FRAMEWORKS FOR LARGE-SCALE RNA STRUCTURE PROFILING IN TRANSCRIPTOMES AND DISEASE

Steven Busan

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry.

Chapel Hill
2015

Approved by:

Kevin M. Weeks

Gary J. Pielak

Eric M. Brustad

Michael B. Jarstfer

Alain Laederach

ABSTRACT

STEVEN BUSAN: Frameworks for large-scale RNA structure profiling in transcriptomes and disease.
(Under the direction of Kevin M. Weeks)

In addition to their role as intermediaries on the route to protein synthesis, RNA molecules have long been known to base-pair into complex structures that serve specific functions. Some structured RNAs play pathogenic roles, especially in viral illnesses and repeat-expansion disorders, and disease-associated RNA structures are potential therapeutic targets. SHAPE is a well-established chemical probing strategy to interrogate RNA flexibility and obtain high-quality structure models. The recent development of an unbiased experimental approach that allows SHAPE to characterize populations of diverse RNAs using massively parallel sequencing presented a challenging data analysis problem.

In this work, I apply SHAPE to study the relevance of huntingtin mRNA structure to Huntington's disease and discover that a classical CAG hairpin is likely absent or short in healthy-length transcripts. The formation of this hairpin correlates with increasing repeat length, which is a predictor of disease severity. I develop a fully-automated data analysis pipeline allowing for the extension of the SHAPE strategy to larger scales using mutational profiling (MaP), an approach that was applied to identify highly-structured elements within an HIV-1 genome. I further pursue a pilot analysis of a bacterial transcriptome MaP dataset obtained in a single experiment, demonstrate the nucleotide accuracy of MaP within this large sample, and apply alignment clustering to identify conserved motifs at the genomic scale. Together, these three projects highlight the power of SHAPE to identify specific RNA structures related to human disease and the value of robust experimental design and careful analysis in large-scale sequencing studies of RNA structure.

"An intellectual is a man who says a simple thing in a difficult way;

an artist is a man who says a difficult thing in a simple way."

- Bukowski

ACKNOWLEDGMENTS

My thanks go to Kevin Weeks for excellent career guidance, and for consistently advancing the RNA structure field. I thank my parents for the moral and monetary support that made my higher education possible.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xiii

1 INTRODUCTION 1

1.1 RNA structure 1

1.1.1 RNA structures as pathogenic agents and therapeutic targets 1

1.1.2 RNA structures within transcriptomes 3

1.2 SHAPE 4

1.2.1 Rationale 4

1.2.2 Chemical and enzymatic probing methods 4

1.2.3 Structure modeling 5

1.2.4 Development of SHAPE 6

1.3 Massively parallel sequencing applied to RNA structure 7

1.3.1 Massively parallel sequencing technologies 7

1.3.2 Previous reports linking parallel sequencing and RNA structure 8

SHAPE-Seq 8

Enzymatic methods 8

DMS probing 9

Hydroxyl radical footprinting 10

Limitations 10

1.4 Research overview 11

REFERENCES 12

2 ROLE OF CONTEXT IN RNA STRUCTURE: FLANKING SEQUENCES RECONFIGURE CAG MOTIF FOLDING IN HUNTINGTIN EXON 1 TRANSCRIPTS 17

2.1 Introduction 17

2.2 Methods 18

2.2.1 Sequences, primers, and antisense oligonucleotides 18

2.2.2 Transcript production and purification 19

2.2.3 *In vitro* transcript folding, SHAPE, and RNase T1 probing 19

2.2.4 Structure disruption using antisense oligonucleotides 20

2.2.5 Electropherogram analysis and structure prediction 20

2.3 Results 21

2.3.1 SHAPE and RNase probing of huntingtin exon 1 21

2.3.2 Structural models of huntingtin transcripts 22

2.3.3 CAG hairpin induction 24

2.4 Discussion 25

REFERENCES 32

3 SOFTWARE FOR THE AUTOMATED ANALYSIS OF SHAPE AND MUTATIONAL PROFILING (SHAPE-MAP) DATA 36

3.1 Introduction 36

3.2 SHAPE-MaP strategy 36
3.3 SHAPE-MaP data analysis pipeline (ShapeMapper) 40
3.3.1 Configuration 40
3.3.2 Quality trimming 40
3.3.3 Read alignment 40
3.3.4 Alignment parsing, ambiguous alignment removal, and mutation counting 42
3.3.5 Reactivity profile creation 45
3.3.6 Final data output 45
3.3.7 Automatic RNA folding and structure drawing by ShapeMapper 48
3.4 Hit level calculation and comparison with other reports 50
3.5 Conclusion 51
REFERENCES 52
4 HIGH-RESOLUTION MAP OF AN *E. COLI* TRANSCRIPTOME 55
4.1 Introduction 55
4.2 Experimental Methods 55
4.3 Software for transcript calling and curation 56
4.3.1 Automated transcript calling 56
4.3.2 Transcript call curation 57
4.3.3 Depth requirements 59
4.3.4 Standard error filter 59
4.4 Validation and global trends 60

4.4.1 Coverage and structure modeling statistics 60

4.4.2 Large-scale trends in *E. coli* transcript flexibility 67

4.5 Transcriptome-wide RNA structure motif discovery by local sequence and MaP clustering 71

4.5.1 Computational methods 71

Low-SHAPE regions 71

Sequence alignment 71

Distance matrix processing 72

Clustering 72

4.5.2 Results 72

4.6 Future improvements 78

4.7 Conclusion 79

REFERENCES 80

LIST OF TABLES

4.1 Sequencing statistics. 60

4.2 Structure modeling statistics for two previously characterized RNA structures. 62

LIST OF FIGURES

1.1 Hierarchy of RNA structure. 2

1.2 Reaction of SHAPE reagent with RNA. 7

2.1 SHAPE profiles for huntingtin exon 1 transcripts as a function of CAG-repeat length. 23

2.2 Structural models for representative normal and disease-associated huntingtin transcripts. 24

2.3 SHAPE analysis of huntingtin transcripts in the presence of antisense oligonucleotides designed to disrupt pairing between CAG sequences and flanking regions. 25

2.4 Role of flanking sequence in defining CAG-repeat RNA structures. 26

2.5 RNA structure probing profiles using RNase T1. 29

2.6 Secondary structure models for 23, 36, and 70-CAG repeat length huntingtin exon 1 transcripts. 30

2.7 Plausible competing structures for long CAG repeat sequences. 31

3.1 SHAPE-MaP overview. 38

3.2 Nucleotide-resolution interrogation of RNA structure. 39

3.3 ShapeMapper software overview. 41

3.4 Ambiguously aligned deletion identification. 43

3.5 Removal of ambiguously aligned deletions. 44

3.6 Example reactivity profiles. 46

3.7 Example histograms and troubleshooting. 47

3.8 Example structure drawing and coloring. 49

4.1 Transcript call curation. 58

4.2 *E. coli* genome coverage. 61

4.3 Transcript lengths and sequencing depths. 63

4.4 Representative SHAPE-MaP reactivity profiles. 64

4.5 Representative secondary structures. 65

4.6 Proposed secondary structure for the transcript encoding major membrane lipoprotein. 66

4.7 SHAPE reactivities across untranslated, protein-coding, and intergenic regions. 69

4.8 AU content and global transcript flexibility trends. 70

4.9 Sequence clusters. 74

4.10 REP elements. 75

4.11 Selected terminators. 76

4.12 Additional terminators. 77

LIST OF ABBREVIATIONS

1M7	1-methyl-7-nitroisatoic anhydride
ASO	antisense oligonucleotide
BAC	bacterial artificial chromosome
cDNA	complementary deoxynucleic acd
CMCT	1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate
ddNTP	dideoxynucleotide triphosphate
DMS	dimethylsulfate
DNA	deoxynucleic acid
DNase	deoxynuclease
E. coli	*Escherichia coli*
HCV	hepatitis C vrus
HD	Huntingtons disease
HIV	human immunodeficiency virus
HRF	hydroxyl radical footprinting
lpp	major membrane lipoprotein
MaP	mutational profiling
MBNL1	muscleblind-like
mRNA	messenger ribonucleic acid
ncRNA	non-coding ribonucleic acid
NMR	nuclear magnetic resonance
nt	nucleotide

PARS	parallel analysis of RNA structure
PCR	polymerase chain reaction
REP	repetitive extragenic palindrome
RNA	ribonucleic acid
RNase	ribonuclease
rRNA	ribosomal ribonucleic acid
SHAPE	selective 2′-hydroxyl acylation analyzed by primer extension
TE	tris ethylenediaminetetraacetic acid
TPP	thiamine pyrophosphate
tRNA	transfer ribonucleic acid
UTR	untranslated region
YAC	yeast artificial chromosome

1 INTRODUCTION

1.1 RNA structure

The description of the iconic DNA double helix in 1953[1] propelled speculation as to the possibility of RNA helices, first experimentally confirmed using X-ray diffraction in 1956[2]. By the early '60s, it was apparent that ribosomal RNA (rRNA), transfer RNA (tRNA), and certain plant virus RNAs contained a high proportion of anti-parallel helical elements[3], but the specific nature of these elements was unknown. In 1964, measurements of the reaction rates of formaldehyde with tRNA supported the idea that native ribonucleotides exist in one of three states: strongly hydrogen bonded (that is, base paired), partially constrained, or flexible[4]. The first full nucleotide sequence of a tRNA (primary structure) was published the following year, along with a set of proposed base pairs (secondary structure)[5]. An atomic-level description of the three-dimensional structure of a tRNA (tertiary structure) would not be obtained until 1974[6] (see Figure 1.1).

Over the last four decades, an array of specialized functional RNA structures have been described. These include small regulatory RNAs, catalytic RNAs, such as ribosomal RNAs and self-cleaving ribozymes, and riboswitches, RNAs that modulate gene expression by binding specific metabolites[8]. In all of these cases, RNA molecules do not simply carry sequence information, but instead fold into specific structural states that allow the chemical interactions necessary for their roles in the cell.

1.1.1 RNA structures as pathogenic agents and therapeutic targets

RNA structures play pathogenic roles in human diseases. Of particular relevance to Chapter 2 of this work are the subset of triplet repeat expansion disorders in which long repeat-expanded RNA gains a toxic function. For example, myotonic dystrophy type 1 is caused by the expansion of a CUG repeat region in a portion of the myotonin

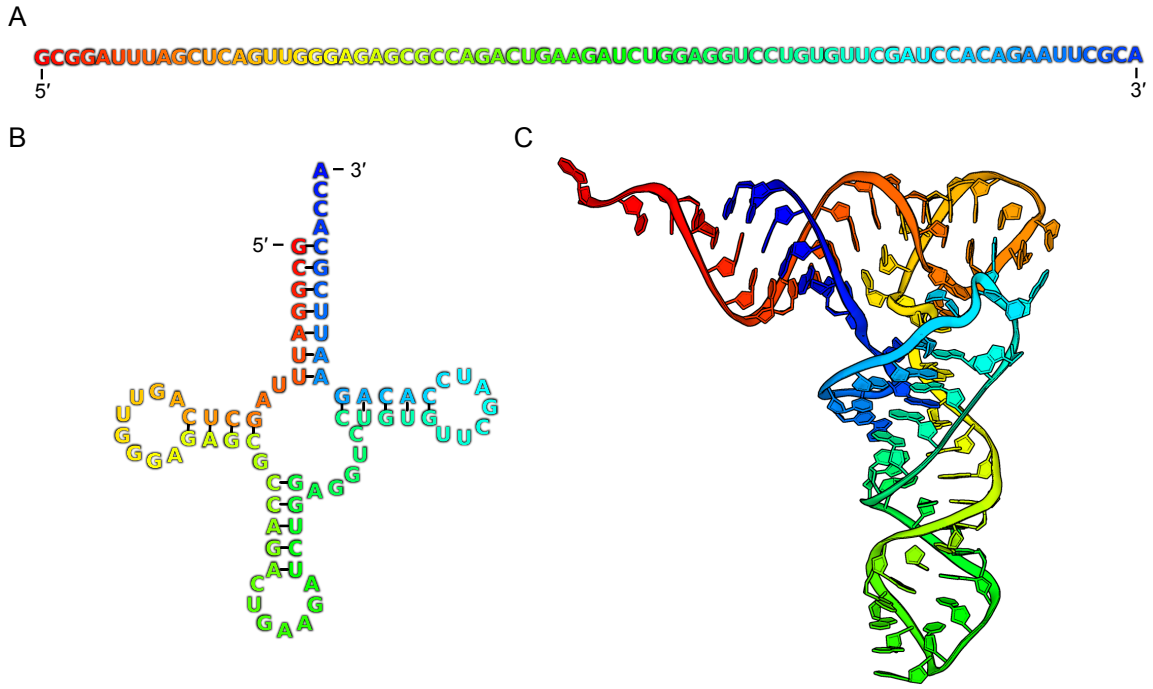

Figure 1.1: Hierarchy of RNA structure. (**A**) Primary, (**B**) secondary, and (**C**) tertiary structures of the yeast phenylalanine tRNA. Post-transcriptional modifications are ignored for simplicity. Three-dimensional model from Protein Data Bank entry 4TNA[7].

protein kinase gene that does not code for protein. Myotonic dystrophy type 1 appears to result from the nuclear sequestration of RNA-binding proteins such as muscleblind-like (MBNL1) by CUG hairpin helices and the improper RNA splicing that results[9]. Repeat-expanded RNAs are proposed to be the primary pathogenic agents in fragile X-associated tremor/ataxia syndrome and spinocerebellar ataxia type 8, and are suspected to cause pathogenic effects in many more triplet-repeat expansion diseases, including Huntington's disease and Huntington's disease-like syndrome 2[10].

Disease-associated RNA structures are potential targets for therapeutic intervention. Small molecule drugs are able to selectively bind certain structured RNAs. For example, several major classes of antibiotics specifically bind the bacterial ribosomal RNA[11]. In repeat-expansion disease research, a recent report identified a small molecule that binds CUG:CUG helices and inhibits the sequestration of MBNL1 protein in a cell culture

model of myotonic dystrophy[12]. Therapeutic development has not been limited to small molecules, however, as many groups have pursued the use of antisense oligonucleotides[13], modular peptoid scaffolds[14], zinc finger proteins[15], and antibodies[16] to bind RNA targets.

Disease-associated RNA structure targets are not limited to repeat expansion disorders and bacterial infection. Illnesses caused by viral infections provide another broad class of valuable RNA structure targets, since many viruses that infect humans use structured RNA at critical stages in their replication cycles. For example, human immunodeficiency virus (HIV) relies on a highly structured RNA element for the nuclear export of its messenger RNAs (mRNAs) and their eventual packaging into virus particles[17]. Hepatitis C virus (HCV) RNA contains a structure that serves as a ribosomal entry site, allowing for the translation of viral proteins[18]. Dengue virus contains two "dumbbell" structures important for RNA replication and translation[19]. The identification and characterization of functional RNA structures are necessary steps preceding nearly all efforts to therapeutically target RNA structure. These remain challenging problems, especially within the larger context of the transcriptome (all the RNAs produced in a cell).

1.1.2 RNA structures within transcriptomes

The development of massively parallel sequencing technologies (to be discussed briefly in Section 1.3.1) gave rise to the genomic era, creating a present in which the identities of billions of nucleotides of DNA and RNA are determined per day globally[20]. This vast landscape heightens the need for strategies to quickly identify specific structured RNAs in the transcriptome, since only a subset of all RNAs have evolved functional structures. A small number of highly-expressed non-coding transcripts have been fully structurally characterized, including the ribosome (the structural core of the protein translation machinery)[21], transfer-messenger RNA (responsible for releasing stalled ribosomes)[22], RNase P (a catalytic RNA cleaving the ends of tRNAs)[23], and 6S RNA (a transcription regulator)[24]. In addition, small stable hairpin structures that terminate

transcription have been identified in numerous locations in bacterial transcriptomes, largely by computational sequence searches[25]. To find functional RNA structures among the full complement of cellular RNAs, strategies are needed to rapidly locate RNAs with low free energies of folding. A first step along this route is to accurately map RNA flexibility at the transcriptome scale, something that many research groups are working toward (to be discussed in Section 1.3.2).

1.2 SHAPE

1.2.1 Rationale

Mapping RNA flexibility is an important tool in developing RNA structure models, and has been used since the early days of RNA structure analysis. For a transcript longer than a few dozen nucleotides, predicting which nucleotides will form base pairs is difficult by visual inspection of the sequence alone. Knowing which nucleotides are in highly constrained versus flexible structural states greatly reduces the magnitude of this problem. Therefore, chemistries and enzymes that react with RNA in a structure-selective manner have been exploited to improve structure models by providing this additional empirical information.

1.2.2 Chemical and enzymatic probing methods

The earliest plausible secondary structure models for RNAs longer than 100 nucleotides were developed by visually attempting to maximize the number of base pairs while leaving single-stranded those nucleotides that showed sensitivity to cleavage by various agents. For example, in 1978, a long rod-like structure model was proposed for a 359-nucleotide potato spindle tuber virus RNA, using bisulphite modification (which preferentially reacts with single-stranded cytosine residues) and a number of structure-specific ribonuclease (RNase) digests[26]. In this and other early studies, both nucleotide sequences and the locations of RNA cleavage or modification were detected by labor-intensive two-dimensional gel eletrophoretic techniques.

More efficient methods to locate sites of cleavage or modification were later developed, the most useful of which has been reverse transcription primer extension. This method relies on the annealing of a labeled DNA primer to the $3'$ end of an RNA, the extension of this primer by a reverse transcriptase enzyme, and the resulting production of DNA fragments whose $3'$ ends correspond to sites of RNA cleavage or modification[27]. These DNA fragments are resolved by gel or capillary electrophoresis to determine the locations and magnitudes of cleavage or modification[28].

Reverse transcription primer extension can be used to quantify the levels of RNA reactivity with a wide variety of chemical probes and nucleases. Common chemical examples are DMS (which reacts with single-stranded adenosines and cytosines), CMCT (which modifies single-stranded uridines and guanosines), kethoxal (which modifies single-stranded guanosines), and in-line probing (the spontaneous cleavage of flexible nucleotides in ionic solution)[29]. Primer extension can also report on chemical probes that measure nucleotide properties other than flexibility. For example, nucleotides cleaved by hydroxyl radical attack tend to be located on the surface of a given molecule and exposed to the surrounding solvent[30]. Popular enzymatic nucleases include RNase I (which cleaves upstream of single-stranded nucleotides of all types) and RNase V1 (which cleaves base-paired helical regions), often used in tandem[31]. However, the comparatively large size of RNase proteins precludes true single-nucleotide measurement[32]. As useful as all these reagents have been, accurately modeling RNA structures using incomplete or nucleotide-biased data is challenging[33].

1.2.3 Structure modeling

Efforts to automate RNA structure prediction have gradually improved over time. The first software to achieve limited success estimated the free energy of each possible structure by summing the modeled free energy contributions of each base pair, loop, and bulge[34]. Data from more extensive thermodynamic experiments were incorporated in a later version

of the software, estimating the energetic contributions of base pairs two at a time instead of alone (referred to as a "nearest-neighbor" energy model)[35]. The development of a dynamic programming algorithm for efficiently computing a minimum free energy RNA structure in 1980 brought RNA structure modeling within practical reach[36]. Thermodynamic models have continued to be refined[37–39], but *de novo* structure modeling accuracies remain modest for RNAs over a few dozen nucleotides in length[40].

1.2.4 Development of SHAPE

The limitations of prior chemical and enzymatic RNA structure probing methods spurred the development of selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE), first reported in 2005[41]. SHAPE reagents are electrophiles that selectively react with the 2′-hydroxyl group on the RNA backbone, a group common to all four ribonucleotides (see Figure 1.2). SHAPE therefore reports on the local flexibility (and by proxy, base pairing status) of nearly all the nucleotides in a given RNA molecule, with very little bias[41]. The covalent adducts produced by the reaction of SHAPE reagents with RNA are quantifiable by primer extension, similar to other chemical modifications or cleavage products previously discussed. SHAPE was first applied to measure the flexibility of tRNA nucleotides[41], and was subsequently applied to study the structures of a range of transcripts, from small riboswitch domains[42], to a bacterial ribosome[43], to an entire HIV-1 genome[44], and many others[45–48].

SHAPE is now perhaps the only RNA structure probing strategy enabling robust high-quality structure modeling for RNAs of realistic length. A three-reagent SHAPE experiment provides sufficient information to allow the creation of structure models with consistently greater than 90% correct base pairs[40]. Structure predictions are obtained by minimizing a nearest-neighbor free energy model using the software RNAstructure with SHAPE reactivities input as additional pseudo-free energies[42].

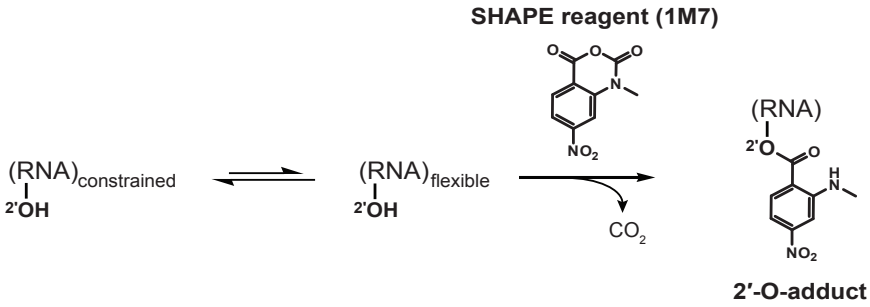

Figure 1.2: Reaction of SHAPE reagent with RNA. Shown is a simplified reaction pathway of the SHAPE reagent 1-methyl-7-nitroisatoic anhydride (1M7) with RNA. Flexible nucleotides sample conformations favorable to reaction with 1M7 more than nucleotides that are conformationally constrained, leading to selective adduct formation on single-stranded flexible nucleotides.

1.3 Massively parallel sequencing applied to RNA structure

1.3.1 Massively parallel sequencing technologies

The chain termination method of DNA sequencing, also called Sanger sequencing, was first published in 1977[49]. A labeled DNA primer is extended complementary to a template strand by DNA polymerase. Small concentrations of chain-terminating dideoxynucleotide triphosphates (ddNTPs) are included in four separate reactions, one for each nucleotide, resulting in the formation of cDNA fragments that are resolved by gel electrophoresis. The chain termination method has been streamlined using fluorescent ddNTPs and capillary electrophoresis[50], but is limited to generating sequences of about one thousand nucleotides at a time. Methods that provide increased throughput by performing hundreds of thousands of primer extension reactions in parallel have been critical to genome-scale sequencing projects[51].

One such approach is pyrosequencing, in which the incorporation of specific dNTPs is detected as a luminescent signal from an enzymatic cascade. Pyrosequencing instruments can generate about 100 million nucleotides of sequence information in a single run in a single day[52]. A second popular approach uses fluorescent, reversible chain terminators to extend primers against template DNA. Illumina instruments using this method are capable

of generating billions of nucleotides of sequence in a single run[53].

1.3.2 Previous reports linking parallel sequencing and RNA structure

Several groups have recently reported methods for probing RNA structure while taking advantage of massively parallel sequencing technologies. These methods have used SHAPE, other chemical probes, and enzymatic and hydroxyl radical footprinting to measure RNA flexibility or solvent accessibility. Several of these methods have been applied to the characterization of large-scale RNA structure trends in transcriptome studies.

SHAPE-Seq

A method combining SHAPE with parallel sequencing, SHAPE-Seq[54], was reported in 2011. Reverse transcription of SHAPE-modified RNA is performed using designed primers, producing cDNA with known $5'$ ends and stops corresponding to SHAPE adduct sites. Known sequence adapters are added on the $3'$ end by DNA-DNA ligation, allowing polymerase chain reaction (PCR) amplification and sequencing. SHAPE-Seq was demonstrated to allow the simultaneous probing of mutant RNAs in complex mixtures in a single experiment. However, because it relies on designed primers for the initial reverse transcription, this method is limited in practice to studying RNAs of several hundred nucleotides, and not easily extended to transcriptome-scale experiments.

Enzymatic methods

Two methods for resolving enzymatic cleavage experiments are FragSeq[55] and parallel analysis of RNA structure (PARS)[56], both reported in 2010. In the FragSeq method, RNAs are treated with RNase P1, a nuclease that preferentially cleaves single-stranded RNA. The resulting RNA fragments are then ligated to DNA adapters of known sequence, followed by reverse transcription, PCR amplification, and parallel sequencing. Obtained sequences are aligned to a reference sequence, and RNase cut site counts are compared with undigested and polynucleotide kinase-treated controls. This method was demonstrated on the transcriptomes of mouse embryonic stem cells and differentiated neural precursor cells,

primarily focusing on the structures of a small number of non-coding RNAs (ncRNAs) shorter than 200 nucleotides. The PARs method is similar, using enzymatic cleavage followed by adapter ligation, PCR, and sequencing. PARs, however, uses two RNases, one that cleaves single-stranded regions, and one that cleaves helical (base-paired) regions. PARs was applied to a yeast transcriptome, showing that untranslated regions (UTRs) tend to be less structured than coding regions, and that translation start and stop codons tend to be single-stranded. A three-nucleotide periodic trend in structure within coding regions was also reported. When applied to human transcriptomes, PARs provided evidence that human UTRs are overall more structured than human coding regions, and that natural sequence variation changes RNA structure at thousands of sites between individuals[57].

DMS probing

Three strategies for coupling DMS probing and massively parallel sequencing have been reported to date: structure-seq[58], DMS-seq[59], and Mod-seq[60]. Structure-seq uses random primers with known adapter sequence for reverse transcription of DMS-modified or control RNA, producing cDNAs with known $5'$ ends. $3'$ adapters are then ligated to the cDNA, allowing PCR amplification and sequencing. Sites of DMS modification are detected as reverse transcription stops. DMS-seq and Mod-seq follow similar protocols, with subtle differences in size selection and PCR steps designed to enrich for DNA fragments resulting from DMS-induced reverse transcription stops. Structure-seq was applied to RNA from *Arabidopsis thaliana*, a model plant species. This showed that *A. thaliana* UTRs are more flexible than coding regions, that a short less-structured region exists upstream of translation start codons, and that the first nucleotide in each codon of highly-translated transcripts is on average more flexible than the second and third positions, an effect not explained by sequence identity[58]. DMS-seq was applied to RNA from yeast and cultured human cells, under *in vivo*, *in vitro* refolded, and denatured conditions. *In vitro* refolded RNA appeared the least reactive (the most structured), while *in vivo*-modified

RNA showed intermediate reactivity between *in vitro* and denatured. This evidence, along with data from follow-up experiments, was interpreted to show the effects of both active RNA helicases and passive RNA-binding proteins on RNA unfolding in cells[59].

Hydroxyl radical footprinting

A recent report described a method for coupling hydroxyl radical footprinting and massively parallel sequencing, called HRF-Seq[61]. This method uses reverse transcription of cleavage products followed by adaptor ligation and PCR. HRF-Seq attempts to rigorously account for PCR biases using randomized barcode primers and a computational correction. HRF-Seq also corrects for variation in sequencing coverage, unlike the previously mentioned methods, which assume uniform coverage. This is a critical difference, as reverse transcription sequencing coverages often greatly vary over the length of a given RNA.

Limitations

The methods described in this section have provided strong evidence for various global trends in RNA structure. All these methods, however, share common weaknesses. First, a single-stranded RNA-DNA or DNA-DNA ligation is one of the first steps after chemical modification or cleavage in all these protocols. These reactions are inefficient and highly biased by structure[62]. Second, with the possible exception of SHAPE-seq, none of these methods provide true single-nucleotide resolution, as they use reagents or enzymes that only react with a subset of the four nucleotides in RNA. Third, none of these methods have been shown to produce consistently accurate secondary structure models, likely as a consequence of the previous two limitations. For example, despite using over 80 million reads mapping to ribosomal sequence, the initial structure-seq publication reported that the inclusion of DMS probing data was unable to improve the modeling of the small subunit of the yeast ribosome above 50% accuracy[58] (for comparison, SHAPE data improve the modeling accuracy of a bacterial ribosomal small subunit to 97%[43]). Furthermore, in the

race to publish large-scale structural studies, some authors have included data in their analyses with extremely low signal above background (see Section 3.4). The authors of Mod-seq and SHAPE-seq have noted the importance of deep sequencing coverage and statistical significance[54,60].

1.4 Research overview

Chapter 2 describes my work applying SHAPE to study the structure of the mRNA associated with the triplet repeat expansion disorder Huntington's disease. I probed the structure of five *in vitro* transcripts covering the first exon of huntingtin, including CAG repeat regions from 17 to as long as 70 triplets. Chapter 3 describes ShapeMapper, a fully-automated software data analysis pipeline that I developed to enable the extension of the SHAPE strategy to larger scales using mutational profiling (MaP). Chapter 4 describes a pilot study of the RNA structures present in the *Escherichia coli* transcriptome using SHAPE-MaP, showing the power of both the mutational profiling strategy and the broad utility of the ShapeMapper software.

REFERENCES

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (Apr. 1953).

2. Rich, A. & Davies, D. R. a new two stranded helical structure: polyadenylic acid and polyuridylic acid. *J. Am. Chem. Soc.* **78,** 3548–3549 (Jan. 1956).

3. Spencer, M, Fuller, W, Wilkins, M. H. F. & Brown, G. L. Determination of the Helical Configuration of Ribonucleic Acid Molecules by X-Ray Diffraction Study of Crystalline Amino-Acid–transfer Ribonucleic Acid. *Nature* **194,** 1014–1020 (June 1962).

4. Marciello, R. & Zubay, G. Quantitative studies on the rate of reaction of adapter RNA with formaldehyde. *Biochem. Biophys. Res. Commun.* **14,** 272–275 (Jan. 1964).

5. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147,** 1462–1465 (Jan. 1965).

6. Kim, S. H. *et al.* Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185,** 435–440 (Aug. 1974).

7. Hingerty, B, Brown, R. S. & Jack, A. Further refinement of the structure of yeast tRNAPhe. *J. Mol. Biol.* **124,** 523–534 (Sept. 1978).

8. Serganov, A. & Patel, D. J. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.* **8,** 776–790 (Sept. 2007).

9. Lee, J. E. & Cooper, T. A. Pathogenic mechanisms of myotonic dystrophy. *Biochem. Soc. Trans.* **37,** 1281–1286 (Dec. 2009).

10. Krzyzosiak, W. J. *et al.* Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.* **40,** 11–26 (Jan. 2012).

11. Lambert, T. Antibiotics that affect the ribosome. *Rev. Sci. Tech.* **31,** 57–64 (Apr. 2012).

12. Hoskins, J. W. *et al.* Lomofungin and dilomofungin: inhibitors of MBNL1-CUG RNA binding with distinct cellular effects. *Nucleic Acids Res.* (Jan. 2014).

13. Kole, R., Krainer, A. R. & Altman, S. RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat. Rev. Drug Discovery* (Jan. 2012).

14. Childs-Disney, J. L., Tsitovich, P. B. & Disney, M. D. Using Modularly Assembled Ligands To Bind RNA Internal Loops Separated by Different Distances. *Chembiochem* **12,** 2143–2146 (Jan. 2011).

15. Font, J. & Mackay, J. P. Beyond DNA: zinc finger domains as RNA-binding modules. *Methods Mol. Bio.* **649,** 479–491 (2010).

16. Sherman, E. M., Holmes, S. & Ye, J.-D. Specific RNA-binding antibodies with a four-amino-acid code. *J. Mol. Biol.* **426,** 2145–2157 (May 2014).

17. Fernandes, J., Jayaraman, B. & Frankel, A. The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biol.* **9,** 6–11 (Jan. 2012).

18. Lukavsky, P. J. Structure and function of HCV IRES domains. *Virus Res.* **139,** 166–171 (Feb. 2009).

19. Manzano, M. *et al.* Identification of cis-acting elements in the 3'-untranslated region of the dengue virus type 2 RNA that modulate translation and replication. *J. Biol. Chem.* **286,** 22521–22534 (June 2011).

20. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta* (July 2014).

21. Jenner, L. *et al.* Crystal structure of the 80S yeast ribosome. *Curr. Opin. Struct. Biol.* **22,** 759–767 (Dec. 2012).

22. Felden, B *et al.* Probing the structure of the Escherichia coli 10Sa RNA (tmRNA). *RNA* **3,** 89–103 (Jan. 1997).

23. Guerrier-Takada, C & Altman, S. Structure in solution of M1 RNA, the catalytic subunit of ribonuclease P from Escherichia coli. *Biochemistry (Mosc.)* **23,** 6327–6334 (Dec. 1984).

24. Trotochaud, A. E. & Wassarman, K. M. A highly conserved 6S RNA structure is required for regulation of transcription. *Nat. Struct. Mol. Biol.* **12,** 313–319 (Apr. 2005).

25. Lesnik, E. A. *et al.* Prediction of rho-independent transcriptional terminators in Escherichia coli. *Nucleic Acids Res.* **29,** 3583–3594 (Sept. 2001).

26. Gross, H. J. *et al.* Nucleotide sequence and secondary structure of potato spindle tuber viroid. *Nature* **273,** 203–208 (May 1978).

27. Ehresmann, C. *et al.* Probing the structure of RNAs in solution. *Nucleic Acids Res.* **15,** 9109–9128 (Jan. 1987).

28. Fekete, R. A., Miller, M. J. & Chattoraj, D. K. Fluorescently labeled oligonucleotide extension: a rapid and quantitative protocol for primer extension. *Biotechniques* **35,** 90–4–97–8 (July 2003).

29. Regulski, E. & Breaker, R. In-Line Probing Analysis of Riboswitches. *Methods Mol. Biol.* (ed Wilusz, J.) 53–67 (2008).

30. Tullius, T. D. & Greenbaum, J. A. Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr. Opin. Chem. Biol.* **9,** 127–134 (Apr. 2005).

31. Brown, T. S. & Bevilacqua, P. C. Method for assigning double-stranded RNA structures. *Biotechniques* **38,** 368–370–372 (Mar. 2005).

32. Gohda, K, Oka, K, Tomita, K & Hakoshima, T. Crystal structure of RNase T1 complexed with the product nucleotide 3'-GMP. Structural evidence for direct interaction of histidine 40 and glutamic acid 58 with the 2'-hydroxyl group of the ribose. *J. Biol. Chem.* **269,** 17531–17536 (July 1994).

33. Quarrier, S., Martin, J. S., Davis-Neulander, L., Beauregard, A. & Laederach, A. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA* **16,** 1108–1117 (June 2010).

34. Tinoco, I. J., Uhlenbeck, O. C. & Levine, M. D. Estimation of secondary structure in ribonucleic acids. *Nature* **230,** 362–367 (Apr. 1971).

35. Tinoco, I. J. *et al.* Improved estimation of secondary structure in ribonucleic acids. *Nat. New Biol.* **246,** 40–41 (Nov. 1973).

36. Zuker, M & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148 (Jan. 1981).

37. Xia, T. *et al.* Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with WatsonCrick Base Pairs. *Biochemistry (Mosc.)* **37,** 14719–14735 (Jan. 2014).

38. Lu, Z. J., Turner, D. H. & Mathews, D. H. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.* **34,** 4912–4924 (Jan. 2006).

39. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **38** (2010).

40. Rice, G. M., Leonard, C. W. & Weeks, K. M. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA* **20,** 846–854 (June 2014).

41. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127,** 4223–4231 (Mar. 2005).

42. Hajdin, C. E. *et al.* Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.* (2013).

43. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 97–102 (Jan. 2009).

44. Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460,** 711–716 (Aug. 2009).

45. Grohman, J. K. *et al.* An immature retroviral RNA genome resembles a kinetically trapped intermediate state. *J. Virol.* **88,** 6061–6068 (June 2014).

46. Archer, E. J. *et al.* Long-range architecture in a viral RNA genome. *Biochemistry (Mosc.)* **52,** 3182–3190 (May 2013).

47. Giguère, T., Raj Adkar-Purushothama, C. & Perreault, J.-P. Comprehensive Secondary Structure Elucidation of Four Genera of the Family Pospiviroidae. *PLoS ONE* **9,** e98655 EP – (Jan. 2014).

48. Grohman, J. K., Kottegoda, S., Gorelick, R. J., Allbritton, N. L. & Weeks, K. M. Femtomole SHAPE reveals regulatory structures in the authentic XMRV RNA genome. *J. Am. Chem. Soc.* **133,** 20326–20334 (Dec. 2011).

49. Sanger, F, Nicklen, S & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (Dec. 1977).

50. Karger, B. L. & Guttman, A. DNA sequencing by CE. *Electrophoresis* **30,** S196–S202 (Jan. 2009).

51. Tucker, T., Marra, M. & Friedman, J. M. Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* **85,** 142–154 (Aug. 2009).

52. Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92,** 255–264 (Nov. 2008).

53. Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. *J.Biomed. Biotechnol.* **2012,** 11 (2012).

54. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11063–11068 (July 2011).

55. Underwood, J. G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* **7,** 995–1001 (Dec. 2010).

56. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (Sept. 2010).

57. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (Jan. 2014).

58. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (Jan. 2014).

59. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505,** 701–705 (Jan. 2014).

60. Talkish, J., May, G., Lin, Y., Woolford, J. L. J. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20,** 713–720 (May 2014).

61. Kielpinski, L. J. & Vinther, J. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic Acids Res.* **42,** e70 (Apr. 2014).

62. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res.* **40,** e54 (Apr. 2012).

2 ROLE OF CONTEXT IN RNA STRUCTURE: FLANKING SEQUENCES RECONFIGURE CAG MOTIF FOLDING IN HUNTINGTIN EXON 1 TRANSCRIPTS[1]

2.1 Introduction

Huntington's disease (HD) is a devastating, ultimately fatal neurodegenerative disorder. In healthy individuals, the first exon of each of the two alleles of the huntingtin gene contains a relatively short region of CAG triplet repeats that encode polyglutamine; the most common allele has 17 repeats. In HD patients, one huntingtin allele is abnormally expanded to contain between 36 and 70 CAG repeats, although patient alleles with shorter or significantly longer repeat regions have also been reported.[1,2] The length of this HD-expanded CAG-repeat region is inversely correlated with patient age at the onset of symptoms, which include involuntary movements and dementia.[2] Pathogenesis is due to the polyglutamine peptides translated from the disease allele, and the expanded CAG repeat-containing RNA transcripts may also be toxic.[3,4]

This study was motivated by the potential for allele-selective therapeutic targeting of the huntingtin mRNA that might result if the RNA structure could be modeled with confidence. Huntingtin is nearly universally expressed and appears to be especially important for correct functioning of the adult nervous system.[5-8] An ideal therapeutic would therefore specifically destroy the disease-expanded huntingtin transcript or block its translation while preserving the function of the healthy length transcript. Recent efforts to selectively target the expanded huntingtin transcript have focused either on targeting single-nucleotide polymorphisms associated with disease alleles[9,10] or on targeting the CAG repeats, taking advantage of the greater number of effective binding sites in the

[1]This chapter previously appeared as an article in Biochemistry. The original citation is as follows: Busan S, Weeks KM. "Role of context in RNA structure: flanking sequences reconfigure CAG motif folding in huntingtin exon 1 transcript," *Biochemistry* 52, no. 46 (November 2013): 8219-25.

expanded transcript.[11] Allele-specific structures within the huntingtin mRNA could provide additional, and more precise, targets for therapeutic development.

Biochemical studies have consistently demonstrated that RNA transcripts containing CAG repeats fold into duplex helices and hairpins.[12-15] CAG-containing duplexes have been examined by NMR[14] and X-ray crystallography.[16] Recent studies have also shown that flanking sequences can modulate triplet-repeat folding. The addition of even a short region of flanking huntingtin sequence to CAG repeats results in the formation of more complex structures.[17] We therefore sought to determine the folded structures of huntingtin transcripts with varying CAG repeat lengths in the context of the sequence of longer transcripts, more closely resembling those found in cells.

We designed five transcripts covering the entire first exon of the huntingtin mRNA. These exon 1 transcripts spanned the 5′ untranslated region (UTR), contained from 17 to 70 CAG repeats, and included the downstream region encoding polyproline repeats (mostly CCG). A combination of SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension), RNase T1 cleavage, and targeted antisense oligonucleotide binding was used to investigate the folded structures of these transcripts. We found that the sequence context had profound effects on the folded structure of the transcript because CAG repeats pair extensively with flanking huntingtin mRNA sequences. A CAG hairpin was absent or short in huntingtin transcripts with repeat lengths typical of healthy individuals (17 and 23 repeats) but was present in transcripts with disease-associated numbers of repeats (36, 41, and 70 repeats). Our data suggest that there are structural differences between healthy and disease-inducing alleles that may be promising targets for therapeutic intervention.

2.2 Methods

2.2.1 Sequences, primers, and antisense oligonucleotides

The sequence of the huntingtin mRNA exon 1 transcript is as follows (n = 17, 23, 36, 41, and 70): GCUGCCGGGA CGGGUCCAAG AUGGACGGCC

GCUCAGGUUC UGCUUUUACC UGCGGCCCAG AGCCCCAUUC AUUGCCC-CGG UGCUGAGCGG CGCCGCGAGU CGGCCCGAGG CCUCCGGGGA CUGC-CGUGCC GGGCGGGAGA CCGCCAUGGC GACCCUGGAA AAGCUGAUGA AG-GCCUUCGA GUCCCUCAAG UCCUUC **(CAG)*n*** CAACAGCCGC CACCGCCGCC GCCGCCGCCG CCGCCUCCUC AGCUUCCUCA GCCGCCGCCG CAGGCACAGC CGCUGCUGCC UCAGCCGCAG CCGCCCCCGC CGCCGCCCCC GCCGCCACCC GGCCCGGCUG UGGCUGAGGA GCCGCUGCAC CGACC. The reverse-transcription primer is GGTCGGTGCAGCG, and the antisense oligonucleotides, listed by the 5′-most target nucleotide in the 70-CAG huntingtin transcript (* indicates a locked nucleotide(18)) are (1) *TCC*CGG*CAG*C, (159) *ATC*AGC*TTT*T, (431) *AGG*AGG*CG*GCG*GCG*G, (464) *GTG*CCT*GCG*G, and (475) *TGA*GGC*AG*CAG*CGG*C.

2.2.2 Transcript production and purification

Plasmids contained huntingtin sequences, a T7 promoter at the 5′ end, and a Bts I restriction site at the 3′ end and were obtained by de novo synthesis (Blue Heron Biotechnology). Cells (SURE 2, Agilent Technologies) were transfected with plasmid, and 500 mL cultures were prepared. Plasmids were extracted, and constructs were verified by sequencing. Plasmids were linearized with Bts I (New England Biolabs), and linearization was confirmed by agarose gel electrophoresis. Linearized template sequences were transcribed using T7 RNA polymerase, and products were separated by polyacrylamide gel electrophoresis, excised from the gel, and recovered by precipitation with ethanol.[19] Transcripts were resuspended at 0.25 M in 1/2 TE buffer, aliquoted for single use, and stored at 20°C.

2.2.3 In vitro transcript folding, SHAPE, and RNase T1 probing

Transcripts were denatured at 95°C for 2 min, snap-cooled on ice for 2 min, and refolded at 37°C for 30 min in 50 mM Tris-HCl (pH 8), 75 mM KCl, and 3 mM MgCl2.

SHAPE probing was performed using 58 mM final concentration 1-methyl-7-nitroisatoic anhydride (1M7) for 5 min at 37°C.[20] Enzymatic cleavage was carried out using RNase T1 (Ambion) at a final concentration of 0.2 U/L for 5 min at 37°C. Transcripts were recovered by ethanol precipitation. SuperScript III reverse transcriptase (Invitrogen) was used to extend the fluorescently labeled reverse-transcription primer (above) for 1 h at 37°C. Fluorescent cDNA fragments were quantified using capillary electrophoresis.[21]

2.2.4 Structure disruption using antisense oligonucleotides

Transcripts were combined with five pooled antisense oligonucleotides (ASOs), containing locked nucleotides (Exiqon) to enhance RNA binding, at a 4-fold excess of each ASO over RNA. Transcripts were then denatured, snap-cooled, folded, and modified as described above. To reduce the concentration of ASOs prior to reverse transcription, transcripts were incubated with three DNA oligonucleotides complementary to ASOs 431, 464, and 475 at a high concentration (200 times that of the RNA) at 95°C for 2 min. Three serial rounds of binding, washing, and elution (RNeasy MinElute columns, Qiagen) were then performed to remove the ASOs and their complements. Structure analysis by reverse transcription was performed as outlined above.

2.2.5 Electropherogram analysis and structure prediction

Electropherograms were analyzed with QuShape.[21] SHAPE data were analyzed as follows: nucleotides with no-reagent signals above the 99th percentile in any trial were excluded from analysis in all transcript data sets. SHAPE reactivity profiles were normalized as described,[22] except that the CAG-repeat region was excluded from the normalization calculation to maintain a consistent SHAPE reactivity distribution across all transcripts. RNase T1 data were analyzed as follows: nucleotides with background signals in the top 3% were excluded. Background and plus-RNase signals were normalized to the median of the plus-RNase signal. After background subtraction, guanosine residues showing normalized intensity values between 1 and 2 were designated low cleavage,

between 2 and 4, medium cleavage, and above 4, high cleavage.

Secondary structures were modeled using the Fold module of RNAstructure,[23] version 5.4, using the latest parameters for incorporating SHAPE data.[24,25] Because the huntingtin mRNA likely forms many noncanonical base pairs and contains multiple regions of repeated sequence, structure modeling was challenging. Without constraining secondary-structure models with SHAPE data, RNAstructure predicted a large number of alternative structures of similar energy. SHAPE constraints brought these predictions into agreement with experimental data and significantly reduced the number of plausible structures. Given the overall similarities in nucleotide reactivities across the five transcripts (Figures 2.1, 2.3, and 2.5), the lowest predicted free-energy structure for the shortest transcript was used as a template to select the most likely structure for each of the CAG-expanded transcripts. In addition, we selected those structural models that showed reactive nucleotides in the CAG-repeat region within two triplets of a CAG hairpin terminus.

2.3 Results

2.3.1 SHAPE and RNase probing of huntingtin exon 1

We used SHAPE[26,27] chemical probing to analyze the structure of five RNA transcripts containing shorter CAG-repeat lengths (17 and 23 repeats) typical of healthy alleles and longer, disease-associated, numbers of repeats (36, 41, and 70 repeats). Little degradation of RNAs was observed as judged by the low peak intensities in reverse-transcription products from the no-reagent controls, as analyzed by capillary electrophoresis. SHAPE reactivity profiles for each of the transcripts are shown split in the center of the CAG-repeat region and aligned at the $5'$ and $3'$ ends (Figure 2.1). Overall, SHAPE reactivity profiles for the five transcripts are highly similar, suggesting that the global secondary structure is not affected by expanded CAG repeats (Figure 2.1). Within the CAG-repeat region in each transcript, most nucleotides were unreactive, consistent with formation of stable base pairing.[26,27] In addition, within each CAG repeat region, there was a short region with

more reactive (conformationally flexible) nucleotides; this region was not centered in the CAG-repeat region but instead was offset in the 3′ direction (Figure 2.1, emphasized with solid arrows). This asymmetry in the CAG-repeat regions was also observed by RNase T1 enzyme probing (Figure 2.5). The group of SHAPE-reactive nucleotides was consistently located six triplets 3′ of the center of the poly-CAG repeat.

2.3.2 Structural models of huntingtin transcripts

We used the SHAPE data to develop experimentally supported[24,25] models for thermodynamically accessible states for each of our huntingtin RNA transcripts. The 5′ UTR and 3′ regions of the RNAs are predicted to form similar or identical structures, independent of CAG-repeat length (Figure 2.2). In general, these structural models are well-defined (Supporting Information Figure 2.2). These models, which are based on RNA transcripts with long flanking sequences, likely capture features relevant to huntingtin mRNA structure in vivo. The 5′ end corresponds to the transcription start site 145 nucleotides from the translation start, although transcripts starting at 135 may also be present in vivo.[28,29] Some end effects are possible because of truncation of the studied transcripts at the 3 exon boundary (155 nucleotides from the CAG-repeat region).

Strikingly, the CAG repeat region forms extensive base-pairing interactions with nucleotides outside the repeat region (Figures 2.2 and 2.6, CAG repeat sequences are highlighted in orange). The 5′ end of the UTR, the CCG-repeat region immediately downstream of the CAG-repeat region, and an 11-nucleotide region with the sequence GCCGCUGCUGC (perfectly complementary to CAG repeats apart from one A:C mismatch) are all predicted to base pair with CAG-repeat nucleotides. The remarkable result of this base pairing is that a hairpin formed only of CAG-repeat nucleotides is entirely absent from the model of the healthy huntingtin transcript that contains 17 CAG repeats (Figure 2.2, left). Moreover, the CAG-repeat hairpin and the three-helix junction from which it extends represent allele-specific structures that occur preferentially in the longer

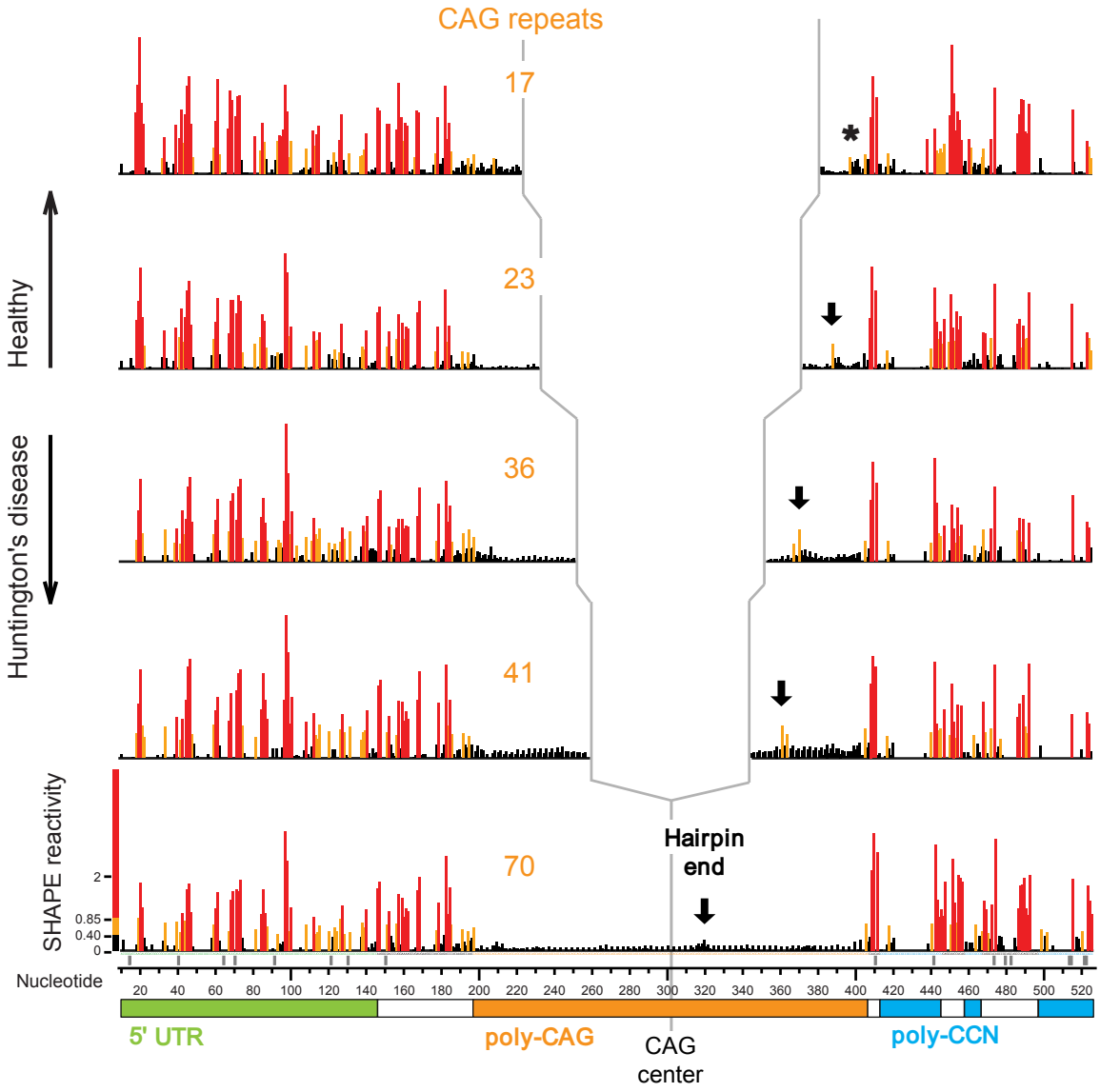

Figure 2.1: SHAPE profiles for huntingtin exon 1 transcripts as a function of CAG-repeat length. Reactivity profiles are shown split in the center of the CAG repeat region and aligned at the 5′ and 3′ ends. The black, yellow, and red scale indicates low, medium, and high SHAPE reactivities, respectively. The most SHAPE-reactive region within the CAG repeat consistently falls six CAG repeats 3′ of the CAG-repeat-region center, as emphasized with solid arrows. The region likely to form an internal loop in the 17-CAG repeat transcript is indicated with an asterisk (top panel). Data shown are the average of three independent experiments. The small number of nucleotides for which no data were obtained (because of strong electropherogram peaks in the no-reagent control, see Methods) are marked with gray boxes on the x axis.

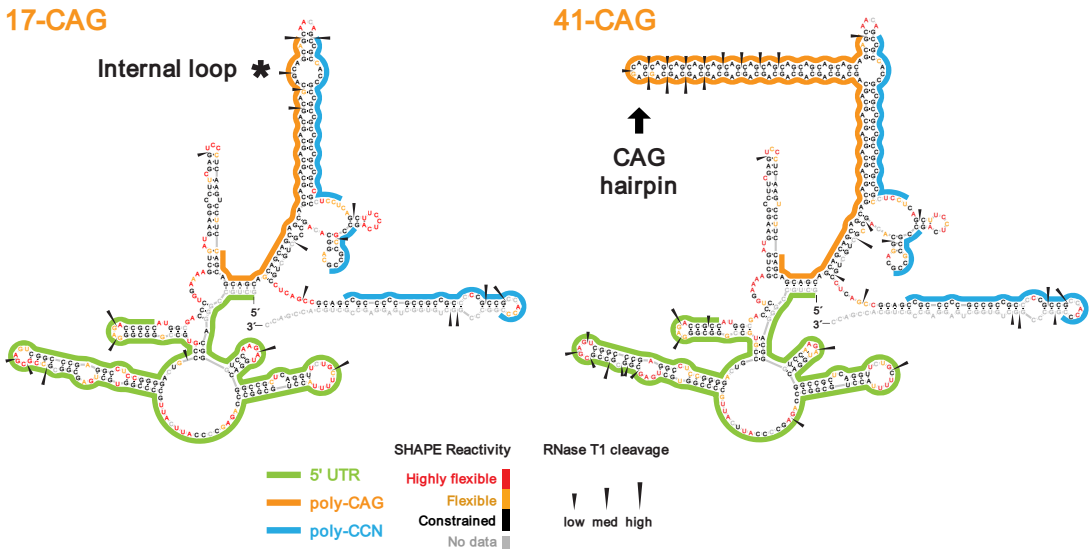

Figure 2.2: Structural models for representative normal and disease-associated huntingtin transcripts. Secondary-structure models for the most common healthy length transcript (17-CAG) and for a strongly disease-associated (41-CAG) RNA are shown. SHAPE and T1 RNase probing are shown with colored nucleotides and arrowheads, respectively. The absence of a CAG hairpin in the 17-CAG repeat RNA is emphasized with an asterisk.

disease-associated alleles.

2.3.3 CAG hairpin induction

If base pairing between CAG repeats and flanking sequences prevents CAG hairpin formation in healthy-length huntingtin transcripts, disrupting this base pairing should allow the RNA to refold and form extended hairpins (Figure 2.3, left). We folded all five huntingtin transcript RNAs in the presence of five antisense oligonucleotides designed to bind sequences flanking the CAG repeats and to compete for base pairing with these non-CAG sequences. Under these conditions, SHAPE-reactive nucleotides occurred at or near the center of the CAG-repeat element in all transcripts, both healthy length and disease expanded (Figure 2.3, right, site of hairpin loop is emphasized with open arrow). Thus, CAG-repeat elements can be forced to form a simple hairpin structure by inhibiting pairing to flanking sequences present in the native transcript.

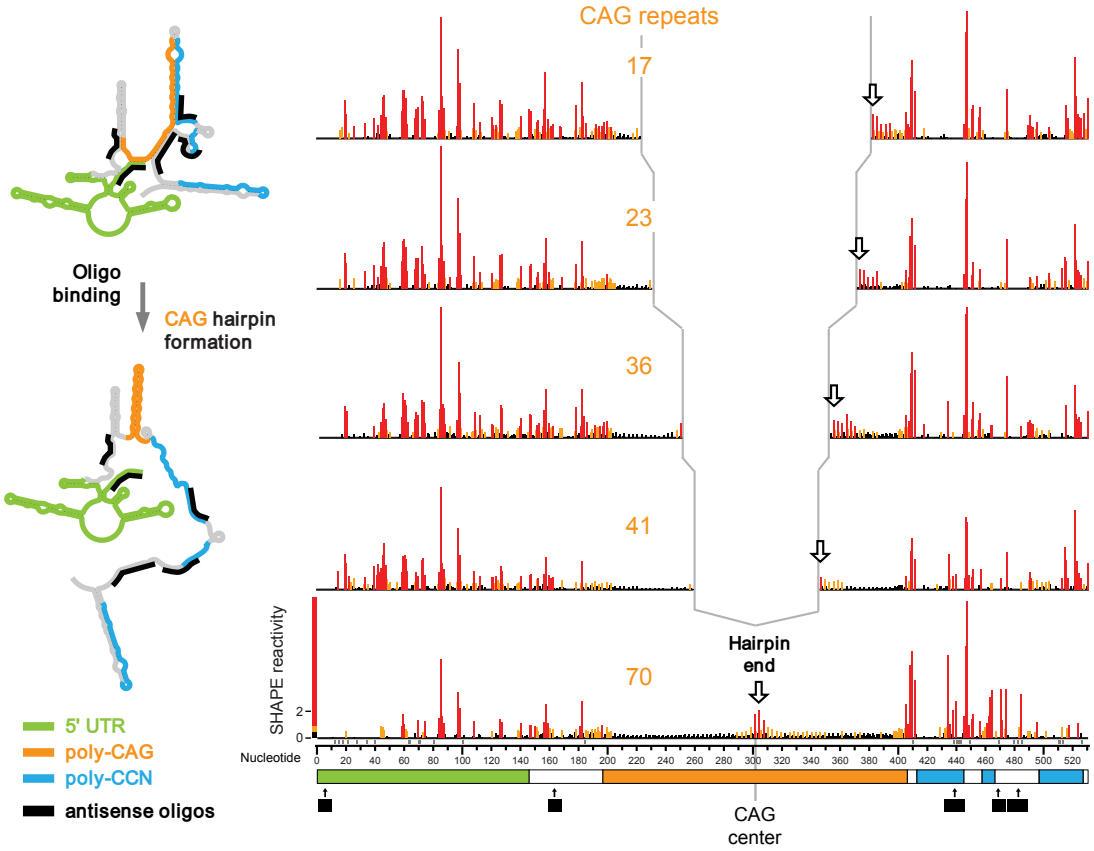

Figure 2.3: SHAPE analysis of huntingtin transcripts in the presence of antisense oligonucleotides designed to disrupt pairing between CAG sequences and flanking regions. Five antisense oligonucleotides were designed to bind specific, non-CAG sequences in the huntingtin mRNA to disrupt base pairing with the CAG-repeat region and to promote formation of a CAG hairpin. Oligonucleotide binding sites are shown with black bars. The center (reactive) region of each CAG-repeat element is emphasized with an open arrow and is consistent with simple hairpin formation by self-paired CAG sequences.

2.4 Discussion

Our work provides the first empirical examination of huntingtin mRNA structure in the context of extended, native flanking sequences (in this case, the entire first exon). Given the GC-rich nature of the huntingtin mRNA, it is not surprising that the transcripts are highly structured (Figure 2.2). The CAG-repeat regions adopt distinct structures that depended on repeat length and on the flanking sequence context (Figure 2.4). In the absence of

interacting flanking sequences, poly-CAG transcripts, which are found in several disease-related contexts,[30,31] fold back on themselves to base pair into simple hairpins.[12,13] In huntingtin exon 1 mRNA sequences, CAG repeats are followed by poly-CCN sequences and a complementary GCCGCUGCUGC sequence, and our analysis indicates that these flanking sequences pair with the poly-CAG element. Because CAG repeats base pair with flanking sequences, a CAG hairpin was not observed in the transcript containing the 17 CAG repeats typical of a healthy individual.

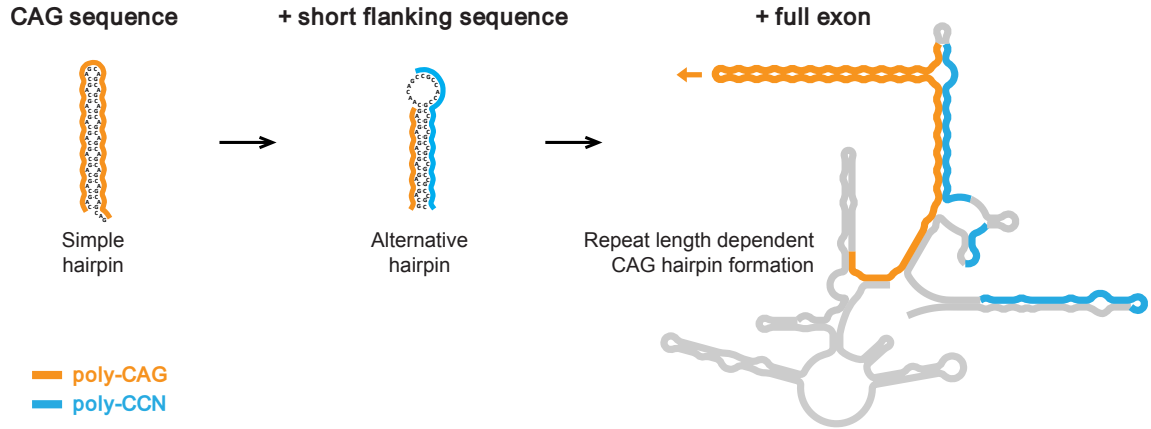

Figure 2.4: Role of flanking sequence in defining CAG-repeat RNA structures. Shown are secondary-structure models for the CAG-repeat sequence,[12] for a CAG repeat with short flanking sequences,[17] and for the full-length huntingtin exon 1 sequence studied in this work. A CAG hairpin begins to form with intermediate-length repeat expansion and preferentially forms a long classical hairpin (shown) with disease-associated CAG expansions.

Cellular and animal models of HD indicate that disease symptoms correlate with several factors including repeat lengths, expression levels, localization of huntingtin transcripts, truncation of the huntingtin sequence, and stoichiometry of native and mutant sequences.[32-34] This work supports the hypothesis that the CAG-repeat-containing transcript itself, and not just its ability to encode polyglutamine, might be important for disease etiology. The two widely used mouse models of HD employ a yeast artificial chromosome (YAC128)[35] or a bacterial artificial chromosome (BACHD).[36]

Despite expressing similar mutant huntingtin mRNAs, BACHD mice do not show aggregate formation or display the transcriptional dysregulation present in YAC128 mice and HD patients.[37] An important distinction between these models is the use of nearly pure CAG repeats in YAC128 versus unnatural, mixed CAA/CAG repeats in BACHD. The presence of CAA triplets disrupts extended hairpin formation and favors branched secondary structures.[38] In addition, CAA sequences will not base pair strongly with CCG sequences and other flanking regions present in the authentic huntingtin transcript sequence. The allele used in the BACHD model will almost certainly lack the striking CAG-repeat-length-dependent hairpin formation found in this study; therefore, some of the phenotypic differences that distinguish pure CAG from mixed-codon HD models may reflect differences in RNA structure.

The secondary-structure models developed in this work also suggest specific roles for huntingtin mRNA structure in splicing and translation. First, expanded CAG repeats within huntingtin transcripts contribute to misregulation of splicing. These defects include sequestration of the splicing factor muscleblind-like protein 1[17,39] and mis-splicing of the huntingtin transcript, possibly because of recruitment of the splicing factor SRSF6.[40] We hypothesize that base pairing by healthy-length CAG repeats to flanking sequences reduces deleterious recognition by splicing factors. Second, the huntingtin 5′ UTR and the region surrounding the primary translation start site form stable RNA structures (Figure 2.2); in general, structured UTRs reduce translation initiation.[41] Taken together with a putative active upstream open reading frame in huntingtin,[29] this work suggests that regulation of huntingtin translation may be complex and involve the interplay of the general translation-initiation machinery, contributions of strong local structure at the translation-initiation site, and the possible presence of multiple initiation sites.

The absence of a CAG hairpin in short, healthy-length huntingtin transcripts and its presence in transcripts with increased numbers of repeats suggests that allele-specific

targeting of huntingtin mRNA structures will be possible. SHAPE-directed structure models suggest that CAG hairpins occur in disease-associated alleles but not in alleles with fewer repeats characteristic of healthy individuals. Molecules that bind specifically to CAG hairpins, especially if they discriminate against duplexes in which CAG sequences pair with CCG repeat sequences (Figure 2.2), are likely to be very selective for disease-causing alleles. The three-helix junction from which the CAG hairpin extends represents another novel RNA target with the potential for both gene and allele selectivity. Broadly, our findings highlight the importance of flanking sequence in RNA folding and hint at the insights to be gained by conducting quantitative, large-scale RNA-structure analyses. Examinations of the effects of context on RNA structure are likely to identify new therapeutic targets in repeat-expansion diseases.

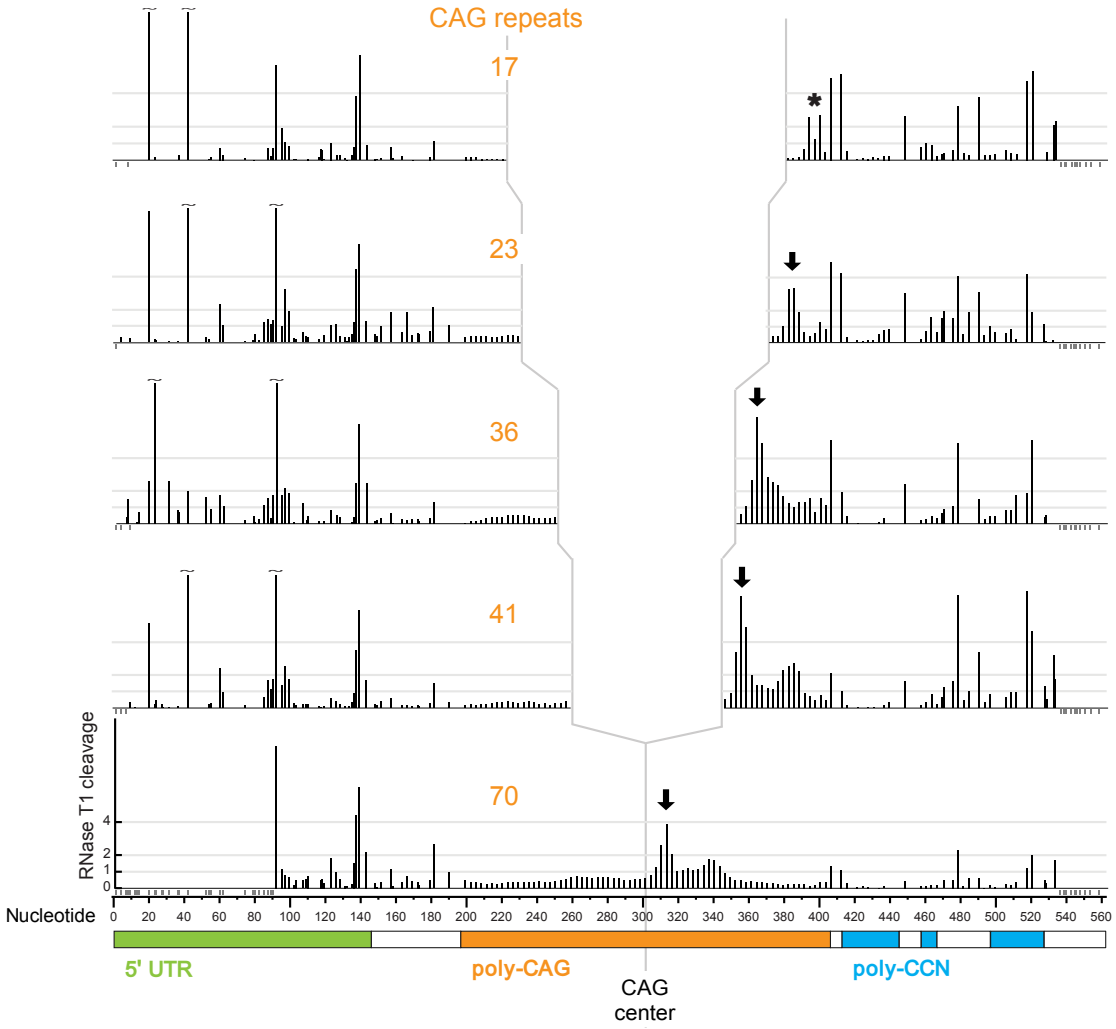

Figure 2.5: RNA structure probing profiles using RNase T1. In the context of complete huntingtin exon 1 sequences, RNase T1 cleavage supports a model in which the CAG hairpin is positioned asymmetrically relative to the center of the sequence (emphasized with solid arrows), consistent with base pairing between CAG sequences and 3′ flanking sequences. Nucleotides for which no data is available are marked with gray boxes at the x-axis.

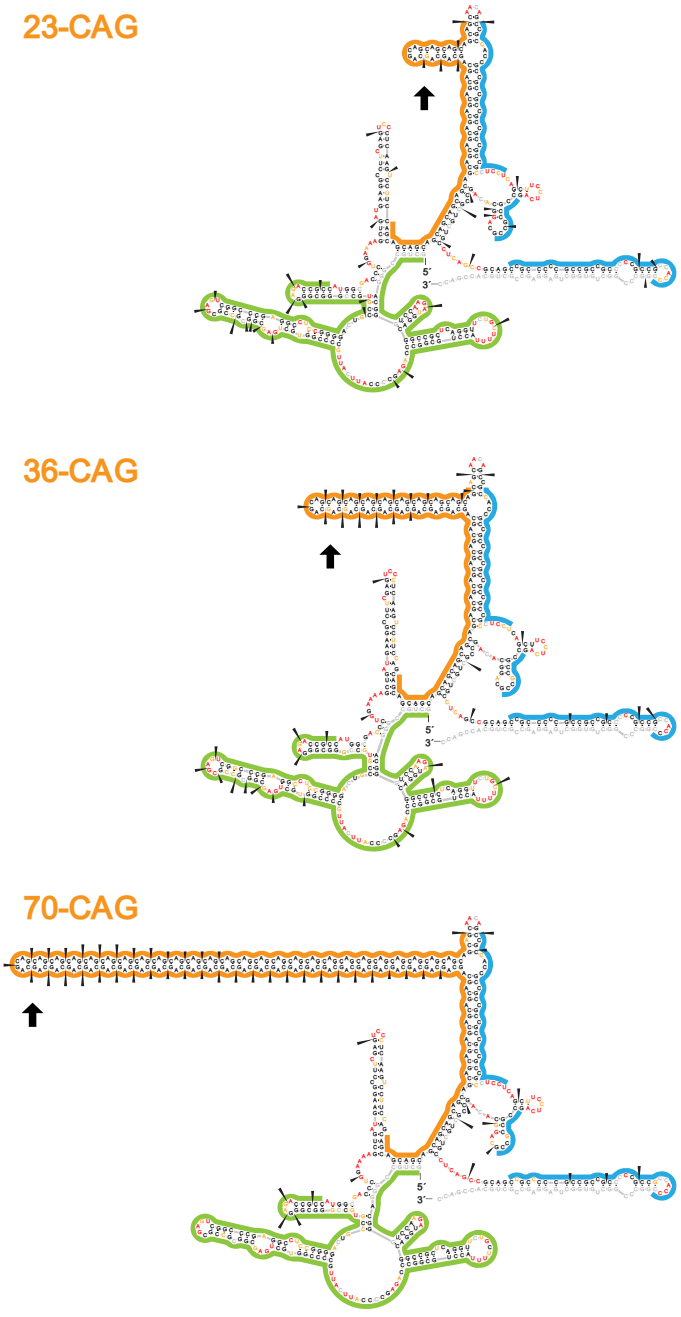

Figure 2.6: Secondary structure models for 23, 36, and 70-CAG repeat length huntingtin exon 1 transcripts. Structure and reactivity annotation scheme is the same as shown in Fig. 2.2.

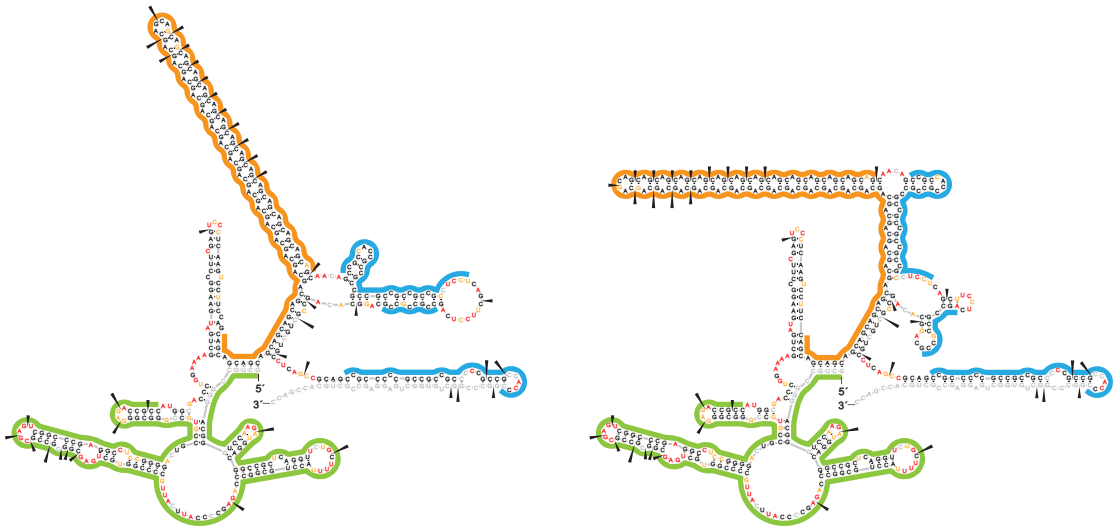

Figure 2.7: Plausible competing structures for long CAG repeat sequences. Representative structures are shown in the context of the 41-CAG transcript.

REFERENCES

1. Kremer B., Goldberg P., Andrew S.E., Theilmann J., Telenius H., Zeisler J., Squitieri F., Lin B., Bassett A., Almqvist E. et al. (1994) A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. *N. Engl. J. Med. 330,* 1401-1406.

2. Lee J.M., Ramos E.M., Lee J.H., Gillis T., Mysore J.S., Hayden M.R., Warby S.C., Morrison P., Nance M., Ross C.A. (2012) CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology 78,* 690-695.

3. Hsu R.J., Hsiao K.-M., Lin M.-J., Li C.-Y., Wang L.-C., Chen L.-K., Pan H. (2011) Long tract of untranslated CAG repeats is deleterious in transgenic mice. *PLoS ONE 6,* e16417.

4. Shieh S.Y., Bonini N.M. (2011) Genes and pathways affected by CAG-repeat RNA-based toxicity in Drosophila. *Hum. Mol. Genet. 20,* 4810-4821.

5. Velier J., Kim M., Schwarz C., Kim T.W., Sapp E., Chase K., Aronin N., DiFiglia M. (1998) Wild-type and mutant huntingtin's function in vesicle trafficking in the secretory and endocytotic pathways. *Exp. Neurol. 152,* 34-40.

6. Gauthier L.R., Charrin B.C., Borrell-Pags M., Dompierre J.P., Rangone H., Cordelires F.P., De Mey J., MacDonald M.E., Lessmann V., Humbert S. et al. (2004) Huntingtin controls neurotrophic support and survival of neurons by enhancing BDNF vesicular transport along microtubules. *Cell 118,* 127-138.

7. Caviston J.P., Holzbaur E.L.F. (2009) Huntingtin as an essential integrator of intracellular vesicular trafficking. *Trends Cell Biol. 19,* 147-155.

8. Zala D., Hinckelmann M.V., Saudou F. (2013) Huntingtin's function in axonal transport is conserved in drosophila melanogaster. *PLoS ONE 8,* e60162.

9. Pfister E.L., Kennington L., Straubhaar J., Wagh S., Liu W., DiFiglia M., Landwehrmeyer B., Vonsattel J.-P., Zamore P.D., Aronin N. (2009) Five siRNAs targeting three SNPs may provide therapy for three-quarters of huntington's disease patients. *Curr. Biol. 19,* 774-778.

10. Carroll J.B., Warby S.C., Southwell A.L., Doty C.N., Greenlee S., Skotte N., Hung G., Bennett C.F., Freier S.M., Hayden M.R. (2011) Potent and selective antisense oligonucleotides targeting single-nucleotide polymorphisms in the Huntington disease gene / Allele-specific silencing of mutant huntingtin. *Mol. Ther. 19,* 2178-2185.

11. Gagnon K.T., Pendergraff H.M., Deleavey G.F., Swayze E.E., Potier P., Randolph J., Roesch E.B., Chattopadhyaya J., Damha M.J., Bennett C.F. et al. (2010) Allele-selective inhibition of mutant huntingtin expression with antisense oligonucleotides targeting the expanded CAG repeat. *Biochemistry 49,* 10166-10178.

12. Sobczak K., de Mezer M., Michiewski G., Kroi J., Krzyzosiak W.J. (2003) RNA structure of trinucleotide repeats associated with human neurological diseases. Nucleic Acids Res. 31, 5469-5482.

13. Michlewski G., Krzyzosiak W.J. (2004) Molecular architecture of CAG repeats in human disease related transcripts. *J. Mol. Biol. 340,* 665-679.

14. Broda M., Kierzek E., Gdaniec Z., Kulinski T., Kierzek R. (2005) Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry 44,* 10873-10882.

15. Yuan Y., Compton S.A., Sobczak K., Stenberg M.G., Thornton C.A., Griffith J.D., Swanson M.S. (2007) Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res. 35,* 5474-5486.

16. Kiliszek A., Kierzek R., Krzyzosiak W.J., Rypniewski W. (2010) Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res. 38,* 8370-8376.

17. de Mezer M., Wojciechowska M., Napierala M., Sobczak K., Krzyzosiak W.J. (2011) Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. *Nucleic Acids Res. 39,* 3852-3863.

18. Vester, B., and Wengel, J. (2004) LNA (locked nucleic acid): High-affinity targeting of complementary RNA and DNA. *Biochemistry 43,* 13233-13241.

19. Wyatt J.R., Chastain M., Puglisi J.D. (1991) Synthesis and purification of large amounts of RNA oligonucleotides. *Biotechniques 11,* 764-769.

20. Mortimer S.A., Weeks K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc. 129,* 4144-4145.

21. Karabiber F., McGinnis J.L., Favorov O.V., Weeks K.M. (2012) QuShape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA 19,* 63-73.

22. Wilkinson K.A., Merino E.J., Weeks K.M. (2006) Selective 2-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc. 1,* 1610-1616.

23. Reuter J.S., Mathews D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics 11,* 116.

24. Deigan K.E., Li T.W., Mathews D.H., Weeks K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A. 106,* 97-102.

25. Hajdin C.E., Bellaousov S., Huggins W., Leonard C.W., Mathews D.H., Weeks K.M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A. 110,* 5498-5503.

26. Merino E.J., Wilkinson K.A., Coughlan J.L., Weeks K.M. (2005) RNA structure analysis at single nucleotide resolution by Selective 2-Hydroxyl Acylation and Primer Extension (SHAPE). *J. Am. Chem. Soc. 127,* 4223-4231.

27. Weeks K.M., Mauger D.M. (2011) Exploring RNA structural codes with SHAPE chemistry. *Acc. Chem. Res. 44,* 1280-1291.

28. Lin B., Nasir J., Kalchman M.A., McDonald H., Zeisler J., Goldberg Y.P., Hayden M.R. (1995) Structural Analysis of the 5 Region of Mouse and Human Huntington Disease Genes Reveals Conservation of Putative Promoter Region and Di- and Trinucleotide Polymorphisms. *Genomics 25,* 707-715.

29. Lee J., Park E.H., Couture G., Harvey I., Garneau P., Pelletier J. (2002) An upstream open reading frame impedes translation of the huntingtin gene. *Nucleic Acids Res. 30,* 5110-5119.

30. Nalavade R., Griesche N., Ryan D.P., Hildebrand S., Krau S. (2013) Mechanisms of RNA-induced toxicity in CAG repeat disorders. *Cell Death Dis. 4,* e752.

31. Galka-Marciniak P., Urbanek M.O., Krzyzosiak W.J. (2012) Triplet repeats in transcripts: structural insights into RNA toxicity. *Biol. Chem. 393,* 1299-1315.

32. Hodgson J.G., Smith D.J., McCutcheon K., Koide H.B., Nishiyama K., Dinulos M.B., Stevens M.E., Bissada N., Nasir J., Kanazawa I. et al. (1996) Human huntingtin derived from YAC transgenes compensates for loss of murine huntingtin by rescue of the embryonic lethal phenotype. *Hum. Mol. Genet. 5,* 1875-1885.

33. Ehrnhoefer D.E., Butland S.L., Pouladi M.A., Hayden M.R. (2009) Mouse models of Huntington disease: variations on a theme. *Dis. Models Mech. 2,* 123-129.

34. Southwell A.L., Warby S.C., Carroll J.B., Doty C.N., Skotte N.H., Zhang W., Villanueva E.B., Kovalik V., Xie Y., Pouladi M.A. et al. (2013) A fully humanized transgenic mouse model of Huntington disease. *Hum. Mol. Genet. 22,* 18-34.

35. Slow E.J., van Raamsdonk J., Rogers D., Coleman S.H., Graham R.K., Deng Y., Oh R., Bissada N., Hossain S.M., Yang Y.Z. et al. (2003) Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease. *Hum. Mol. Genet. 12,* 1555-1567.

36. Gray M., Shirasaki D.I., Cepeda C., Andre V.M., Wilburn B., Lu X.H., Tao J., Yamazaki I., Li S.H., Sun Y.E. et al. (2008) Full-length human mutant huntingtin with a stable polyglutamine repeat can elicit progressive and selective neuropathogenesis in BACHD mice. *J. Neurosci. 28*, 6182-6195.

37. Pouladi M.A., Stanek L.M., Xie Y., Franciosi S., Southwell A.L., Deng Y., Butland S., Zhang W., Cheng S.H., Shihabuddin L.S. et al. (2012) Marked differences in neurochemistry and aggregates despite similar behavioral and neuropathological features of Huntington disease in full-length BACHD and YAC128 mice. *Hum. Mol. Genet. 21,* 2219-2232.

38. Sobczak K., Krzyzosiak W.J. (2005) CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J. Biol. Chem. 280,* 3898-3910.

39. Mykowska A., Sobczak K., Wojciechowska M., Kozlowski P., Krzyzosiak W.J. (2011) CAG repeats mimic CUG repeats in the misregulation of alternative splicing. *Nuc. Acids Res. 39,* 8938-8951.

40. Sathasivam et al. (2013) Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. U.S.A. 110,* 2366-2370.

41. Pickering B.M., Willis A.E. (2005) The implications of structured 5′ untranslated regions on translation and disease. *Semin. Cell. Dev. Biol. 16,* 39-47.

3 SOFTWARE FOR THE AUTOMATED ANALYSIS OF SHAPE AND MUTATIONAL PROFILING (SHAPE-MAP) DATA[1]

3.1 Introduction

SHAPE is unique among RNA structure probing strategies because it reports on the flexibility of all four ribonucleotides and enables highly accurate secondary structure modeling[1–3]. However, in its original capillary electrophoresis version it is limited by signal fall-off to about 500 nucleotides in a single experiment[4], and often requires skilled users to process the data and create reactivity profiles[5,6]. The mutational profiling (MaP) strategy was developed to allow the streamlined application of SHAPE to large RNAs and multiple RNAs in single experiments and the full automation of data analysis.

3.2 SHAPE-MaP strategy

In the SHAPE-MaP strategy, reverse transcription is performed in the presence of a high (6 mM) concentration of Mn^{2+}. This causes a slight decrease in reverse transcription fidelity overall, but a highly useful reduction in fidelity specifically at SHAPE adduct sites. As a result, SHAPE adduct locations are encoded as sequence mutations in the cDNA library (Figure 3.1). In a SHAPE-MaP experiment, two control libraries are also prepared, one from RNA exposed to solvent but no SHAPE reagent (a background control), and one from RNA exposed to SHAPE reagent under highly denaturing conditions (an adduct detection rate control). Mutation rates from the three total samples are compared to produce final reactivity profiles (Figure 3.2, panels A and B), which agree closely with known secondary structures (Figure 3.2, panel C). Importantly, directed primers or random primers[7] can be used in reverse transcription, allowing the probing of large RNAs in single

[1]This chapter previously appeared in extended form as an article in Nature Methods. The original citation is as follows: Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E. & Weeks, K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11,** in press (2014).

experiments.

SHAPE-MaP's ability to accurately report nucleotide-resolution RNA flexibility and to enable accurate secondary structure prediction was extensively validated. SHAPE-MaP was also applied to the discovery of new structured motifs in an HIV-1 genome. This work was primarily performed by Nate Siegfried and Greggory Rice, and is described in detail in the first SHAPE-MaP publication[8].

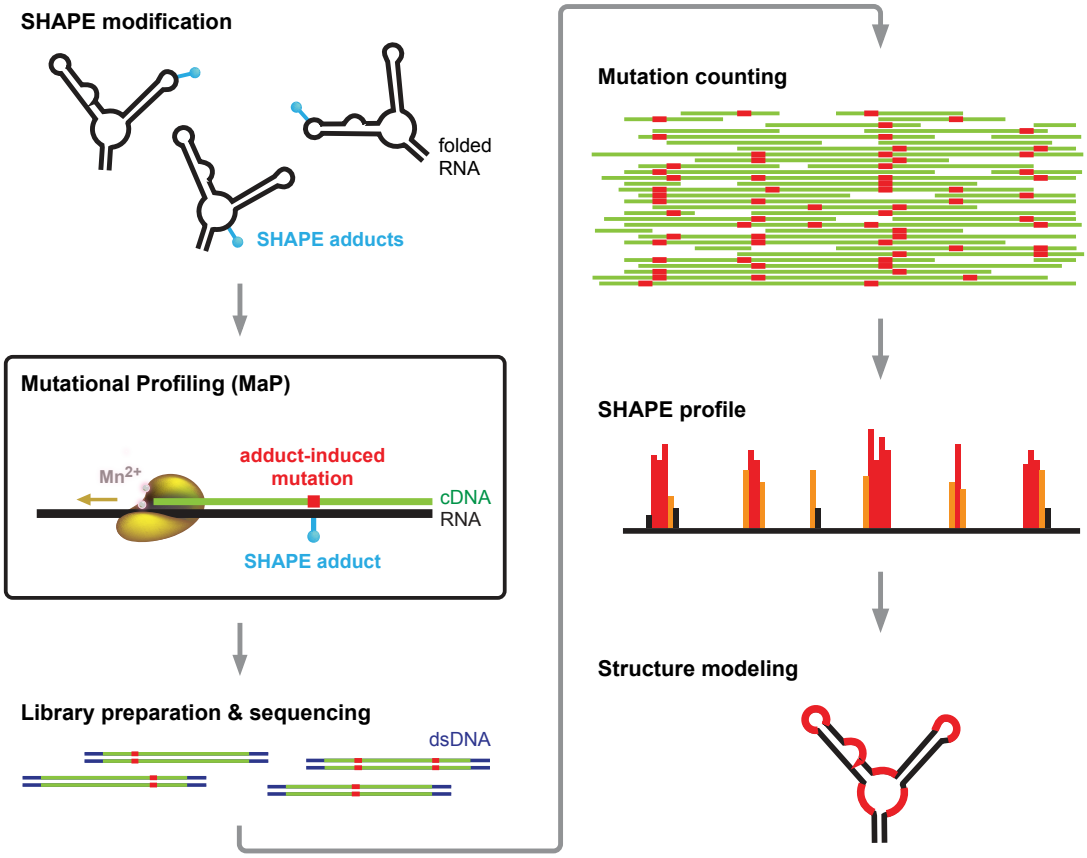

Figure 3.1: SHAPE-MaP overview. RNA is treated with a SHAPE reagent that reacts at conformationally dynamic nucleotides[9]. Reverse transcription is carried out under conditions such that the polymerase reads through chemical adducts in the RNA and incorporates a nucleotide non-complementary to the original sequence (in red) into the cDNA. The resulting cDNA is sequenced using any massively parallel approach to create mutational profiles (MaP). Sequencing reads are aligned to a reference sequence, and nucleotide-resolution mutation rates are calculated, corrected for background and normalized, producing a standard SHAPE reactivity profile. SHAPE reactivities can then be used to model secondary structures, visualize competing and alternative structures, or quantify any process or function that modulates local ribonucleotide dynamics.

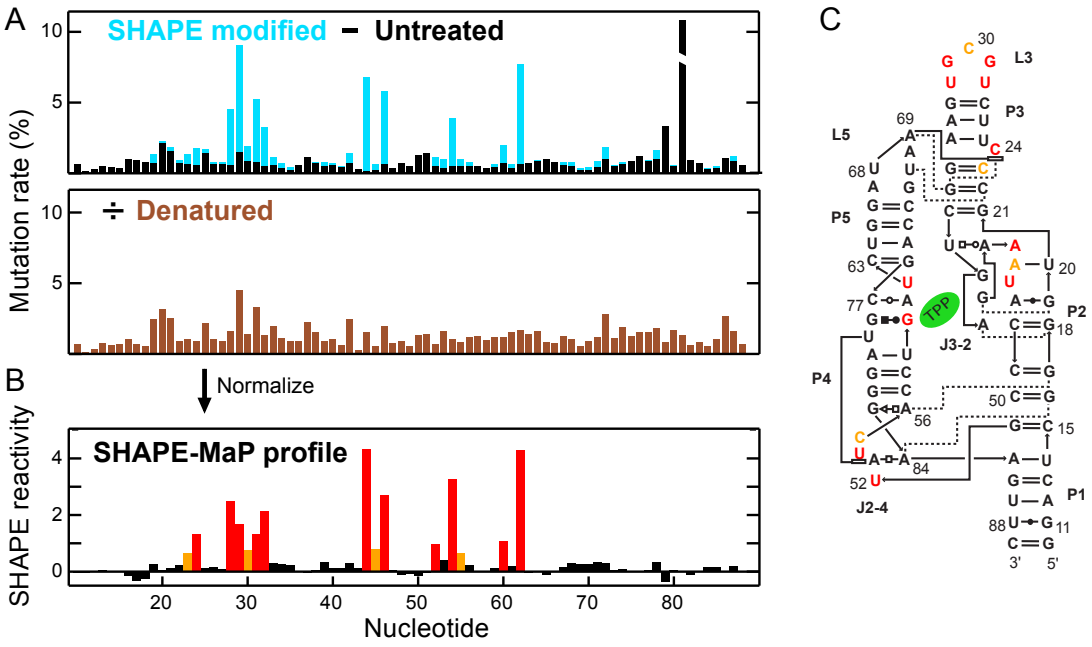

Figure 3.2: Nucleotide-resolution interrogation of RNA structure. (**A**) Mutation rate profiles for the SHAPE modified and untreated thiamine pyrophosphate (TPP) riboswitch RNA in the presence of ligand (top) and for SHAPE modification performed under denaturing conditions (bottom). (**B**) Quantitative SHAPE profile obtained after subtracting the data from the untreated sample from data for the treated sample and normalizing by the denatured control. (**C**) SHAPE reactivities plotted on the accepted secondary structure of the ligand-bound TPP riboswitch[10]. Red, orange, and black correspond to high, moderate, and low reactivities, respectively.

3.3 SHAPE-MaP data analysis pipeline (ShapeMapper)

I created a data analysis pipeline, called ShapeMapper, that can be executed on most unix-based platforms and accepts as input sequencing read files in FASTQ format, reference sequences in FASTA format, and a user-edited configuration file. Without further user intervention, the software creates a SHAPE reactivity profile and standard error estimates for each reference sequence (Figure 3.3). Other useful outputs are provided including mutation counts, sequencing depths, and predicted secondary structures. The analysis software incorporates several third-party programs. Python 2.7 is required[11]; Bowtie 2 is used for read alignment[12]; reactivity profiles are generated using the python library matplotlib[13]; secondary structure prediction uses RNAstructure[14]; and secondary structure drawing uses the Pseudoviewer web service[15].

3.3.1 Configuration

A configuration file is used to specify the reference sequences present in each sample and which samples should be combined to create reactivity profiles (Figure 3.3, panel A). The format is flexible, allowing the alignment of each sample to multiple sequence targets as well as the treatment of multiple samples in unified analyses. Important parameters for each stage of analysis may also be customized.

3.3.2 Quality trimming

Input reads are separated into files by sequencing barcode (this step is integrated into most sequencing platforms)[16]. The first analysis stage trims reads by base-call quality. Each read is trimmed downstream of the first base-call with a phred quality score below 10, corresponding to 90% expected accuracy[17]. Reads with 25 or more remaining nucleotides are copied to new FASTQ files for alignment.

3.3.3 Read alignment

Reads are locally aligned to reference sequences using Bowtie 2[12] (Figure 3.3, panel B). Parameters were chosen to provide high sensitivity, to detect single nucleotide

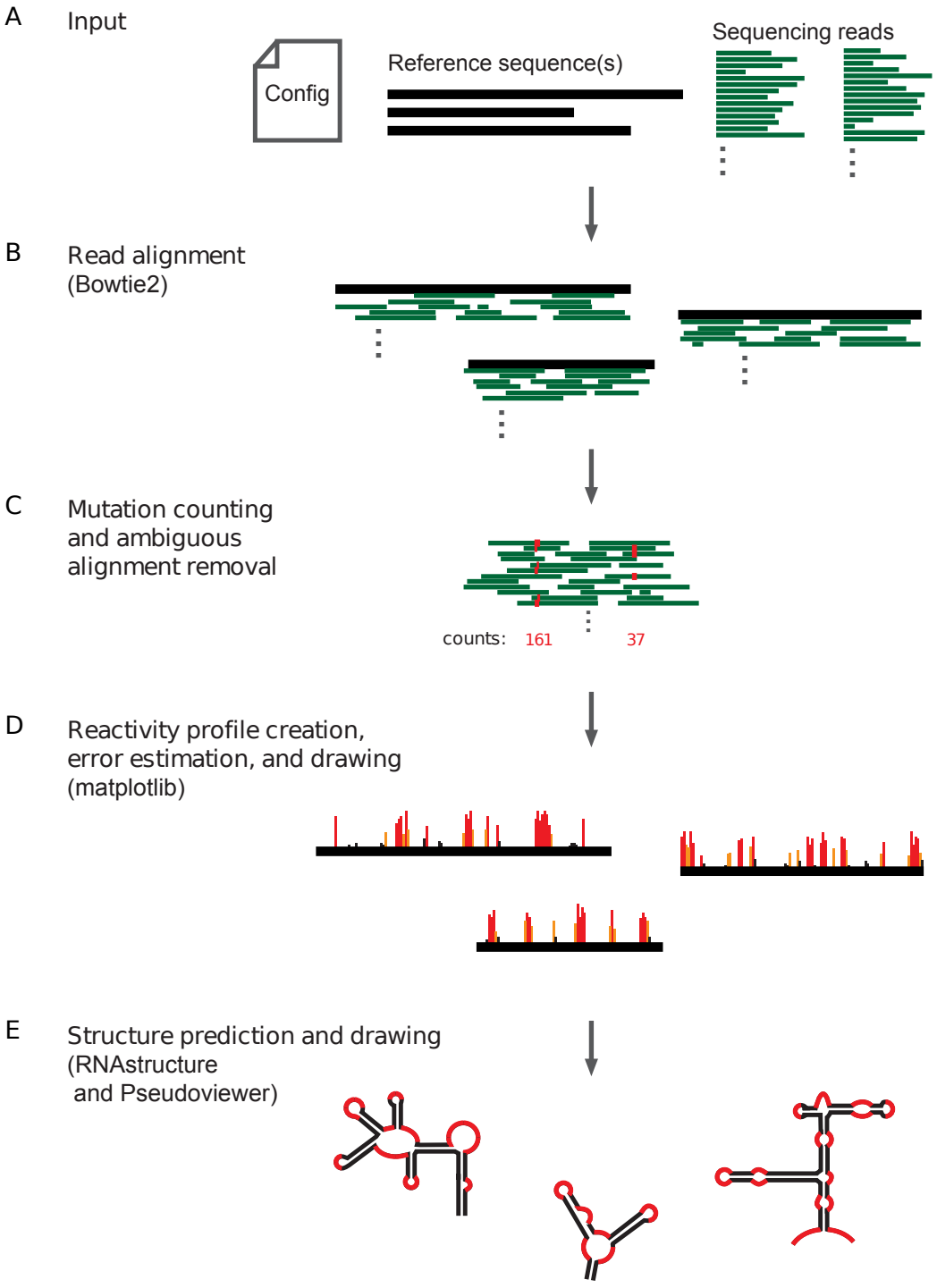

Figure 3.3: ShapeMapper software overview. Outline of software pipeline that fully automates calculations of per-nucleotide mutation rates, SHAPE reactivities, and standard error estimates given massive parallel sequencing data and at least one reference sequence. The software is executable on most unix-based platforms.

mismatches, and to allow deletions of up to about 200 nucleotides. Seed length (-L) is 15 nucleotides. One mismatch is allowed per seed (-N). Maximum seed attempts (-D) is set at 20. Maximum re-seed attempts (-R) is set at 3. Dynamic programming padding (--dpad) is set at 100 nucleotides. The match bonus (--ma) is 2. The maximum and minimum mismatch penalties (--mp) are 6 and 2, respectively. Gap open and extend parameters (--rdg, --rfg) are 5 and 1, respectively. The default minimum alignment score function is used. Soft-clipping is turned on. Paired-end alignment is used by default. Bowtie 2 outputs aligned reads as SAM files.

3.3.4 Alignment parsing, ambiguous alignment removal, and mutation counting

In this stage, aligned reads are ultimately processed into mutation counts (Figure 3.3, panel C). Paired-end reads in SAM files are combined, and higher-quality base-calls are selected where read pairs disagree. Mismatches and deletions contribute to mutation counts; insertions are ignored. Since error-prone reverse transcription generates most of the mutations in each read[18], I treat a sequence change covering multiple adjacent nucleotides as a single mutation event located at the 3′-most nucleotide. If random primers are used, a region one nucleotide longer than the length of the primer is excluded from the 3′end of each read. Reads with reported mapping qualities less than 30 are excluded, corresponding to an estimated probability of greater than 0.1% that a given read originated from a different location[19].

Deletions are an important part of the mutation signal, but deletions that are ambiguously aligned can blur this signal, preventing single-nucleotide resolution. To resolve this problem, a simple local realignment is performed to identify and remove ambiguously aligned deletions. The reference sequence surrounding a deletion is stored (Figure 3.4, panel A). The deletion is then slid upstream or downstream one nucleotide at a time to a maximum offset equal to the deletion length (Figure 3.4, panel B). At each offset, the surrounding reference sequence is compared to the stored sequence. If any

offset sequence matched, this indicates a possible alternate alignment, and the deletion is excluded. This algorithm correctly identifies ambiguous deletions in homopolymeric regions as well as repeated sequences (Figure 3.5).

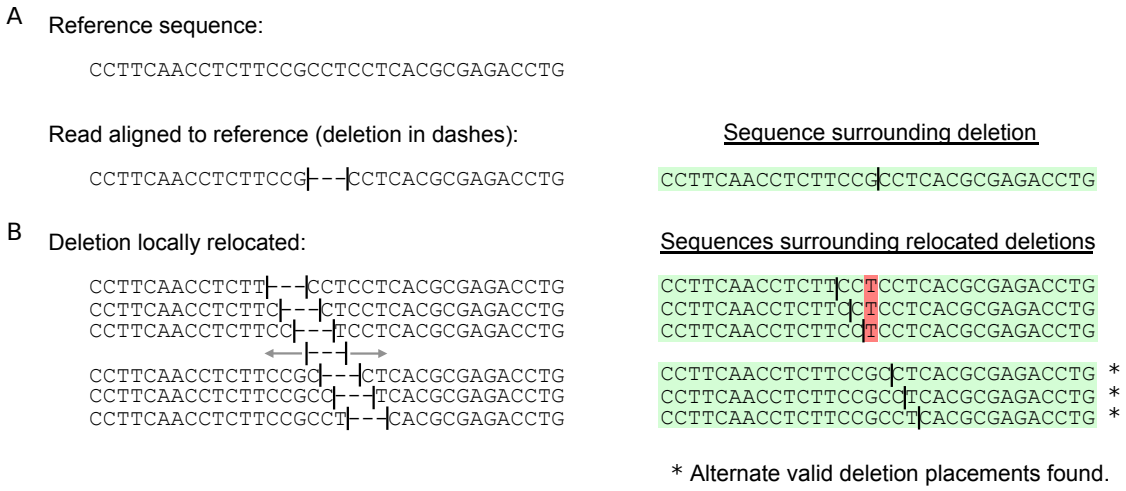

Figure 3.4: Ambiguously aligned deletion identification. Demonstration of the ambiguously aligned deletion detection algorithm applied to a single read. Nucleotides in red show sequence differences compared to the initial sequence surrounding the deletion. All-green sequences are identical to the initial sequence surrounding the deletion, indicating alternate valid deletion placements. In this case, the existence of a duplicated CCT sequence means four alignment locations are possible for a triplet deletion, making unambiguous placement impossible. This deletion would be excluded from mutation counting.

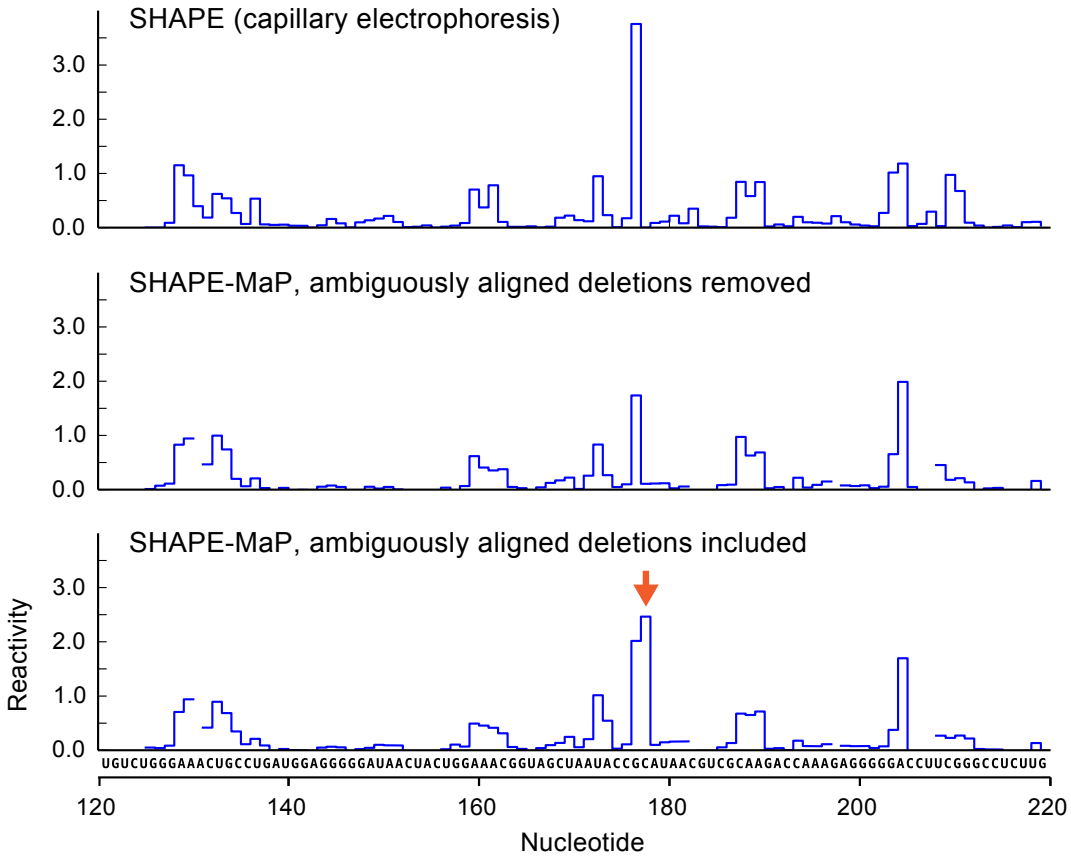

Figure 3.5: Removal of ambiguously aligned deletions. Representative reactivity profiles showing the effects of including or removing ambiguously aligned deletions. The nucleotide highlighted with an orange arrow shows an increased reactivity when ambiguous deletions are included. Data shown is from the 16S rRNA[3] (MaP experiments performed by Greggory Rice.)

3.3.5 Reactivity profile creation

The mutation rate (*mutr*) at a given nucleotide is simply the mutation count (mismatches and unambiguously aligned deletions) divided by the read count at that location. Raw reactivities are generated for each nucleotide using the following expression, where *S* corresponds to a SHAPE modified sample, *U* to untreated, and *D* to reaction under denaturing conditions:

$$R = \frac{mutr_S - mutr_U}{mutr_D} \tag{1}$$

The standard error (stderr) associated with the mutation rate at a given nucleotide in the *S*, *U*, or *D* samples is calculated as:

$$stderr = \frac{\sqrt{mutr}}{\sqrt{reads}} \tag{2}$$

The final standard error of the reactivity at a given nucleotide is:

$$SE = \sqrt{(\frac{stderr_S}{mutr_D})^2 + (\frac{stderr_U}{mutr_D})^2 + (stderr_D \times \frac{(mutr_S - mutr_U)}{mutr_D^2})^2} \tag{3}$$

Reactivities are normalized to a standard scale that spans zero (no reactivity) to 2 (high SHAPE reactivity) as described[20]. Nucleotides with mutation rates greater than 5% in untreated control samples are excluded from analysis, as are nucleotides with sequencing depths less than 10 in any sample.

3.3.6 Final data output

ShapeMapper automatically produces figures showing SHAPE reactivity profiles and standard errors (Figure 3.3, panel D, and Figure 3.6). SHAPE reactivity profiles are also output as tab-delimited text files (.shape) with the first column indicating nucleotide number and the second reactivity. A SHAPE-MaP reactivity file is also output (.map). This file is in the SHAPE file format with the addition of two columns: standard error and nucleotide sequence. Another file (.tab) containing mutation counts, read depths, mutation rates, raw reactivities, normalized reactivities, and standard errors for SHAPE modified, untreated,

and denatured samples is also created. Files containing figures showing mutation rate histograms, sequencing depths, and reactivity profiles are generated (.pdf). These are useful in diagnosing potential experimental problems, including insufficient sequencing depth or low mutagenesis efficiency (Figure 3.7).

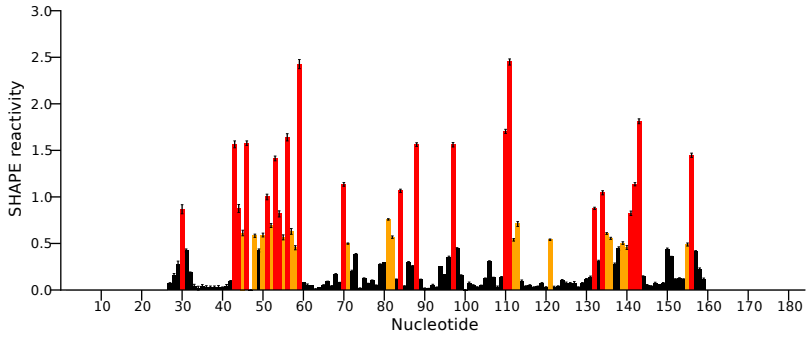

Figure 3.6: Example reactivity profile. Shown are the SHAPE reactivities obtained for the 6S RNA from *E. coli* – see chapter 4 for more details on this experiment. Error bars show standard error as calculated from equation (3).

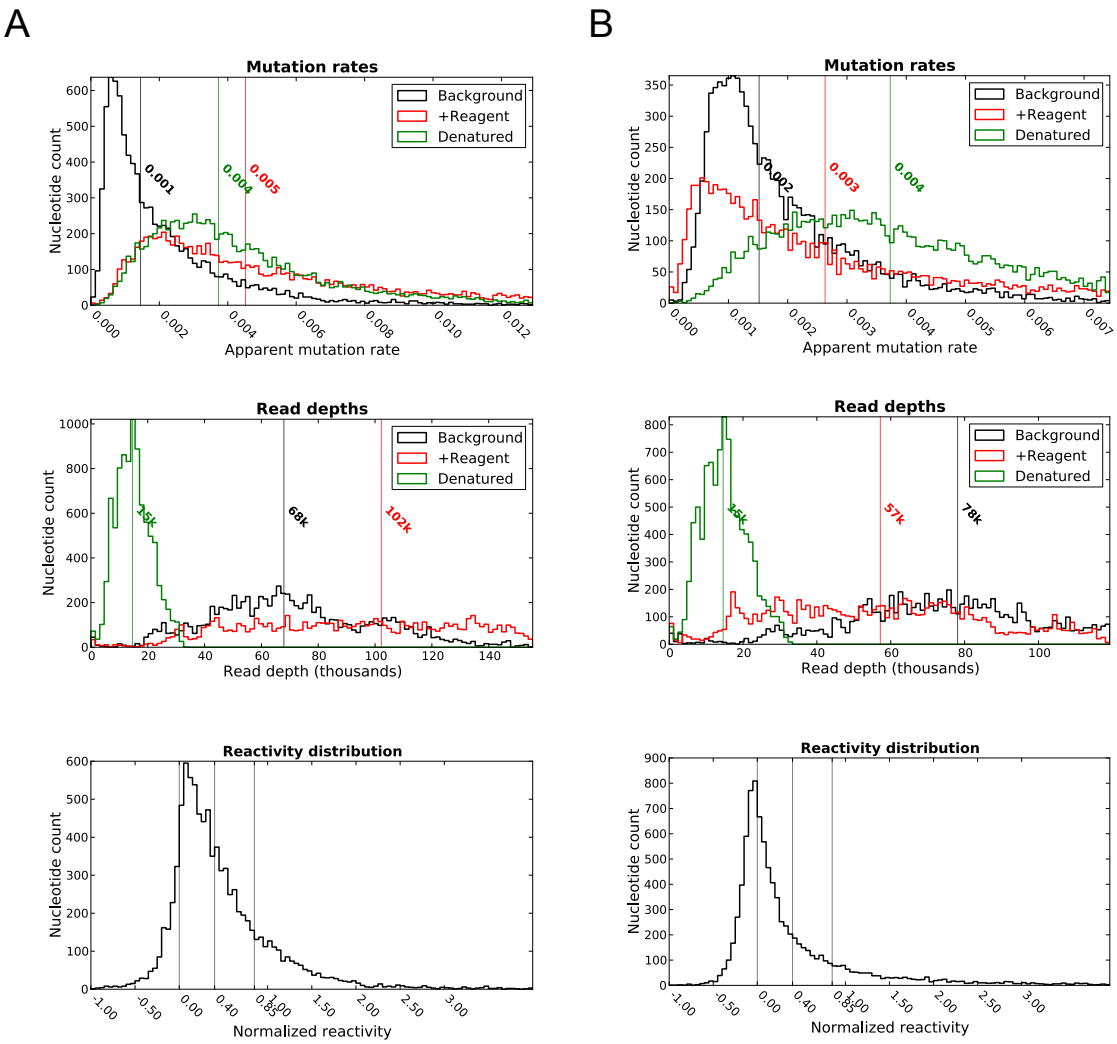

Figure 3.7: Example histograms and troubleshooting. (**B**) Successful MaP experiment. Read depths for all samples are largely above the nominal recommended level of 5000. Mutation rates in the +reagent condition are above background. The majority of the reactivities are positive. (**B**) Failed MaP experiment. Background mutation rates are unusually high, and mutation rates in the +reagent condition are not above background.

3.3.7 Automatic RNA folding and structure drawing by ShapeMapper

For sequences shorter than 4000 nucleotides and with sufficient read depth, the automated pipeline allows secondary structures to be automatically modeled using RNAstructure (Figure 3.3, panel E). FASTA sequence files are converted to SEQ files required by RNAstructure. SHAPE reactivities are incorporated into RNAstructure as pseudo-free energies using standard parameters for the 1M7 reagent[20] [slope (-sm) 1.8, intercept (-si) -0.6]. Predicted structures are written to .ct files. The lowest energy predicted secondary structures can be drawn and annotated by SHAPE reactivity (Figure 3.8). This stage queries the Pseudoviewer web service[15,21] over an active internet connection. A custom client (pvclient.py) internally converts connect-table[22] (.ct) structures to dot-bracket notation, submits server requests, and retrieves responses. This client also handles coloring of nucleotides by reactivity. Colored structure drawings are vector .eps files. Structures are also automatically converted to .xrna files[23] for optional manual editing. Additional options for rendering multiple structures and customizing coloring are available if this client is executed manually.

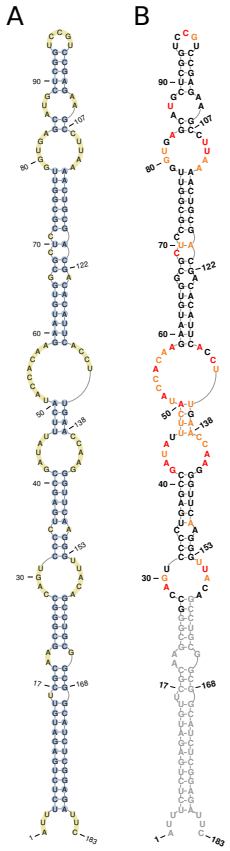

Figure 3.8: Example structure drawing and coloring. The Pseudoviewer web service[15] is automatically queried with a secondary structure. (**A**) Postscript[24] image generated by Pseudoviewer. (**B**) Postscript image with automatic annotation by SHAPE reactivity coloring.

3.4 Hit level calculation and comparison with other reports

SHAPE-MaP structure analysis as read out by massively parallel sequencing presents a valuable tool for structural interrogation of RNA bases at a single nucleotide level. Several other techniques have been developed with similar goals[25–30]. To compare the read depth requirement of SHAPE-MaP (and its mutational profiling readout) with other approaches, we calculated a "hit level". The hit level metric quantifies the total background-subtracted signal per nucleotide of transcript:

$$hit\ level = \frac{total\ events_S - \frac{read\ depth_S}{read\ depth_B} \times total\ events_B}{transcript\ length} \tag{7}$$

where the subscripts *S* and *B* indicate the experimental sample and background control, respectively; *events* are either ligation-detected sequence stops or mutations, depending on readout method, and *read depth* corresponds to the median number of reads overlapping each nucleotide in the transcript. A hit level of 15 is required to fully recover RNA structure information as interrogated by SHAPE, although highly useful structure models were consistently obtained at hit levels as low as 5[8]. In SHAPE-MaP experiments, we often obtained hit levels greater than 100. For example, we obtained a hit level of 160 for the 16S rRNA (experiment performed by Greggory Rice).

High-resolution RNA structure probing and modeling requires that most or all of an RNA be interrogated at a high hit level. Individual regions probed at low hit levels, even if the overall average hit level is 5, are likely to contain notable errors. In PARS experiments, a minimum threshold of 1 average read stop per nucleotide of transcript was required[25,26] corresponding to hit level of 1, assuming zero background for enzymatic cleavage data. Similarly, a report describing DMS chemical probing, structure-seq, used a similar threshold of 1 average stop per A or C nucleotide[27]; this corresponds to an estimated hit level (by our definition) of 0.2, assuming a signal:background ratio of 1.7 (estimated from Extended Data Fig. 1, panel D in ref. 26) and that half of all transcript nucleotides

are A or C. A minimum of 15 reads per A or C on average was required by the creators of DMS-seq[28]. This corresponds to a hit level of 3.3, assuming a signal:background ratio of 1.8 (estimated from Fig. 1, panel C in ref. 27). The authors of SHAPE-seq[29] and Mod-seq[30] have independently noted the importance of read depth in obtaining quantitative RNA structure probing information.

This hit level analysis emphasizes that, although several prior studies have been performed in which the full complement of RNAs in a given transcriptome were present during the probing phase of the experiment, only a few thousand nucleotides in each case were sampled at a depth consistent with recovery of the underlying structure information obtainable using DMS or enzyme probes.

3.5 Conclusion

SHAPE-MaP and the software ShapeMapper have now been extensively validated on RNAs of known structure. *E. coli* ribosomal RNAs have been structurally probed, as well as small structured RNAs including RNase P, the HCV internal ribosomal entry site, a group I intron, a group II intron, a phenylalanine tRNA, and the adenine, glycine, lysine, Mbox, thiamine pyrophosphate, and cyclic-di-GMP riboswitches[2]. The SHAPE-MaP strategy has been applied to characterize the structures of diverse viral RNAs, including HIV-1[8], hepatitis C, satellite tobacco mosaic, and Dengue. ShapeMapper allows any lab to generate accurate SHAPE data, eliminating user bias and reducing data analysis workload. SHAPE-MaP yields accurate and high-resolution secondary structure models and will ultimately democratize RNA structure analysis.

REFERENCES

1. Wilkinson, K. A. *et al.* Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *RNA* **15,** 1314–1321 (2009).

2. Rice, G. M., Leonard, C. W. & Weeks, K. M. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA* **20,** 846–854 (June 2014).

3. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 97–102 (Jan. 2009).

4. Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460,** 711–716 (Aug. 2009).

5. Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M. & Giddings, M. C. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* **14,** 1979–1990 (2008).

6. Karabiber, F., McGinnis, J. L., Favorov, O. V. & Weeks, K. M. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19,** 63–73 (Jan. 2013).

7. Stangegaard, M., Dufva, I. H. & Dufva, M. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. *Biotechniques* **40,** 649–657 (2006).

8. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11,** in press (2014).

9. Mortimer, S. A. & Weeks, K. M. A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry. *J. Am. Chem. Soc.* **129,** 4144–4145 (2007).

10. Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R. & Patel, D. J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441,** 1167–1171 (2006).

11. Rossum, G. v. *The Python Language Reference* <http://docs.python.org/2.7/reference/index.html> (2014).

12. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (Apr. 2012).

13. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9,** 90–95 (Jan. 2007).

14. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11,** 129 (2010).

15. Byun, Y. & Han, K. PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res.* **34,** W416–22 (July 2006).

16. *Torrent Suite Use Cases* <http://mendel.iontorrent.com/ion-docs/Use-DNA-Barcodes-with-the-Ion-Sequencers.html> (2014).

17. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* **8,** 251–259 (Mar. 1998).

18. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* (Jan. 2011).

19. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (Aug. 2009).

20. Hajdin, C. E. *et al.* Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.* (2013).

21. Byun, Y. & Han, K. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics* **25,** 1435–1437 (Jan. 2009).

22. Mathews, D. H. RNA secondary structure analysis using RNAstructure. *Curr. Protoc. Bioinformatics* **46,** Unit 12.6 (Mar. 2014).

23. Weiser, B & Noller, H. *XRNA: Auto-interactive program for modeling RNA.* <http://rna.ucsc.edu/rnacenter/xrna/> (1995).

24. Taft, E. & Walden, J. *PostScript language reference manual* (eds Walden, J. & Engstrom, P.) (Adobe Systems Incorporated, 1990).

25. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (Sept. 2010).

26. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (Jan. 2014).

27. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (Jan. 2014).

28. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505,** 701–705 (Jan. 2014).

29. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11063–11068 (July 2011).

30. Talkish, J., May, G., Lin, Y., Woolford, J. L. J. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20,** 713–720 (May 2014).

4 HIGH-RESOLUTION MAP OF AN *E. COLI* TRANSCRIPTOME

4.1 Introduction

The successful development of SHAPE-MaP made accurately mapping transcriptome-scale RNA flexibility a feasible experiment, with the ultimate goals of locating functional RNA structures with low free energies of folding and determining global trends in RNA structure. Several groups have reported RNA structure probing results from transcriptomes[1–4], but very few or none have achieved adequate signal over background over large numbers of nucleotides (see Section 3.4). In addition, none have demonstrated consistently accurate structure modeling. In contrast, the SHAPE-MaP strategy both estimates the variation in the signal and enables accurate structure modeling. For these reasons, SHAPE-MaP was an ideal strategy to apply to the structures formed in a bacterial transcriptome.

4.2 Experimental Methods

Cell growth, SHAPE probing, and sequencing library preparation were performed by Christopher Leonard. SHAPE probing and library preparation were performed according to the strategies detailed in (Siegfried, N. *et al.*, 2014)[5], with the addition of a ribosomal RNA depletion step.

Briefly, *E. coli* DH5α cells were grown to mid-log phase in Luria broth. RNA was extracted by three serial phenol:chloroform extractions, followed by DNase I treatment (Ambion) according to the manufacturer's recommended protocol. Ribosomal RNAs were depleted using a Ribo-Zero kit (Epicentre). RNA was purified and concentrated using an RNeasy kit (silica membrane spin columns, Qiagen).

RNA for the untreated and SHAPE samples was refolded *in vitro* in the presence of 100 mM HEPES, pH 8.0, 100 mM NaCl, and 10 mM $MgCl_2$. RNA in the SHAPE

sample was modified in the presence of 10 mM 1M7 at 37 °C for 3 minutes. The untreated sample was incubated with solvent at 37 °C for 3 minutes. The denatured sample RNA was modified with 1M7 under strongly denaturing conditions: 50 mM HEPES (pH 8.0), 4 mM EDTA, and 50% formamide at 95 °C.

RNA was fragmented[5], and reverse transcription primer extension was performed using random dodecamers with SuperScript II (Invitrogen) in the presence of 6 mM Mn^{2+}. Sequencing library preparation was performed using TruSeq adapters, and the library was sequenced on a HiSeq instrument (Illumina), using 100-nucleotide paired end reads.

4.3 Software for transcript calling and curation

I developed a simple software program for locating transcript boundaries in the transcriptome dataset. Transcript boundaries were first called automatically, then manually curated in a graphical environment (Figure 4.1).

4.3.1 Automated transcript calling

The automated calling of transcript bounds based on sequencing read depths required an estimate of the distribution of read depths over any given transcript in the absence of transcript edges. The distribution of read depths over a single transcript was ultimately modeled as a normal distribution centered at 1 with a standard deviation of 0.431. This distribution was estimated by extracting read depth profiles for all coding regions with median depths greater than 2000. Coding regions (genes) were chosen as a proxy for transcripts, since coding regions are unlikely to contain discontinuities from transcription start and stop sites. The per-gene depth profiles were normalized to their respective median depths, and combined to give the final distribution.

For individual transcript calls, a local maximum depth was chosen. The upstream and downstream boundaries on the transcript were simultaneously incremented until the depths at both edges met the condition:

$$localMedian - depth_{edge} > m * stdev_{model} * localMedian \quad (1)$$

where $localMedian$ is the median of the depths in a 150-nucleotide window nearest to the current boundary (upstream or downstream), $depth_{edge}$ is the depth at the current boundary, $stdev_{model}$ is 0.431, and m is a multiplier allowing the selection of various confidence intervals. For automated calls, m was set to 1.96, corresponding to a confidence interval of 95%.

Initial transcript calls were generated by selecting the nucleotide with the highest depth, choosing transcript boundaries by the algorithm above, and repeating until no uncalled nucleotides existed with depths above 1000.

4.3.2 Transcript call curation

Transcript calls were manually curated using a graphical environment (Figure 4.1) designed and implemented in python using the matplotlib library. The program displays four rows of information (from top to bottom): the boundaries of existing automated transcript calls, the base-ten logarithm of the untreated sample read depth, the boundaries and strands of annotated genes, and the boundaries and strand of EcoCyc-annotated transcripts[6].

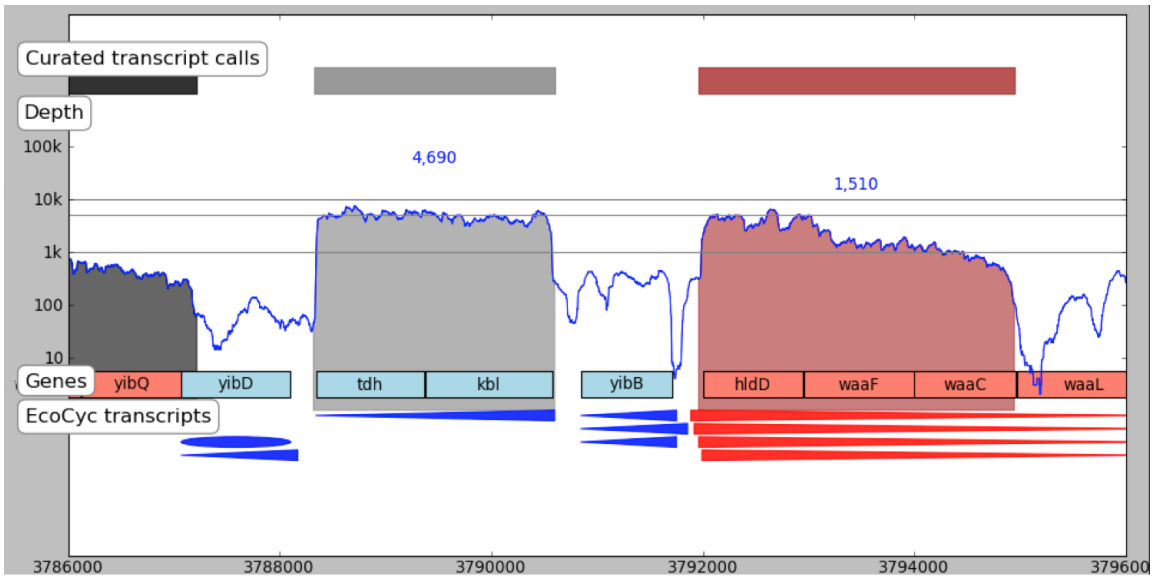

Figure 4.1: Transcript call curation. Screenshot of user interface for transcript boundary curation. Existing transcript annotations from EcoCyc[6] are shown on the bottom track. Vertical edges on these shapes indicate known transcription start sites or terminators. Transcripts and genes on the sense strand are shown in red, and those on the reverse strand in blue. Median coverage depths for curated transcripts are shown in blue text.

Transcription units were extracted from the EcoCyc database using queries in the Lisp language and written to simplified text files for loading and display within the graphical environment. Existing database transcript annotations are helpful, but incomplete[7]. Semi-automated transcript calling is therefore available within the program by clicking on a nucleotide in the depth profile. The value of m in equation (1) above can be manually set to any value. Any visible boundary (from EcoCyc annotations, gene coordinates, existing transcript calls, or interactive transcript calls) may be selected and written to file.

Reactivity files were created for each transcript call, containing genome coordinates, nucleotide sequence, SHAPE reactivities, and reactivity standard errors. These files also contain automatically generated warning messages for potential overlap with nearby transcripts, unclear transcript boundaries, or errors of transcript sense. A total of 562 transcripts were called.

4.3.3 Depth requirements

A sequencing read depth of 5000 was required for consistently accurate structure prediction in bootstrapped simulations of 16S rRNA modeling[5]. In transcript calling, I required that transcripts have a median untreated read depth of above 1000, since lower-quality reactivity profiles might still be useful for assessing global RNA flexibility trends. For transcriptome-wide reactivity profile normalization, I required an untreated read depth of 5000.

4.3.4 Standard error filter

For individual nucleotides, I imposed a filter based on the standard error of the SHAPE reactivity signal, requiring that:

$$stderr \leq |SHAPE| * 0.5 + 0.4 \tag{2}$$

At unreactive positions, this filter requires that the standard error be no greater than the reactivity at which the SHAPE pseudo-free energy term in RNAstructure[8] is zero. At highly reactive positions, this filter allows comparatively larger standard errors, while still

rejecting highly noisy positions.

4.4 Validation and global trends

4.4.1 Coverage and structure modeling statistics

HiSeq 2x100 rapid run

Sample	Paired-end reads	Reads mapping to *E. coli*
SHAPE treated	75,579,880	73,698,260
Untreated	62,957,520	61,536,398
Denatured	61,953,160	60,747,140

Table 4.1: **Sequencing statistics.**

A single lane of a HiSeq sequencing run produced nearly 2 billion reads mapping to the *E. coli* genome (Table 4.1). 1.89 million, or 41% of nucleotides in the *E. coli* genome passed the data quality filter given in equation (2). However, these nucleotides are not continuously distributed. Only 940 thousand, or 20% of genomic nucleotides are located in transcripts with median untreated sequencing depths above 1000, for a total of 562 transcripts. Only 260 thousand, or 6% of genomic nucleotides pass a more stringent median depth requirement of 5000, in a total of 166 transcripts. Given that the majority of nucleotides in the *E. coli* genome appear to be transcribed, the present study provides accurate structural information for a subset of the transcriptome. A plot of genome coverage makes this apparent (Figure 4.2).

Accurate structural information is available for a number of transcripts, primarily highly expressed "housekeeping" RNAs[9] (Figure 4.3), from short ncRNAs of about 200 nucleotides to long mRNAs of up to 15,000 nucleotides. In contrast to other massively parallel sequencing approaches for RNA structure, SHAPE-MaP reports single-nucleotide resolution reactivity information including all four nucleotide types (Figure 4.4). SHAPE

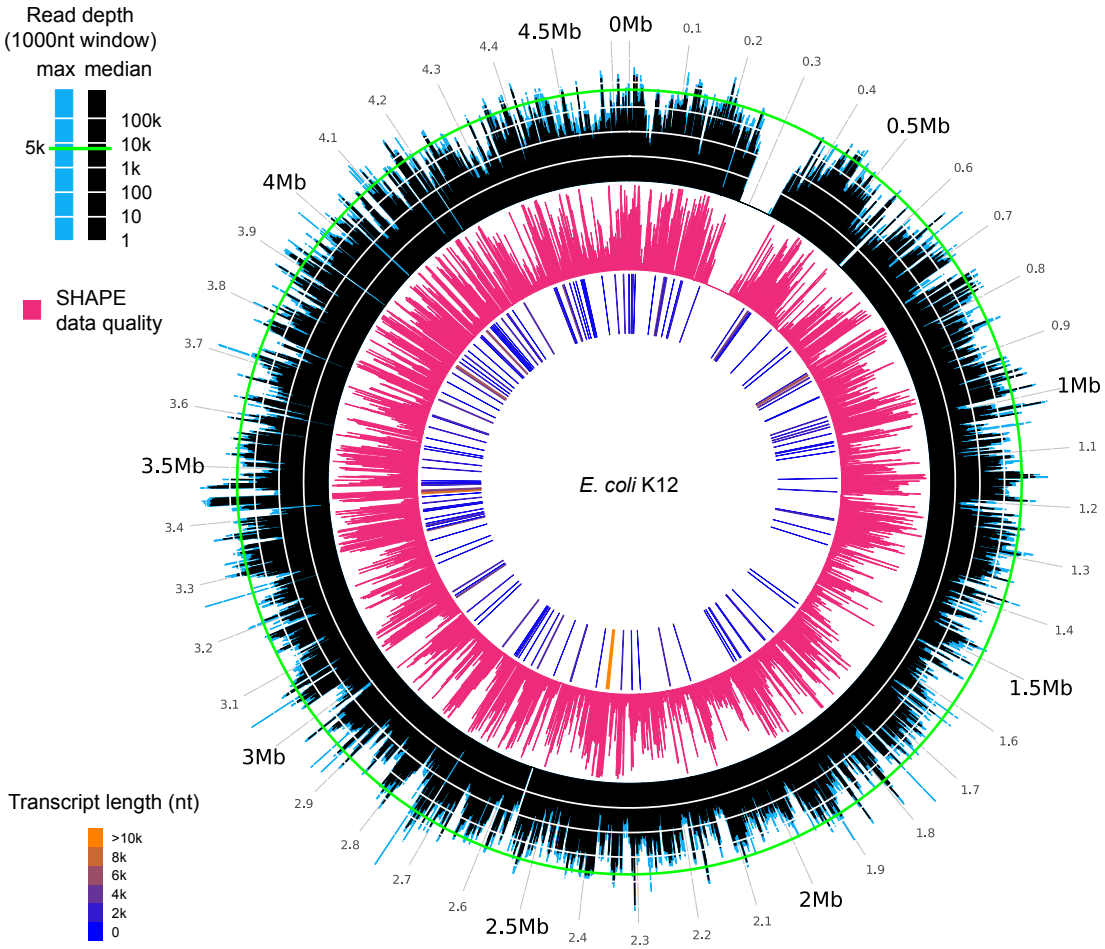

Figure 4.2: *E. coli* genome coverage. *Outer track:* depth of sequencing coverage in the untreated control sample, shown as both a max and median over 1000-nucleotide windows. A nominal threshold for the sequencing depth required to obtain high-quality SHAPE reactivity profiles and structure models is shown in green. Gaps in coverage (for example, at 0.3 Mb) reflect prophage deletions in the DH5α strain probed compared to the K12 reference strain. *Middle track:* "SHAPE data quality" is the count of nucleotides passing a filter based on the standard error of the SHAPE reactivity signal (see equation (2)) over 1000-nucleotide windows, and complements the data quality assessment based on depth alone. *Inner track:* Transcripts with untreated sample median depths above 5000, colored by transcript length.

reactivities closely matched known patterns of base pairing for three highly-structured ncRNAs (Figure 4.5). SHAPE-constrained models created using RNAstructure and ShapeKnots[10] recovered 93% and 85% of base pairs for the 6S rRNA and the RNA component of RNase P, respectively (Table 4.2). The unusually highly pseudoknotted structure of tmRNA precludes its accurate prediction using current methods.

Name	Length	Pseudoknots	Sensitivity (%)	PPV (%)
6S ncRNA	183	0	93.0	91.4
RNase P	377	2	85.5	86.9

Table 4.2: **Structure modeling statistics for two previously characterized RNA structures.**

Highly-structured RNAs are easily discoverable by identifying transcripts with long regions of low median reactivity (as described in section 4.5.1, "Low-SHAPE regions"). Unlike previous approaches, SHAPE-MaP allows the modeling of novel structures. An intriguing example is the mRNA encoding major membrane lipoprotein (*lpp*) (Figure 4.6). SHAPE reactivities clearly indicate the *lpp* transcript is highly base paired, but this structure has been uncharacterized until now, with the exception of a transcription terminator hairpin containing 13 base pairs[11]. The *lpp* mRNA is highly abundant, largely poly-adenylated, and relatively long-lived, with a half-life of 12 minutes in cells[12]. The translation start site is positioned in a large single-stranded loop, the translation stop codon is located in a smaller loop, and no pseudoknots are predicted. The function of the *lpp* RNA structure is unknown, but it could act to resist exonuclease degradation or to aid recognition by protein factors involved in poly(A) polymerization, such as Hfq[11].

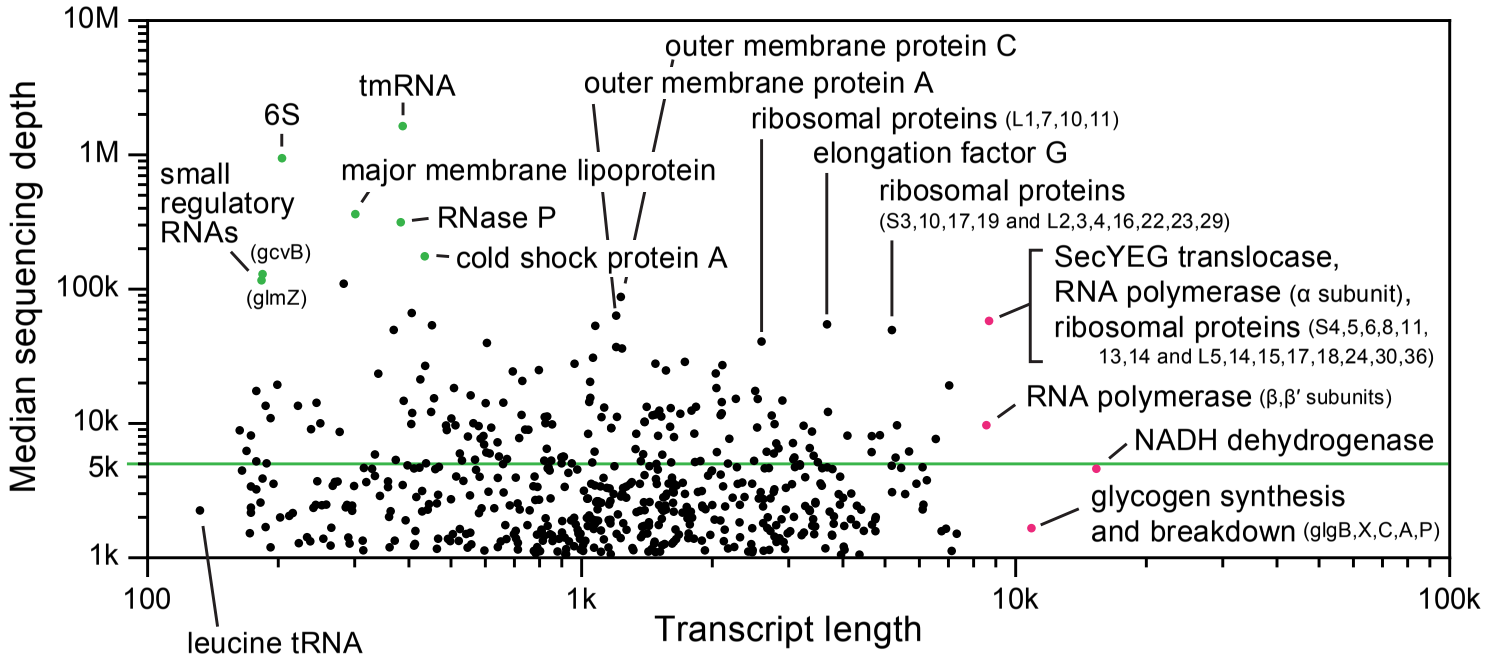

Figure 4.3: Transcript lengths and sequencing depths. Median depths are from the untreated control sample. Several transcripts with especially high read depths or lengths are colored green or magenta, respectively.

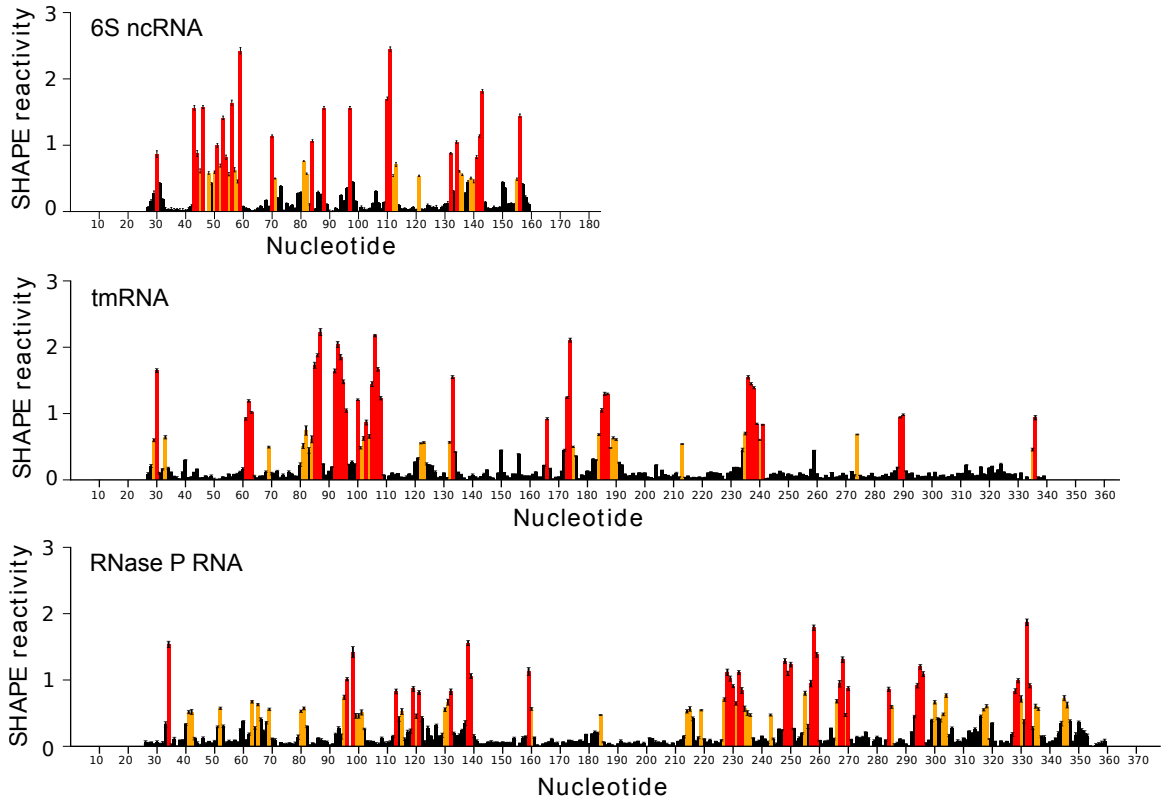

Figure 4.4: Representative SHAPE-MaP reactivity profiles. Reactivity profiles for three highly expressed ncRNAs. Error bars indicate standard errors.

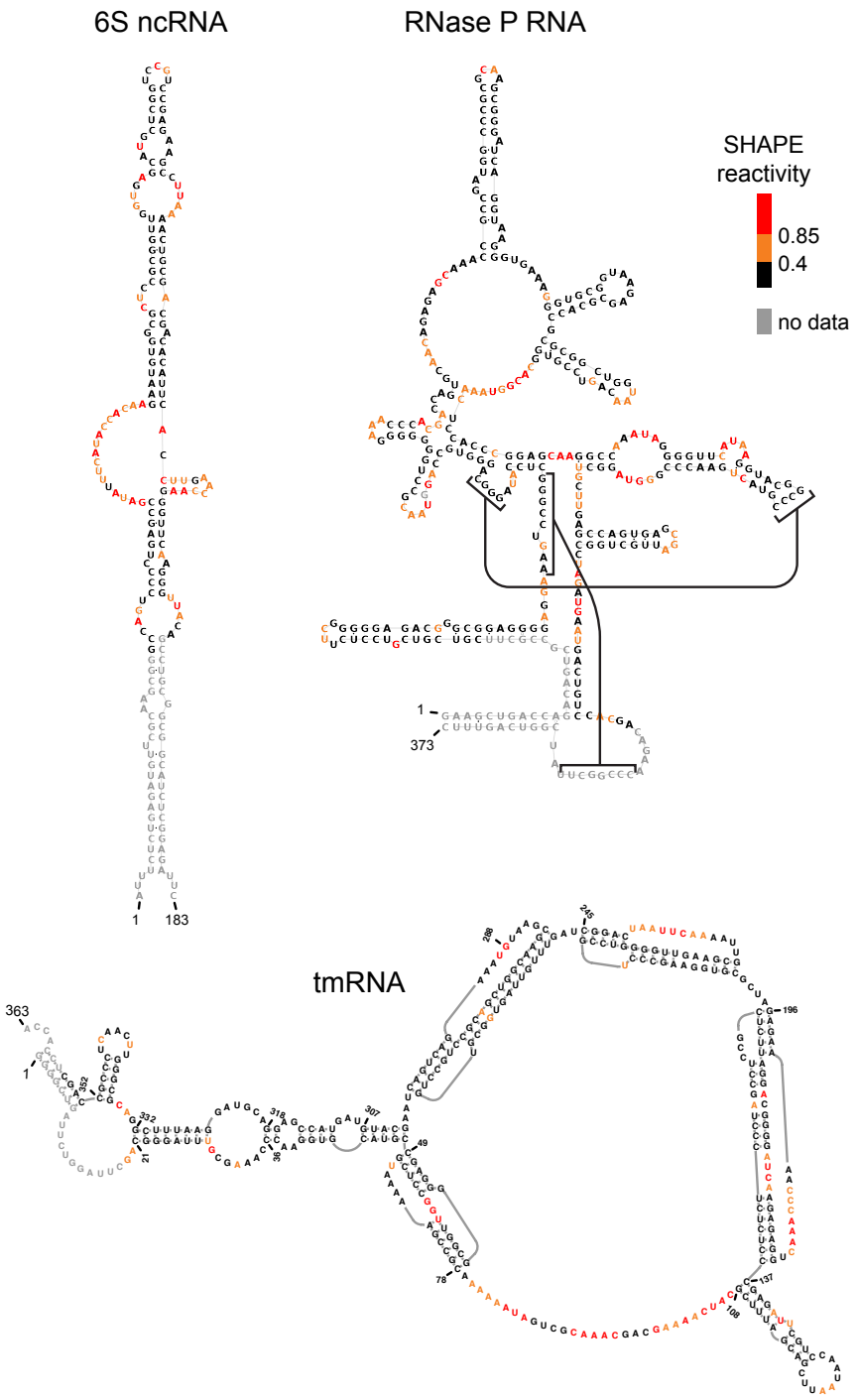

Figure 4.5: Representative secondary structures. SHAPE-MaP reactivity colorings superimposed on accepted secondary structures for three RNAs.

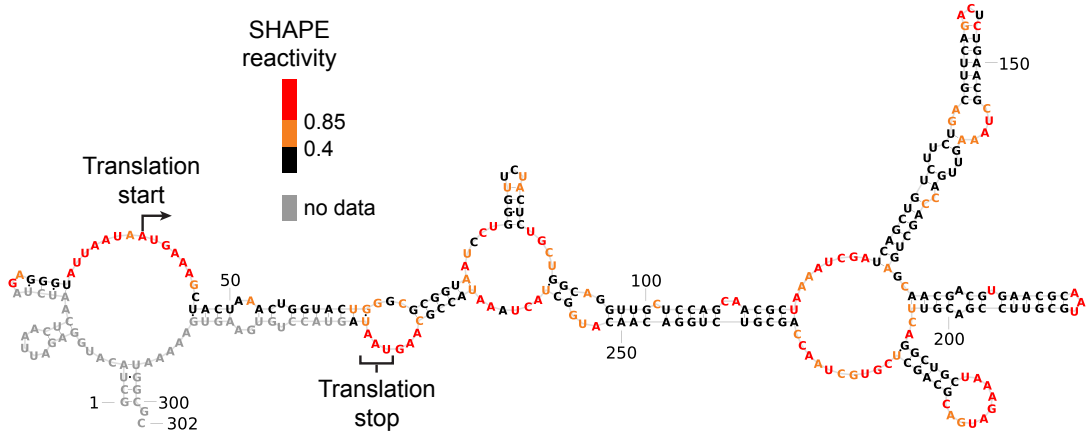

Figure 4.6: Proposed secondary structure for the transcript encoding major membrane lipoprotein. Possible poly-adenylation of this mRNA is not shown, as the current randomly-primed data do not provide information about the very ends of transcripts. A previously reported terminator hairpin[11] is not included in this model—its addition would likely change the base pairing pattern of 8 nucleotides on the 5′ end of the RNA.

4.4.2 Large-scale trends in E. coli transcript flexibility

Several previous reports of transcriptome-scale RNA structure probing using massively parallel sequencing have included average reactivity or cleavage profiles over the regions surrounding start codons, over the regions surrounding stop codons, and over coding region interiors for several eukaryotic organisms[1–4]. A similar analysis was performed over the bacterial transcripts in the current study (Figure 4.7, panel A). The large standard deviations emphasize the wide variation in local transcript flexibility, and suggest that there are no universal RNA structure features in relation to translation start and stop sites.

However, several trends are apparent. The 50 nucleotides surrounding start codons exhibit a distinctive pattern of reactivity (Figure 4.7, panel C). This trend is present in start codons both near the 5′ ends of transcripts and 3′ of intergenic regions (Figure 4.7, panel A). Stop codons display a simple increased reactivity, regardless of whether they precede intergenic regions or 3′ UTRs.

If the trend for increased stop codon flexibility reflected a functional RNA structural feature, out-of-frame stop codons could be expected to show a different range of reactivities than in-frame codons, reflecting the effects of selection. A comparison of in-frame and out-of-frame stop codons (Figure 4.7, panel D) shows no evidence for selection of stop codon flexibility, suggesting that the average reactivity trends are largely a result of local sequence content. Since base pairs containing adenosine and uridine usually participate in one less hydrogen bond than base pairs containing guanosine and cytidine, sequences in an RNA with high AU content will on average display greater flexibility than regions with high GC content[13]. Indeed, the three-nucleotide centered mean AU content closely follows the mean SHAPE reactivity (Figure 4.8, panels B and C), with a linear correlation R value of 0.67. In contrast, the single-nucleotide AU content correlates more poorly with mean SHAPE reactivity, especially over coding regions. In coding regions, a periodic sequence trend is

clearly present, but it does not bias the reactivity profile (Figure 4.8, panel A). This analysis provides a strong demonstration that SHAPE-MaP accurately reports nucleotide flexibility with little nucleobase bias. Previous reports of a periodic flexibility or base pairing trend within coding regions[1–4] are likely showing a side effect of using probes or enzymes that only report on a subset of the four nucleotides.

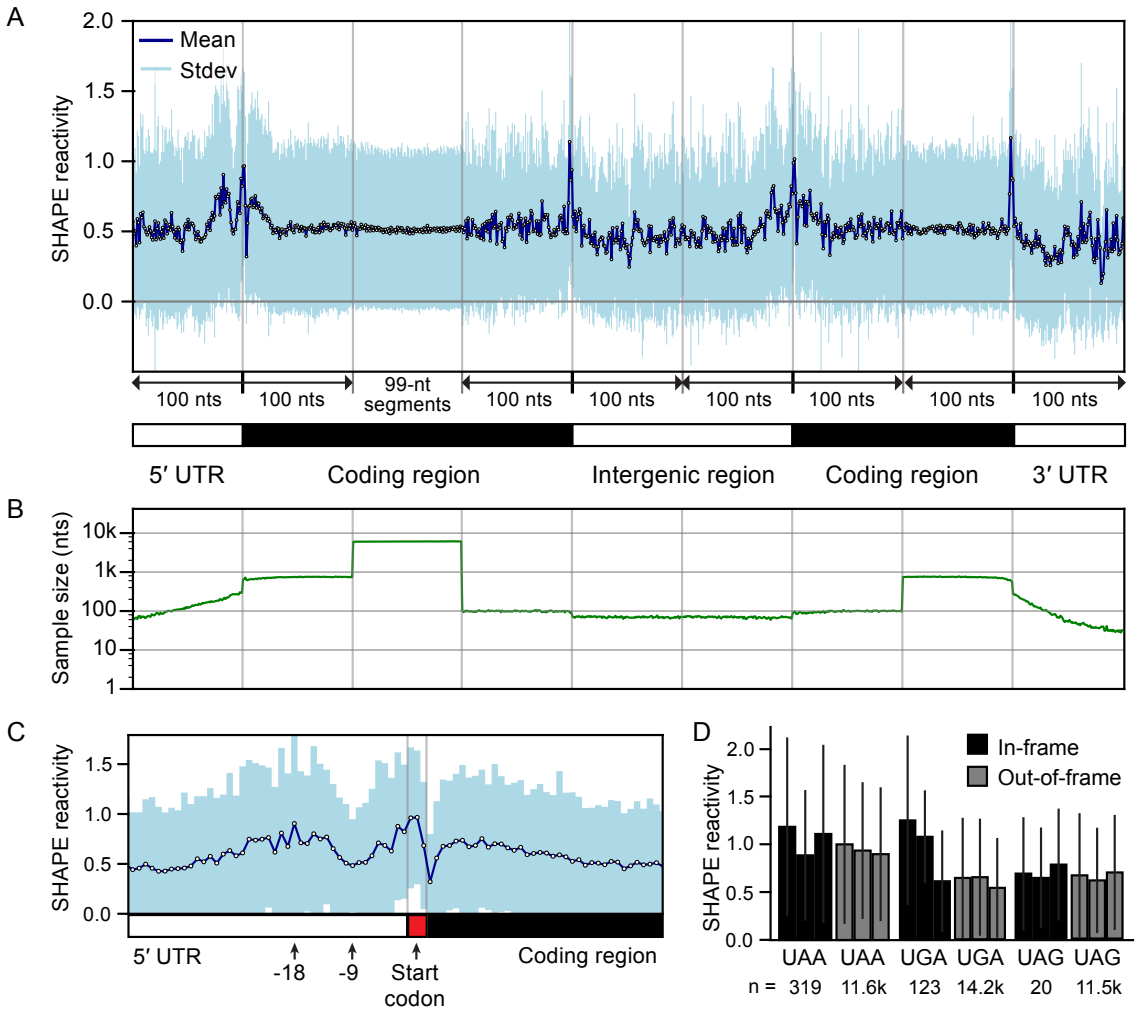

Figure 4.7: SHAPE reactivities across untranslated, protein-coding, and intergenic regions. (**A**) Mean and standard deviation of SHAPE reactivities surrounding various genome features. Short vertical black lines below the top panel indicate aligned locations. (**B**) Number of high-quality data points (transcripts) at each position. Noisy locations in the mean reactivity profile generally fall in regions with low representation, for example, in the intergenic regions, to which less than 100 transcripts contribute. (**C**) Close-up of the trend surrounding start codons. Start codons tend to be highly reactive, as are nucleotides centered around 18 nt upstream. The nucleotide immediately downstream of start codons tends to be lowly reactive. A general (**D**) Comparison of the three major stop codons, both in-frame and out-of-frame. Bars show mean SHAPE reactivity, and lines show standard deviation. No significant differences between in-frame and out-of-frame codons exist, providing no evidence for selection for or against stop codon flexibility.

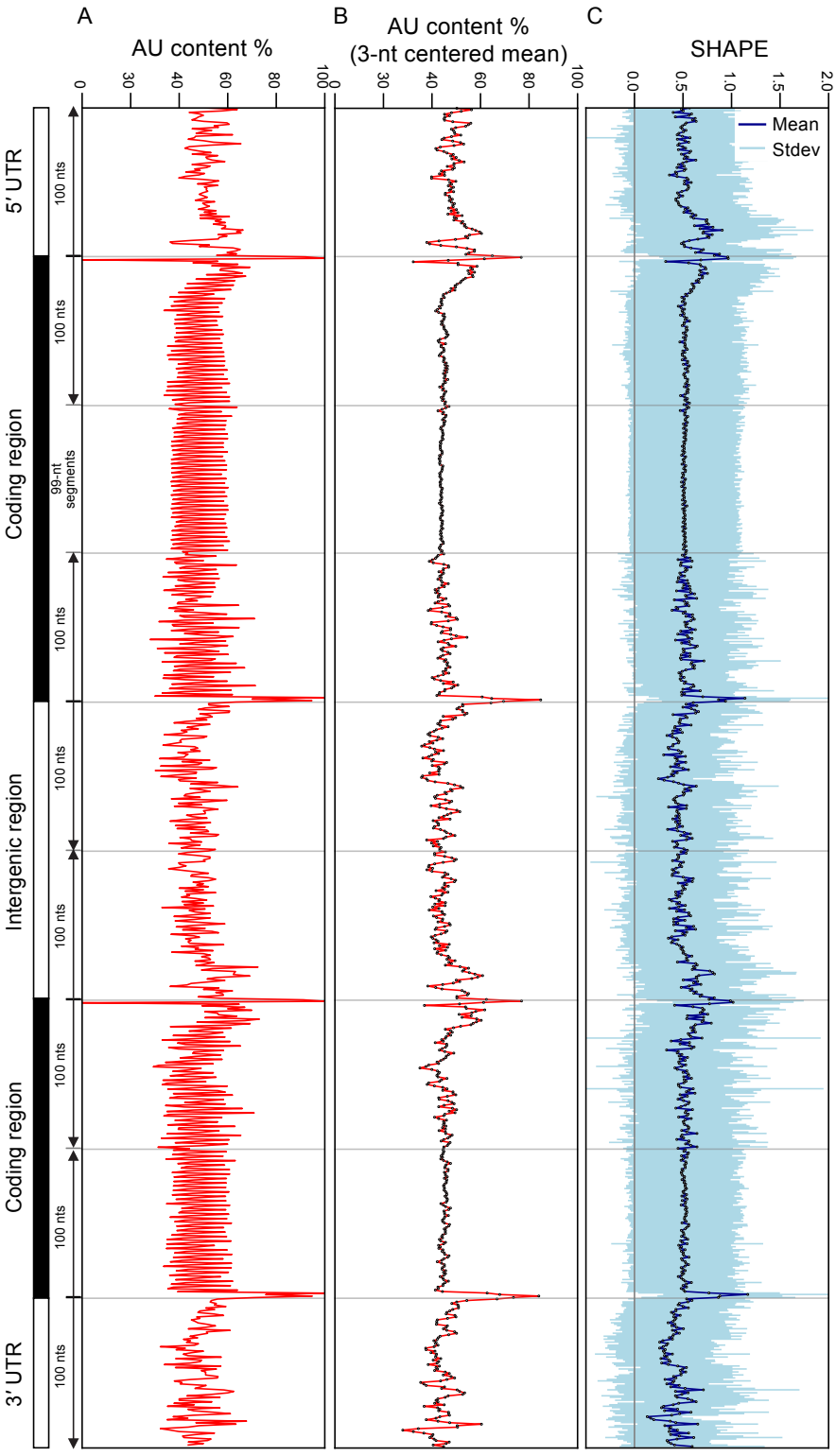

Figure 4.8: AU content and global transcript flexibility trends. The 3-nt centered mean AU content correlates with mean SHAPE reactivity with an R-value of 0.67.

4.5 Transcriptome-wide RNA structure motif discovery by local sequence and MaP clustering

The high levels of variability in SHAPE reactivities across genomic features (see Figure 4.7) suggested that pairwise alignment of local transcriptome regions and clustering using similarity scores could detect conserved or duplicated RNA structures where alignment to genomic features failed. Regions with relatively low SHAPE reactivities are good candidates for highly structured RNAs.

4.5.1 Computational methods

Low-SHAPE regions

Low-SHAPE regions were selected as follows. The standard error filter described in Section 4.3.4 was applied to each nucleotide, excluding noisy positions. A windowed 50-nucleotide centered median SHAPE reactivity was calculated for each transcript, excluding 50-nt regions with more than 25 nucleotides thrown out in the previous step. Nucleotides with windowed medians less than or equal to 0.25 (a low reactivity) were collected into contiguous regions. These regions were expanded by 25 nucleotides on both ends, and overlapping regions merged. The resulting regions were segmented into 150-nucleotide windows, with a step size interval of 50 nucleotides. Windows with more than 25 excluded nucleotides were not included.

Sequence alignment

Pairwise sequence alignments were performed using an approach developed by Andy Lavender and others in the Weeks lab (submitted). Briefly, the Needleman-Wunsch dynamic programming method for finding the optimal global sequence alignment was performed, with the addition of an optional SHAPE reactivity match score. Standard alignment parameters were as follows: gap open penalty: -8.5, gap extension penalty: 0, sequence match bonus: 2, sequence mismatch penalty: -0.5. The optional SHAPE reactivity score was given for each pair of nucleotides in the scoring matrix by:

$$SHAPEscore = b + n_0 * e^{-l*|SHAPE_A - SHAPE_B|} \quad (3)$$

with n_0=4, b=-1, and l=1. For speed, the alignment algorithm was re-implemented in the C programming language (from its original implementation in python). For the alignment and clustering shown in Figure 4.9, alignments were performed with sequence alone, without the optional SHAPE match score.

Distance matrix processing

A distance matrix was calculated, recording the alignment scores for all possible pairs of 150-nucleotide low-SHAPE regions. Scores were discarded for which less than 30 nucleotides overlapped in the alignment. Scores were normalized to the maximum alignment score.

Because sliding windows contain regions of identical sequence, a given region may align to contiguous windows with similar scores, confounding visualization. For each 150-nucleotide low-SHAPE region, alignment scores to contiguous low-SHAPE regions were therefore pruned by iteratively selecting the highest score and removing the scores with indices within +2 and -2 of the maximum scoring region. After this process, scores in the top 99.9th percentile were retained for clustering.

Clustering

Clustering was performed using the Markov Cluster Algorithm[14], using default parameters and the distance matrix described above.

4.5.2 Results

676 low-SHAPE regions of 150 nucleotides each, with a step size of 50 nucleotides, were extracted from the overall dataset for this analysis, as detailed in Section 4.5.1. Attempts to identify conserved RNA structures without respect to sequence (clustering using pairwise SHAPE profile correlations) were inconclusive, but should be revisited with access to a more exhaustive dataset. This section will instead focus on a simpler approach—the identification of RNA structures repeated in the transcriptome by clustering

using sequence alignment scores between low-SHAPE regions.

A schematic of the results of this analysis is shown in Figure 4.9. The detection of pre-tRNAs provides a validation of this approach, since tRNAs are highly-structured RNAs with a high degree of sequence conservation[15]. The other RNA structure elements identified by this approach are repetitive extragenic palindromic (REP) elements and transcriptional terminators (in some cases previously unannotated).

REP elements are mobile genomic elements in bacteria composed of short inverted complementary regions of about 10–30 nucleotides separated by a spacer of about 2–6 non-conserved nucleotides[16]. REP elements are often located near each other in tandem groups of 2 to 4, and usually fall in intergenic (non-coding) regions[17]. Currently, the best-supported model explaining the distribution of these elements describes REPs as selfish DNA replicators within the genome[18], in some cases using specific transposases to catalyze replication[19,20]. If this is the case, the folded RNA structure formed by these elements is likely to be a side effect of the sequences necessary for their propagation, and not a cause, although there are limited examples of REP hairpins that modulate RNA transcription or degradation[21]. Even so, these elements provide another clear demonstration of the accuracy of SHAPE-MaP data, since highly reactive nucleotides in these elements occur precisely at the non-conserved spacers between inverted repeats (Figure 4.10).

Transcriptional terminators are RNA hairpins that serve to terminate transcription without the requirement for external protein factor binding (commonly termed “intrinsic termination”). These elements have typically been described as short hairpins, containing 13–23 total nucleotides[22], although extended structures have been described that appear to enhance transcription termination[23]. The current analysis identified transcriptional terminators likely to fold into longer extended hairpins, which surprisingly contained up to 127 nucleotides (Figures 4.11 and 4.12).

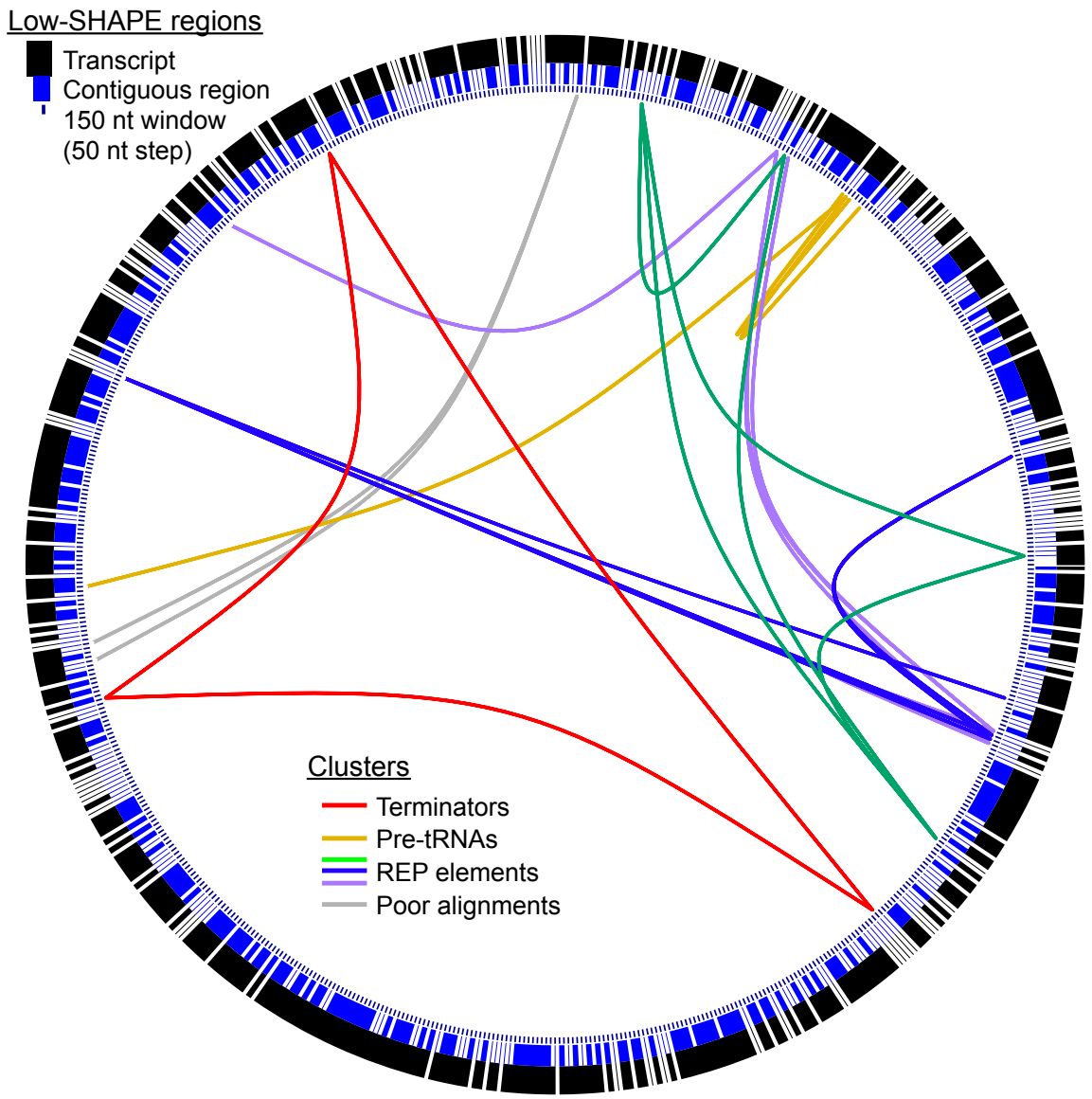

Figure 4.9: Sequence clusters. Known conserved, repeated sequences are clustered, including pre-tRNAs, transcription terminators, and repetitive extragenic palindromic (REP) elements. Arcs between regions indicate pairwise alignment scores above the threshold described in Section 4.5.1. Arcs between regions in the same cluster use the same color. Clusters with less than three member regions were excluded from this plot.

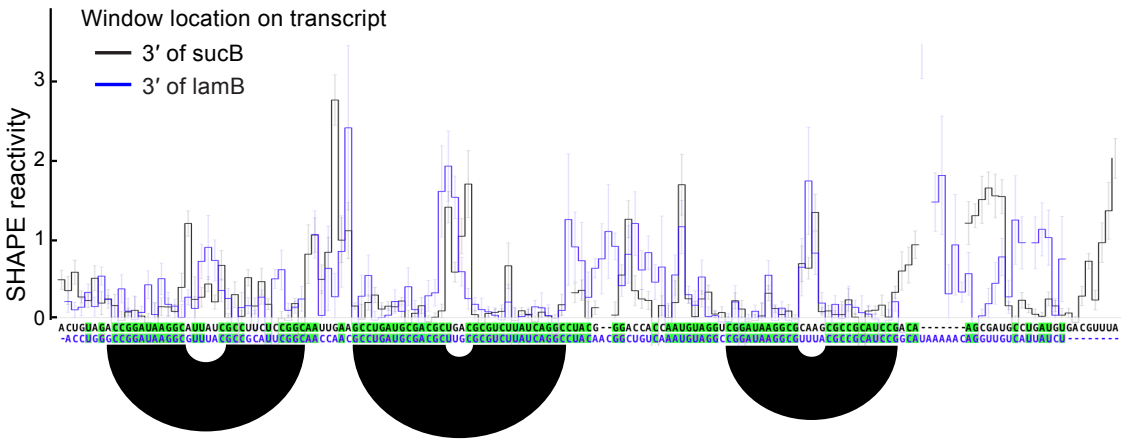

Figure 4.10: REP elements. Two low-SHAPE regions containing REP elements are shown aligned by sequence. Nucleotides highlighted in green are conserved between the two windows. Error bars indicate standard error.

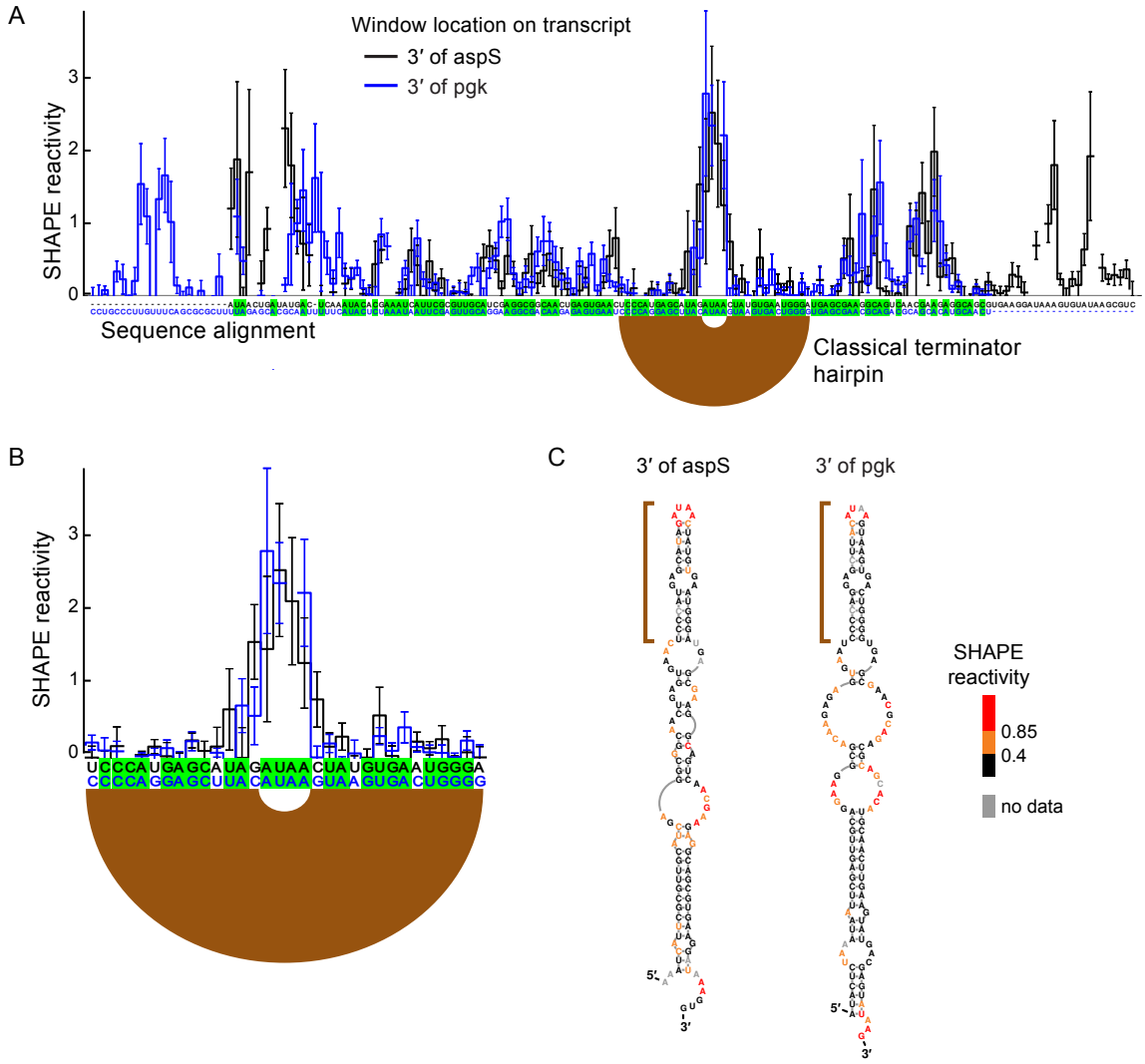

Figure 4.11: Selected terminators. (**A**) Sequence alignment of two 150-nucleotide windows and SHAPE reactivities. Nucleotides highlighted in green are conserved between the two windows. Error bars indicate standard error. Brown arcs indicate previously annotated base pairing. (**B**) Close-up of classical terminator hairpin. (**C**) Structure modeling of the conserved region surrounding each hairpin indicates more extensive pairings are likely present. Brown brackets indicate the extent of each classical stem.

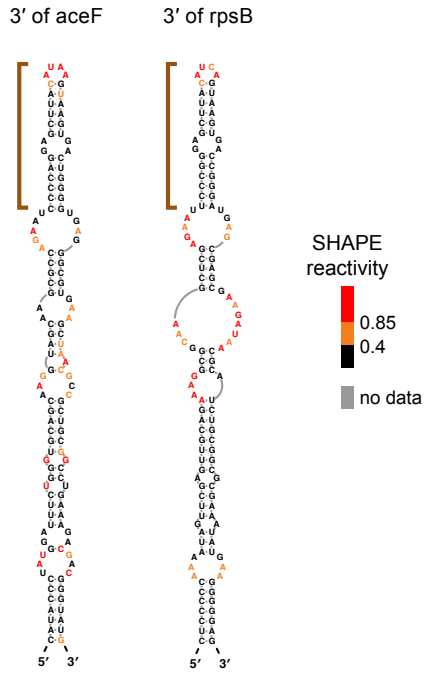

Figure 4.12: Additional terminators. Two previously unannotated terminator-like hairpins identified by low SHAPE reactivity and clustering by sequence alignment. Brown brackets indicate the extent of each classical stem.

4.6 Future improvements

As a pilot study, this dataset points to the need for increased data quality, either by increasing the total read depth, by increasing the signal above background, or by leveling the distribution of cDNAs in the sequencing library. Increasing the sequencing depth is the simplest approach to improve data quality, but is currently prohibitively expensive[24,25]. This approach is also likely to provide diminishing returns, since additional sequencing improves the coverage of highly-abundant transcripts to a greater degree than rare transcripts.

Increased signal above background could be obtained in at least four ways: increased SHAPE adduct formation, increased adduct detection rate, decreased background signal, and revised sequencer sample loading ratios. Increased adduct formation would require either multiple rounds of modification or the creation of newer, more highly soluble SHAPE reagents. Increased adduct detection rates could in theory be obtained by engineering improved reverse transcriptases or by changing reverse transcription conditions, although the current adduct detection rate is already estimated at 50%. Decreased background signal could also be obtained by reverse transcriptase engineering[26–28]. The contributions of the three samples (SHAPE-modified, untreated, and denatured) to the standard error of the SHAPE reactivity signal (ch. 3 equation 3) suggest that decreased noise for the same total sequencing depth could be obtained by reducing the concentration of untreated sample cDNA loaded on the sequencer relative to the other samples.

Leveling the abundance distribution of cDNAs in the library to be sequenced would improve signal quality for rare transcripts. This could be performed using a method called cDNA normalization[29], in which double-stranded cDNA is denatured, renatured, and treated with a double-stranded DNA nuclease. This process selectively digests highly abundant cDNAs, since DNAs with rare sequences are more likely to be unpaired at the time of DNase treatment[30].

4.7 Conclusion

This is the first study applying SHAPE-MaP to a bacterial transcriptome. SHAPE-MaP accurately reported nucleotide-resolution structural information in this large-scale experiment. For well-studied RNAs, SHAPE-MaP data agreed closely with known structures, while for poorly characterized RNAs, these data enabled the accurate modeling of novel structures. Profiling SHAPE reactivity over short windows identified regions of high structure, and clustering these structured regions by sequence alignment score identified repeated structured elements in the *E. coli* genome, including pre-tRNAs, transcription terminators, and REP elements. With improvements in experimental efficiency, SHAPE-MaP will allow the comprehensive structural characterization of nearly all the RNAs produced in a bacterium.

REFERENCES

1. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505,** 701–705 (Jan. 2014).

2. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (Jan. 2014).

3. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (Jan. 2014).

4. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (Sept. 2010).

5. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11,** in press (2014).

6. Keseler, I. M. *et al.* EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.* **39,** D583–90 (Jan. 2011).

7. Karp, P. D. *et al.* Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res.* **35,** 7577–7590 (2007).

8. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11,** 129 (2010).

9. Gil, R., Silva, F. J., Pereto, J. & Moya, A. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68,** 518–37–table of contents (Sept. 2004).

10. Hajdin, C. E. *et al.* Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.* (2013).

11. Mohanty, B. K. & Kushner, S. R. Bacterial/archaeal/organellar polyadenylation. *Wiley Interdiscipl. Rev. RNA* **2,** 256–276 (Mar. 2011).

12. Taljanidisz, J, Shen, P & Sarkar, N. Half-life of Escherichia coli polyadenylated lipoprotein mRNA. *Biochem. Mol. Biol. Int.* **42,** 211–215 (June 1997).

13. Zheng, H. & Wu, H. Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species. *BMC Bioinformatics* **11,** S7 (Dec. 2010).

14. Enright, A. J., Van Dongen, S & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (Apr. 2002).

15. Widmann, J., Harris, J. K., Lozupone, C., Wolfson, A. & Knight, R. Stable tRNA-based phylogenies using only 76 nucleotides. *RNA* **16,** 1469–1477 (Aug. 2010).

16. Di Nocera, P. P., De Gregorio, E. & Rocco, F. GTAG- and CGTC-tagged palindromic DNA repeats in prokaryotes. *BMC Genomics* **14,** 522 (2013).

17. Bachellier, S., Clément, J.-M. & Hofnung, M. Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol.* **150,** 627–639 (Nov. 1999).

18. Bertels, F. & Rainey, P. B. Within-Genome Evolution of REPINs: a New Family of Miniature Mobile DNA in Bacteria. *PLoS Genet.* **7,** e1002132 EP – (2011).

19. Nunvar, J., Huckova, T. & Licha, I. Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. *BMC Genomics* **11,** 44 (2010).

20. Messing, S. A. J. *et al.* The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. *Nucleic Acids Res.* **40,** 9964–9979 (Oct. 2012).

21. Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J. L. & Marqués, S. Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida. *Nucleic Acids Res.* **30,** 1826–1833 (Jan. 2002).

22. Wilson, K. S. & von Hippel, P. H. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 8793–8797 (Sept. 1995).

23. Cambray, G. *et al.* Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Res.* **41,** 5139–5148 (Jan. 2013).

24. Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. *J.Biomed. Biotechnol.* **2012,** 11 (2012).

25. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52,** 413–435 (2011).

26. Alvarez, M., Matamoros, T. & Menendez-Arias, L. Increased thermostability and fidelity of DNA synthesis of wild-type and mutant HIV-1 group O reverse transcriptases. *J. Mol. Biol.* **392,** 872–884 (Oct. 2009).

27. Yasukawa, K., Mizuno, M., Konishi, A. & Inouye, K. Increase in thermal stability of Moloney murine leukaemia virus reverse transcriptase by site-directed mutagenesis. *J. Biotechnol.* **150,** 299–306 (Nov. 2010).

28. Xie, J. *et al.* Mechanistic insights into the roles of three linked single-stranded template binding residues of MMLV reverse transcriptase in misincorporation and mispair extension fidelity of DNA synthesis. *Gene* **479,** 47–56 (June 2011).

29. Zhulidov, P. A. *et al.* Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* **32,** e37–e37 (Jan. 2004).

30. Bogdanov, E. A. *et al.* Normalizing cDNA libraries. *Curr. Protoc. Mol. Biol.* **Chapter 5,** Unit 5.12.1–27 (Apr. 2010).