Modelling toeholder switches *in silico* for the detection of miRNA signatures for the diagnosis of colorectal cancer

Medha Shridharan

Under the direction of

Dr. Elizabeth Wood
Founder
JURA Bio, Inc.

Research Science Institute
August 1, 2022

Abstract

Colorectal cancer (CRC) is an umbrella term encompassing cancers originating in the colon or rectum and is the third most common form of cancer in the world with a very high disease burden, but may be treated effectively if detected early. CRC is characterised by certain circulating microRNA (miRNA) biomarkers in the bloodstream, namely miR-203a-3p, miR-145-5p, miR-375-3p and miR-200c-3p. These miRNA targets can be detected with high specificity by carefully designed toehold switches, which quantify the target molecules via fluorescence. In this study, we use the Toeholder 2.0 software to model toehold switches for each miRNA target and analyse the biophysical properties of the switch-target complexes to identify the optimal switch for each target. We evaluate the viability of these switches and find promising candidate switches for miR-203a, miR-200c and miR-375-3p for further development and implementation in paper-based diagnostic devices for CRC. However, we rejected the switch for miR-145 due to its low specificity. Based on these results, for a patient to be diagnosed with CRC, they must have a significant increase in miR-203a concentration and significant decrease in miR-375-3p and miR-200c-3p concentrations in blood serum. These results can be implemented in paper-based diagnostic devices, allowing for the development of affordable, stable, minimally-invasive and point-of-care diagnostic devices for CRC.

Summary

Colorectal cancer (CRC) is a class of cancers originating in the colon or rectum and is the third most common form of cancer in the world, but can be treated effectively if detected early. Current state-of-the-art diagnostic procedures are expensive, highly invasive and may have limited accessibility. A promising new approach for detecting CRC is to quantify genetic biomarkers in patient blood samples. Toehold switches are synthetic biomolecules that can detect the presence of such biomarkers in the bloodstream with great precision. In this study, we design toehold switches for CRC biomarkers and evaluate their theoretical viability. We identified three such switches that are promising candidates for further development in implementation for early detection of CRC.

1 Introduction

Colorectal cancer (CRC) is an umbrella term encompassing carcinomas that develop in the colon or rectum. CRC is the third most common form of cancer in the world, with 1.9 million cases in 2021 [1]. Even among cancers, CRC has notoriously high disease burden, with symptoms including nonstop abdominal aches, unexplained weight loss and blood in stool [2]. However, CRC progression can be mitigated with early diagnosis, with a 5-year survival rate of 91% if the cancer is detected in its localized stage, as compared to 72% when the cancer spreads to surrounding tissues, organs or regional lymph nodes[3].

The current state of the art for diagnosing CRC is colonoscopy biopsy, where superficial biopsy samples are obtained from the thin intestinal wall. However, this procedure is highly invasive and costly [4]. Colonoscopy is associated with significant complications, with an electronic record review of diagnostic colonoscopies revealing that one in 200 patients experienced a severe complication such as bleeding or perforation of the intestinal wall within 30 days of the procedure [5]. Additionally, the sensitivity of colonoscopy for advanced neoplasia is only approximately 88% [6].

One method to circumvent the issues with biopsy is to detect microRNA (miRNA) signatures from the cancer circulating in the blood. *miRNA molecules* are short, noncoding RNA molecules that typically regulate translation in cells. Aberrant expression of miRNA has been shown to be not only be causal in carcinogenesis but may also exert a causal role in different steps of the tumorigenic process [7, 8]. Certain miRNAs are able to circulate in the blood as they are commonly found in lipid or lipoprotein complexes (such as microvesicles, exosomes or apoptotic bodies), making them highly stable [9, 10, 11]. Given that these miRNA molecules are highly specific, stable and easily accessible from blood samples, they prove to be excellent biomarkers for distinguishing types of cancer and even stages of cancer progression [12].

The panel of miRNA signatures identified for CRC diagnosis in particular is as follows: miR-203a-3p, miR-145-5p, miR-375-3p and miR-200c-3p. Huang *et al.* showed that miR-203a-3p is significantly upregulated in CRC patients as compared to healthy controls, while miR-145-5p, miR-375-3p and miR-200c-3p are significantly downregulated [13]. Hence, we hypothesize that in order to diagnose a patient with CRC using this method, we must find that miR-145-5p, miR-375-3p and miR-200c-3p are downregulated in patient serum, while simultaneously finding that miR-203a-3p is significantly upregulated.

Each miRNA target can be detected by binding to an mRNA molecule known as a toehold switch sensor.

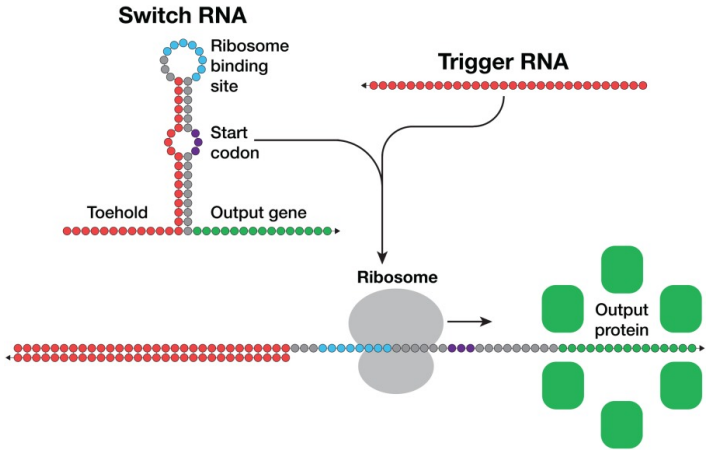

Figure 1: Composition and activation of toehold riboswitches. All the necessary elements for translation of a reporter gene are present in a toehold riboswitch. However, the ribosome binding sequence (RBS) is only accessible to the ribosome when the riboswitch binds to its trigger RNA.

Toehold switches are a class of synthetic riboregulators with excellent specificity, modularity and orthogonality, which can sense virtually any RNA sequence [14]. Toehold-switch riboregulators are mRNA elements which function by occluding the ribosome from translating a downstream gene with a hairpin structure. Prior to base pairing, the switch is considered to be in the OFF state. Upon base pairing with the trigger RNA sequence, the toehold structure of the switch unfolds and sequestration of both the ribosome binding site (RBS) and

start codon is relieved, allowing for the translation of the output gene and expression of a reporter gene such as Green Flourescent Protein (GFP). The magnitude of fluorescence is proportional to the concentration of target present in the sample. This switch-target complex is also known as ON state of the switch (Figure 1) [15, 16]. Toehold switches have been extensively used in programmable point-of-care diagnostic devices, such as paper-based tests for Zika virus [15, 17].

This study aims to design toehold switches *in silico* to detect miR-203a-3p, miR-145-5p, miR-375-3p and miR-200c-3p for the diagnosis of CRC and evaluate their viablity for further development.

2 Methods

2.1 Design of toehold switches *in silico* with Toeholder software

Toeholder 2.0, an open-source software package [18] was used with the NUPACK v3.2.2 [19] and BLAST+ v2.9.0 [20] packages to design the toehold switches.

The gene sequences for the target miRNAs miR-203a-3p, miR-145-5p, miR-375-3p and miR-200c-3p were obtained from the NCBI Gene database [21] and confirmed with miRBase[22]. One unusual limitation of the Toeholder software is that the target gene sequence input must have length between 35 and 100bp. As the specific corresponding DNA sequences for miR-203a-3p, miR-200c-3p and miR-145-5p are only 22–23bp long, we instead used the full genes miR-203a, miR-200c and miR-145 respectively, which include the aforementioned DNA sequences. However, as the DNA sequence for miR-375-3p is 64bp long, this target sequence was maintained.

Toeholds were also aligned to several reference genomes to test their predicted specificity and versatility using the BLASTN tool for short sequences [20]. These reference genomes, consisting of various phage and bacterial genomes, were selected based on their high likeli-

hood of being present in the same blood sample as the target miRNA and were extracted from the NCBI Genome database [23]. An example of a genome tested was *Escherechia coli*, with the full list of genomes in Appendix A.6.

Toeholder was run on a Linux server with the Jupyter Lab IDE. The software was run under the default parameters, as listed in Appendices A.1 to A.4.

The Toeholder workflow for designing toehold switches is as follows. Toeholder receives a target gene and other parameters [1] as input that will be used to perform a sliding window scan of the target sequence. The sliding window is used to determine the trigger sequence, that is, the complement of the intended target sequence. Afterwards, the sequence that will close the hairpin is added as the complement of the second part of the trigger sequence [18]. The detailed procedure is illustrated in Appendix A.5 (Figure S1).

Toehold switches produced by Toeholder are tested automatically using NUPACK. The minimum free energy (MFE) secondary structures of the proposed toehold switch and the target mRNA were generated separately, as well as the MFE secondary structure for the proposed toehold switch bound to the target mRNA.

ΔG, refers to the change in Gibbs free energy over the course of a reaction. $\Delta G_{binding}$ refers to the change in energy of the switch-target system due to binding. This is calculated by finding the change in energies between the final bound state of the complex ($\Delta G_{bound/ON}$) and the initial unbound states ($\Delta G_{unbound/OFF}$ and ΔG_{target}), as shown in Equation 1.

$$\Delta G_{binding} = \Delta G_{bound/ON} - (\Delta G_{unbound/OFF} + \Delta G_{target}) \tag{1}$$

The switch-target complex with the most negative $\Delta G_{binding}$ is said to have the lowest MFE and hence its formation is most favoured.

[1] These parameters include the length of the trigger region bound to the target, the length of the trigger in the hairpin and the reporter gene sequence.

2.2 Analysis of data generated

In our analysis, we focused on three main biophysical properties of the toehold switches: Guanine-Cytosine content (GC content), minimum Gibbs free energy (MFE) and pair probability graphs.

GC content refers to the proportion of two nitrogenous bases, guanine and cytosine, which might be in any domain of a gene, single gene, gene clusters, or even non-coding regions. In this case, it refers to the proportion of G and C in the switch RNA. GC-rich regions tend to facilitate base stacking, making them more stable than sequences with low GC content. Additionally, secondary structures formed by high GC content regions also tend to be stable and more resistant to denaturation. Sequences containing many guanine repeats can also generate complicated inter-strand folding due to hydrogen bonds between adjacent guanines. Hence, high GC content is important for the strength of the ON and OFF state stabilities [14]. However, synthesis of high GC content switches can be troublesome due to issues with secondary structure, mispriming, or mis-annealing [24]. Successful toeholds have been found to have a range of acceptable GC content between 20–60% [25].

The MFE of RNAs increases at a linear rate with sequence length. Simple indices, obtained by dividing the MFE by the number of nucleotides, have been used for a direct comparison of the folding stability of RNAs of various sizes [26]. The MFE of the secondary structure of toehold switches is calculated by a series of equations as shown in Appendix B. These calculations are run by the NUPACK software within Toeholder. For a switch to be successful, the switch-target complex must have a significantly lower MFE than that of the unbound switch and target. This implies that the formation of the complex is energetically favoured, hence maximising the number of such complexes formed and maximising detection of the target miRNAs.

Pair probability graphs represent the equilibrium base-pairing probabilities for the ordered complexes, treating all strands as distinct. It illustrates the probability of each base

pair binding in the switch-target complex at equilibrium. By definition, this data is independent of concentration and of all other ordered complexes in solution.

3 Results

3.1 Secondary structures and biophysical parameters of switch-target complexes

We used NUPACK to visualise the secondary structures and equilibrium probabilities based on sequences generated by Toeholder (Figure 2).

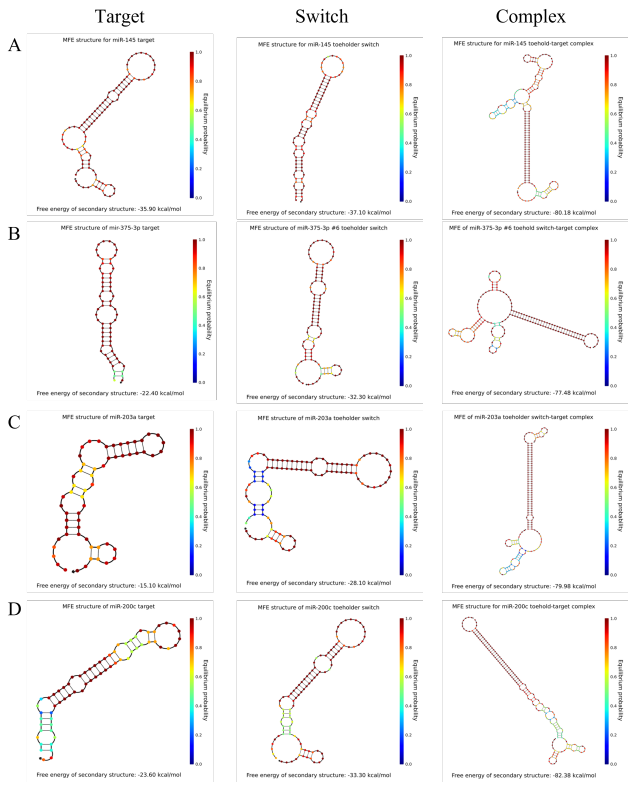

Figure 2: Secondary structures generated for (from left to right) the switch, target, and switch-target complex for each miRNA target respectively: **(A)** miR-145, **(B)** miR-375-3p, **(C)** miR-203a,**(D)** miR-200c. The coloured gradient on the right of each graph refers to the equilibrium probabilities, which is the probability that each base adopts the predicted paired or unpaired state at chemical equilibrium.

Toeholder summary statistics for each generated sequence fall in the desired ranges for GC content and MFE of switch-target complex (Table 1).

	GC content in %	MFE in kcal/mol	MFE of target in kcal/mol	MFE of switch in kcal/mol
miR-145	53.33	-80.18	-35.90	-37.10
miR-375-3p	60.0	-77.48	-22.40	-32.30
miR-203a	40.0	-79.98	-15.10	-28.10
miR-200c	56.67	-82.38	-23.60	-33.30

Table 1: Summary of GC content and MFE for each of the four switch-target complexes, in addition to the MFE for the unbound target and switch. The acceptable range of values for is 20-60%. The acceptable range of values for the MFE of the switch-target complex is any value more negative than the MFEs of the unbound target and switch.

For miR-145, the strongest binding occurs between bases 25–55 of the target miR-145 and bases 3–33 of the switch, characterised by a series of equilibrium probabilities of 1.0. There is also strong binding with equilibrium probability of 0.8 between bases 5–19 of the target miR-145 and bases 70–81 of the switch. The probability for the formation of this complex is 0.018 (Figure 3A).

For miR-375-3p, the strongest binding occurs between bases 6–39 of the target miR-375-3p and bases 3–36 of the switch, with a series of equilibrium probabilities of 1.0. The probability for the formation of the selected complex is 0.013. (Figure 3B).

For miR-203a, the strongest binding occurs between bases 0–40 of the target miR-203a and bases 0–40 of the switch, with a series of equilibrium probabilities of 1.0. The probability for the formation of this complex is 0.022 (Figure 3C).

For miR-200c, the strongest binding occurs between bases 20–60 of the target miR-200c and bases 0–40 of the selected switch, with a series of equilibrium probabilities of 1.0. The equilibrium probability of binding decreases between bases 40–60 of the switch and 0–20 of the target miR-200c. The probability for the formation of this complex is 0.019. (Figure 3D)

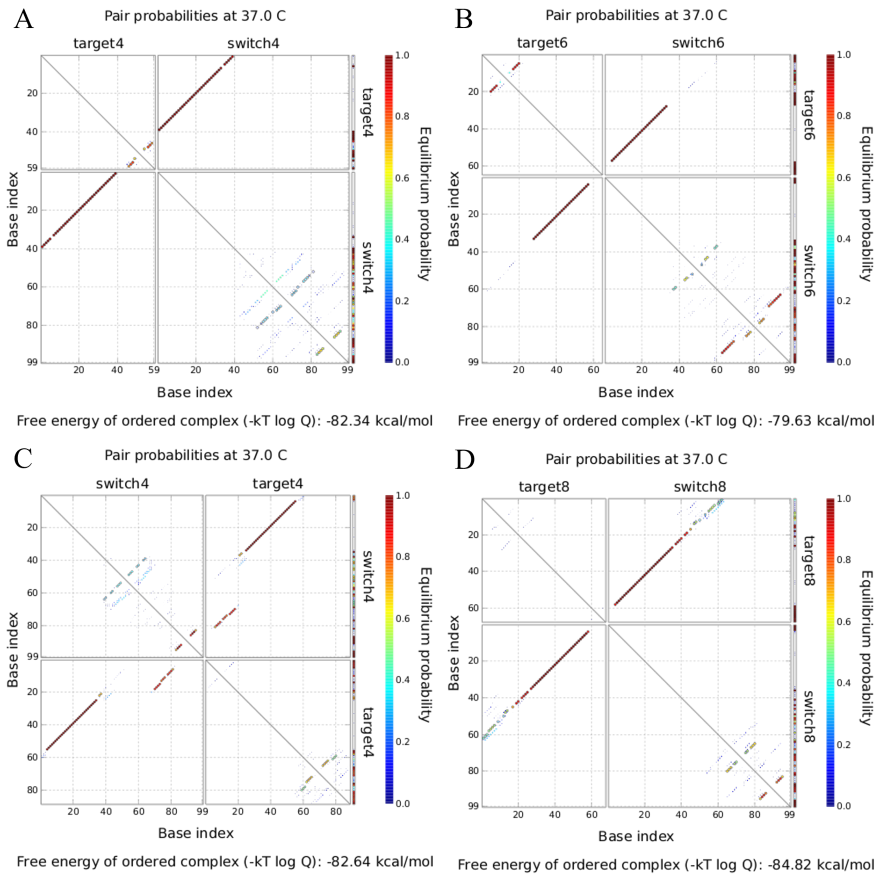

Figure 3: Graph of pair probabilities at equilibrium for the switch-target complexes for each miRNA target respectively: **(A)** miR-145, **(B)** miR-375-3p, **(C)** miR-203a, **(D)** miR-200c. Base index on the axes refers to the numerical index of the base with respect to the 5' end of the molecule (either switch or target). Any given point on the graph corresponds to a base pair between switch-target, switch-switch or target-target. The area and colour of each dot represents the probability of the corresponding base pair at that point on the graph forming at equilibrium. Probabilities below 0.001 are not depicted on the graph. With this convention, the plot is symmetric, with the upper and lower triangles separated by a diagonal line. The area and colour of each dot in the column on the right scale represents the equilibrium probability that the corresponding base is unpaired.

3.2 Low signal to noise of miR-145 switch

While running the toehold switches through NUPACK, the concentrations of the switch and target miRNA molecules were both fixed at 1μM, and it was observed that for miR-375-3p, miR-203a-3p and miR-200c-3p the concentrations of switch-target complexes formed were 1μM, implying complete binding of all target and switch molecules.

However, it was observed that for miR-145, the concentration of switch-target complexes was significantly lower than 1μM. Hence, we investigated how the formation of this switch-target complex could be optimized by changing the concentration of the switch. (Figure 4)

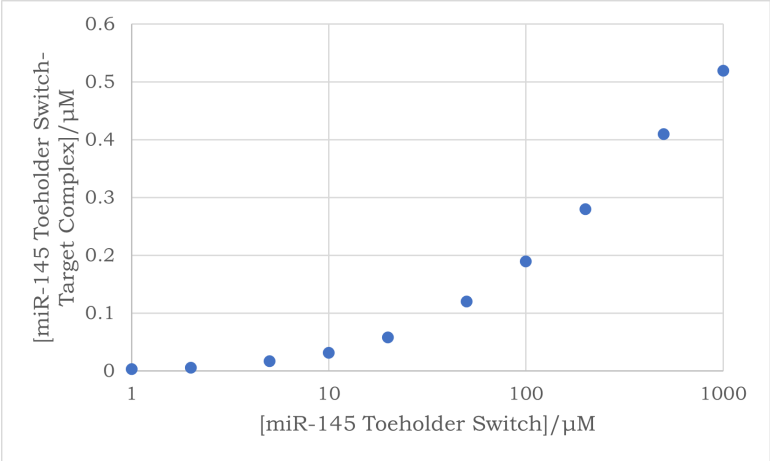

Figure 4: Graph of the concentration of switch-target complexes/μM against the concentration of the generated switch. The x-axis is plotted on a logarithmic scale.

The binding of the switch-target complex appears to increase as the concentration of switch in the reaction mixture increases. However, it should be noted that using 1000μM of switch only results in 0.52μM of switch-target complexes, as compared to 350μM of switch-switch complexes. This trend can be observed across all datapoints, with the noise from switch-switch binding consistently being significantly higher than signal from switch-target complexes.

4 Discussion

Based on secondary structure stability and other desired thermodynamic properties, the switches for all four miRNA targets have been shown to be theoretically viable. The GC content and MFE of all four switches fall within the predetermined acceptable ranges.

An interesting point to note is that although the equilibrium probabilities for most bases in all 4 switch-target complexes are close to 1.0, the probability of each of the complexes forming is very low. This may potentially be due to the the fact that the concentrations of the switch and target were very low (1μM). Hence, the concentrations of the switches and targets may need to be adjusted to optimize switch-target complex formation and biomarker detection for accurate diagnosis.

miR-203a-3p is significantly upregulated in CRC patients as compared to healthy controls, while miR-145-5p, miR-375-3p and miR-200c-3p are significantly downregulated [13]. Hence, for a patient to be diagnosed with reasonable certainty with CRC, they must have both a positive test result for miR-203a-3p and negative test results for miR-145-5p, miR-375-3p and miR-200c-3p when applied to a paper-based diagnostic test for CRC.

However, the low signal to noise ratio of the miR-145 switch has major implications in the development of paper-based diagnostic tests for CRC. There is a high probability of false positives due to the formation of switch-switch complexes, which renders the test ineffective for diagnosing CRC. As a result, we can exclude the miR-145 switch as a viable sensor to be applied in a paper-based test format.

Although many studies have recommended studying the biophysical properties of toehold switches, this analysis may lack sufficient predictive power due to the broad range of acceptable values. An alternative method is to study nucleotide representation in the toehold switches. This involves stratifying a database of sequences into high- and low-performing sequences and analysing the characteristics of both, then evaluating the generated switches

based on the previously identified criteria. This is believed to be another method of comprehensively analysing the viability of a toehold switch, but this technique is beyond the scope of this research.

5 Future Work

In this study, we identify toehold switches that should have the sensitivity and specificity to detect CRC miRNA biomarkers. Further proof-of-concept requires development and testing of a paper-based test using our toehold switches, which would improve and expand point-of-care diagnostics for CRC. As an intermediate step, we intend to evaluate our proposed switches against the alternate criteria from Valeri *et al.* that characterise viable toehold switches in practice [25].

6 Conclusion

This research acts as a foundation for the design of a paper-based test for CRC. We designed viable toehold switches to bind to four distinct CRC miRNA biomarkers using the Toeholder 2.0 software. We selected the optimal toehold switch for each target and evaluated the feasibility of applying them on a paper-based diagnostic test, hence eliminating the toehold switch for miR-145. We developed logic for the diagnostic test to be developed as part of future work, namely that for a patient to be diagnosed with CRC, they must have a higher concentration of miR-203a-3p and lower concentrations of miR-375-3p and miR-200c-3p.

7 Practical Takeaways

Colorectal cancer (CRC) is the third most common form of cancer in the world with high disease burden, but can be treated effectively if detected early. In this study, we design the molecular framework for a paper-based diagnostic test for CRC based on synthetic gene networks. Once fully developed, this paper-based diagnostic test would significantly expand access to effective and minimally invasive early testing to detect CRC.

8 Acknowledgments

Special thanks to Dr Elizabeth Wood, Dr Matt Cain, Dr André Nguyen, Tutor Joy Xu, Tutor Catherine Xue, Ali Christine Yang, my RSI peers, my parents and my teachers for their guidance and assistance with this project. Thank you to the Singapore Ministry of Education, RSI, Massachusetts Institute of Technology (MIT) and The Center for Excellence in Education (CEE) for this opportunity.

References

[1] American Society of Clinical Oncology. What is colorectal cancer?: How does colorectal cancer start? https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html, 2020.

[2] Centers for Disease Control and Prevention. What are the symptoms of colorectal cancer? https://www.cdc.gov/cancer/colorectal/basic_info/symptoms.htm, Feb 2022.

[3] American Society of Clinical Oncology. Colorectal cancer – statistics. https://www.cancer.net/cancer-types/colorectal-cancer/statistics, May 2022.

[4] S. Wang, L. Wang, N. Bayaxi, J. Li, W. Verhaegh, A. Janevski, V. Varadan, Y. Ren, D. Merkle, X. Meng, et al. A microRNA panel to discriminate carcinomas from high-grade intraepithelial neoplasms in colonoscopy biopsy tissue. *Gut*, 62(2):280–289, 2013.

[5] T. R. Levin, W. Zhao, C. Conell, L. C. Seeff, D. L. Manninen, J. A. Shapiro, and J. Schulman. Complications of colonoscopy in an integrated health care delivery system. *Annals of Internal Medicine*, 145(12):880–886, 2006.

[6] R. M. Hoffman, D. Espey, and R. L. Rhyne. A public-health perspective on screening colonoscopy. *Expert Review of Anticancer Therapy*, 11(4):561–569, 2011.

[7] A. J. Schetter, S. Y. Leung, J. J. Sohn, K. A. Zanetti, E. D. Bowman, N. Yanaihara, S. T. Yuen, T. L. Chan, D. L. Kwong, G. K. Au, et al. microRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. *Jama*, 299(4):425–436, 2008.

[8] M. V. Iorio and C. M. Croce. microRNA involvement in human cancer, Apr 2012.

[9] M. M. H. Sohel. Circulating microRNAs as biomarkers in cancer diagnosis. *Life Sciences*, 248:117473, 2020.

[10] R. Singh, B. Ramasubramanian, S. Kanji, A. R. Chakraborty, S. J. Haque, and A. Chakravarti. Circulating microRNAs in cancer: Hope or hype? *Cancer Letters*, 381(1):113–121, 2016.

[11] K. Nakamura, K. Sawada, A. Yoshimura, Y. Kinose, E. Nakatsuka, and T. Kimura. Clinical relevance of circulating cell-free microRNAs in ovarian cancer. *Molecular Cancer*, 15(1):1–10, 2016.

[12] Y. Wu, X. Wang, F. Wu, R. Huang, F. Xue, G. Liang, M. Tao, P. Cai, and Y. Huang. Transcriptome profiling of the cancer, adjacent non-tumor and distant normal tissues from a colorectal cancer patient by deep sequencing. 2012.

[13] G. Huang, B. Wei, Z. Chen, J. Wang, L. Zhao, X. Peng, K. Liu, Y. Lai, and L. Ni. Identification of a four-microRNA panel in serum as promising biomarker for colorectal carcinoma detection. *Biomarkers in Medicine*, 14(9):749–760, 2020.

[14] A. Green. a, Silver, P. a, Collins, JJ, and Yin, P.(2014) Toehold switches: de-novo-designed regulators of gene expression. *Cell*, 159:925–39.

[15] K. Pardee, A. A. Green, M. K. Takahashi, D. Braff, G. Lambert, J. W. Lee, T. Ferrante, D. Ma, N. Donghia, M. Fan, et al. Rapid, low-cost detection of Zika virus using programmable biomolecular components. *Cell*, 165(5):1255–1266, 2016.

[16] P. R. S. Baabu, S. Srinivasan, S. Nagarajan, S. Muthamilselvan, R. R. Suresh, T. Selvi, and A. Palaniappan. Rapid, low-cost detection of Zika virus using programmable biomolecular components. *bioRxiv*, 2021.

[17] R. A. Hall and J. Macdonald. Synthetic biology provides a toehold in the fight against Zika. *Cell Host & Microbe*, 19(6):752–754, 2016.

[18] C. C. AF, F. Rouleau, C. Bautista, P. Lemieux, and N. Dumont-Leblond. Toeholder: a software for automated design and in silico validation of toehold riboswitches. 2021.

[19] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, 32(1):170–173, 2011.

[20] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, and K. Bealer. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[21] G. R. Brown, V. Hem, K. S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. D. Pruitt, D. R. Maglott, et al. Gene: A gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42, 2015.

[22] S. Griffiths-Jones, R. J. Grocock, S. Van Dongen, A. Bateman, and A. J. Enright. BLAST+: Architecture and applications. *Nucleic Acids Research*, 34(suppl_1):D140–D144, 2006.

[23] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Research*, 45(Database issue):D37, 2017.

[24] GC-rich gene synthesis. `https://www.synbio-tech.com/gc-rich-gene-synthesis/`.

[25] J. A. Valeri, K. M. Collins, P. Ramesh, M. A. Alcantar, B. A. Lepe, T. K. Lu, and D. M. Camacho. Sequence-to-function deep learning frameworks for engineered riboregulators. *Nature communications*, 11(1):1–14, 2020.

[26] E. Trotta. On the normalization of the minimum free energy of RNAs by sequence length. *PloS One*, 9(11):e113380, 2014.

[27] A. N. Burnett-Hartman, P. A. Newcomb, and J. D. Potter. Infectious agents and colorectal cancer: a review of Helicobacter pylori, Streptococcus bovis, JC virus, and human papillomavirus. *Cancer Epidemiology Biomarkers & Prevention*, 17(11):2970–2979, 2008.

A Material Parameters for Toeholder software

A.1 Nucleic acid material

The parameter files defining the nucleic acid material are specified via the argument parameters which represents either a filename prefix or a shorthand identifier for an included parameter set. If the filename does not contain a relative or absolute path, then the program will look for the files first in the current directory, and then in the directory $NUPACKHOME/parameters. Available filename prefixes currently include:

- rna1995 (default; shorthand: rna)

  Parameter files *.dG and *.dH for RNA allowing calculations at different temperatures (Serra and Turner, 1995; Zuker, 2003); includes pseudoknot parameters from (Dirks and Pierce, 2003)

- dna1998(shorthand: dna)

  Parameter files *.dG and *.dH for DNA allowing calculations at different temperatures (SantaLucia,1998; Zuker, 2003); there are no pseudoknot parameters.

- rna1999

  Parameter file *.dG for RNA for calculations at 37 ∘C (Mathews et al., 1999; Zuker, 2003); includes pseudoknot parameters from (Dirks and Pierce, 2003).

- custom

  Custom parameter files *.dG and *.dH allowing calculations at different temperatures; or custom parameter file *.dG allowing calculations at one temperature. Custom parameter files must be placed in the same location as the default parameter files (/urs/local/share for a default installation).

DNA/RNA hybrids are not allowed.

A.2 Sodium concentration

The Na+ concentration of the solution in units of molar (default: 1.0, range: [0.05,1.1]) is specified by concentration (SantaLucia and Hicks, 2004). This flag is only valid when the -material dna is also selected because no RNA salt correction parameters are available.

A.3 Magnesium concentration

The Mg++ concentration of the solution in units of molar (default: 0.0, range: [0.0,0.2]) is specified by concentration (Koehler and Peyret, 2005). This flag is only valid when the -material dna is also selected.

A.4 Dangles treatment

The way in which dangle energies are incorporated is specified by treatment, which may have the following values:

- none: No dangle energies are incorporated.
- some: (default) A dangle energy is incorporated for each unpaired base flanking a duplex (a base flanking two duplexes contributes only the minimum of the two possible dangle energies).
- all: A dangle energy is incorporated for each base flanking a duplex regardless of whether it is paired

A.5 Detailed Toeholder Workflow

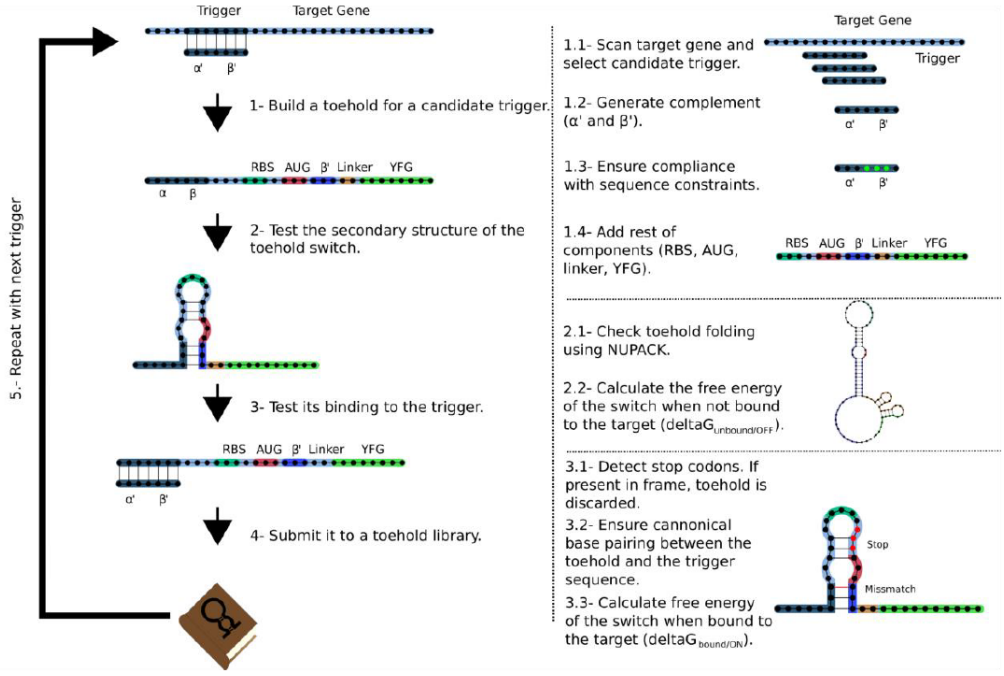

Figure S1: Workflow used by Toeholder 2.0 to design toehold riboswitches. [18]

A.6 Reference genomes selected for specificity checking

Valeri et al. included a set of genomes which have a high likelihood of being cross-referenced in the human body, including *Escherichia coli*, *Homo sapiens*, MS2 phage, PM2 phage, Human Papillomillivirus (HPV), JC Polyomavirus (JCV), *Streptococcus bovis* and *Helicobacter pylori*)[18]. Burnett et al. identified JCV, HPV, H. pylori and S. bovis as potential etiologic agents for CRC, and hence these genomes are likely to be found in the bloodstream. [27]

B Thermodynamic analysis of interacting nucleic acid strands from Zadeh et al.

B.1 Secondary structure model

A polymer graph for a secondary structure is constructed by ordering the strands around a circle, drawing the backbones in succession from 50 to 30 around the circumference with a nick between each strand, and drawing straight lines connecting paired bases. A secondary structure is pseudoknotted if every strand ordering corresponds to a polymer graph with crossing lines. A secondary structure is connected if no subset of the strands is free of the others. An ordered complex corresponds to the unpseudoknotted structural ensemble Γ, comprising all connected polymer graphs with no crossing lines for a particular ordering of a set of strands. [2]

For a secondary structure $s \in \Gamma$, the free energy,

$$\Delta G(\phi, s) = (L-1)G^{assoc} + \sum_{loop \in s} \Delta G(\phi, loop) \tag{2}$$

is calculated using nearest-neighbor empirical parameters for RNA in 1M Na+ [9, 10] or for DNA in userspecified Na+ and Mg++ concentrations [11–13], of all loops in that structure. Here, L is the number of strands in the complex, G^{assoc} is the penalty for strand association [14], and secondary structure loop classification is depicted in Figure 1.1. This physical model provides the basis for rigorous analysis and design of equilibrium base-pairing in the context of the free energy landscape defined over ensemble Γ.

B.2 Characterizing equilibrium secondary structure

By calculating the partition function,3 [17]

[2] Pseudoknotted structures are excluded from the ensemble Γ for computational expediency

$$Q(\phi) = \sum_{s \in \Gamma} e^{(-\Delta G(\phi,s)/k_B T)} \tag{3}$$

over Γ, it is possible to evaluate the equilibrium probability,

$$p(\phi, s) = \frac{1}{Q(\phi)} e^{(-\Delta G(\phi,s)/k_B T)} \tag{4}$$

of any secondary structure $s \in \Gamma$. [3] The secondary structure with the highest pribability at equilibrium is the *minimum free energy* (MFE) structure, [4] satisfying

$$s^{MFE}(\phi) = \arg min_{s \in \Gamma} \Delta G(\phi, s) \tag{5}$$

[3] Here, k_B is the Boltzmann constant and T is the temperature.

[4] For simplicity of exposition, we assume that there is a unique MFE structure; only superficial changes are required if this is not the case.

C Toehold switch data generated by Toeholder

C.1 miR-145

Index	Sequence	Structure in DU+ notation	Binding_energy_toehold_mRNA	Binding_energy_toehold	Binding_energy_mRNA	GC content
4	CCCAGGAAUCCCUUAGAUGCUAAGAUGGGG	.((((((((((.....))))	-80.181	-35.9	-37.1	53.33
1	CAGUUUUCCCAGGAAUCCCUUAGAUGCUAA	((((.((.((((((((((.....	-78.981	-33.9	-37.1	43.33
0	UCCAGUUUUCCCAGGAAUCCCUUAGAUGCU	..((((.((.((((((((((.....	-78.181	-34.5	-37.1	46.67
3	UCCCAGGAAUCCCUUAGAUGCUAAGAUGGG	(.((((((((((.....)))	-77.281	-30.4	-37.1	50
10	CCCUUAGAUGCUAAGAUGGGGAUUCCUGGA	((((.....)))))))))))).	-76.881	-32.2	-37.1	50
6	CAGGAAUCCCUUAGAUGCUAAGAUGGGGAU	((((((((((.....))))))	-76.881	-27.9	-37.1	46.67
13	UGCUAAGAUGGGGAUUCCUGGAAAUACUGU	.....)))))))))))).)).)))).	-75.881	-34.8	-37.1	43.33
11	CCUUAGAUGCUAAGAUGGGGAUUCCUGGAA	(((.....)))))))))))).)	-74.881	-30.1	-37.1	46.67
2	UUUCCCAGGAAUCCCUUAGAUGCUAAGAUG	.((.((((((((((.....)	-74.081	-27.2	-37.1	43.33
7	GAAUCCCUUAGAUGCUAAGAUGGGGAUUCC	(((((((((.....)))))))))	-72.381	-29	-37.1	46.67
12	GAUGCUAAGAUGGGGAUUCCUGGAAAUACU	.....)))))))))))).)).)))	-71.881	-31.1	-37.1	43.33
5	CCAGGAAUCCCUUAGAUGCUAAGAUGGGGA	((((((((((.....))))))	-70.981	-33.8	-37.1	50
9	UCCCUUAGAUGCUAAGAUGGGGAUUCCUGG	(((((.....)))))))))))	-70.081	-32.9	-37.1	50
8	AUCCCUUAGAUGCUAAGAUGGGGAUUCCUG	((((((.....))))))))))	-68.481	-31.1	-37.1	46.67

Table S1: miR-145 switch data from Toeholder, with selected switch highlighted

C.2 miR-203a

Index	Sequence	Structure	Binding_energy_toehold_mRNA	Binding_energy_toehold	Binding_energy_mRNA	GC content
4	CGCAAUUGUGAAAUGUUUAGGACCACUAGA	(((..((.((....((((((.....)))))	-79.981	-28.1	-15.1	40
9	UGUUUAGGACCACUAGACCCGGAGGGCGCG	.((((((.....)))))).)).)).)))).	-77.581	-48	-15.1	63.33
8	AUGUUUAGGACCACUAGACCCGGAGGGCGC	..((((((.....)))))).)).)).))))	-75.181	-36	-15.1	60
6	AAAUGUUUAGGACCACUAGACCCGGAGGGC	....((((((.....)))))).)).)).))	-74.881	-32.8	-15.1	53.33
0	CUGUAGCGCAAUUGUGAAAUGUUUAGGACC	.....((((..((.((....(((((....	-73.081	-29.2	-15.1	43.33
2	AGCGCAAUUGUGAAAUGUUUAGGACCACUA	.((((..((.((....(((((.....)))	-71.381	-23.7	-15.1	40
1	UGUAGCGCAAUUGUGAAAUGUUUAGGACCA	....((((..((.((....(((((.....	-69.981	-23.7	-15.1	40
5	AAUUGUGAAAUGUUUAGGACCACUAGACCC	..((.((....((((((.....))))))).)	-67.681	-19.1	-15.1	40

Table S2: miR-203a switch data from Toeholder, with selected switch highlighted

C.3 miR-200c

Index	Sequence	Structure in DU+ notation	Binding_energy_toehold_mRNA	Binding_energy_toehold	Binding_energy_mRNA	GC content
8	GCGGUUGGGAGUCUCUAAUACUGCCGGGUA	((.....)))))..)))))))))..))))))	-82.381	-33.3	-23.6	56.67
5	UGGGUGCGGUUGGGAGUCUCUAAUACUGCC	.(((.((.....)))))..)))))))).	-79.981	-34.9	-23.6	56.67
7	UGCGGUUGGGAGUCUCUAAUACUGCCGGGU	.((.....)))))..)))))))..))))	-79.181	-29.9	-23.6	56.67
1	UCUUACCCAGCAGUGUUUGGGUGCGGUUGG	(..((((((.(((((((((.(((.((.....	-78.981	-28.7	-23.6	56.67
6	GGUGCGGUUGGGAGUCUCUAAUACUGCCGG	((.((.....)))))..)))))))..))	-78.581	-38.5	-23.6	60
0	GUCUUACCCAGCAGUGUUUGGGUGCGGUUG	((..((((((.(((((((((.(((.((.....	-76.781	-38	-23.6	56.67
2	UUACCCAGCAGUGUUUGGGUGCGGUUGGGA	.((((((.(((((((((.(((.((.....	-75.081	-29.2	-23.6	56.67
4	GUGUUUGGGUGCGGUUGGGAGUCUCUAAUA	(((((.(((.((.....)))))..))))	-73.981	-30.6	-23.6	50
3	AGUGUUUGGGUGCGGUUGGGAGUCUCUAAU	((((((.(((.((.....)))))..)))	-72.981	-33.2	-23.6	50

Table S3: mir-200c switch data from Toeholder, with selected switch highlighted

C.4 miR-375-3p

Index	Sequence	Structure	Binding_energy_toehold_mRNA	Binding_energy_toehold	Binding_energy_mRNA	GC content
0	GACGAGCCCCUCGCACAAACCGGACCUGAG	(.(((((((...((.((((((.(((.....	-83.381	-40.1	-22.4	66.67
1	ACGAGCCCCUCGCACAAACCGGACCUGAGC	.((((((...((.((((((.(((.....)	-81.081	-36.4	-22.4	66.67
2	GAGCCCCUCGCACAAACCGGACCUGAGCGU	((((((...((.((((((.(((.....)).	-80.881	-33	-22.4	66.67
6	GGACCUGAGCGUUUUGUUCGUUCGGCUCGC	(.....)).)))))).))...))))))))	-77.481	-32.3	-22.4	60
3	GCACAAACCGGACCUGAGCGUUUUGUUCGU	(.((((((.((.....)).)))))).)).	-76.981	-34.7	-22.4	53.33
7	GACCUGAGCGUUUUGUUCGUUCGGCUCGCG	.....)).)))))).))...))))))))	-75.681	-38	-22.4	60
8	ACCUGAGCGUUUUGUUCGUUCGGCUCGCGU	.....)).)))))).))...))))))))	-74.881	-31.1	-22.4	56.67
9	UGAGCGUUUUGUUCGUUCGGCUCGCGUGAG	....)).)))))).))...)))))))))).)	-74.581	-32.9	-22.4	56.67

Table S4: miR-375 switch data from Toeholder. While Switch 0 (highlighted in blue) had the lowest MFE value, its GC-content (66.67%) was out of the acceptable range of 20–60%. Hence, Switch 6 (highlighted in yellow) was selected instead as it had both acceptable MFE and GC-content. selected switch highlighted in yellow. (>60).