Big picture stuff

Reviewers:

Important correspondence:

1.) We don't want to mess with published data

Brent: What melt data should we use? My main thought here is that I don't want to draw negative attention to any of our previous melt data...where people begin to question the data we and others have already published using Meltwin. I'm not sure how to handle this. Some possibilities would include (1) using unpublished data but not provide sequences (maybe just refer to them as Duplex 1, not provide raw data, etc.) so we could publish later in a thermo paper (not sure if that's legit), (2) using previously published data if the free energy values are very similar and we're just focusing on smaller error (again...not drawing too much negative attention to previously published work), or (3) collect new data that will not be part of another project (which will take more time and have some cost associated with it). I can see how a direct comparison to Meltwin could be a good thing for MeltR, but I'd hate to do that at the expense of all of the data previously published using Meltwin. This is my major concern. Is there a way to tout the data analysis of MeltR without bringing down Meltwin? Thoughts?

Phil: I think we can say only the nicest things about Meltwin. Nothing disparaging at all. At the same time, we can talk about reducing error and improving accuracy going forward. We would be very careful to say that prior data is still valuable.

Jacob: I agree with Phil. We say, "Meltwin is great. It enabled a generation of nearest neighbor parameters and all the tools that rely on them. Nobody can get Meltwin anymore and its a black box." We can leave it at that.

On not drawing negative attention to published data, this is important. Any differences between the analysis below and the published analysis are do to idiosyncrasies. I want to proceed with the data we have. However, this is ultimately, not our data. If we are nearing a complete draft and all authors are not comfortable, I will collect a new data set to use for this paper.

2.) How do justify baseline trimming versus?

Brent: If a lot of the focus is on the baseline trimming tool, we need to be able to convince people (who think like Phil) that there are good reasons to trim.

Phil: I think we should allow different options for trimming but my philosophy is the baselines as long as possible but not longer option. This might do something recursive like look for the first five data points away from the transition that fit y=mx+b with random residuals and then lengthen 1 point at a time, stopping when the residual lose randomness.

Jacob: There are good reasons to trim. First of all, everyone does it. Second, linear baselines are an approximation and this approximation gets less valid the longer the baseline is. The question is "how to trim?" and everyone has to make their own idiosyncratic rules, which are always hard to justify because every reviewer can say "what about this?". My feeling is that people would be happy for us to take this out of their hands if we present a sufficiently clever method.

I really liked Phil's idea for the recursive residual analysis. Unfortunately, it does not work. (1) The baselines of ideal melting curves are never perfectly linear because it is a sigmoid. (2) On real data, this strategy just starts grabbing random parts of the melting curve.

3.) Meltwin exists and is trusted, how do we justify MeltR

Brent: I'm not sure what you had in mind for the paper. At first, I was thinking about all of the comparisons that could be made to Meltwin. But now, the more I think about it, I'm not sure that is necessary. Not many people use Meltwin, so a point-by-point comparison is probably not needed, as MeltR would be the only software available for this purpose (unless you know of others). But, I could work with Sebastian to come up with a list of limitations related to Meltwin that aren't related to the data analysis (not available for distribution, some features don't work on newer operating systems, etc.) that could be discussed in the paper.

Phil: This is more tricky. Again, I think we should cherish Meltwin and the many important papers published with it.

Jacob: I want a direct comparison because it will make Meltwin users feel safe and I want to use the word "Meltwin" a lot because it will make our paper pop up on google when somebody does a melting curve, reads a Znosko paper, and starts looking for a copy of Meltwin.

SI to do list

- Supplemental_file_1.pdf: Supplemental figures and tables
- Supplemental_file_2.pdf: meltR.A help documentation
- Supplemental_file_3.pdf: BLtrimmer help documentation

Manual to do list

- Theory section for the baseline trimmer

Coding to do list

- Add an easy outlier exclusion protocol to meltR.A
- Generate a library of modeled absorbance data and fit it using the BLtrimmer to make sure it works
- Add a clearer pre-canned output to the BLtrimmer for the two-state folding test.

Facile Determination of RNA Folding Energies from Absorbance Data Using MeltR

Jacob P. Sieg[1,2], Sebastian Arteaga, Brent Znosko, Philip C. Bevilacqua[1,2,3]

[1]Department of Chemistry, Pennsylvania State University, University Park, PA 16802.

[2]Center for RNA Molecular Biology, Pennsylvania State University, University Park, PA 16802.

[3]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802.

Keywords: Absorbance melting curves, folding energies, RNA, DNA

Abstract

UV-detected absorbance melting curves of RNA are used to determine helix folding energies, providing the thermodynamic insight into RNA secondary structure, which underlies a plethora of structure prediction tools. Appropriate analysis of absorbance melting curves is not trivial, requiring many data cleaning, regression, and error analysis steps. The absorbance melting curve-fitting software *Meltwin*, originally introduced in 1996, provided researchers with the consistent and facile absorbance melting curve-analysis platform that underlies a generation of RNA folding parameters. Unfortunately, *Meltwin* software is no longer maintained. Herein, we provide *MeltR*, a modern curve-fitting package in the popular R statistical programming language. In particular, the *meltR.A* function provides the same facile conversion of absorbance data to folding energies provided by *Meltwin*, with additional useful features for modern experiments. Likewise, the *BLtrimmer* function provides a consistent protocol that improves folding energy estimation and uncertainty analysis, based on analysis of an ensemble of trimmed baselines. We believe that *MeltR* will be a useful tool for analyzing another generation of absorbance melting curve experiments.

Introduction

Paragraph 1: Absorbance melting curves are important

Paragraph 2: Fitting is tedious and Meltwin is nice

Paragraph 3: Meltwin was used in a generation of research that did amazing things but it's gone now.

Paragraph 4: Present MeltR

Results

Facile fitting of raw absorbance data with meltR.A

Paragraph 1: MeltR is an R package and its advantages

Paragraph 2: meltR.A usage (Figure 1A)

Paragraph 3: meltR.A output (Figure 1B)

Paragraph 4: Describe precanned figures (SI figure 1)

Paragraph 4: Other useful features of meltR.A (SI file 2-help file)

Fit parameters are sensitive to baseline trimming

Paragraph 1: Baseline trimming is necessary but forces users to make idiosyncratic choices

Paragraph 2: Dependence of fit parameters on baseline length for modeled data (Figure 2A)

Paragraph 3: Dependence of fit parameters on baseline length for ideal, real data (Figure 2B)

Paragraph 3: Dependence of fit parameters on baseline length for noisy, real data (Figure 2C)

Paragraph 4: MeltR provides a auto-base line trimmer that works on three principles

(1) Baselines should be between 5 and 25 °C (minimizes violation of linear assumption while still accurate)

(2) The best baselines are produce the most internally consistent thermodynamic parameters (Deals with baseline noise)

(3) Beyond (1) and (2), optimum baseline combinations should be treated as an essemble (converts a systematic error, error in trimming choices, that cannot be treated statistically, into a random error which can be treated statistically).

The BLtrimmer treats fit results from an ensemble of baseline combinations statistically

Paragraph 1: Baseline trimmer usage (Figure 2A)

Paragraph 2: This section and Figure 3 may be more appropriate in the methods.

Paragraph 3: Other useful BLtrimmer features (SI file 3, help file)

MeltR reproduces thermodynamic parameters from Meltwin

Paragraph 1: meltR.A and the BLtrimmer work on modeled data (SI figure 3).

Paragraph 2: Compare method 1, 2, and 3 for MeltR (SI table 1). Describe method 1 MeltR versus Method 1 Meltwin (SI table 1) and method 2 MeltR versus Method 2 Meltwin (SI table 2)

Paragraph 3: Describe method 3 MeltR versus Method 1 Meltwin and method 2 Meltwin (SI table 3, Figure 4)

The BLtrimmer provides a two-state folding test

Paragraph 1: Two-state folding test based on analysis of the distribution of randomly chosen baselines

Paragraph 4: Why the BLtrimmer does not force a two state interpretation on non-two state RNA.

Discussion

Materials and Methods

Data transparency and reproducibility

Absorbance melting curves

Paragraph 1: How Brent and Sebastian compiled Adams data

Paragraph 2: Aquisition in the Bevilacqua lab

Fitting absorbance melting curves

This will be copied and cleaned up from the theory section of the MeltR manual

BLtrimmer

This will be copied and cleaned up from the theory section of the MeltR manual

Definitions

R specific definitions will go here.

Acknowledgments

References

Tables

Table 1 Regression equations used to determine folding energies in meltR.A.

Figures and legends

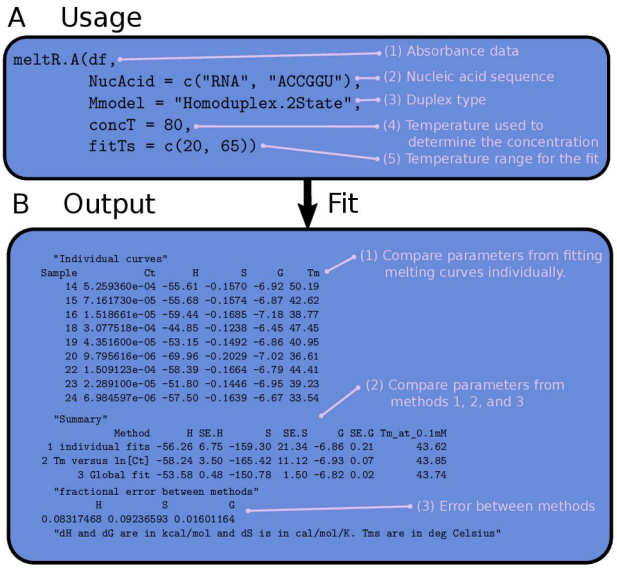
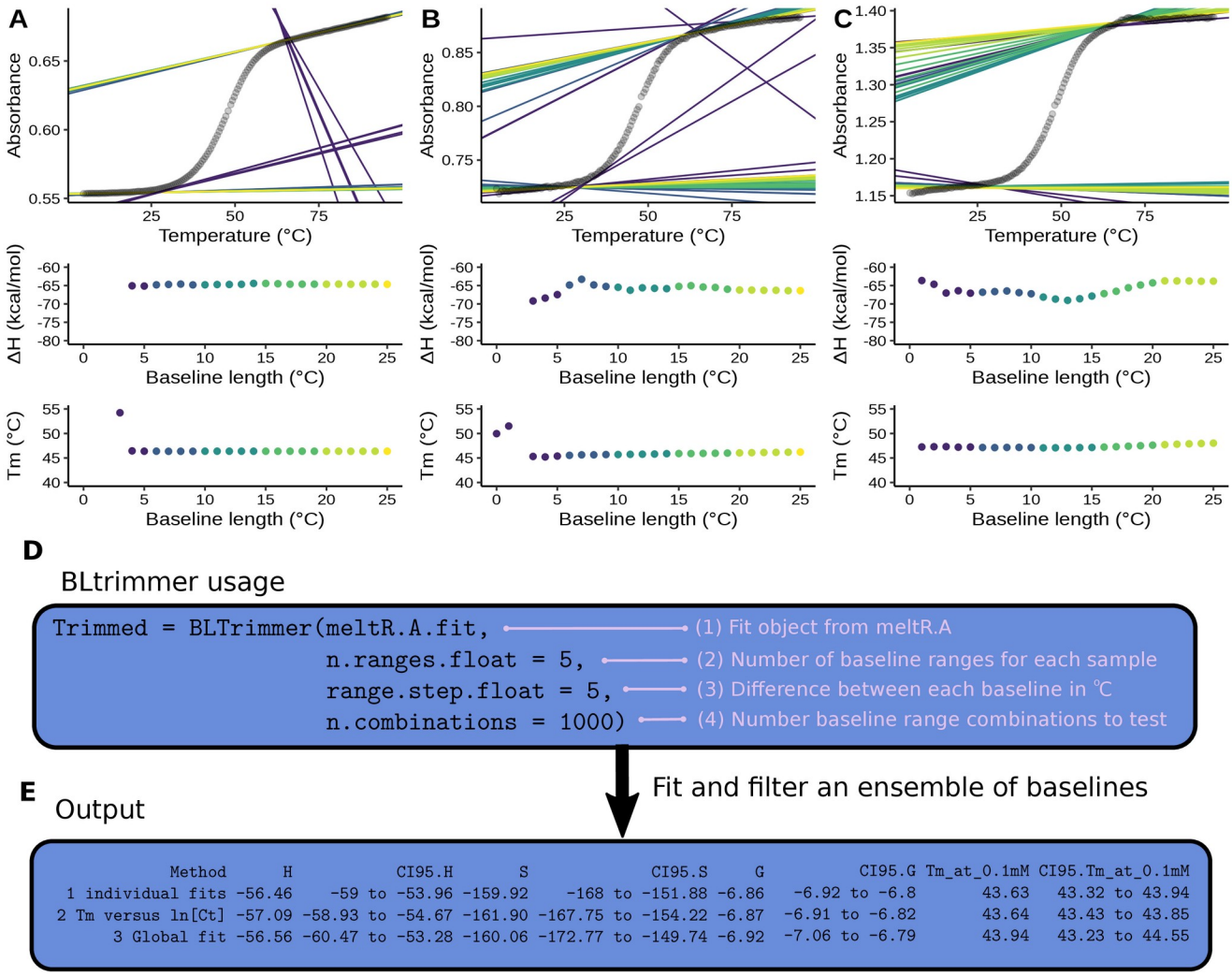


Figure 1 The usage and output of meltR.A in the MeltR package. **(A)** The usage of meltR.A in an R script. (1) The data frame (df) containing the absorbance data, created by reading a flat text file. (2) A vector containing the nucleic acid type and sequence, used to calculate extinction coefficients. Alternatively, custom extinction coefficients can be provided (Supplemental file 2). (3) Duplex type, either “"Monomolecular.2State", "Heteroduplex.2State", or "Homoduplex.2State". (4) The temperature used to calculate the RNA concentration using Beer’s law. (3) The temperature range that is fit. Used for manual baseline trimming. Either a single range for the whole data set, or a list of ranges, one for each sample. **(B)** The output of meltR.A in an R console. (1) Thermodynamic parameters generated by fitting each curve in a data set individually. (2) Comparison of the results for the different methods that meltR.a used to determine folding energies. Method 1 averaged the results from fitting each melting curve individually. Method 2 fit the Van’t Hoff relationship between the melting temperature and the RNA strand concentration. Method 3 combined Method 1 & 2, by using the Van’t Hoff equation to globally fit the entire data set to one regression model. (3) Fractional error between methods used for a test of two-state folding. Non two-state folding generates a fractional error between methods that is greater than 0.15.



Method	H	CI95.H	S	CI95.S	G	CI95.G	Tm_at_0.1mM	CI95.Tm_at_0.1mM
1 individual fits	-56.46	-59 to -53.96	-159.92	-168 to -151.88	-6.86	-6.92 to -6.8	43.63	43.32 to 43.94
2 Tm versus ln[Ct]	-57.09	-58.93 to -54.67	-161.90	-167.75 to -154.22	-6.87	-6.91 to -6.82	43.64	43.43 to 43.85
3 Global fit	-56.56	-60.47 to -53.28	-160.06	-172.77 to -149.74	-6.92	-7.06 to -6.79	43.94	43.23 to 44.55

Figure 2 Auto-baseline trimming absorbance data using the BLtrimmer in MeltR. **(A)** Dependence of fit parameter accuracy on baseline length for fitting modeled absorbance data assuming a ΔH° of -64.76, a T_m of 46.4 °C, and a C_t of 8 μM. Lower and upper base lines start where 90% and 10% (molar basis), respectively, of the RNA strands are in the helical state. (Top panel) Linear baseline fits superimposed on raw absorbance data. Colors correspond to baseline length specified in middle and bottom panels. (Middle panel) Dependence of ΔH° on baseline length. (Bottom panel) Dependence of T_m on baseline length. **(B)** Dependence of fit parameter accuracy on baseline length for fitting a nearly-ideal, real absorbance melting curve collected on a CGAAAGGU/ACCUUUCG at a C_t of 8 μM.

(C) Dependence of fit parameter accuracy on baseline length for fitting a non-ideal, real absorbance melting curve collected on a CGAAAGGU/ACCUUUCG at a C_t of 12 μM. **(D)** BLtrimmer usage in a R script. (1) A MeltR fit object produced by fitting raw data with meltR.A. (2) The number of baseline ranges the BLtrimmer will produce for each sample. (3) The temperature difference for each baseline produced on an absorbance melting curve. (4) The number of baseline combinations to test in a given BLtrimmer run. **(E)** Output of the BLtrimmer in the R console. Data were fit using 1000 baseline combinations and then a subset of the most internally consistent combinations were used to estimate folding energies and 95% confidence intervals.

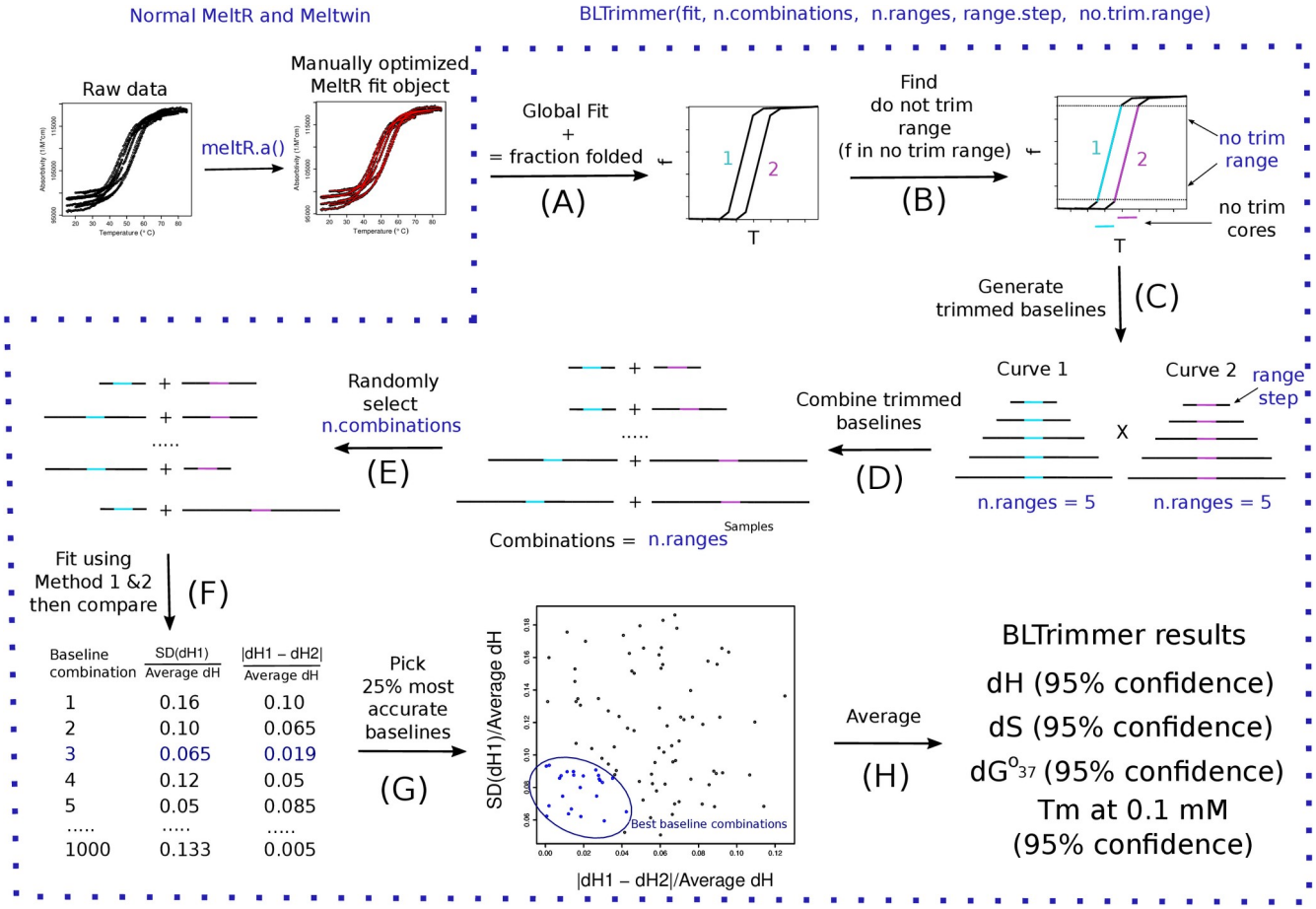


Figure 3 The MeltR BLtrimmer applies an ensemble analysis to calculate optimum helix association energies for an absorbance data set. **(A)** A MeltR fit object, produced by fitting raw absorbance data with meltR.A, is used to calculate the fraction of strands in the helical state (*f*) as a function of temperature. **(B)** The “no trim” temperature range is identified where *f* is greater than 0.1 and less than

0.9. **(C)** Baselines are added to the now trim range in 5 °C steps. **(D)** Ranges from each sample are exhaustively permuted to generate n.ranges^{Samples} baselines. **(E)** Thermodynamic parameters are determined from a subset of the baseline combinations using method 1 (average of fitting samples individually) and method 2 (Van't Hoff analysis). Baseline combinations are assessed for enthalpy agreement for method 1 and enthalpy agreement between method 1 and method 2 to identify the baseline combinations that have the highest internal consistency. **(G)** The top 25% most internally consistent baseline combinations are selected for ensemble analysis. **(H)** The ensemble of internally consistent baseline combinations is passed to meltR.A and fit. The results are treated statistically to determine folding energies and 95% confidence intervals.

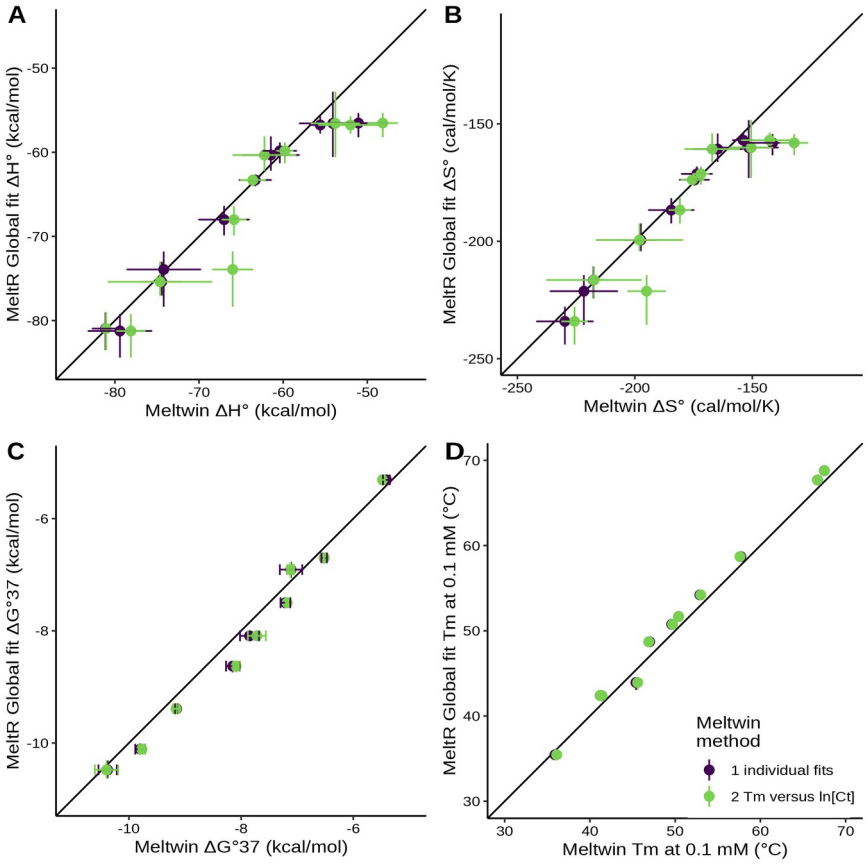


Figure 4 MeltR functions meltR.A and BLtrimmer reproduce folding energies calculated using Meltwin with minimal user input. **(A)** Helix association ΔH° determined from the method 3, global fitting, using

meltR.A followed by the BLtrimmer, versus the ΔH° determined from method 1 (purple) and method 2 (green) using Meltwin to fit the same absorbance data set. Horizontal error bars represent precision in the parameters calculated by Meltwin and vertical error bars represent 95% confidence intervals calculated using the BLtrimmer. **(B)** Helix association ΔS° determined from the method 3, global fitting, using meltR.A followed by the BLtrimmer, versus the ΔS° determined with Meltwin. Colors and error bars are the same as A. **(C)** Helix association $\Delta G°_{37}$ determined from the method 3, global fitting, using meltR.A followed by the BLtrimmer, versus the $\Delta G°_{37}$ determined with Meltwin. Colors and error bars are the same as A. **(D)** T_m at 0.1 mM determined from the method 3, global fitting, using meltR.A followed by the BLtrimmer, versus the T_m at 0.1 mM determined with Meltwin. Colors and error bars are the same as A.

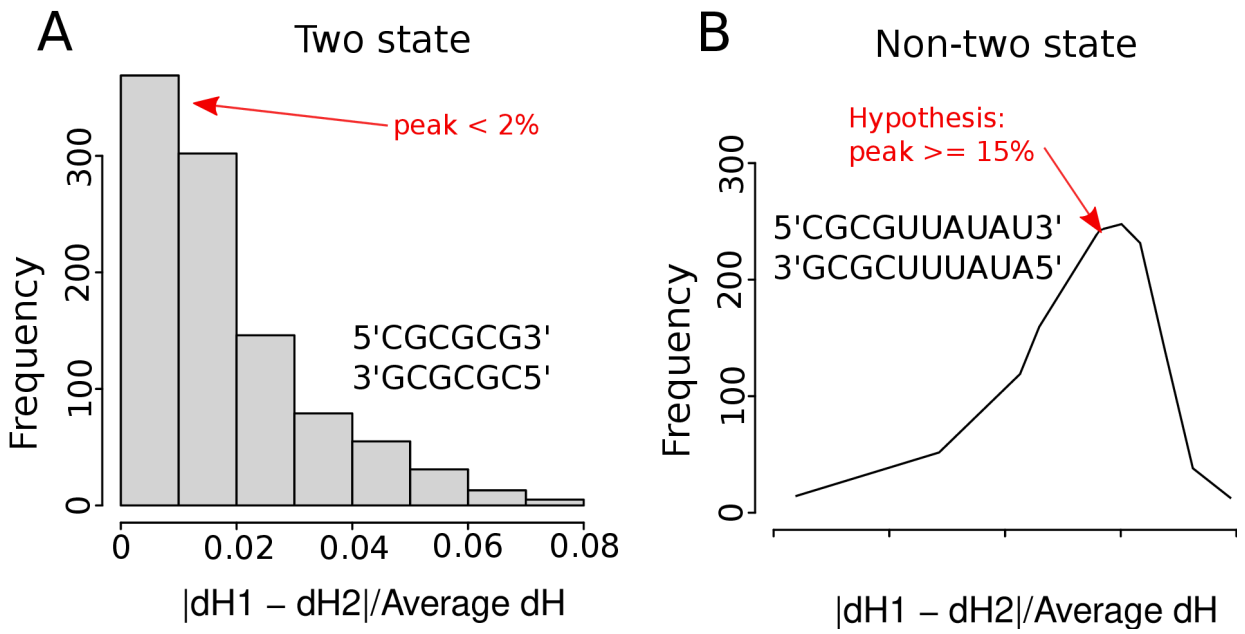


Figure 5 Two-state folding test using the BLtrimmer in MeltR. **(A)** The distribution of the error between the ΔH° using Method 1 (average of fitting samples individually) and Method 2 (Van't Hoff analysis) using 1000 randomly determined baseline combinations peaked at 2% for two-state folding sequences. **(B)** The distribution of the error between the ΔH° using Method 1 (average of fitting samples individually) and Method 2 (Van't Hoff analysis) using 1000 randomly determined baseline combinations peaked at >15% for non-two-state folding sequences.