Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2

STEVEN BUSAN and KEVIN M. WEEKS

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA

ABSTRACT

Mutational profiling (MaP) enables detection of sites of chemical modification in RNA as sequence changes during reverse transcription (RT), subsequently read out by massively parallel sequencing. We introduce ShapeMapper 2, which integrates careful handling of all classes of adduct-induced sequence changes, sequence variant correction, basecall quality filters, and quality-control warnings to now identify RNA adduct sites as accurately as achieved by careful manual analysis of electrophoresis data, the prior highest-accuracy standard. MaP and ShapeMapper 2 provide a robust, experimentally concise, and accurate approach for reading out nucleic acid chemical probing experiments.

Keywords: SHAPE; RING; 1M7; NMIA; 1M6; 5NIA; NAI; dimethyl sulfate; single molecule; correlated chemical probing; mutational profiling; RNA structure modeling

INTRODUCTION

The ability to detect chemical adducts in RNA is the foundation for powerful technologies that enable analysis of RNA secondary and tertiary structure (Ehresmann et al. 1987; Mortimer and Weeks 2007; Tijerina et al. 2007; Weeks 2010, 2015) and detection of some epigenetic modifications (Behm-Ansmant et al. 2011). Several RNA structure probing technologies, including those using SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) and dimethyl sulfate, have been implemented in high-throughput formats (Kwok et al. 2013; Incarnato et al. 2014; Loughrey et al. 2014; Rouskin et al. 2014; Siegfried et al. 2014; Talkish et al. 2014; Poulsen et al. 2015; Smola et al. 2015b; Spitale et al. 2015). Most methods for reading out the results of RNA chemical probing experiments use primer extension–truncation and adapter-ligation to create libraries for analysis by massively parallel sequencing. These strategies are experimentally demanding to implement and can result in significant biases (Jackson et al. 2014; Fuchs et al. 2015). Truncation–adapter-ligation based strategies often fail to recover structural information with the accuracy intrinsic to the original chemical probing or epigenetic modification detection experiment (Kwok et al. 2013; Smola et al. 2015a; Weeks 2015). The shortcomings of truncation–adapter-ligation approaches are often accepted as an intrinsic cost of high-throughput readouts and are taken to be an acceptable trade-

off in which collecting an extensive catalog of data compensates for the low quantitative accuracy of some individual RNA reactivity measurements.

Mutational profiling (MaP) takes a different strategy. Under specialized conditions, some reverse transcriptase enzymes will extend cDNA synthesis through the site of a nucleotide containing a chemical modification on the base or ribose backbone (Fig. 1A), recording the site of the chemical adduct as a variation relative to the sequence complementary to the RNA being copied (Fig. 1C,D). MaP thus records the site of a chemical adduct directly at an internal position in the cDNA. These DNAs can be amplified using methods that ultimately introduce little bias relative to a no-modification control (Siegfried et al. 2014; Smola et al. 2015a,b), and sequenced using massively parallel methods (Fig. 1A). Most users also find that MaP is extremely straightforward to implement. For some chemical probes, MaP also appears to be more sensitive to low-level modifications than truncation–adapter-ligation methods (Krokhutin et al. 2017). The MaP step can be implemented to enable analysis of both short and long RNAs, of entire transcriptomes, and of rare transcripts in complex transcriptomes (Smola et al. 2015b, 2016). Because multiple chemical adducts can be read out in a single sequencing read (Fig. 1A), MaP can detect correlated chemical

Corresponding author: weeks@unc.edu

Article is online at <http://www.rnajournal.org/cgi/doi/10.1261/rna.061945.117>.

© 2018 Busan and Weeks This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

FIGURE 1. MaP experiment and analysis overview. (A) Quantification of chemical probing reactivities by MaP, based on massively parallel sequencing. (B) Algorithmic steps implemented in ShapeMapper. (C) Types of observed mutations and their frequencies in MaP-based analysis of *E. coli* 16S and 23S ribosomal RNA data sets collected previously under protein-free conditions using the 1M7 SHAPE reagent (Deigan et al. 2009; Siegfried et al. 2014). (D) Examples of simple and complex mutations detected in reads from the *E. coli* rRNA data set.

modifications in the same RNA strand. This feature makes it possible to examine through-space interactions (RNA interaction groups or RINGs) and corresponds to a single molecule experiment read out by sequencing: the RING-MaP experiment (Homan et al. 2014; Larman et al. 2017). MaP can also be used to detect sites of certain epigenetic modifications in RNA.

In the MaP strategy, chemical adducts are inferred through the location of (often multiple) mutations in sequence reads, and extracting this information from a MaP experiment presents unique analysis challenges. These challenges include

(i) accounting for diverse classes of sequence variations introduced during reverse transcription, (ii) correctly inferring chemical modification sites from ambiguously aligned mutations, and (iii) accounting for mutations that result from sequencing errors rather than the chemical probing experiment.

SHAPEMAPPER 2 ACCURACY

Here we introduce ShapeMapper 2 for analysis of MaP data (Fig. 1B) to achieve high levels of accuracy, usability, and

empirical performance. ShapeMapper 2 correctly detects and makes comprehensive use of all types of mutations generated during reverse transcription including mismatches, simple and complex deletions and insertions, and complex sequence changes (Fig. 1C,D). All types of mutations contribute positively to the recovery of base-pairing information, and the highest accuracy is obtained by including all mutation types (Fig. 2). ShapeMapper 2 handles multinucleotide mutations with an empirically optimized separation threshold (Supplemental Fig. S1) and interprets ambiguously aligned mutations that result from partial dissociation and reannealing of cDNA and RNA during reverse transcription (Supplemental Fig. S2). ShapeMapper 2 achieves high read coverage without sacrificing mutation rate accuracy by applying windowed read quality trimming and a post-alignment basecall quality filter on mutation counts and effective read depths (Supplemental Fig. S3).

Combined, these new features mean that ShapeMapper 2 calculates reactivity profiles that are often more accurate than those generated by the prior version of ShapeMapper and that are as accurate or are more accurate than those produced by careful manual analysis of capillary electropherogram data (Deigan et al. 2009), the prior highest-accuracy standard. For short RNAs, SHAPE-MaP data sets analyzed by ShapeMapper 2 recover information about base-pairing with an accuracy comparable to manually curated electrophoresis SHAPE (Supplemental Fig. S4). For long RNAs, randomly primed SHAPE-MaP is more accurate than manual electrophoresis analyses using multiple primers (Fig. 3A,B), and enables RNA secondary structure modeling with comparable accuracy (Fig. 3C).

SHAPEMAPPER 2 USABILITY

Although the major goal of ShapeMapper 2 was primarily to achieve high accuracy in RNA modification detection, the software also implements extensive new usability features, runs roughly twice as fast as prior draft software, and uses less than 1% of the disk space. ShapeMapper emphasizes straightforward command-line execution and arguments for simple use cases and flexibility for varied experiments and data formats, such as experiments with multiple RNA targets, multiple sets of input files, compressed input files, and regions of masked sequence. STAR aligner is supported as an alternative to Bowtie2 for improved speed with long target sequences (Langmead and Salzberg 2012; Dobin et al. 2013). ShapeMapper also automatically detects sequence variants in input target sequences and makes corrections to these sequences (see Materials and Methods). This feature is especially useful for performing MaP on incompletely characterized RNAs or RNAs subject to moderate levels of mutation or evolution.

ShapeMapper 2 calculates and plots per-nucleotide estimates for standard errors in SHAPE reactivities and histograms of sequencing depths and mutation rates, which are highly useful for troubleshooting and determining data and experiment quality (Supplemental Fig. S5). In addition, ShapeMapper 2 performs quality-control checks (see Materials and Methods) and integrates these into an overall PASS/FAIL message for the user. These checks are necessarily heuristic, since downstream analyses require different levels of data quality and since individual RNAs have different overall signal levels above background as a function of their extent

FIGURE 2. Importance of counting all mutation types. Receiver operating characteristic (ROC) curves for SHAPE-MaP reactivity profiles calculated using all mutation types or only certain types from the *E. coli* ribosomal RNA data set. SHAPE reactivity profiles were evaluated against reported Watson–Crick base-pairing interactions identified from crystal structures (Bernier et al. 2014). True positive rate: fraction of unpaired nucleotides with SHAPE reactivity above a given threshold; false positive rate: fraction of paired nucleotides with SHAPE reactivity above a given threshold. True positive and false positive rates were evaluated at all possible SHAPE reactivity thresholds from the lowest value in the data set to the highest. Inserts are far less frequent than other mutation types (see Fig. 1C), which accounts for low recovery of base-pairing information when analyzed alone.

FIGURE 3. Recovery of base-pairing information by the MaP experiment analyzed by ShapeMapper. (A) ROC curves for both subunits of the *E. coli* ribosome comparing accuracy of ShapeMapper 2 with ShapeMapper 1 and the prior high-accuracy standard, analysis by capillary electrophoresis. SHAPE-MaP data analyzed by ShapeMapper 2 gives both a higher true positive rate and a lower false positive rate than both ShapeMapper 1 and manually curated electrophoresis SHAPE for any given reactivity threshold, reflected by a statistically significant increase in area under the curve (*inset*). Capillary electrophoresis data were collected previously (Deigan et al. 2009; Siegfried et al. 2014). Shaded area shows the 95% confidence interval for the true positive rate calculated with 2000 bootstrap samples at 400 evenly spaced false positive rates. Error bars in the *inset* show a 95% confidence interval for the area under the curve calculated with 2000 bootstrap samples. (*) $P = 0.05$, (****) $P \leq 5 \times 10^{-12}$. (B) SHAPE reactivity data obtained by manually curated electrophoresis (*left*) and by SHAPE-MaP (*right*) superimposed on a representative region of the *E. coli* large ribosomal subunit RNA domain II. (C) Structure modeling accuracy using Superfold (Smola et al. 2015b) and RNAstructure (Reuter and Mathews 2010) with no SHAPE data or with SHAPE data from electrophoresis or MaP readouts as soft constraints, as previously described (Deigan et al. 2009). Models were evaluated against nonconflicting canonical Watson-Crick base-pairing interactions identified from crystal structures (Bernier et al. 2014), allowing base pairs offset by up to one nucleotide in either direction, and excluding base pairs separated by more than 600 nt in primary sequence. Sensitivity: the fraction of base pairs in the reference structure present in a modeled structure. PPV: positive predictive value, the fraction of modeled pairs present in the reference structure. These sensitivity and ppv values underestimate the likely true values by $\geq 5\%$, because regions where experimental SHAPE data are inconsistent with the reference structure have not been excluded (see Deigan et al. 2009).

of internal structure. In general, more sequencing read depth is always helpful, as are higher modification rates.

PERSPECTIVE

ShapeMapper 2 and MaP are a comprehensive solution to analysis of nucleic acid chemical modification data as read out by massively parallel sequencing. ShapeMapper 2 runs efficiently, yields reactivity profiles that are as accurate as highly validated low-throughput electrophoresis-based methods, and includes multiple features that facilitate application to diverse analysis problems. It is our intent that the availability of high-quality standardized software for MaP analysis will allow researchers to focus their efforts on science and discovery rather than bioinformatics pipelines. Quality-control checks and sequence-variant correction will encourage the use of well-designed MaP experiments and reduce the burden on nonexpert users, and standardized file formats will encourage data sharing between groups. We hope that the successful development of MaP technologies and ShapeMapper 2 will inspire additional easily implemented approaches enabling routine structural analyses of complex transcriptomes, using massively parallel sequencing approaches, that are as accurate as highly curated and focused studies of RNA model systems.

MATERIALS AND METHODS

ShapeMapper 2 implementation

The core components of ShapeMapper 2 have been rewritten using C++11 and modern libraries including Boost. Open-source third-party components can be installed manually or automatically downloaded in pre-compiled binary form using the Conda package manager (<https://conda.io/docs/>). A Python3.5 framework controls execution of individual components, handles the locations of their outputs, and allows parallelization through the use of named pipes for passing intermediate data, inspired by existing workflow software (Berthold et al. 2008). These design elements in ShapeMapper 2 yielded substantial speed gains and reductions in hard drive usage. For the *E. coli* ribosomal RNA data set, ShapeMapper 2 ran 40% faster and used less than 1% of the disk space compared to draft software (Smola et al. 2015b). The addition of unit and end-to-end tests ensures that ShapeMapper 2 produces the expected outputs and will do so through continued development.

Documentation

Software documentation is packaged with ShapeMapper and includes overall installation and execution instructions. Also included are file format descriptions, argument descriptions for component executables, and detailed explanations of quality-control checks. In-source documentation is provided for high-level Python module source code, and browseable documentation in HTML format is provided for low-level C++ components.

Data quality-control checks

The following quality-control checks are automatically implemented in ShapeMapper 2: (i) read-depth check, at least 80% of nucleotides meet a minimum sequencing depth of 5000 in all samples; (ii) positive mutation rates above background check, at least 50% of good-depth nucleotides have a higher mutation rate in the SHAPE-modified sample than in the untreated sample; (iii) high background mutation rates check, no more than 5% of good-depth nucleotides have an untreated mutation rate above 0.05 (an unusually high number of high-background nucleotides can indicate the presence of native modifications, sequence variants, or instrument run failure); and (iv) number of highly reactive nucleotides check, at least 8% of good-depth nucleotides have a modified mutation rate above 0.006 after background subtraction. Failure to pass these checks indicates close user scrutiny is merited.

Sequence variant correction

Small sequence changes are often present in studied RNAs when compared to expected target sequences. ShapeMapper 2 provides an optional preliminary stage that aligns reads to target sequences, identifies mutations occurring with above 60% frequency, and generates corrected target sequences including all identified sequence changes. This is appropriate for many situations in which polymorphisms are present within a single major RNA species, but is insufficient for mixtures of very similar RNAs. ShapeMapper 2 attempts to warn the user of the presence of conflicting or subthreshold variants. In these cases, more focused sequence characterization experiments and sequence assembly with other software may be required.

Choice of sequence aligner

ShapeMapper supports both Bowtie2 and STAR software for sequence alignment stages (Langmead and Salzberg 2012; Dobin et al. 2013). Read mapping percentages are typically comparable between the two aligners, and calculated reactivity profiles are virtually identical (Supplemental Fig. S6A). For long RNA targets, STAR is much faster than Bowtie2 (about three times as fast for the *E. coli* ribosomal RNA data set, and even faster for longer target sequences). However, the performance of STAR degrades when faced with reads from RNAs not present in reference sequences. Therefore, we do not recommend its use for experiments involving mixtures of unknown RNAs, unless directed gene-specific RT-PCR is performed to enrich for desired targets.

Use of a denatured control

Obtaining a denatured control (Siegfried et al. 2014; Smola et al. 2015b) for a MaP experiment can be challenging (and in some cases infeasible), uses valuable sequencing bandwidth, and can even hurt calculated reactivity profile accuracy if RNAs are degraded or over-amplified. For these reasons, ShapeMapper 2 does not require the use of a denatured control. Most background mutations are accounted for using mutation rates from a no-reagent control, but when the very highest accuracy is desired, a denatured control can provide an approximate mutation detection rate correction that improves recovery of base-pairing information (Supplemental Fig. S6B).

DATA DEPOSITION

ShapeMapper 2 and the *E. coli* ribosomal RNA data set used here are available from the corresponding author's website, www.chem.unc.edu/rna.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by a sponsored research agreement from Moderna Therapeutics and by the National Institutes of Health, Office of Extramural Research (AI068462 to K.M.W.). We are indebted to David Mauger and Iain McFadyen for continuous feedback on the software. We thank Nate Siegfried for collecting SHAPE-MaP data for the RNAs in Supplemental Figure S4.

Received April 28, 2017; accepted November 5, 2017.

REFERENCES

- Behm-Ansmant I, Helm M, Motorin Y. 2011. Use of specific chemical reagents for detection of modified nucleotides in RNA. *J Nucleic Acids* **2011**: 408053.
- Bernier C, Petrov AS, Waterbury C, Jett J, Li F, Freil LE, Xiong B, Wang L, Le A, Milhouse BL, et al. 2014. RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss* **169**: 195–207.
- Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinel T, Ohl P, Sieb C, Thiel K, Wiswedel B. 2008. KNIME: the Konstanz Information Miner. In *Data analysis, machine learning and applications: proceedings of the 31st annual conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*, pp. 319–326. Springer, Berlin, Germany.
- Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106**: 97–102.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.
- Ehresmann C, Baudin F, Mougél M, Romby P, Ebel J-P, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res* **15**: 9109–9128.
- Fuchs RT, Sun Z, Zhuang F, Robb GB. 2015. Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PLoS One* **10**: e0126049.
- Homan PJ, Favorov OV, Lavender CA, Kursun O, Ge X, Busan S, Dokholyan NV, Weeks KM. 2014. Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci* **111**: 13858–13863.
- Incarnato D, Neri F, Anselmi F, Oliviero S. 2014. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol* **15**: 491.
- Jackson TJ, Spriggs RV, Burgoyne NJ, Jones C, Willis AE. 2014. Evaluating bias-reducing protocols for RNA sequencing library preparation. *BMC Genomics* **15**: 569.
- Krokhoutin A, Mustoe AM, Weeks KM, Dokholyan NV. 2017. Direct identification of base-paired RNA nucleotides by correlated chemical probing. *RNA* **23**: 6–13.
- Kwok CK, Ding Y, Tang Y, Assmann SM, Bevilacqua PC. 2013. Determination of in vivo RNA structure in low-abundance transcripts. *Nat Commun* **4**: 2971.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.
- Larman BC, Dethoff EA, Weeks KM. 2017. Packaged and free satellite tobacco mosaic virus (STMV) RNA genomes adopt distinct conformational states. *Biochemistry* **56**: 2175–2183.
- Loughrey D, Watters KE, Settle AH, Lucks JB. 2014. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res* **42**: e165.
- Mortimer SA, Weeks KM. 2007. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc* **129**: 4144–4145.
- Poulsen LD, Kielpinski LJ, Salama SR, Krogh A, Vinther J. 2015. SHAPE selection (SHAPE-S) enrich for RNA structure signal in SHAPE sequencing-based probing data. *RNA* **21**: 1042–1052.
- Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**: 129.
- Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**: 701–705.
- Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**: 959–965.
- Smola MJ, Calabrese JM, Weeks KM. 2015a. Detection of RNA–protein interactions in living cells with SHAPE. *Biochemistry* **54**: 6867–6875.
- Smola MJ, Rice GM, Busan S, Siegfried NA, Weeks KM. 2015b. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc* **10**: 1643–1669.
- Smola MJ, Christy TW, Inoue K, Nicholson CO, Friedersdorf M, Keene JD, Lee DM, Calabrese JM, Weeks KM. 2016. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc Natl Acad Sci* **113**: 10322–10327.
- Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, Kuchelmeister HY, Batista PJ, Torre EA, Kool ET, et al. 2015. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**: 486–490.
- Talkish J, May G, Lin Y, Woolford LJ, McManus CJ. 2014. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20**: 713–720.
- Tijerina P, Mohr S, Russell R. 2007. DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* **2**: 2608–2623.
- Weeks KM. 2010. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol* **20**: 295–304.
- Weeks KM. 2015. Toward all RNA structures, concisely. *Biopolymers* **103**: 438–448.

RNA

A PUBLICATION OF THE RNA SOCIETY

Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2

Steven Busan and Kevin M. Weeks

RNA 2018 24: 143-148 originally published online November 7, 2017

Access the most recent version at doi:[10.1261/rna.061945.117](https://doi.org/10.1261/rna.061945.117)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2017/11/07/rna.061945.117.DC1>

References

This article cites 27 articles, 6 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/24/2/143.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>