample preparation and structure probing for human renatured RNAs. Human lymphoblastoid cell lines GM12878, GM12891 and GM12892 were obtained from Coriell. Total RNA was isolated using TRIzol reagent (Invitrogen) and polyA selected as described previously[2]. Two micrograms of Poly(A)$^+$ RNA was structure probed at 37 °C using RNase V1 (Life Technologies, final concentration of 10^{-5} units per μl) or S1 nuclease (Fermentas, final concentration of 0.4 units per μl) at 37 °C for 15 min.

Sample preparation and structure probing for human native deproteinized RNAs. GM12878 cells were lysed in lysis buffer (150 mM NaCl, 10 mM $MgCl_2$, 1% NP40, 0.1% SDS, 0.25% Na deoxycholate, Tris pH 7.4) on ice for 30 min. The lysate was deproteinized by phenol chloroform extractions. Total RNA (1 μg per 90 μl) was incubated in 1 × RNA structure buffer at 37 °C for 15 min and structure probed using RNase V1 (final concentration of 2×10^{-5} units per μl) and S1 nuclease (final concentration of 0.2 units per μl) at 37 °C for 15 min.

Library construction and analysis. The structure probed RNA was cloned using Ambion RNA-Seq Library Construction Kit (Life Technologies)[2], and sequenced using Illumina Hi-seq. The reads were trimmed and mapped to UCSC RefSeq and the Gencode v12 databases (hg19 assembly) using the software Bowtie2 (ref. 17). Double (V1) and single-stranded reads (S1) for each sequencing sample were normalized by sequencing depth.

RiboSNitch analysis. Data normalization for each sample was performed by calculating standard deviation (s.d.) for each transcript and dividing the PARS score per base by the s.d. of that transcript. We defined a structure difference of the *i*th base of transcript *j* between conditions *m* and *n* in this formula, where PARS represents the normalized PARS score, abs represents absolute value, and *k* represents the *k*th base of the transcript:

$$\mathrm{StrucDiff}_{i,j,m,n} = \sum_{k=i-2}^{k=i+2} \frac{\mathrm{abs}(\mathrm{PARS}_{k,j,m} - \mathrm{PARS}_{k,j,n})}{5}$$

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 April; accepted 16 December 2013.

1. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nature Rev. Genet.* **12,** 641–655 (2011).
2. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
3. Li, F. *et al.* Global analysis of RNA secondary structure in two metazoans. *Cell. Rep.* **1,** 69–82 (2012).
4. Shabalina, S. A., Ogurtsov, A. Y. & Spiridonov, N. A. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* **34,** 2428–2437 (2006).
5. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
6. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–233 (2009).
7. Skalsky, R. L. *et al.* The viral and cellular microRNA targetome in lymphoblastoid cell lines. *PLoS Pathog.* **8,** e1002484 (2012).
8. Marín, R. M., Voellmy, F., von Erlach, T. & Vanicek, J. Analysis of the accessibility of CLIP bound sites reveals that nucleation of the miRNA:mRNA pairing occurs preferentially at the 3′-end of the seed match. *RNA* **18,** 1760–1770 (2012).
9. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27,** 91–105 (2007).
10. Ritz, J., Martin, J. S. & Laederach, A. Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics* **13,** (Suppl. 4) S6, (2012).
11. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* **6,** e1001074 (2010).
12. Ouyang, Z., Snyder, M. P. & Chang, H. Y. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.* 377–387 (2013).
13. Salari, R., Kimchi-Sarfaty, C., Gottesman, M. M. & Przytycka, T. M. Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. *Nucleic Acids Res.* **41,** 44–53 (2013).
14. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7,** 1009–1015 (2010).
15. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–1593 (2012).
16. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nature Chem. Biol.* **9,** 18–20 (2013).
17. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Chang laboratory, S. Rouskin, and J. Weissman, A. Mele and R. Darnell for discussion. This work is supported by NIH R01-HG004361 (H.Y.C. and E.S.). H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute.

Author Contributions H.Y.C. conceived the project; Y.W. and H.Y.C. developed the protocol and designed the experiments; Y.W. and R.A.F. performed experiments; Y.W., K.Q., Q.C.Z., O.M., Z.O., J.Z., R.C.S., M.P.S., E.S., and H.Y.C. planned and conducted the data analysis; Y.W., K.Q. and H.Y.C. wrote the paper with contributions from all authors.

Author Information Data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE50676. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.Y.C. (howchang@stanford.edu) and Y.W. (wany@gis.a-star.edu.sg).

METHODS

Sample preparation for renatured RNA structure probing. Human lymphoblastoid cell lines GM12878, GM12891 and GM12892 were obtained from Coriell. Total RNA was isolated from lymphoblastoid cells using TRIzol reagent (Invitrogen). Poly(A)$^+$ RNA was obtained by purifying twice using the MicroPoly(A)Purist kit (Life Technologies). The Tetrahymena ribozyme RNA was *in vitro* transcribed using the T7 RiboMax Large-scale RNA production system (Promega) and added into 2 μg of poly(A)$^+$ RNA (1% by mole) for structure probing and library construction.

Structure probing of renatured poly(A)$^+$ RNA. Two micrograms of Poly(A)$^+$ RNA in 160 μl of nuclease free water was heated at 90 °C for 2 min and snap-cooled on ice for 2 min. Twenty microlitres of 10 × RNA structure buffer (150 mM NaCl, 10 mM $MgCl_2$, Tris, pH 7.4) was added to the RNA and the RNA was slowly warmed up to 37 °C over 20 min. The RNA was then incubated at 37 °C for 15 min and structure probed independently using RNase V1 (Life Technologies, final concentration of 10^{-5} units per μl) or S1 nuclease (Fermentas, final concentration of 0.4 units per μl) at 37 °C for 15 min. The cleavage reactions were inactivated using phenol chloroform extraction.

Structure probing and ribosomal RNA depletion for native deproteinized RNA structure probing. GM12878 cells were lysed in lysis buffer (150 mM NaCl, 10 mM $MgCl_2$, 1% NP40, 0.1% SDS, 0.25% Na deoxycholate, Tris, pH 7.4) on ice for 30 min. The chromatin pellet was removed by centrifugation at 16,000*g* for 10 min at 4 °C. The lysate was deproteinized by passing through two phenol followed by one chloroform extractions. The concentration of RNA in the deproteinized lysate was measured using the Qubit fluorometer (Invitrogen). We diluted the RNA to a concentration of 1 μg per 90 μl using 1 × RNA structure buffer (150 mM NaCl, 10 mM $MgCl_2$, Tris, pH 7.4) and incubated the RNA at 37 °C for 15 min. The native deproteinized RNA was structure probed independently using RNase V1 (final concentration of 2×10^{-5} units per μl) and S1 nuclease (final concentration of 0.2 units per μl) at 37 °C for 15 min.

To compare structural differences between renatured and native deproteinized RNAs, we independently prepared an RNA sample that was similarly lysed and deproteinized. After removal of proteins, we ethanol precipitated the RNA and dissolved it in nuclease free water. We diluted the RNA to a concentration of 1 μg per 80 μl in water and heated the RNA at 90 °C for 2 min before snap-cooling the RNA on ice. We added 10 × RNA structure buffer and renatured the RNA by incubating it at 37 °C for 15 min and performed structure probing similarly as in native deproteinized RNAs.

The cleavage reactions were inactivated using phenol chloroform extraction and DNase treated before undergoing ribosomal RNA depletion using Ribo-Zero Ribosomal RNA removal kit (Epicentre).

Validation of riboSNitches by manual footprinting. We cloned approximately 200 nucleotide fragments of both alleles of *MRPS21, WSB1, HLA-DRB1, HLA-DQA1, hnRNP-AB, HLA-DRA, LDHA, XRCC5* and *FNBP1* from GM12878, GM12891 and GM12892 using a forward-T7-gene-specific primer and a reverse-gene-specific primer. All constructs were confirmed by sequencing using capillary electrophoresis. DNA from each of the different clones was then *in vitro* transcribed into RNA using MegaScript Kit from Ambion, following manufacturer's instructions.

Two picomoles of each RNA is heated at 90 °C for 2 min and chilled on ice for 2 min. 3.33 × RNA folding mix (333 mM HEPES, pH 8.0, 20 mM MgCl2, 333 mM NaCl) was then added to the RNA and the RNA was allowed to fold slowly to 37 °C over 20 min. The RNA was then structure probed with either DMS (final concentration of 100 mM) or 2-methylnicotinic acid imidazolide (NAI) (final concentration of 100 mM)[16] at 37 °C for 20 min or structure probed with S1 nuclease (final concentration of 0.4 units per μl) or RNase V1 (final concentration of 0.0001 units per μl) at 37 °C for 15 min. The DMS structure probed samples were quenched using 2-mercaptoethanol before phenol chloroform extraction. The NAI and nuclease treated samples were phenol chloroform extracted directly after structure probing. The structure probed RNA was then recovered through ethanol precipitation. The RNA structure modification/cleavage sites were then read out using a radiolabelled RT primer by running onto denaturing PAGE gel as described previously[18].

Library construction. The structure-probed RNA was fragmented at 95 °C using alkaline hydrolysis buffer (50 mM Sodium Carbonate, pH 9.2, 1 mM EDTA) for 3.5 min. The fragmented RNA was then ligated to 5′ and 3′ adapters in the Ambion RNA-Seq Library Construction Kit (Life Technologies). The RNA was then treated with Antarctic phosphatase (NEB) to remove 3′ phosphates before re-ligating using adapters in the Ambion RNA-Seq Library Construction Kit (Life Technologies). The RNA was reverse-transcribed using 4 μl of the RT primer provided in the Ambion RNA-Seq Library Construction Kit and polymerase chain reaction (PCR)-amplified following the manufacturer's instructions. We performed 18 cycles of PCR to generate the complementary DNA library.

Illumina sequencing and mapping. We performed paired end sequencing on Illumina's Hi-Seq sequencer and obtained approximately 400-million reads for each paired end lane in an RNase V1 or S1 nuclease library. Obtained raw reads were truncated to 50 bases, (51 bases from the 3′ end were trimmed). Trimmed reads were mapped to the human transcriptome, which consists of non-redundant transcripts from UCSC RefSeq and the Gencode v12 databases (hg19 assembly), using the software Bowtie2 (ref. 17). We allowed up to one mismatch per seed during alignment, and only included reads with perfect mapping or with Bowtie2 reported mismatches on positions annotated as SNVs in genetically modified cells. We obtained 166- to 212-million mapped reads for an RNase V1 or S1 nuclease sample.

PARS-score calculation. After the raw reads were mapped to the transcriptome, we calculated the number of double-stranded reads and single-stranded reads that initiated on each base on an RNA. The number of double (V1) and single stranded reads (S1) for each sequencing sample were then normalized by sequencing depth. For a transcript with N bases in total, the PARS score of its ith base was defined by the following formula where V1 and S1 are normalized V1 and S1 scores, respectively. A small number 5 was added to reduce the potential over-estimating of structural signals of bases with low coverage:

$$\mathrm{PARS}_{i=1\ldots N} = \log_2(\mathrm{V1}_i + 5) - \log_2(\mathrm{S1}_i + 5)$$

To identify structural changes caused by SNVs, we applied a 5-base average on the normalized V1 and S1 scores to smoothing the nearby bases' structural signals; therefore, the PARS score is defined as:

$$\mathrm{PARS}_{i=1\ldots N} = \log_2\left(\sum_{j=i-2}^{j=i+2}\frac{\mathrm{V1}_j+5}{5}\right) - \log_2\left(\sum_{j=i-2}^{j=i+2}\frac{\mathrm{S1}_j+5}{5}\right)$$

Bases with both high V1 and S1 scores, and transcripts with multiple conformations. Bases with both strong single- and double-strand signals are potentially present in multiple conformations. We first normalized all bases with detectable S1 or V1 counts by their sequencing depth. We then calculated an S1 ratio and a V1 ratio by normalizing S1 (and V1) counts to the transcript abundance. S1 and V1 ratios indicate the relative strength of single and double signals respectively. We then ranked all the bases by their S1 ratio and V1 ratio independently, and used the top one-million S1 ratio bases and the top one-million V1 ratio bases as high S1 ratio bases and high V1 ratio bases, respectively. We defined a base as being in multiple conformations if the base has both high S1 and high V1 ratios. If a transcript contains more than five multi-confirmation bases, this transcript is defined as a multi-confirmation transcript.

V1 replicates correlation analysis. Pearson correlation of RNase V1 replicates on GM12878 was performed using a parsV1 score (a value that uses the V1 score only to represent secondary structure) defined as:

$$\mathrm{parsV1}_{i=1..N} = \log_2(\mathrm{V1}_i + 5)$$

Structure differences between AGO PAR-CLIP bound and not bound transcripts. Predicted conserved and non-conserved miRNA target sites of conserved miRNA families were obtained from TargetScan[19]. AGO PAR-CLIP (photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation) data set in Epstein–Barr virus (EBV)-transformed lymphoblastoid cells was obtained from ref. 7. For 11 of the most abundant miRNAs that were expressed in the 4 lines of EBV transformed lymphoblastoid cells, we asked whether the predicted target site fell within the AGO CLIP clusters. Predicted target sites that resided within the PAR-CLIP clusters were considered as AGO-bound, whereas the rest were considered as non-AGO-bound. The non-AGO-bound transcripts are further controlled to fall within 25 and 75% of 3′ UTR length, mRNA abundance and CpG dinucleotide content of the AGO-bound transcripts. The PARS scores for AGO-bound and non-bound transcripts were aligned to the start (either −7 or −8 position of the miRNA) of the miRNA:target binding site and averaged. *P* values of structural changes were calculated using the Wilcoxon rank-sum test.

AGO-iCLIP library generation. AGO iCLIP was performed as described previously[20] with the following modifications: 2×10^7 genetically modified cells (per biological replicate) were collected under log-phase growth and washed once in ice-cold 1 × PBS. The pellet was resuspended in 10 × pellet volumes of ice-cold 1 × PBS and plated out on 10-cm tissue-culture dishes. Cells were crosslinked with ultraviolet radiation at 254nm for 0.3 J cm^{-2}, collected in ice-cold PBS and cell pellets were frozen on dry ice. Lysate preparation, RNaseA, and immunoprecipitation of AGO were performed as described previously[21] using the anti-AGO antibody (clone 2A8, Millipore). To produce iCLIP libraries, on-bead enzymatic steps and off-bead final-library preparation was performed as described previously[21]. AGO-iCLIP libraries were produced in biological duplicates for each individual (GM12891, GM12892 and GM12878), barcoded, and pooled for sequencing. Samples were single-end-sequenced for 75 bases on an Illumina HiSeq2500 machine.

Processing of AGO-iCLIP data. Raw sequencing reads were preprocessed using FASTX-Tookit before alignment was performed. Sequencing adaptor was trimmed off using fastx_clipper and low-quality reads were filtered using fastq_quality_filter. PCR duplicates were further removed using the program fastq_collapser. Preprocessed reads were aligned to hg19 genome assembly using Bowtie[22], and AGO–RNA crosslinking positions were obtained through self-generated script passing through the sequence alignment/map (SAM) file. AGO-RNA binding signal was smoothed by extending ± 10 bases around the crosslinking position, and signals from both replicates were normalized by sequencing depth. AGO–RNA per-base enrichment was defined as the minimum signal of the replicates divided by the corresponding RNA abundance.

To identify miRNA predicted sites for miRNAs that are expressed in GM12878 cells, we downloaded the small RNA sequencing data from the ENCODE consortium (GEO accession number GSM605625), and aligned the raw reads to the human miRNA database using Blastn. We estimated the amount of miRNA expression by counting the Blastn perfect matches for each miRNA. Predicted miRNA target sites from the top 100 highest expressed miRNA were then aligned to the miRNA–target binding sites and were separated into two groups: 0 predicted sites with an average PARS score of less than −1 (from −3 to 1 of the miRNA–target pair) were classified as single-stranded sites, whereas those with an average PARS score of greater than 1 (from −3 to 1 of the miRNA–target pair) were classified as double-stranded sites. We then calculated the average AGO-iCLIP enrichment score for the two groups of miRNA binding sites (from −25 to 25 bases), and estimated the significance of their difference using the Student's *t*-test.

miRNA-target downregulation in Hela cells. Average gene expression changes upon expression of miR142 or miR148 in HeLa cells were obtained from Grimson *et al.* by averaging the gene expression changes induced by the miRNA at 12 h and 24 h of overexpression[9]. For the miR142 or miR148 Targetscan predicted miRNA sites, we calculated the average PARS score across −3 to +1 (from the start of the miRNA–target pair) and sorted the predicted sites according to their structural accessibility. The *P* value for difference in downregulation of transcripts that contain the top 100 accessible sites versus transcripts that contain the bottom 100 accessible sites was calculated using Wilcoxon rank-sum test.

RiboSNitch analysis. RNAs with known secondary structures were doped into the initial RNA pool as positive controls to estimate the baseline changes in RNA structure in PARS. We calculated the PARS scores for all the bases in the transcripts and performed data normalization in order to compare directly secondary structures between different individuals. To normalize the data, we calculated the standard deviation (s.d.) for each transcript and divided the PARS score per base by the s.d. of that transcript. This resulted in a normal distribution of PARS scores for each transcript in each individual and enabled us to calculate the change in PARS scores due to SNVs by subtraction of PARS scores between the individuals. Since a true structure change is likely to extend beyond a single base, we define a structure difference of the *i*th base of transcript *j* between conditions *m* and *n* in this formula, where *PARS* represents the normalized PARS score:

$$\mathrm{StrucDiff}_{i,j,m,n} = \sum_{k=i-2}^{k=i+2} \frac{\mathrm{abs}(\mathrm{PARS}_{k,j,m} - \mathrm{PARS}_{k,j,n})}{5}$$

We calculated the StrucDiff for all the bases in all the transcripts between each pair of individuals: GM12891 and GM12892, GM12891 and GM12878, GM12892 and GM12878. To identify riboSNitches, we downloaded SNV annotations from HapMap project[23], and then converted SNV annotations from hg18 assembly to hg19 assembly using UCSC executable LiftOver. We then overlaid the hg19 SNV coordinates with our transcriptome annotation, a non-redundant combination of RefSeq and Gencode v12 transcriptome assembly, to identify the positions in the transcriptome that have SNVs. For highly confident detection of structural changes, we require that the sequencing coverage around SNV is dense, such that: first, the SNV is located on a transcript whose average coverage is greater than 1 (on average one read per base); and second, the average coverage in a 5-base window centred around the SNV is greater than 10 (average S1 + V1 ≥ 5). We exclude bases that fall within 100 nucleotides from the 3′ end of all the transcripts due to the blind tail of 100 nucleotides.

To identify SNVs with statistically significant changes in structure, we estimated a global baseline of structural change by calculating the fold differences between the doping control and SNV cumulative frequencies. We calculated a *z*-score for each detected SNV: $z = (\mathrm{StrucDiff} - \mathrm{mean})/(\text{s.d. of doped in controls})$. We used the Tetrahymena ribozyme as the doped in control. We noticed that a StrucDiff ≥ 1 is equivalent to a *z*-score ≥ 4.5 and a 100-fold difference between the SNV and doping control cumulative frequencies. To calculate the *P* value for the structural change at each detected SNV, we performed 1,000 permutations on the absolute values of the non-zero δ PARS scores within each transcript that contains SNV. This *P* value is an estimate of the likelihood that a 5-base average of the permutated PARS structural change is greater than the 5-base average of the SNV base's structural change. The false discovery rate (FDR) of the significance of the structural change at the SNV site is estimated by a multi-hypothesis testing performed using the p.adjust function in R. A SNV is defined as a riboSNitch if: first, its StrucDiff is greater than 1 (equivalent to *z*-score ≥ 4.5 and 100-fold cumulative frequency difference); second, its *P* value is less than 0.05 and FDR less than 0.1; and third, local read coverage greater than 10 and at least 3 out of 11 bases contain S1 or V1 signals in an 11-base sliding window centred by the SNV site. We also permutated the structural changes between the trio by shuffling the StrucDiffs within every transcript. After structural PARS scores were permutated, we identified only 16 riboSNitches based on the exact same aforementioned methods and thresholds. This number is less than 1% of the original number of riboSNitches found, indicating that most of the discovered riboSNitches are not random noise.

RiboSNitch noise and signal estimation. We estimated the amount of structural change between two replicates with the same sequence and compared it to the change in two replicates with differing sequences. For example, the father may have heterozygous alleles A and C at a particular locus, whereas the mother has the alleles C and C and the child has alleles A and C at the same locus. As the local genotype of the father is the same as that of the child, we can calculate the amount of structure change between that of the father and child (δ_1, noise). If this SNP was predicted to be a riboSNitch, then the local structural change between the father and mother (δ_2, signal) should be significantly greater than the noise. We took all the heterozygous riboSNitches we predicted that satisfy the above-mentioned pattern (861, 558 and 519 SNVs between three pairs of individuals in the trio), and calculated the absolute structure change in a 21-nucleotide window centred on the riboSNitch. Plotting signal (δ_2) and noise (δ_1) windows across these riboSNitches demonstrated that on average, the signal plot has threefold greater structure changes than that of the noise plot ($P = 7.94 \times 10^{-177}$, Student's *t*-test), indicating that the riboSNitches that we identified clearly distinguishes from the biological noise.

As a further control, we generated two additional biological replicates of PARS with RNase V1 from refolded RNA of the child, and obtained 70–110-million mapped reads for each sample. As expected, biological replicates of the same individual are better correlated than between individuals. No difference in variance was detected at riboSNitch neighbourhoods versus other sites, or when 5′ UTRs and CDSs were compared against 3′ UTRs. These results indicate that riboSNitches are not simply passenger mutations residing in structurally flexible or poorly measured regions.

Estimation of structural disruption at the gene level. The extent of structural disruption of a transcript is estimated by an eSDC (experimental structural disruption coeffiency) score that is defined as:

$$\mathrm{eSDC} = (1 - \mathrm{cc}) \times \sqrt{l}$$

where cc is a Pearson correlation of the transcript between two samples, and *l* is the length of that transcript[10]. The greater the eSDC is, the more disrupted the transcript is.

RiboSNitch allele-specific cross-validation. We first generated an allele-specific sequence reference for the lymphoblastoid cells by compiling 150-base sequence fragments (50 bases upstream and 100 bases downstream of the SNV) of both wild-type and mutant alleles. We then built Bowtie indexes using this reference, and mapped trimmed raw reads from GM12878 (child) to the indexes. We only accepted reads with perfect match to the wild-type or mutant sequences and calculated S1, V1 and PARS score as described above. We examined riboSNitches that were homozygous in both GM12891 (father) and GM12892 (mother), and that had both alleles detected as expressed in GM12878 (child). A riboSNitch is considered as cross-validated if the structural change between the two detected alleles in the child follows the same direction as the structural changes between the two alleles in the parents. Out of 184 homozygous riboSNitches in the parents, 117 of these riboSNitches can be cross-validated in the child (63.6%). Allele-specific cross-validation using the child's native deproteinized data was also performed as above.

RiboSNitch and microRNA RBP and splicing. Predicted miRNA-target sites (both conserved and nonconserved targets of conserved miRNA families) were downloaded from Targetscan. RBP clip data sets were downloaded from the doRiNA database[24]. In addition, CLIP sequencing data sets for *LIN28* were from ref. 25, and for *DGCR8* were from ref. 26.

RiboSNitch and splicing analysis. We defined a percent inclusion (percentage spliced in, PSI) value similarly to a previous paper[15]. We considered every internal exon in each annotated transcript as a potential 'cassette' exon. Each cassette alternative-splicing event is defined by three exons (C1, A and C2, where A is the alternative exon, C1 is the 5′ constitutive exon and C2 is the 3′ constitutive exon); two constitutive junctions (C1A (connecting exons C1 and A) and AC2 (connecting exons A and C2)); and one alternative (or 'skipped') junction (C1C2 (connecting

exons C1 and C2)). First, we constructed a reference library containing unique, non-redundant constitutive and alternative junction sequences that are based on exon annotations and their RNA sequences. These junction sequences were constructed such that there is a minimum five-nucleotide overlap between the mapped reads and each of the two exons involved. Each junction sequence was annotated with a gene name and exon indexes for downstream analysis. As we trimmed the sequencing raw reads to 50 bases, we created a junction sequence library, indexed using Bowtie-build[22], using junction sequences of 90 bases. We downloaded independent RNA sequencing data from the ENCODE consortium (GM12878, GM12891 and GM12892) to estimate the PSI differences between samples. Raw reads were trimmed to 50 bases and then aligned to the non-redundant junction sequences using Bowtie[22], with unique mapping (the -m option in Bowtie = 1) and allowing a maximum of two mismatches. The number of reads that were uniquely mapped to a junction sequence, corresponding to the junction's effective number of mappable reads, was calculated by an in-house generated script. We then counted the number of reads that were uniquely mapped to each junction C1A, AC2 and C1C2, respectively. The PSI value for each internal exon was defined as:

$$\mathrm{PSI} = 100 \times \frac{\mathrm{average(C1A,AC2)}}{\mathrm{C1C2} + \mathrm{average(C1A,AC2)}}$$

where C1A, AC2 and C1C2 are the normalized read counts for the associated junctions.

We calculated PSIs for all of the internal exons in the samples GM12891, GM12892 and GM12878 and calculated the change in PSI between each pair of samples. Out of 12,233 transcribed SNVs, 498 SNVs were found in internal exons with PSI differences in the trio, and 169 SNVs were located within 20 nucleotides of the splicing sites. We ranked these 169 SNVs by the degree of their structural changes (StrucDiff score), and found that the exons containing SNVs with higher StrucDiff scores (StrucDiff > 1) show greater PSI differences than those exons containing SNVs with lower StrucDiff scores (StrucDiff < 1).

RiboSNitch and local structure environments. We defined bases of PARS scores greater than 1 as double-stranded (D), PARS scores of less than -1 as single stranded (S), and PARS scores between -1 and 1 as poised region (.). Using these cutoffs, we classified local structures around a SNV site into different categories (for example, S.D, DDD), and the average PARS-score changes for riboSNitches under different local structure categories were analysed.

RiboSNitch and SNV densities in flanking regions. We calculated the average number of SNVs within a certain distance to a riboSNitches using SNV annotation from the 1000 Genome Project. We also made the same calculation on 2,450 non-structural changing SNV sites as negative control. We used the Kolmogorov–Smirnov test to determine whether the two distributions are significantly different.

RiboSNitches predicted by SeqFold using PARS scores. For each SNV we used SeqFold to predict RNA secondary structure for a transcript fragment of 151 nucleotides (50 nucleotides upstream to 100 nucleotides downstream of the SNV sites). We used the PARS scores from allele-specific mapping as input to SeqFold. We then compared the SeqFold predicted structures for the different alleles at the SNV site. Green and red circles indicate bases with PARS scores ≤ -1 and ≥ 1, respectively.

Enrichment of SNVs in genomic features. We compared different genomic features or annotations of 993 unique riboSNitches to 1,009 control SNVs. For each genomic annotation, the fraction of riboSNitches that are inside the genomic region covered by the annotation (for example, histone mark) was compared to the fraction of control SNVs by Student's *t*-test. The different genomic annotations were downloaded and compiled from various online resources (Supplementary Table 5). A cutoff value of $P = 0.05$ was used.

18. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1,** 1610–1616 (2006).
19. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
20. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* **460,** 479–486 (2009).
21. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Struct. Mol. Biol.* **17,** 909–915 (2010).
22. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
23. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
24. Anders, G. *et al.* doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.* **40,** D180–D186 (2012).
25. Wilbert, M. L. *et al.* LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Mol. Cell* **48,** 195–206 (2012).
26. Macias, S. *et al.* DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nature Struct. Mol. Biol.* **19,** 760–766 (2012).