2021-04-15

Understanding the effect of adaptive mutations on the three-dimensional structure of RNA

Justin Cook

Understanding the effect of adaptive mutations on the three-dimensional structure of RNA

Justin L. Cook[1]; Lauren Sugden, PhD[2]

[1]Department of Engineering, John G. Rangos, Sr. School of Health Sciences, Duquesne University

[2]Department of Mathematics And Computer Science, McAnulty College and Graduate School of Liberal Arts, Duquesne University

Submitted for Publication in the Duquesne University Scholarship Collection on May 7, 2021

ABSTRACT

Single-nucleotide polymorphisms (SNPs) are variations in the genome where one base pair can differ between individuals.[1] SNPs occur throughout the genome and can correlate to a disease-state if they occur in a functional region of DNA.[1] According to the central dogma of molecular biology, any variation in the DNA sequence will have a direct effect on the RNA sequence and will potentially alter the identity or conformation of a protein product. A single RNA molecule, due to intramolecular base pairing, can acquire a plethora of 3-D conformations that are described by its structural ensemble. One SNP, rs12477830, which was previously shown to harbor signatures of positive selection by Sugden et. al,[3] was passed through multiple RNA folding algorithms. The results of *SNPfold* [2] demonstrate that the SNP significantly alters the structural ensemble, and the significance of this change offers a potential explanation of SWIF(r)'s result.[3] Furthermore, the *RNAfold Webserver* [4-6] reveals that the mutant RNA molecule is more stable than the wild-type with a more negative free energy and a higher frequency. These loci of variation should be studied in order to understand the potentially induced conformational changes that could significantly alter the functional capacity of an RNA molecule. Future work aims to assess conformational changes elicited by SNPs previously shown to harbor signatures of positive

selection using ancestry-specific reference genomes to better understand motivations behind a locus experiencing positive selective pressure.

INTRODUCTION

The Central Dogma of Molecular Biology describes the hierarchy that the identity of a protein product is dictated by the sequence of the corresponding RNA, which is determined by the DNA. Thus, any alteration to the DNA sequence will alter the RNA sequence and has the potential to alter the protein product and lead to disease. Unlike DNA, RNA is single-stranded, and can adopt a 3-D conformation as a result of intramolecular base pairing.[7] This "fold" can exert a direct effect on the synthesis of a protein product by sequestering or exposing the binding sites of translational machinery. RNA molecules are composed of four nucleotides (Adenine, Guanine, Cytosine, and Uracil) that are able to form specific bonds with one another.[7] Many types of RNA molecules exist, and each type has a specific function ranging from catalysis, carriage of amino acids, regulation of protein synthesis, etc.[7]

SNPs are a type of variation in the genome where one base pair can differ between individuals.[1,7] SNPs occur throughout the genome and can correlate to a disease-state if they occur in a functional region of DNA.[1] Mutations in RNA can lead to altered protein concentrations and abnormal protein products.[7]

A multitude of structures exist that a single RNA molecule can acquire; a structural ensemble.[8] Given an RNA sequence, existing algorithms are capable of predicting the "minimum free energy" structure and the "centroid" structure, which represents the "mean" structure of the ensemble.[4-6] RNA folding algorithms exist according to various theories of RNA folding, including covariation/compensatory changes,[9] energy minimization,[10] and maximization of the total number of base pairs.[11] The locus of variation introduced by an SNP should be studied in order to understand the potentially induced conformational changes that could significantly alter the functional capacity of an RNA molecule.

We investigate a collection of SNPs previously shown to harbor signatures of positive selection according to the SWIF(r) algorithm presented by Sugden et al.[3] Genome-wide scans, such as SWIF(r),[3]

can predict SNPs that are adaptive within populations but interpreting why these SNPs are adaptive is difficult.

METHODS

After completing a comprehensive literature review to identify available RNA folding algorithms and quantitative metrics capable of describing RNA three-dimensional structure, the SNPs contained within 5'UTRs and 3'UTRs that were tagged by SWIF(r) [3] as being potentially under selective pressure were isolated. Using the *UCSC Genome Browser* [12] with the hg19 build of the human genome, the genomic coordinates of the region bounding the SNP were calculated with consideration for the sign of the DNA strand containing that SNP. Using those coordinates, the RNA sequences corresponding to the UTRs were extracted from the *UCSC Human Genome Browser*.[12] Two RNA sequences were saved for each UTR, one with the wild-type sequence and another with the mutant sequence defined by the substitution of the SNP.

The RNA sequences for each UTR were passed into the *SNPfold* [2] algorithm to obtain a preliminary assessment of the potential conformational change induced by the SNP. SNPs that elicited significant structural change were marked by a significant p-Value.[2] The RNA sequences for any UTRs containing an SNP found to significantly alter the conformation of the RNA molecule were passed into the *RNAfold Webserver*,[4-6] *MutaRNA*,[13-15] and *IPKnots* [16-21] in order to examine the specific structural differences predicted between the wild-type and the mutant.

RESULTS

SNP of Interest

The preliminary passage of RNA sequences containing SNPs tagged by SWIF(r)[3] through the *SNPfold*[2] algorithm identified one SNP that significantly altered the conformation of the RNA molecule corresponding to the UTR. This research examines one SNP, rs12477830, which is a mutation that substitutes an adenine with a guanine at the 42nd position (Table 1).[12] This SNP is found within the 5-prime UTR of a pseudogene.[12] SWIF(r) identified this SNP as harboring signatures of positive selection with a probability that the site contains an adaptive mutation of approximately 40%.[3]

SNP of Interest	
SNP Identifier[12]	rs12477830
SNP[12]	A42G
SNP Position in Genome[12]	Chromosome 2 ; 108938735
Population[3]	CHB + JPT
Gene[12]	SULT1C2P1
SNP Type[3,12]	5' UTR
Strand[12]	Positive
Genomic Coordinates of UTR[12]	108938694 – 108938839
Length of UTR[12]	145
SNP Position within UTR[12]	42
SWIF(r) Calibrated P(sweep)*[3]	0.40095325
* Value can be "interpreted directly as the probability that a site contains an adaptive mutation."[3]	

Table 1. Details of SNP rs12477830.[3,12]

SNPfold [2] *Algorithm*

The RNA sequence of interest was initially passed into the *SNPfold* [2] algorithm to assess the significance of any conformational changes to the RNA introduced by the SNP. This algorithm generates partition function matrices where the color of each point within the plot represents the base-pairing probability according to the heatmap legend (Figure 1).[2] The summary statistics outputted by the algorithm are presented in Table 2.[2]

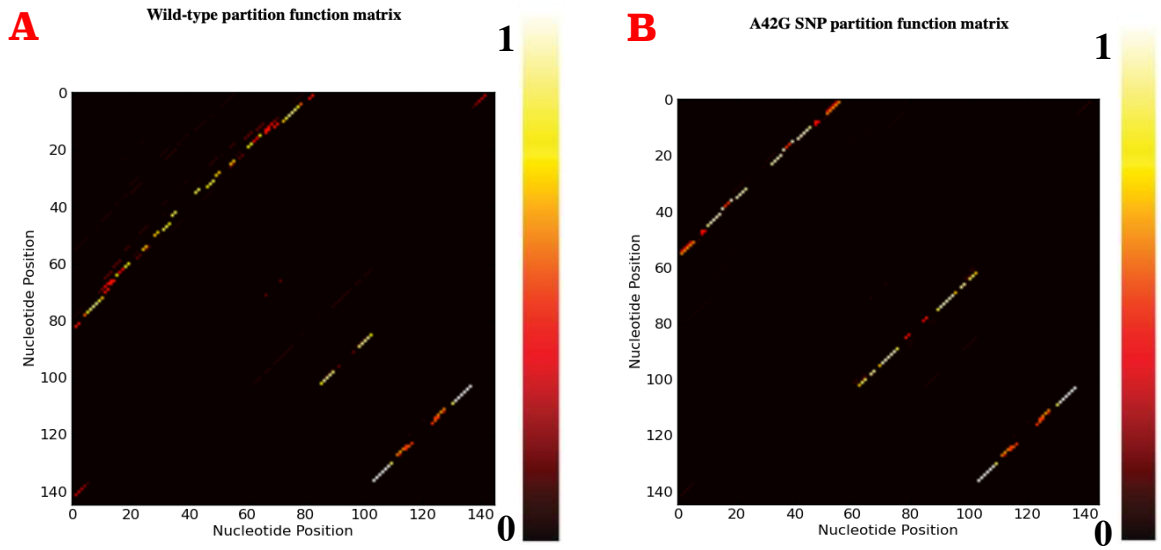

Figure 1. Partition function matrices. Point color represents the base-pairing probability according to the heatmap.[2] (**A**) Wild-type RNA molecule.[2] (**B**) Mutant RNA molecule.[2]

Correlation Coefficient [2]	0.3852074
Rank [2]	2/450
p-Value [2]	0.0044

Table 2. Summary statistics from *SNPfold* algorithm.[2]

RNAfold Webserver [4-6]

The RNA sequence of interest was passed into the *RNAfold Webserver*.[4-6] This algorithm predicts minimum free energy (MFE) secondary structures (Figure 2) for RNA sequences.[4-6] The color of each nucleotide corresponds to its base-pairing probability.[4-6]

The *RNAfold Webserver* also outputs mountain plots for RNA molecules (Figure 3).[4-6] According to the supporting documentation for the algorithm, "These plots represent a secondary structure in a plot of height versus position where height is a function of the number of base pairs enclosing a base. Loops are represented by plateaus, hairpin loops are represented by peaks, and helices are represented by slopes." [4-6] The summary statistics outputted by the algorithm are presented in Table 3.[4-6]

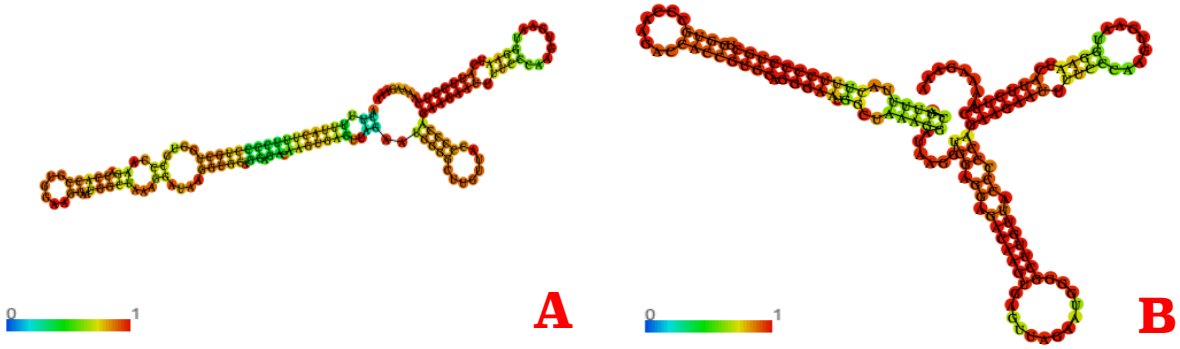

Figure 2. Predicted MFE secondary structures for RNA.[4-6] Nucleotide color corresponds to the base-pairing probability according to the heatmap.[4-6] (**A**) Wild-type RNA molecule.[4-6] (**B**) Mutant RNA molecule.[4-6]

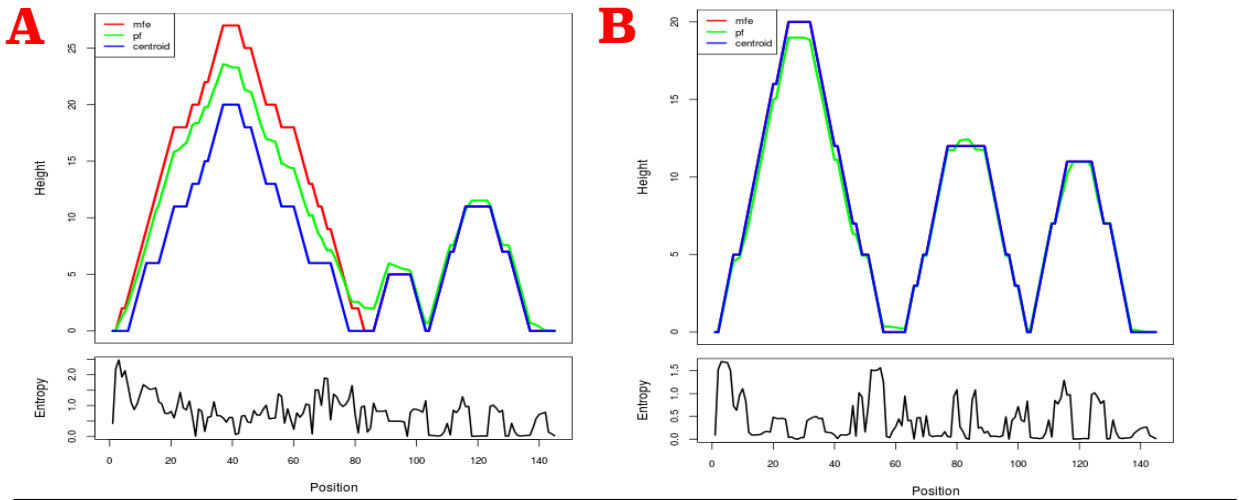

Figure 3. Mountain plots of RNA three-dimensional structures.[4-6] (**A**) Wild-type RNA molecule.[4-6] (**B**) Mutant RNA molecule.[4-6]

	Free Energy of Thermodynamic Ensemble (kcal/mol)	Frequency of MFE Structure (%)	Ensemble Diversity
Wild-type [4-6]	-38.68	1.53	26.73
Mutant [4-6]	-42.13	6.00	14.02

Table 3. Summary statistics from *RNAfold Webserver*.[4-6]

MutaRNA [13-15]

The RNA sequence of interest was passed into *MutaRNA* to display another form of representing RNA folding and the changes potentially elicited by a mutation.[13-15] This algorithm utilizes circular plots (Figure 4) to graphically represent the conformation of an RNA molecule.[13-15] In this plot-type, darker lines correlate to higher base pairing probabilities.[13-15] Additionally, the red rectangle in (**B**) designates the genomic position of the SNP.[13-15]

MutaRNA also outputs an arc plot (Figure 5) that describes the change in base-pairing probabilities elicited by the SNP mutation with weakened interactions on the top and strengthened interactions on the bottom.[13-15]

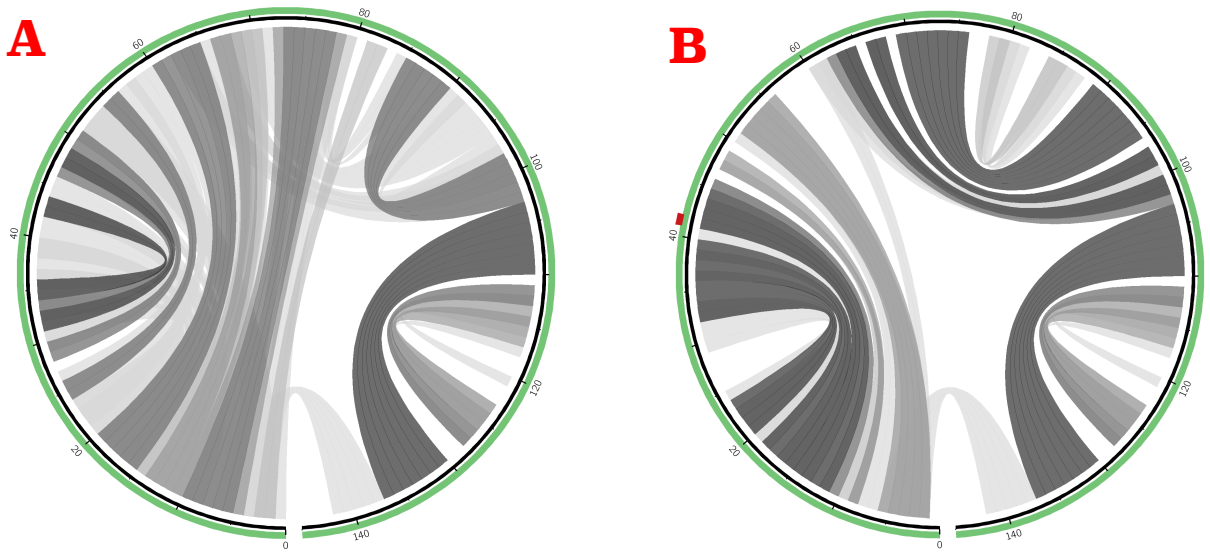

Figure 4. (**A** and **B**) Circular plots of wild-type RNA and mutant RNA molecules, respectively.[13-15] Higher base pairing probabilities are represented by darker lines.[13-15] The red rectangle in (**B**) designates the SNP's position.[13-15]

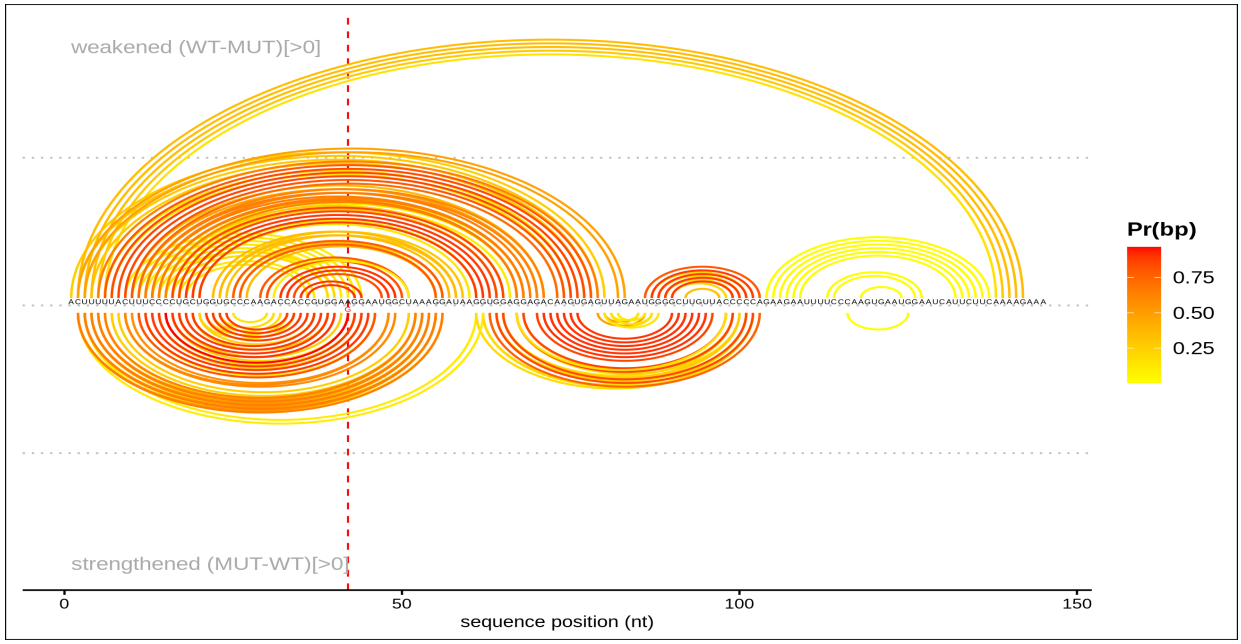

Figure 5. Arc plot describing the change in base-pairing probabilities elicited by the SNP mutation with weakened interactions on the top of the plot and strengthened interactions on the bottom of the plot.[13-15]

IPKnots [16-21]

In recognizing that many algorithms prohibit pseudoknots, the RNA sequence was passed into *IPKnots*, which is a folding algorithm that allows for this specific type of structural motif.[16-21] The results are shown in Figure 6.[16-21]

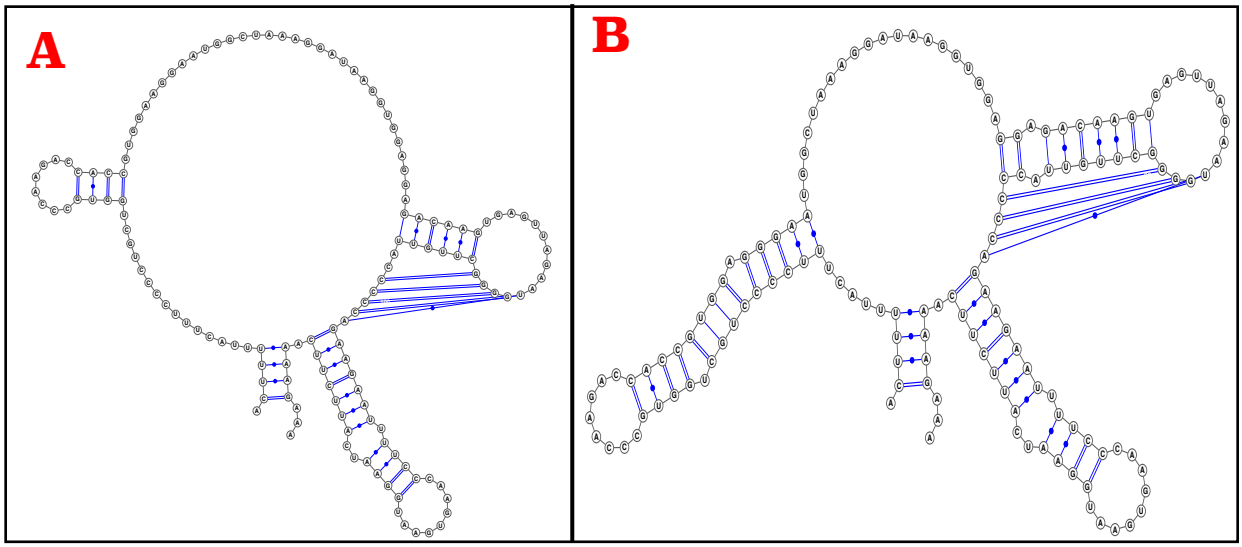

Figure 6. Predicted RNA secondary structures generated with consideration for pseudoknots.[16-21] (**A**) Wild-type RNA molecule.[16-21] (**B**) Mutant RNA molecule.[16-21]

ANALYSIS

Looking at the overall statistics outputted by the *RNAfold Webserver,* it is apparent that this mutation is predicted to make the RNA more stable, as the free energy decreases and the MFE structure represents a larger percentage of the ensemble, thus decreasing the ensemble diversity.[4-6] Note the differences in the predicted conformation; specifically the lengthening of the stem-loop motif inferior to the central loop.

To better understand the results presented in the partition matrices of *SNPfold*, we can look to the algorithm's summary statistics.[2] This algorithm yields a small p-Value,[2] which can be interpreted as a significant change in the structure was elicited by the mutation. Moreover, the correlation coefficient supports this finding, as a lower value corresponds to a greater disruption in folding.[2]

Overall, *MutaRNA* offers a unique means of visualizing and understanding RNA three-dimensional structure,[13-15] and *IPKnots* reveals that the pseudoknot predicted is relatively constant between the WT and mutant.[16-21]

CONCLUSION

The SNP, rs12477830, harbors signatures of positive selection according to SWIF(r) [3] and significantly alters the structural ensemble of RNA secondary structures as evidenced by the p-Value < 0.05 calculated by *SNPfold*.[2] The significance of the structural change elicited by the SNP offers a potential explanation of SWIF(r)'s result.[2,3] The *RNAfold Webserver* reveals that the mutant RNA molecule is more stable than the wild-type with a more negative free energy and a higher frequency.[4-6]

The SNP of interest is located within the sulfotransferase family 1C member 2 pseudogene 1, SULT1C2P1, which is responsible for the phenotype of eyebrow thickness [22] and has been implicated in the development of a primitive neuroectodermal tumor/medulloblastoma in the central nervous system post translocations.[23-26] An understanding of the RNA structural change elicited by an SNP can further medical knowledge and offer insight into potential treatment approaches.

FUTURE WORK

Investigative efforts are ongoing to explore whether SNPs that were previously found to not significantly alter the structural ensemble of an RNA molecule will have different predicted effects when considered in the context of the appropriate ancestral reference genome. Given that the "background" genome of an individual from Africa, for example, can vary from that of an individual from Europe, the influence of an SNP could be much more pronounced when the locus is surrounded by varying nucleotides.

REFERENCES

1. Single Nucleotide Polymorphisms (SNPs). National Human Genome Research Institute. http://www.genome.gov/genetics-glossary/Single-Nucleotide-Polymorphisms. Accessed May 8, 2021.
2. Halvorsen M, Martin JS, Broadaway S, Laederach A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet*. 2010;6(8):e1001074. Published 2010 Aug 19. doi:10.1371/journal.pgen.1001074
3. Sugden LA, Atkinson EG, Fischer AP, Rong S, Henn BM, Ramachandran S. Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nat Commun*. 2018;9(1):703. Published 2018 Feb 19. doi:10.1038/s41467-018-03100-7
4. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*. 2011;6:26. Published 2011 Nov 24. doi:10.1186/1748-7188-6-26
5. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA websuite. *Nucleic Acids Res*. 2008;36(Web Server issue):W70-W74. doi:10.1093/nar/gkn188
6. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA

secondary structure. *Proc Natl Acad Sci U S A*. 2004;101(19):7287-7292. doi:10.1073/pnas.0401799101

7. Iwasa J, Marshall W. *Karp's Cell and Molecular Biology: Concepts and Experiments*. 8th ed. Wiley; 2016.
8. Ritz J, Martin JS, Laederach A. Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics*. 2012;13(Suppl 4):1-11. doi:10.1186/1471-2164-13-S4-S6
9. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. *Nat Rev Genet*. 2011;12(9):641-655. Published 2011 Aug 18. doi:10.1038/nrg3049
10. Barash D, Churkin A. Mutational analysis in RNAs: comparing programs for RNA deleterious mutation prediction. *Brief Bioinform*. 2011;12(2):104-114. doi:10.1093/bib/bbq059
11. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*. 1980;77(11):6309-6313. doi:10.1073/pnas.77.11.6309
12. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006. doi:10.1101/gr.229102
13. Miladi M, Raden M, Diederichs S, Backofen R. MutaRNA: analysis and visualization of mutation-induced changes in RNA structure. *Nucleic acids research*. 2020;48(W1):W287-W291. doi:10.1093/nar/gkaa331
14. Raden M, Ali SM, Alkhnbashi OS, et al. Freiburg RNA tools: a central online resource for RNA-focused research and teaching. *Nucleic Acids Res*. 2018;46(W1):W25-W29. doi:10.1093/nar/gky329
15. *Freiburg RNA Tools Version (4.8.4)*, rna.informatik.uni-freiburg.de/.
16. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*. 2011;27(13):i85-i93. doi:10.1093/bioinformatics/btr215

17. Kato Y, Sato K, Asai K, Akutsu T. Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming. *Nucleic Acids Res*. 2012;40(Web Server issue):W29-W34. doi:10.1093/nar/gks412
18. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. *RNA*. 2010;16(12):2304-2318. doi:10.1261/rna.1950510
19. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*. 2006;22(14):e90-e98. doi:10.1093/bioinformatics/btl246
20. Zadeh JN, Steenberg CD, Bois JS, et al. NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem*. 2011;32(1):170-173. doi:10.1002/jcc.21596
21. Kévin Darty, Alain Denise, Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*. 2009;25(15):1974. doi:10.1093/bioinformatics/btp250
22. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120
23. Huret JL, Ahmad M, Arsaban M, et al. Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic Acids Res*. 2013;41(Database issue):D920-D924. doi:10.1093/nar/gks1082
24. Huret J-L, Dessen P, Bernheim A. An internet database on genetics in oncology. *Oncogene*. 2003;22(13). doi:10.1038/sj.onc.1206225
25. Pearson H. Lifelines: browsing the cancer catalogue. *Nature Science Updates*. May 2001.
26. Kaiser J. Fingerprinting a killer. *Science*. 2001;292(5523):1803-1803. doi:10.1126/science.292.5523.1803b