Experimental Confirmation of Multiple Co-Existent DNA Secondary Structures using Low-Yield Bisulfite Sequencing

Jin H. Bae*,[1,2] Jiaming Li*,[1,3] Boyan Yordanov*,[4,5] Michael X. Wang,[1] Javier González,[4] Andrew Philips+,[4] and David Yu Zhang+[1]

[1] *Department of Bioengineering, Rice University, Houston, TX*
[2] *Current affiliation: Broad Institute of MIT and Harvard, Cambridge, MA*
[3] *Department of Systems, Synthetic, and Physical Biology, Rice University, Houston, TX*
[4] *Microsoft Research, Cambridge, UK*
[5] *Current affiliation: Scientific Technologies, Ltd, London, UK*
* *Authors contributed equally*

(Dated: March 21, 2021)

The prediction of DNA secondary structures from DNA sequences using thermodynamic models is imperfect for many biological sequences, both due to insufficient experimental data for training and to the kinetics of folding that lead to metastable structures. Here, we developed low-yield bisulfite sequencing (LYB-seq) to query the secondary structure states of cytosine (C) nucleotides in thousands of different DNA oligonucleotides on a single-molecule level. We observed that the reaction kinetics between bisulfite and C nucleotides is highly dependent on the secondary structure state of the C nucleotides, with the most accessible C nucleotides (those in small hairpin loops) reacting 70-fold faster than those in stable duplexes. Next, we developed a statistical model to evaluate the likelihood of an NGS read being consistent with a particular proposed secondary structure. By analyzing thousands of NGS reads for each DNA species, we can infer the distribution of secondary structures adopted by each species in solution. We find that 84% of 1,057 human genome subsequences studied here adopt 2 or more stable secondary structures in solution.

Secondary structures of nucleic acids are of both biological and biomedical importance because the chemical accessibility of the nucleotides impact a variety of biological processes including transcriptional efficiency [1], splicing patterns [2], translational efficiency [3, 4], degradation rates [5], mutation rates [6], and ribozyme catalytic activity [2]. Software tools for predicting secondary structures of both DNA and RNA from sequence have been in use for over 30 years [7–12], but the accuracy of these tools is limited, particularly for longer biological nucleic acid sequences [13]. A significant bottleneck for the development of better predictive computational models of nucleic acid secondary structure prediction is the low number of nucleic acid sequences for which experimental secondary structure data is available, due to the low-throughput nature of melting temperature experiments. Additionally, the stochastic folding process of nucleic acids may result in an ensemble that differs from the equilibrium distribution predicted by thermodynamics.

RNA chemical probing and sequencing methods, such as SHAPE-MaP [14–17] and DMS-seq [18, 19], have been developed for assisting the determination of RNA secondary structures. These methods first make chemical modifications to the RNA molecules, and then reverse transcribes the modified RNA into DNA with special enzymes that generates cDNA with truncations or mutations at the chemically reacted sites, which are then assayed through high-throughput sequencing (NGS). These approaches, however, are limited to RNA due to their reliance on specialized reverse transcriptase enzymes, and are also not generally applicable to all RNA sequences, due to the large number of known RNA epigenetic modifications that can have unpredictable interactions with the chemical probing reagents and with reverse transcriptase. As far as we are aware, there have not been any reported approaches for high-throughput structural determination of DNA molecules.

Here, we present low-yield bisulfite sequencing (LYB-seq), a massively parallel method for chemically probing the secondary structure of many single-stranded DNA species. Importantly, multiple co-existing states of the same DNA species can be identified through this approach, allowing biophysical characterization of the ensemble of molecules in solution, rather than a simple minimum free energy (mfe) structure. In LYB-seq, a heterogeneous mixture of multiple DNA species is reacted with a dilute solution of sodium bisulfite to convert a small fraction of cytosine nucleotides to uracils (Fig. 1c). DNA nucleotides base paired in hairpin stems have significantly lower yield of conversion than unpaired DNA nucleotides. Because all cytosine nucleotides in a single NGS correspond to haplotype-phased information about a single molecule of DNA, the NGS data produced reflects a high-throughput single-molecule assay of DNA secondary structures. We developed a conditional probability framework to compute the likelihood any particular NGS read being generated by a proposed secondary structure. By comparing the likelihoods of two or more different secondary structures, we can determine the most likely secondary structure for many of our NGS reads, and infer a solution distribution of secondary structures. Applying LYB-seq to 1,057 subsequences of the human genome, each 100 nucleotides (nt) long, we find that 84% of them adopt 2 or more secondary structures in solution.

Results

Bisulfite conversion yield depends on secondary structure. Bisulfite conversion is a chemical reaction in

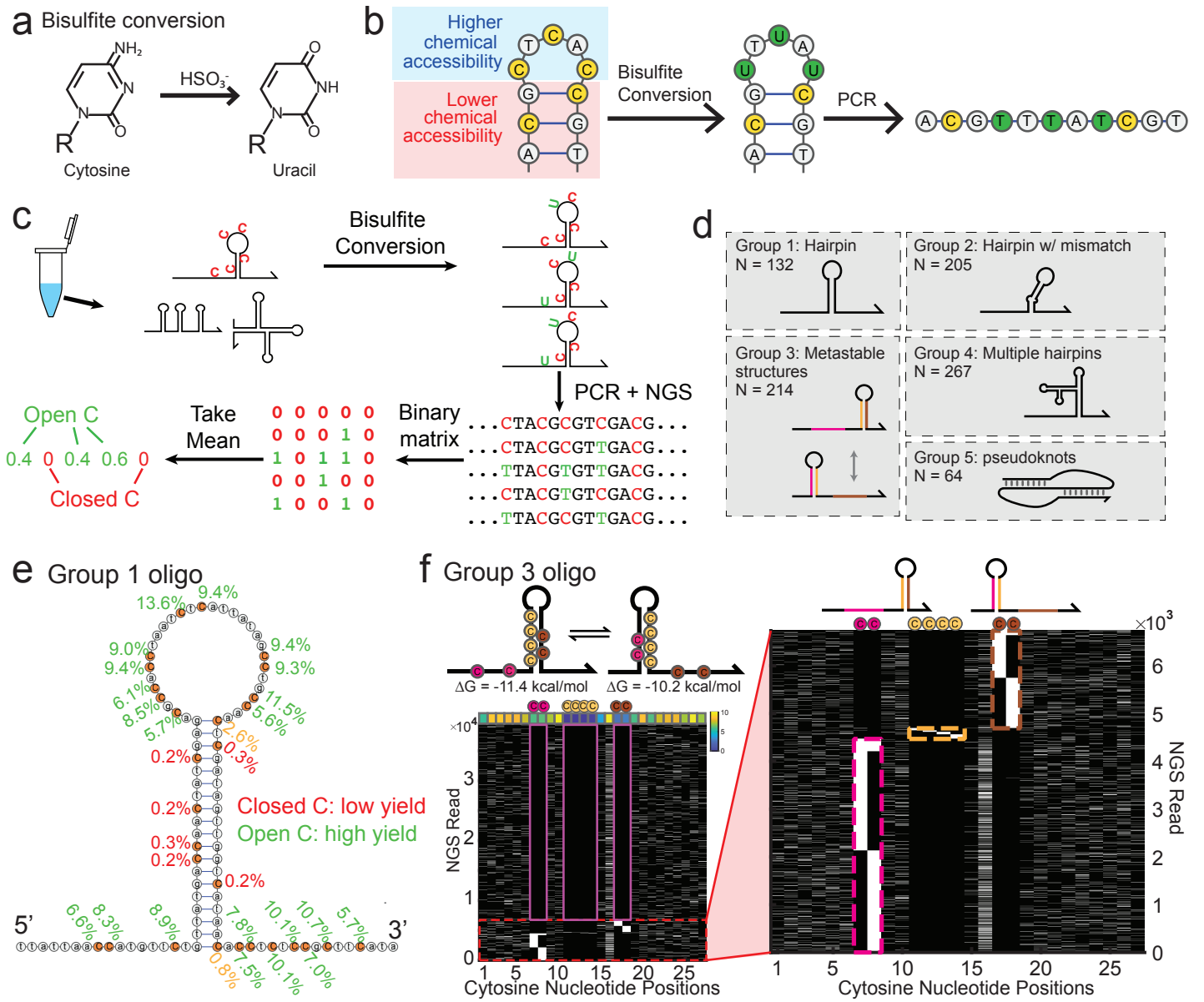

Figure 1: The concept, workflow, and proof of feasibility of LYB-seq. **(a)** Bisulfite conversion. R stands for the ribose or deoxyribose in RNA/DNA. **(b)** DNA secondary structure information are reflected as DNA sequence changes through the course of bisulfite conversion. C nucleotides on the hairpin loop are converted into U with higher probability than C nucleotides on the hairpin stem. **(c)** Low-yield bisulfite sequencing (LYB-seq) workflow overview. **(d)** 882 rationally designed control oligos for charcterizing LYB-seq, divided into 5 structural groups. **(e)** The conversion rates of all C positions on a group 1 oligo. Group 1 oligos were designed to have a stable single-hairpin structure, and all C nucleotides were colored orange for easier viewing. The conversion rate of each C was marked by its side, green for open and red/yellow for closed. The two C nucleotides marked by yellow were at the ends of the duplex stem. **(f)** Analysis of a group 3 oligo (DNA with two competing metastable hairpin secondary structures). The bottom left panel shows the conversion patterns of all reads mapped to this oligo. Each row represents an NGS read and each column represents a different C nucleotide in the DNA sequence ordered from 5′ to 3′. Black denotes unconverted (C) and white denotes converted (T). The upper row of color-mapped grids shows the overall conversion rates of each C position. The purple boxes marked the eight C positions shown on the oligo sketch.

which unmethylated cytosine (C) nucleotides are converted into uracil (U) nucleotides when the nucleic acid is treated with sodium bisulfite solution (Fig. 1a). The efficiency of the bisulfite conversion reaction is lower when the C nucleotide is base paired to a complementary guanine (G) nucleotide, as compared to when the C nucleotide is in a single-stranded state. Base-paired C's are henceforth referred to as "closed", and unpaired C's are referred to as "open". To overcome this efficiency bias for epigenetic profiling purposes, typical bisulfite conversion reactions use very high sodium bisulfite concentrations (e.g. 5M) and multiple thermal annealing steps to ensure near 100% conversion yield (CY) for all cytosines [20–22]. In our case, the bias in CY becomes a promising feature that can be leveraged to provide information regarding the state of each C nucleotide in a DNA molecule without involving any enzyme or other macromolecules (Fig. 1b).

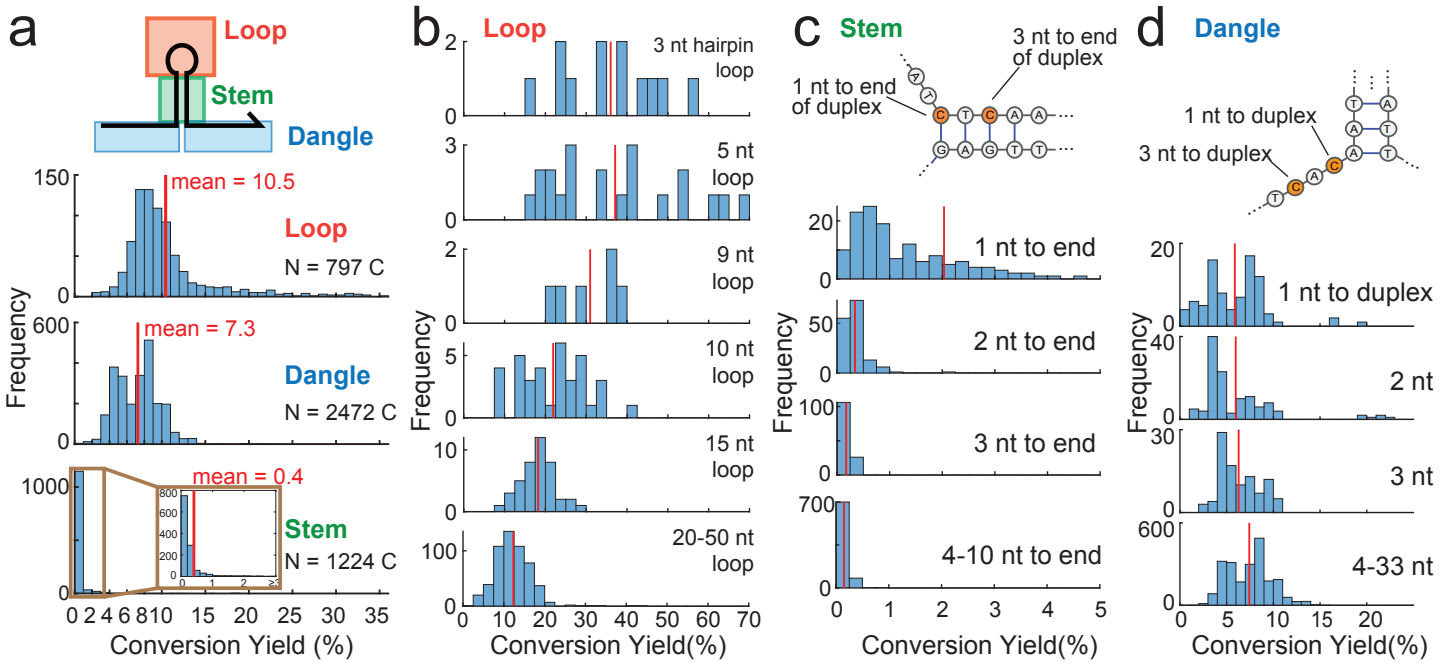

Figure 2: Distribution of bisulfite conversion yields (CY) for C nucleotides at different positions on a hairpin DNA sequence. **(a)** C nucleotides in hairpin loops and dangles have higher CY than C nucleotides in hairpin stems. **(b)** Smaller hairpin loops that are topologically constrained are observed to exhibit higher CY values. **(c)** The first nucleotide in a hairpin stem typically has higher CY compared to other hairpin stem nucleotides, possibly due to base pair breathing. **(d)** The position of C nucleotides in dangles does not appear to have a significant effect on CY.

To systematically characterize and validate the LYB-seq method, we purchased a pool of 1,939 DNA oligo species from Twist Bioscience for experimental studies. Among these DNA oligo species, 882 were rationally designed to have known and well-defined secondary structures, serving as positive controls. These rationally designed oligos are further subdivided into 5 groups, based on the designed secondary structure (Fig. 1d). The other 1,057 oligos were biological subsequences selected from 212 human genes. The entire DNA oligo pool simultaneously underwent the LYB-seq library preparation procedure (Fig. 1c, see also Supplementary Section S0). The NGS FASTQ reads were first aligned to the 1,939 DNA oligo sequences, and each oligo was individually analyzed.

We first analyzed the Group 1 oligos, DNA molecules designed and predicted to have a single, highly stable hairpin structure (Fig. 1e). From the analysis of C nucleotides in these oligos, a clear pattern emerged that open C nucleotides exhibit significantly higher CYs than closed C nucleotides (Fig. 2a, Fig. S2), consistent with our expectations. We also note that the very low (≈0.4%) CY of C nucleotides in hairpin stems may be an artifact of DNA sequencing. It is well known that sequencing-by-synthesis possesses an intrinsic error rate of between 0.2% and 1%, which is consistent with the observed CY of closed C nucleotides.

We also identified two interesting phenomena in addition to confirming the base pairing state impact on CY: First, C nucleotides in smaller hairpin loops, which are expected to be more topologically constrained, exhibited 3-fold higher CY than those in larger hairpin loops and dangles (Fig. 2b). One possible explanation for this behavior is that the C nucleotides in larger hairpin loops and dangles can transiently base pair with G nucleotides on other DNA species in solution, but C nucleotides in small hairpin loops are physically incapable of base pairing. Second, the C nucleotides in a hairpin stem closest to the end of the stem exhibit higher CY than C nucleotides near the middle of a hairpin stem (Fig. 2c), suggesting that base pair breathing may have resulted in higher conversion yield. In contrast, no such position dependence were observed for the distance of open C nucleotides to the hairpin stem (Fig. 2d). Similar analysis for Group 2 and Group 4 DNA oligos are shown in Supplementary Section S4.

Group 3 oligo sequences were designed to adopt multiple metastable secondary structures (Fig. 1f), and was intended to demonstrate that LYB-seq can identify distinct subpopulations of molecular states. In Fig. 1f, the CY of the C nucleotides in the 3 pairing regions (as marked by the magenta, yellow, and brown dashed line boxes) are nearly mutually exclusive, corresponding nicely to the two prospective single-hairpin structures. The reads in the magenta box with C nucleotides converted in the left pairing region correspond to molecules folded into the left structure, and the reads in the brown box with C nucleotides converted in the right pairing region correspond to the structure shown on the right. The four yellow C nucleotides in the middle paring region had very few conversions as marked by the yellow box because they were always at closed states for both dominant structures. These mutually exclusive conversions demonstrate that LYB-seq identifies co-existing secondary structures at a single molecule level.

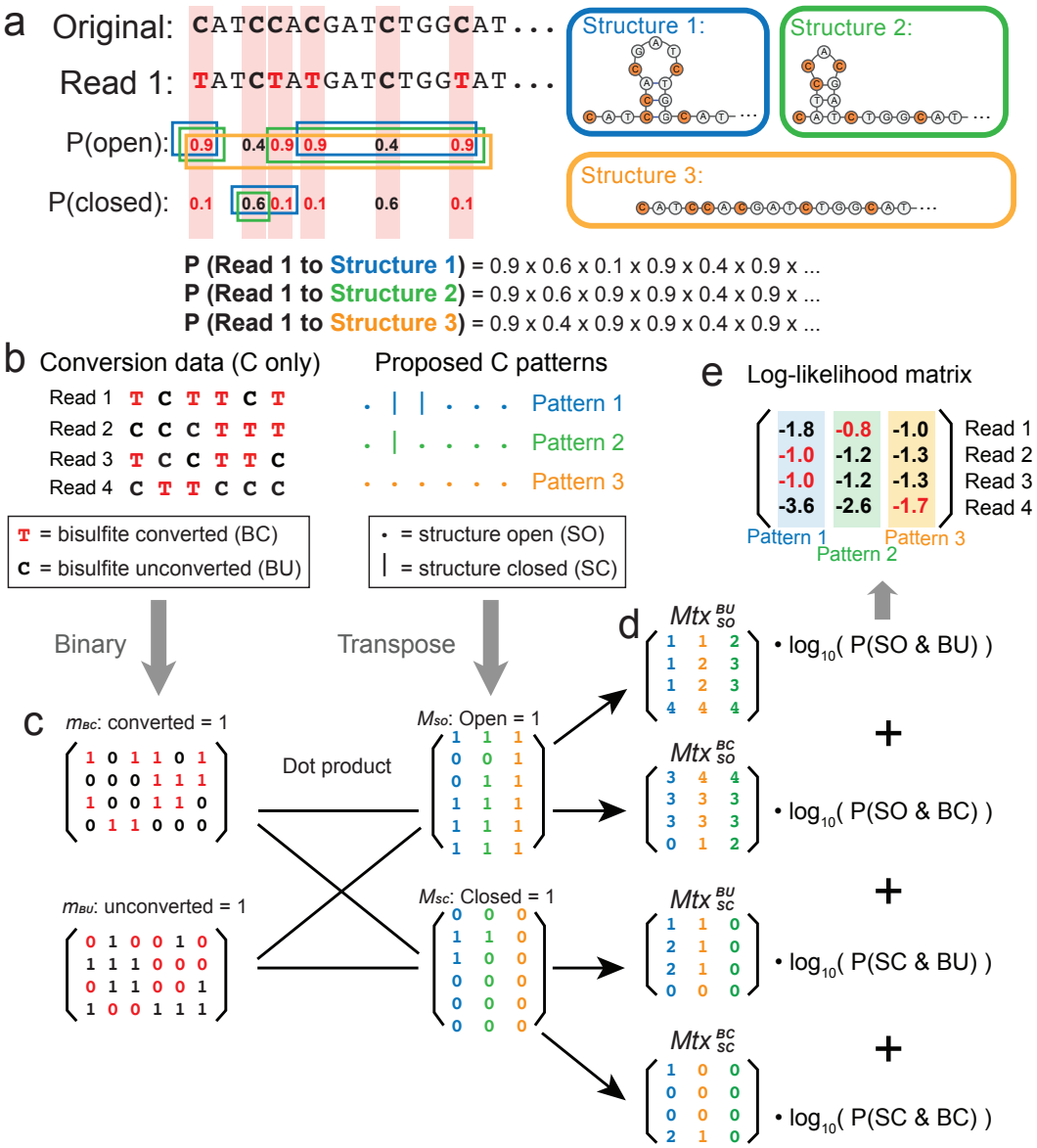

Figure 3: Framework for calculating likelihood of secondary structures in generating observed NGS reads. **(a)** Illustration of likelihood calculation based on observed NGS read. **(b)** Conversion state data from NGS reads and predicted structure open/closed patterns serve as the two input matrices. **(c)** Four binary matrices are generated, and the pairwise dot products are calculated. **(d)** The dot product matrices are multiplied to the experimentally observed probabilities of Open and Unconverted, Open and Converted, Closed and Unconverted, and Closed and Converted. **(e)** The sum of the matrices is our output, the log-likelihood of each read corresponding to each proposed structure.

Likelihood model. With the insights gained so far and the supports for the reproducibility of the experimental workflow of LYB-seq (see Supplementary Sections S1-4), we made a simplifying assumption that the the probability of C nucleotide conversion is dependent only on the base pairing state of the nucleotide, and otherwise independently and identically distributed. For ease of expression, we denote these two conditional probabilities as $p(BC|SO)$ and $p(BC|SC)$, where p stands for probability of one C nucleotide, BC stands for bisulfite converted, and SO and SC stands for structure open and closed, respectively. From the experimental data of the positive-control oligos from Groups 1, 2, and 4, we measured the specific values of $p(BC|SO)$ and $p(BC|SC)$ (Supplementary Section S5). The value of $p(BC|SO)$ is significantly larger than that of $p(BC|SC)$, because an open C nucleotide has higher CY.

Given experimental LYB-seq data, and we would like to infer its secondary structure. If a C nucleotide was converted, it indicates this C position is more likely to be open than closed because $p(BC|SO)$ is much larger than $p(BC|SC)$. Mathematically we could reverse the previously calculated conditional probabilities and obtain the probabilities of the C nucleotide being at an open ($P(SO|BC)$) or closed state ($P(SC|BC)$), given that it is converted, where P(SO|BC) is

larger than $P(SC|BC)$. Similarly, we obtain the probabilities $P(SC|BU)$ and $P(SO|BU)$ for an unconverted C nucleotide (See Supplementary Section S5 for more details). For any oligo whose conversion data we possess, we could calculate the probability of each read corresponding to any possible structure by multiplying the probabilities of each C nucleotide being open or closed based on its conversion status, since we assume all C nucleotides are independent (Fig. 3a).

Based on this concept, we developed a conditional probability model (Fig. 3) to utilize the conversion data of a target oligo and to calculate the likelihood of each NGS read being folded at any candidate structure during the reaction. For each target oligo, the model starts with two matrices as input: (1) the conversion data and (2) the open and closed states of all C positions from the proposed candidate structures that the oligo is possibly at (Fig. 3b). We obtained the conversion data exclusively from experiments, and we used the NUPACK software [9, 10] to generate candidate structures for oligos. These two inputs were each processed into two binary matrices: a converted matrix and an unconverted matrix from the conversion data (denoted as m_{BC} and m_{BU}), as well as an open matrix and a closed matrix from the structural patterns (denoted as M_{SO} and M_{SC}, Fig. 3c). Then we applied dot products to associate conversion data with structural patterns and gained four matrices (Fig. 3d):

$$Mtx_{SO}^{BU} = m_{BU} \cdot M_{SO}$$

$$Mtx_{SO}^{BC} = m_{BC} \cdot M_{SO}$$

$$Mtx_{SC}^{BU} = m_{BU} \cdot M_{SC}$$

$$Mtx_{SC}^{BC} = m_{BC} \cdot M_{SC}$$

The element on the i_{th} row, j_{th} column of matrix Mtx_{SO}^{BU} represents the number of C nucleotides in the i_{th} read being both unconverted in the reaction and open in the j_{th} proposed structure. These four matrices covered all C nucleotides in all four orthogonal conditions whether converted or not, open or not. By multiplying the four matrices to the four corresponding conditional probabilities and summing them together, we obtained the probabilities of each NGS read corresponding to each candidate structure (Fig. 3e). Note that the probabilities were calculated in the log scale because these conditional probabilities of all C positions are independent and should be multiplied for the overall probability, but the matrix dot product manipulates by summing. Log scale converts multiplying probability to adding, thus making it compatible with the matrix manipulation. Now for each NGS read, the structure with the highest (least negative) log-likelihood represents how the molecule was most likely folded during the reaction.

Model validation with positive control oligos. To analytically validate LYB-seq and the likelihood model, we first examined the NGS data from the Group 1, 2, and 4 oligos. For each oligo, we feed two candidate structures into the model: the designed minimum-free-energy (MFE) structure as structure 1 and a thermodynamically unfavored structure as structure 2. Details on the generation of these thermodynamically unfavored structures can be found in Supplementary Section S6.

Fig. 4a shows the model output of an example oligo from group 1, with the log probabilities of each read corresponding to each structure plotted as a heatmap. The two columns represent structure 1 (Fig. 4b) and structure 2 (Fig. 4c), and the rows represent all the NGS reads from this oligo. The color of each grid represents the log-scaled probability of this particular molecule (NGS read) folded at the corresponding structure. The heatmap has been sorted in the vertical axis for easier viewing. About 7,000 reads are consistent with both structures, as clustered on the upper half of the heatmap. This is understandable since the overall reaction rate is low and there are a number of molecules without sufficient conversions to specifically distinguish between the two structures. Ignoring these non-specific reads should not affect the inference we get on the relative concentrations of the structures from the data. The MFE structure was significantly supported by more than 4,000 reads, resulting in a supporting rate of 98.2% (fraction of reads with distinct likelihoods that support the structure). The remaining 1.8% of reads supporting structure 2 may be the result of sequencing errors, as described earlier. Fig. 4d shows the supporting rate of all designed oligos in Groups 1, 2, and 4; almost all of these exhibited high supporting rates for the Nupack-predicted mfe structure.

To test whether LYB-seq and our statistical model is capable of contradicting incorrect structures, we consider the Group 5 rationally designed oligos, that all have a pseudoknot secondary structure. A pseudoknot structure occurs when nucleotides in the loop of a hairpin participate in hybridization to a set of nucleotides outside of the loop. Due to the technical details of the Nussinov algorithm for mfe structure prediction, DNA secondary structures with pseudoknots cannot generally be predicted. Thus, the Nupack-predicted mfe structures for Group 5 oligos do not have pseudoknots, and are expected to be incorrect. For the Group 5 oligos, we compared the Nupack-predicted mfe structure (Structure 1) with the expected pseudoknot structure (Structure 2), and observed significantly lower supporting rate for the Nupack-predicted mfe than for Groups 1, 2, and 4 (Fig. 4e).

LYB-seq results on biological DNA sequences. We next analyzed the LYB-seq results on the 1,057 human DNA sequences. The selection process of these sequences from 212 human genes, and their chromosome locations, exon/intron status, GC content, and NUPACK predicted mfe structures are shown in the Supplementary Section S7. Due to technical limitations of NGS library preparation, 33 of these 1,057 sequences (3.1%) had too few NGS reads for reliable statistical analysis, and were discarded.

We first constructed a number of potential secondary struc-

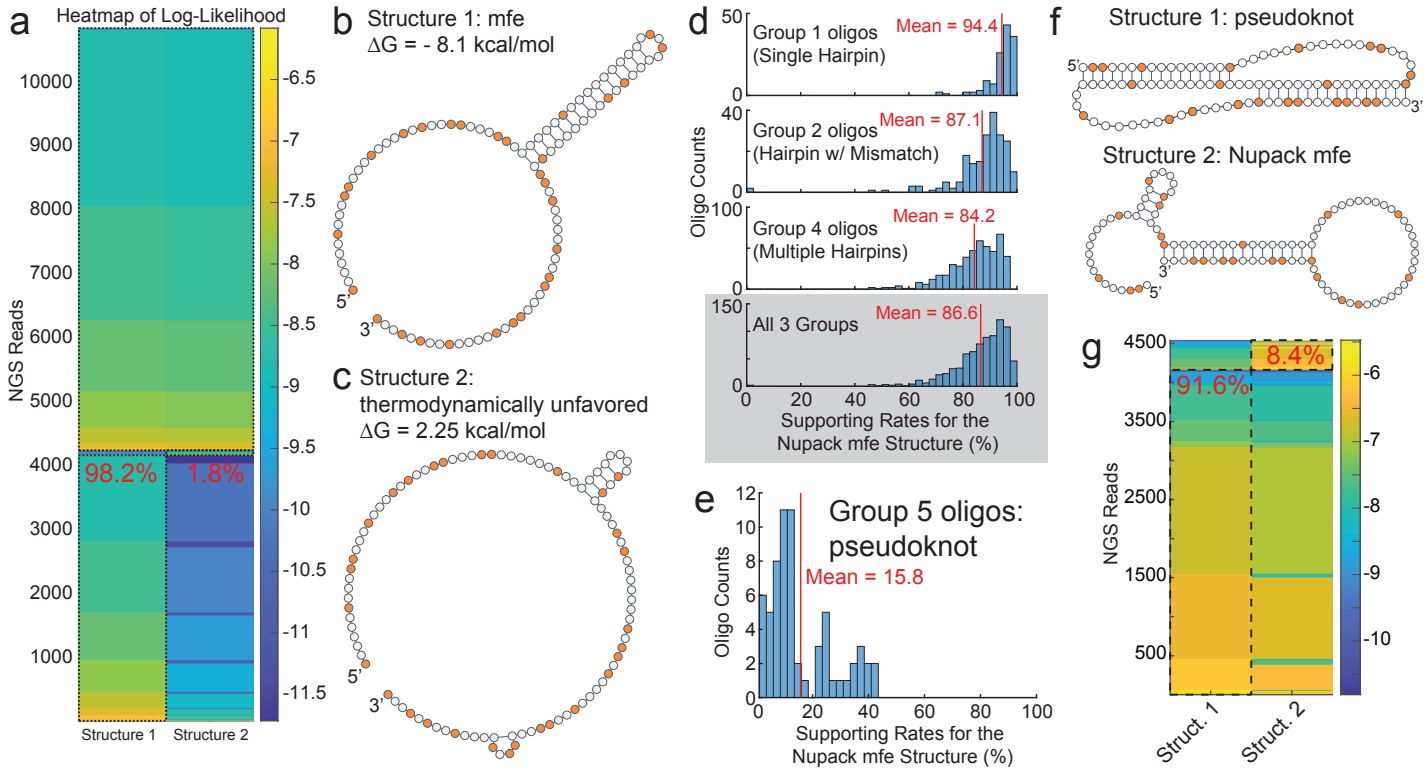

Figure 4: Predicted structures and computed likelihoods from LYB-seq data. **(a)** Heatmap of log-likelihood for 2 structures. The black dashed-line boxes mark reads supporting structure 1, reads supporting structure 2, and the non-specific reads that are consistent with both structures. The two percentage values colored in red are normalized supporting rates for the two structures. **(b)** Structure 1, the mfe structure of the oligo computed by Nupack. **(c)** Structure 2, a manually generated structure that is expected to be significantly less thermodynamically favorable. **(d)** Experimental summary of the fraction of NGS reads that support the mfe structures for oligos in Groups 1, 2, and 4. **(e)** Group 5 oligos comprise DNA sequences expected to form pseudoknots, which currently cannot be predicted by software such as Nupack. Consistent with our expectations, the fraction of NGS reads that support the mfe structures as much lower than in panel (d). **(f)** Expected pseudoknot structure and Nupack mfe structure for a Group 5 oligo. **(g)** The log-likelihood heatmap of the oligo shown in panel (f); the pseudoknot structure is strongly supported by LYB-seq.

tures for each DNA sequence using the Nupack suboptimal function [9, 10]. Initially, all structures with predicted folding energies within 2 kcal/mol of the minimum free energy (mfe) structure were retained. Because many structures obtained for a particular sequence were highly similar with only minor base pairing changes (e.g. base pair breathing at the end of hairpin stems), these do not really represent distinct metastable structural states that the molecule can adopt in solution. Thus, we next clustered these potential structures by pairwise Hamming distances. Within each cluster, the structure with lowest folding energy was selected as the representative structure (Fig. 5a).

For each sequence, the LYB-seq reads are compared against the representative structures two at a time (Fig. 5bc). An NGS read with significantly higher log-likelihood for one proposed structure over the other is viewed as supporting the former. The fraction of NGS reads supporting each structure can then be tallied. Because the overall bisulfite CY is low, there is a significant fraction of reads that exhibit similar log-likelihoods for both structures, these are excluded when considering the relative support for pairs of structures. In this way, we can determine both the number of distinct structures that co-exist in solution, and the fraction of DNA molecules that adopt the mfe structure predicted by Nupack (see also Supplementary Section S5).

We next considered how to distinguish between two scenarios of when a particular DNA sequence adopts (A) two distinct co-existing secondary structures in solution, vs. (B) a single structure that is not consistent with Nupack predictions. Because open (unpaired) C nucleotides have higher CY than closed C nucleotides (Fig. 2a), the quality of a secondary structure prediction for a particular set of NGS reads can be quantitated as R^{OC}, the ratio of the CY for the C nucleotides predicted to be open divided by the CY of the C nucleotides predicted to be closed. For our rationally designed DNA sequences, the value of R^{OC} is roughly 20 (8% vs. 0.4%). In contrast, a random secondary structure with no grounding in reality would presumably result in $R^{OC} \approx 1$. The higher the value of R^{OC}, the more confident that we can be of the correctness of the structure.

Consequently, the LYB-seq data for DNA sequences that adopt two distinct metastable secondary structures should be able to be divided into two subsets of reads, R1 and R2, with very different R^{OC} values when compared against two proposed structures S1 and S2. Fig. 5c and Fig. 6ab confirm that this is the case for many human DNA sequences. For

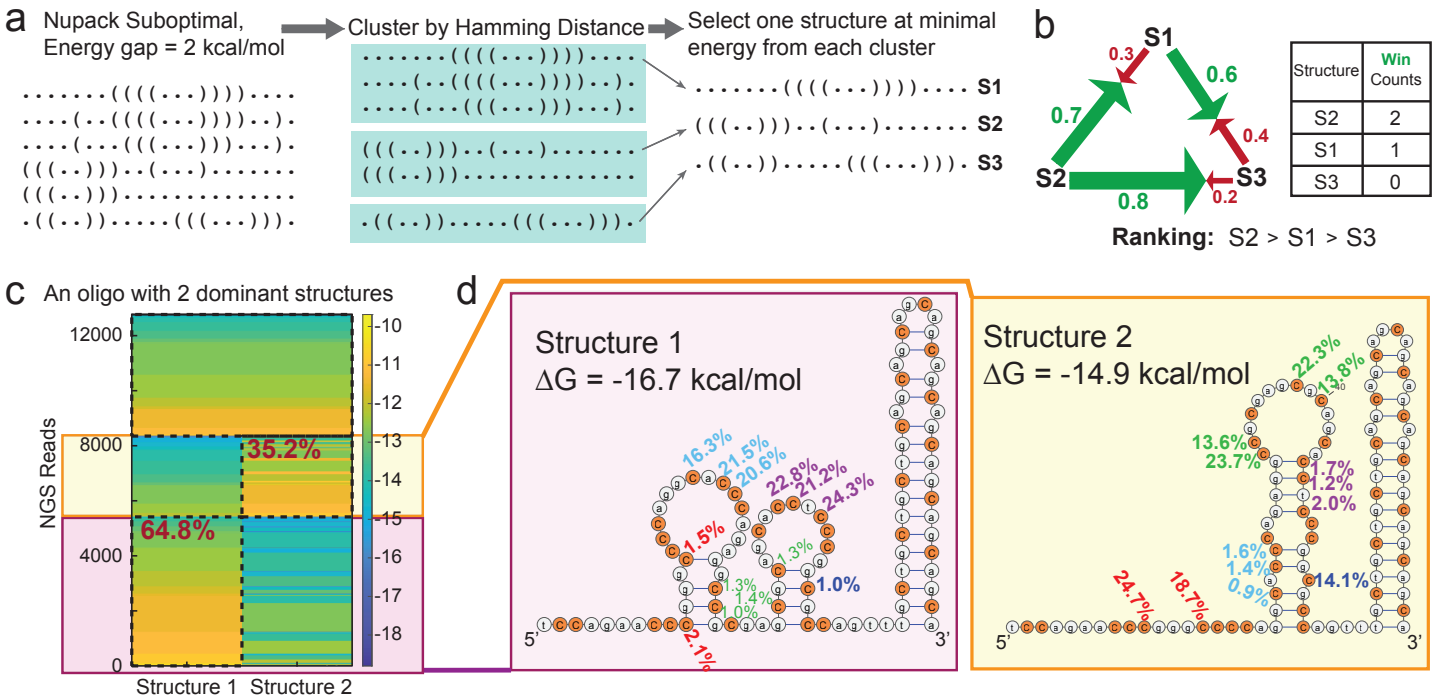

Figure 5: Using LYB-seq to assess the possibility and concentration distributions of multiple co-existing secondary structures for each DNA oligonucleotide. **(a)** Generation of the candidate structures for biological oligos. **(b)** Pairwise comparison method for determining the dominant secondary structure when more than 2 potential secondary structures exist. Supporting rates were calculated using the log-likelihood model for each pair of structures, and the number of winning match-ups are tallied for each structure. **(c)** In a significant fraction of biological oligo sequences, our analyses suggest that multiple secondary structures exist in solution. Shown here is one biological DNA sequence with strong evidence for co-existence of 2 DNA structures. **(d)** The two co-existing secondary structures from panel (c). The same colors represent the same C nucleotide loci. The purple, cyan, and green C nucleotide have highly distinctive conversion yields within the reads supporting the two different structures.

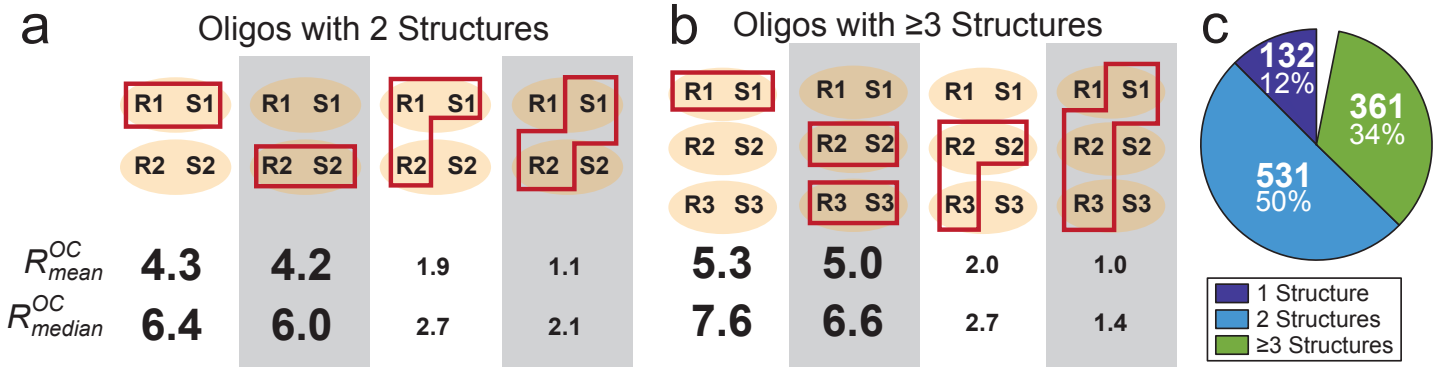

Figure 6: LYB-seq data shows that most biological DNA sequences can adopt multiple co-existing secondary structures. **(a)** Verification of multiple co-existing structures. R^{OC}_{mean} and R^{OC}_{median} correspond to the mean and median relative ratio CY of yields for C nucleotides predicted to be open vs. those predicted to be closed by the secondary structure. When the secondary structure is supported by the LYB-seq reads, we would expect the R^{OC} values to be high. In contrast, when the secondary structures are inconsistent with LYB-seq reads, the R^{OC} values should be low. R1 corresponds to the NGS reads that we group to be consistent with structure S1, and R2 corresponds to the NGS reads that we group to be consistent with structure S2. The R^{OC} values are for R1 reads compared to S1 and low for R2 reads compared with S1, as expected. **(b)** Verification of DNA oligos with 3 or more co-existing structures. Here, we only performed R^{OC} analysis for the top 3 structures for each oligo. **(c)** Distribution of the number of co-existing secondary structures observed in human DNA sequences based on LYB-seq. The remaining 33 sequences (out of 1,057) had too few NGS reads for confident secondary structure determination.

oligos with 2 co-existing structures, the R1 reads produce high R^{OC} values when compared to S1, but when the entire R1 and R2 datasets are combined, there is low R^{OC} value to S1. Of the 1,024 human DNA sequences, we find that 84% of them adopt 2 or more co-existing DNA secondary structures in solution (see also Supplementary S5, S8, and S9 for more

details).

Discussion

With accelerating advances in machine learning, it is now highly possible that complex problems such as nucleic acid secondary structure prediction from sequence can be more accurately solved using neural networks, rather than rules-based biophysical models. Machine learning approaches, however, require a massive amount of labeled instances to serve as training data. It has been heretofore difficult to obtain large and accurate secondary structure datasets, due to the low-throughput nature of X-ray crystallography and other techniques. LYB-seq leverages the secondary structure dependence on bisulfite reaction kinetics, and uses NGS to generate large secondary-structure informing datasets on thousands of different DNA species. In principle, this approach can be further scaled up to millions of distinct nucleic acid species, and thus can enable the development of machine learning based predictors of secondary structure.

The primary limitation of LYB-seq is that we can only observe the state of C nucleotides. Probing C nucleotides already gives us enough information to generate differential likelihoods for different proposed secondary structure for most NGS reads and most DNA sequences. However, it does mean that LYB-seq is not applicable to A/T-rich DNA sequences, and that a fraction of the NGS reads do not provide sufficient information to distinguish between two proposed secondary structures. Biological DNA with methylated cytosines would also be resistant to bisulfite conversion, results in lower confidence of base pair state inference for CpG loci.

From our LYB-seq NGS data, we observed that C nucleotides in small, topologically constrained hairpins have higher CY than those in dangles and large loops, by a factor of 3 to 5 (Fig. 2ab). For model simplicity and to maintain single-molecule resolution, we did not consider the differences in CY for different types of open C nucleotides. In principle, a more sophisticated model that attempts to infer hairpin loop size and relative position of paired C's to hairpin stem ends can provide more accurate and detailed information to confirm or disconfirm proposed secondary structures. Furthermore, given that the kinetic nature of bisulfite conversion bias based on C nucleotide state, it is possible that integrating data from multiple longitudinal LYB-seq timepoints can give more information to further ascertain secondary structure.

New species and classes of single-stranded DNA viruses [23, 24] and circular DNA [25–27] have recently been discovered shotgun sequencing analysis of microbiomes. The secondary structure of DNA molecules can be a critical link between the identified DNA sequences and their biological function and mechanisms. LYB-seq, in conjunction with DNA folding algorithms and software, can help determine the structures that these nucleic acid molecules adopt *in vitro* and *in vivo*.

Acknowledgements. This work was funded by NIH grants R01HG008752 and R01HG011356 to DYZ. The authors thank Lijin Zeng for editorial assistance.

Code Availability. NGS data alignment was processed by a tool written in F♯. The script for the generation of thermodynamically unfavored structures was written in Python. All other analysis was done in MATLAB. The F♯, Python, and MATLAB scripts are available at https://github.com/Gavin-J-Li/LYB-seq.

Author contributions. JB, JL, BY, AP, and DYZ conceived the project. JB designed the oligo sequences and developed the experimental workflow. JB and JL performed the experiment. BY aligned the NGS data to target oligos. JL performed statistical analysis and developed the model. JL and BY validated the model. JL, MXW, and JG performed analysis on biological oligos. JL and DYZ wrote the manuscript with input from all authors.

Additional information. Correspondence may be addressed to AP (Andrew.Phillips@microsoft.com) or DYZ (dyz1@rice.edu). DYZ is a co-founder and significant equity holder of Nuprobe Global, Torus Biosystems, and Pana Bio. JL is a consultant for Torus Biosystems. BY was a full-time employee of Microsoft during this research. MXW is a consultant for Nuprobe USA. JG and AP are full-time employees of Microsoft. There is a patent pending on the low-yield bisulfite sequencing.

[1] Aslanzadeh, V., Huang, Y., Sanguinetti, G., & Beggs, J. D. (2018). Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Research,* 28(2), 203,Äì213.

[2] Vandivier, L. E., Anderson, S. J., Foley, S. W., & Gregory, B. D. (2016). The Conservation and Function of RNA Secondary Structure in Plants. *Annual Review of Plant Biology,* 67(1), 463,Äì488.

[3] Burkhardt, D. H., Rouskin, S., Zhang, Y., Li, G.-W., Weissman, J. S., & Gross, C. A. (2017). Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *ELife,* 6, e22037.

[4] Chiaruttini, C., & Guillier, M. (2020). On the role of mRNA secondary structure in bacterial translation. *WIREs RNA,* 11(3).

[5] Trinquier, A., Durand, S., Braun, F., & Condon, C. (2020). Regulation of RNA processing and degradation in bacteria. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms,* 1863(5), 194505.

[6] Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences,* 73(23), 4433,Äì4448.

[7] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research,* 31(13), 3406-3415.

[8] Dirks, R. M., & Pierce, N. A. (2003). A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry,* 24(13), 1664-1677.

[9] SantaLucia, J., & Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics and Biomolecular Structure,* 33(1), 415-440.

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

[11] Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research,* 31(13), 3429-3431.

[12] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chemie Chemical Monthly,* 125(2), 167-188.

[13] Zhao, Y., Wang, J., Zeng, C., & Xiao, Y. (2018). Evaluation of RNA secondary structure prediction for both base-pairing and topology. *Biophysics Reports,* 4(3), 123-132.

[14] Kubota, M., Tran, C., & Spitale, R. C. (2015). Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology,* 11(12), 933-941.

[15] Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods,* 11(9), 959-965.

[16] Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., & Weeks, K. M. (2015). Selective 2'-hydroxyl acylation analyzed by primer exten-

sion and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature Protocols,* 10(11), 1643-1669.

[17] Li, H., & Aviran, S. (2018). Statistical modeling of RNA structure profiling experiments enables parsimonious reconstruction of structure landscapes. *Nature Communications,* 9(1), 606.

[18] Umeyama, T., & Ito, T. (2018). DMS-seq for In Vivo Genome-Wide Mapping of Protein-DNA Interactions and Nucleosome Centers. *Current Protocols in Molecular Biology,* 123(1), e60.

[19] Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., & Rouskin, S. (2017). DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods,* 14(1), 75-82.

[20] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B., & Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics,* 98(4), 288-295.

[21] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences,* 89(5), 1827-1831.

[22] Moran, S., Arribas, C., & Esteller, M. (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics,* 8(3), 389-399.

[23] Tisza, M. J., Pastrana, D. V., Welch, N. L., Stewart, B., Peretti, A., Starrett, G. J., & Buck, C. B. (2020). Discovery of several thousand highly diverse circular DNA viruses. *Elife,* 9, e51971.

[24] Noell, K., & Kolls, J. K. (2019). Further defining the human virome using NGS: identification of Redondoviridae. *Cell host & microbe,* 25(5), 634-635.

[25] Koche, R. P., Rodriguez-Fos, E., Helmsauer, K., Burkert, M., MacArthur, I. C., Maag, J., & Henssen, A. G. (2020). Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nature genetics,* 52(1), 29-34.

[26] Toptan, T., Abere, B., Nalesnik, M. A., Swerdlow, S. H., Ranganathan, S., Lee, N., & Chang, Y. (2018). Circular DNA tumor viruses make circular RNAs. *Proceedings of the National Academy of Sciences,* 115(37), E8737-E8745.

[27] Moller, H. D., Mohiyuddin, M., Prada-Luengo, I., Sailani, M. R., Halling, J. F., Plomgaard, P., & Regenberg, B. (2018). Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nature communications,* 9(1), 1-12.