SEQualyzer: **S**tructure-profiling **E**xperiment **Qua**lity Ana**lyzer**

September 25, 2016

Choudhary, K., Ruan, L., Deng, F., Shih, N., Aviran, S., SEQualyzer: Interactive tool for quality control and exploratory analysis of high-throughput RNA structural profiling data. (submitted)

Contents

1 Introduction

The SEQualyzer (**S**tructure-profiling **E**xperiment **Qua**lity Ana**lyzer**) platform allows users to visualize and quality control data from a wide range of structure profiling experiments. The user uploads a dataset containing read stop/mutation counts and coverage from sequencing-based profiling experiments at each residue for one or multiple transcripts. These counts are converted into reactivity scores using a reconstruction scheme chosen by the user. The quality and variability of these scores are then evaluated using several quantitative and visual metrics. Users can save all analyses to a local drive before exiting the application.

1.1 Features

SEQualyzer allows users to perform exploratory analysis of structure-profiling data in an easy and efficient manner. Available features include:

- Transcriptome-wide applicability for a variety of structure-profiling methods
- Complete flexibility to choose reactivity scoring scheme
- Fast bootstraps to get error bars and other quality measures
- Flexibility to zoom into regions of interest
- Single-click save of all quality measures
- Use sequence data to get more informative plots
- Filter for well-covered transcripts for closer inspection
- Compatibility with StructureFold platform
- Insights from residue-level to transcriptome-level
- Utilize parallel computing facilities in R to speed up analysis

1.2 Downloading SEQualyzer

A compressed folder containing SEQualyzer can be downloaded from the following link.

`http://bme.ucdavis.edu/aviranlab/sequalyzer`

It contains necessary R codes and application design files to run SEQualyzer. All codes must be present for the application to execute. Besides R code, the compressed folder contains SEQualyzer manual and example datasets. Files can be moved to user's desired location but it is recommended that contents of the compressed folder be kept at one place.

1.3 Quality control

Available quality control metrics include:

- Signal-to-noise (SNR) ratio from replicates
- Bootstrap to get SNR for individual replicates
- Window level SNR
- SNR distributions
- Pairwise SNR comparisons
- Correlation matrix for replicates
- Error bars for reactivity scores
- Coverage Quality Index

1.4 Other data summaries

SEQualyzer provides additional data summaries for exploring systematic biases, if any. These include:

- Plot of read termination/mutation counts per residue
- Plot of local coverage per residue for paired-end data
- Plot of reactivity score distribution
- Plot of detected modification counts at A, C, G and U separately, if sequence information is provided
- Plot of Lorenz curve to gauge degree of uniformity in coverage across transcriptome

The quality metrics and data summaries made available by SEQualyzer not only enable comparisons of datasets but can also guide future experiments.

2 Requirements and Installation

2.1 Hardware requirements

SEQualyzer is built in native R and is platform independent. It has no specific hardware requirements and should run on all laptop/desktop devices as long as software requirements are met.

2.2 Launch SEQualyzer

2.2.1 Software requirements

SEQualyzer requires R 3.2.2 or higher and is best run with RStudio (`https://www.rstudio.com/products/rstudio/download/`), an integrated development environment for R (`https://www.rstudio.com/`). While there is a paid, premium version of RStudio, the open source, free version of RStudio is sufficient. The application has been tested on MacOS X and Windows 7, and installers for these operating systems can be found at provided link.

Users must install R 3.2.2 or higher to use SEQualyzer. To install R visit `https://www.r-project.org`. To install RStudio visit `https://www.rstudio.com/products/rstudio/`. To upgrade R in RStudio, simply install the right version of R. Next, change the version of R in RStudio. To read how to do that, visit `https://support.rstudio.com/hc/en-us/articles/200486138-Using-Different-Versions-of-R`. Users must install shiny package in R. It can be installed by typing

```
install.packages(‘‘shiny")
```

and pressing enter in R Studio console. The installation needs to be done only once. For MacOS users, SEQualyzer will attempt to install *shiny* on its own.

Users should be able to run SEQualyzer even without RStudio, in which case SEQualyzer will launch in default web browser. SEQualyzer will attempt to install *shiny* and other packages in R. All R packages needed for SEQualyzer are automatically installed if found missing and loaded when the application is launched. These packages are- *ggplot2, reshape2, tools, corrplot, foreach, parallel, doParallel*, *plyr*, *seqinr*, *rmngb* , *zoo* and *ineq*. If users experience an error when attempting to launch SEQualyzer, these packages may need to be installed/loaded manually, using the command *install.packages()*. The packages can be installed by typing

```
install.packages(‘‘<insert_package_name>’’)
```

and pressing enter in R Studio console. The installation needs to be done only once. For best experience, it is recommended that a new session of SEQualyzer be launched for every new dataset.

2.2.2 General method (works on both Windows and Mac OS)

Launch RStudio and set working directory to the SEQualyzer folder. To launch SEQualyzer itself from RStudio, just click the button “Run App” on top right of RStudio source pane. Alternatively, it can be launched by using

```
shiny::runApp()
```

command. For other ways to launch SEQualyzer, including launching in a web browser such as Firefox, Chrome, Safari, etc. see `http://shiny.rstudio.com/reference/shiny/latest/runApp.html`.

2.2.3 Standalone version (Windows)

Download “SEQualyzer_windows_standalone.zip". Unzip the folder and double-click on “run.vbs" to launch SEQualyzer.

2.2.4 Deploying from desktop (Mac OS)

In the downloaded folder, users will find a file named "SEQualyzer.command" inside the "Code" subdirectory. Double-clicking this file will launch SEQualyzer. Users can make an *alias* for this file and move it to desktop or any desired location. The alias will launch SEQualyzer with a simple double-click. Users might face troubles using this method if the path to SEQualyzer has a white space. Additionally, it might be required to make the SEQualyzer.command file in "Code" folder an executable file. This can be done in Terminal by changing the current directory to "Code" folder and using the following command.

```
chmod +x SEQualyzer.command
```

2.3 Coupling SEQualyzer with other softwares

SEQualyzer can take output of StructureFold [1] platform as input. StructureFold takes FASTA/FASTQ files from structure-profiling experiments and outputs read stop counts. These files can be fed to SEQualyzer after renaming as described below. The files to input can also be obtained from custom softwares/programs and reformatted to be compatible with SEQualyzer using any programming language.

3 Input format

3.1 StructureFold format

SEQualyzer can take input as data for multiple transcripts and replicates thereof simultaneously. The direct file input to SEQualyzer is a file containing name (with file path) of the files with read counts and sequences for all transcripts. There is one file per experiment and another file for transcript nucleotide sequence, which is optional. For e.g., in case of an experiment with two replicates, there will be five files, one for the sequences and other four corresponding to two channels (*plus* and *minus*) each for two replicates. Words “sequence”, “minus”, “plus” and “rep#” (where “#” represents any number) are reserved keywords. They must be used only to describe the contents of the file. Example, “minus” in name indicates the file corresponds to minus channel of replicate #, which in turn is indicated by “rep#” in the name. All the keywords must be preceeded and followed by two underscores (“__”). For example, data for two replicates of yeast can be stored in five files named -

1. Sac_cer__minus__rep1__.txt (contains read counts for minus channel of replicate 1)
2. Sac_cer__minus__rep2__.txt (contains read counts for minus channel of replicate 2)
3. Sac_cer__plus__rep1__.txt (contains read counts for plus channel of replicate 1)
4. Sac_cer__plus__rep2__.txt (contains read counts for plus channel of replicate 2)
5. Sac_cer__sequence__.fasta

All these file names will be listed together in separate lines of a single text file, with any name, that will be input to SEQualyzer. All files must have the same order of transcripts.

The RNA sequence file is provided in a fasta format and is optional. The other files have read counts and/or local coverage information. There must be three lines per transcript in case that local coverage information is not available and four otherwise. The first line should be name of the transcript; second line should be tab-delimited read stop/mutation counts per nucleotide; third line (only in case local coverage information is available) should be tab-delimited local coverage and fourth (or third line in absence of local coverage information) should be blank. It goes without saying that the lines corresponding to counts must have as many entries as the length of the transcript.

The file format mentioned above is consistent with output from StructureFold pipeline [1]. StructureFold pipeline can be used for a variety of chemical-based or nuclease-based protocols and allows users to save count summaries for both plus and minus channels. Output from StructureFold can be directly fed to SEQualyzer after renaming the files appropriately and compiling file addresses in a single file as detailed above. For mutation based approaches, such as SHAPE-MaP [2], users can utilize ShapeMapper [2] for summarizing per residue mutation counts. ShapeMapper stores counts in a folder named *output/counted_mutations/*. However, users may have to slightly reformat the output from ShapeMapper. Users can input data from plus-reagent and minus-reagent experiments to SEQualyzer. Furthermore, most computational tools allow users to save summarized per-residue modification counts and a simple reformatting will render the outputs from such tools compatible with SEQualyzer. Additionally, adhering to these file formats should be easy if raw reads are processed and summarized using a custom pipeline.

3.2 RMDB format

SEQualyzer can take input in the RNA Mapping database format [3] which has the extension ".rdat". In this format, the first line specifies the RDAT version. After a line skip, the data annotations are given. Annotation names follow the same rules as names for StructureFold format. For example, data for two replicates of yeast can be annotated as named -

1. Sac_cer__minus__rep1__ (contains read counts for minus channel of replicate 1)
2. Sac_cer__minus__rep2__(contains read counts for minus channel of replicate 2)
3. Sac_cer__plus__rep1__ (contains read counts for plus channel of replicate 1)
4. Sac_cer__plus__rep2__ (contains read counts for plus channel of replicate 2)

The sequence information should be provided as a data lane with header “SEQUENCE". Local coverage data can be included as a data lane with header “LOC_COV”. For an example, refer to sample datasets. Additionally, we recommend that users of SNRNASM Isa-tab format use RMDB web repository to convert their files to RDAT format.

4 Interface layout

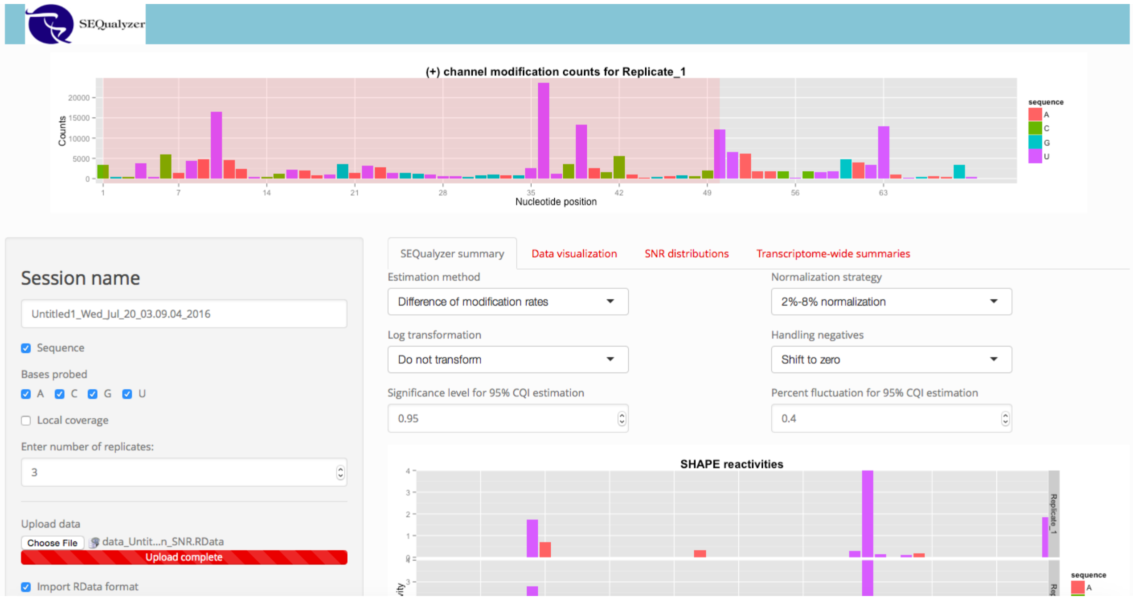
SEQualyzer interface contains a top panel, a side panel and a main panel. Contents of the main panel depend on choice of tab from the navigation bar.

4.1 Top panel

The top panel consists of a plot of the modification counts per residue in plus channel of the first replicate of the transcript selected from the side panel. The region highlighted in red represents the region of transcript chosen for examination in the main panel.

4.2 Side panel

The side panel consists of various options in the form of checkboxes, inputs to be chosen from drop-down menus, buttons for execution, options to save all analyses to a local drive. Together they facilitate data description, data selection and analysis choices.



Session name

Untitled1_Wed_Jul_20_03.09.04_2016

☑ Sequence

Bases probed

☑ A ☑ C ☑ G ☑ U

☐ Local coverage

Enter number of replicates:

3

Upload data

Choose File data_Untit...n_SNR.RData

Upload complete

☑ Import RData format

Estimation method: Difference of modification rates

Normalization strategy: 2%-8% normalization

Log transformation: Do not transform

Handling negatives: Shift to zero

Significance level for 95% CQI estimation: 0.95

Percent fluctuation for 95% CQI estimation: 0.4

Figure 1: SEQualyzer screenshot displaying the layout.

4.3 Navigation bar

The navigation bar contains four tabs-

1. *SEQualyzer summary* which allows user to check the most relevant quality control metrics,
2. *Data visualization* which allows user to check various secondary level summaries of data such as reactivity distribution, plots of residue level local coverages, modification counts and tally of modifications at A, C, G and U if sequence information is provided,
3. *SNR distributions*, which allows users to quickly examine the various forms in which we present SNR to reveal quality of data when tested on different criteria, and
4. *Transcriptome-wide summaries*, which provides transcriptome-wide insights and analysis options for dataset that presumably has information about a transcriptome.

4.4 Main panel

The main panel displays a set of results given the users' tab-selection. Details of plots contained in the main panel are given in Section 6.

5 Side panel layout and controls

The side panel is divided into three sections, one each for data description, data selection and data analysis. At the top is a field to give a name to current session. The input data will be restructured in R *list* format, transcriptome-wide summaries will be appended to the RData and saved in the "Saved_results" folder under current session name in RData format. This can be later loaded in R and used for miscellaneous purposes at a later time.

5.1 Data description

The experiments need to be described properly by choosing the appropriate probed bases, presence/absence of local coverage information and number of replicates for each transcript. Users must also indicate the presence or absence of sequence data in the dataset. The number of replicates must be the same for all the transcripts and sequence data if provided, must be present for all the transcripts.

5.2 Data selection

After the data file has been chosen by clicking the 'Choose file' button, pushing the 'Load' button will run a program that reads all files. If the file descriptions do not match file input or if the files are not properly formatted, the user is prompted with an error message. Selecting a new file automatically refreshes the results on display. So, all analysis results should be saved before uploading a new file. The 'Analyze' button needs to be pushed once for the first dataset uploaded in any session. For subsequent datasets, loading them will automatically initiate data processing.

Once plots and results have been rendered in main panel for study, users can switch between transcripts by selecting them by name from a drop-down menu. Options available in the drop-down menu can be filtered based on mean coverage, overall coverage or length of transcripts. For each transcript, users can zoom into regions of interest by selecting range of nucleotides from a slider tool. Thus, SEQualyzer allows users to look at the data from nucleotide-level resolution to whole transcriptome level.

5.3 Data processing choices

Pushing the 'Evaluate replicates separately' button initiates bootstrapping of the data (for chosen transcript in case of multiple transcripts) and produces additional plots to address various quality related questions. Bootstrapping for error estimation can be replaced with a routine using a theoretical formula, if local coverage information is provided.

5.4 Saving all analyses

All the analyses can be saved by clicking the 'Save analyses' button. It prompts the user for a download location. Given a valid download location, all the plots generated by SEQualyzer including reactivity scores with error bars in the form of a tab-delimited file are saved to user-specified location.

6 Main panel layout

All plots and summaries in the main panel are for the region of transcript chosen in the side panel.

6.1 SEQualyzer summary

On this tab, users can customize their reactivity scoring pipeline, and specify parameters to calculate CQI. SEQualyzer summaries include CQI report, plots of reactivity scrores with error bars (if available), SNR from replicate comparisons, correlation matrix for replicates, pairwise mean SNR matrix, individual replicate SNR with window SNR (if user checks the checkbox for including window SNR) and in addition, a summary of individual replicate SNR in the form of mean and median of SNR for all residues.

6.2 Data visualizations

This tab allows users to check various secondary level summaries of data such as reactivity distribution, plots of residue level local coverages, modification counts and tally of modifications at A, C, G and U if sequence information is provided. These can be useful to identify systematic biases in data, if any are present. These visualizations are available only if the experiment contains the relevant information. For example, local coverage information is not accurately captured by single-end data and so, the corresponding plot would be missing for single-end data.

6.3 SNR plots and distributions

In addition to SNR from replicate comparisons, pairwise mean SNR matrix and individual replicate SNR, this tab contains overlaid reactivity scores for direct comparison of replicate agreement, window SNR from replicates overlaid for comparisons, and SNR distributions for individual replicates. These plots and summaries highlight various quality aspects of the dataset.

6.4 Transcriptome-wide summaries

This tab contains plots revealing transcriptome-wide data characteristics. A plot of Lorenz curve describes the degree of uniformity of distribution of overall library among transcripts. A histogram of transcripts by mean local coverage indicates the distribution of per-nucleotide sequencing depth for the transcripts.

7 Definitions and interpretations

- Local coverage : Local coverage of a residue is the total number of reads that mapped to a site or in case of termination-based assays, the total number of reads that mapped to a site or terminated one residue downstream. Higher local coverages are desired for high precision and accuracy of reactivity estimation. For example, experiments with paired-end sequencing can be used to recover local coverage information.
- Overall coverage : Overall coverage of a transcript is the total number of reads that map to a transcript.
- Modification counts : Modification counts for a residue refer to the total number of reads that are indicative of noise or modification at the site.
- Reactivity scores: Using the per-residue coverage and modification count information, reactivity scores are calculated for every residue of a transcript. If local coverage information is available, it is used for reactivity scoring while in cases it is not available, overall coverage is divided by length of transcript and used as a proxy for local coverage. The rate of modifications in plus channel and noise detections in minus channel are calculated as the ratio of modification counts to coverages for respective channels. The rates for plus and minus channels are combined with four options, as described below, are available to the user.
  - The plus and minus channel rates may be combined by taking ratio or difference. If ratio is selected, 1 is deducted from the ratio of plus to minus channel rates, to peg the reactivity of unreactive sites at zero.
  - The modification counts and coverages may be log transformed before calculating rates. Alternatively, users may opt to log transform reactivities. log of non-positive quantities is set to zero.
  - Negative reactivites (if any) may be shifted to zero.
  - The resulting scores may be normalized using 2%-8% normalization or, box plot normalization or, not be normalized. In 2%-8% normalization strategy, the highest 2% of reactivities are removed and a normalization constant is calculated as average of next 8% of highest reactivities. All reactivities are then scaled by the normalization constant. For box plot normalization, the normalization constant is calculated as average of top 10% of reactivities left after filtering out outliers using Tukey box plots.

  The defaults are set to using differences for reactivity calculation, no log transformation, 2%-8% normalization and shift negative reactivities to zero.
- Signal-to-Noise Ratio (SNR) : SNR is defined as the ratio of sample mean to standard deviation of reactivity scores. [4] The samples used for calculating SNR may be experimental replicates or bootstrap resamples of each replicate. Bootstrapping is done by resampling modification counts only while local coverage is held fixed for all resamples. Alternatively, in cases where accurate local coverage information

is available from an experiment, a highly accurate formula [4] can be used to calculate standard deviation. Furthermore, based on characteristics of datasets [4], the upper limit of SNR is fixed to 35. All values higher than 35 are clipped to 35. While SNR based on replicates is informative of replicate agreement, SNR based on resamples of a replicate inform about quality of the replicate. Residue SNR values greater than 5 are considered good quality, between 3 and 5 are considered ambiguous quality and <3 are poor quality.

- Coverage Quality Index (CQI) : CQI for a residue is the ratio of observed local coverage of a site to desired local coverage [4]. The desired value of local coverage is calculated based on user specified significance level and a target range of variation. Details of calculations are provided in the Supplementary Information with the SEQualyzer manuscript and in Choudhary *et al.*, in press. The reported values are 95th percentile of CQIs for reactivities belonging to low (< 0.3), medium (≥ 0.3 & < 0.7) and high (≥ 0.7) reactivities. CQI provides assessment of how poorly the error bars in reactivity scores compare with the desired error bars, assuming that the estimated mean reactivities are close to true means. CQIs <1 are desirable as they indicate well-probed transcripts. Note that CQI results provided in SEQualyzer are for selected range of nucleotides of transcript. Users can get CQI for complete transcript by selecting entire length of the transcript from side panel.

- Mean SNR : Mean SNR is calculated as mean of the residue-level SNRs. This can be used to quickly sift huge datasets for good quality transcripts as transcripts found to have high mean SNR have been demonstrated to have good quality information. The subset of transcripts with high-quality information can later be examined more closely with SEQualyzer.

- Pairwise mean SNR : While mean SNR is calculated by utilizing information from all replicates, in situations where there are more than two replicates, pairwise mean SNR is calculated by comparing all possible combinations of two replicates at a time. This along with correlation matrix can be used to identify discordant replicate(s). While correlation matrix is indicative of only Pearson correlation of replicates, mean SNR is also indicative of depth of coverages.

- Correlation matrix : The correlation matrices are general Pearson correlation matrices from pairwise comparison of replicates.

- Rolling mean SNR (or window SNR) : Window SNR is calculated as center-aligned rolling mean of residue SNR in user-defined window sizes The default is set to 20 residues. It is indicative of quality over region of transcript.

- Lorenz curve : Lorenz curve for an experiment illustrates the proportion of all reads that mapped to the bottom $x\%$ of the transcripts. It is used to present the degree of non-uniformity in overall coverages of transcripts. The diagonal is the line of perfectly homogeneous coverage. Sag below the diagonal line in curves corresponding to replicates are indicative of degree of non-uniformity in coverages. See case studies for example. The degree of non-uniformity is also quantified as Gini index, which can range from 0 (all transcrips have equal overall coverage) to 1 (almost all of the reads pertain to 1 or very few transcripts), and reported along with figure legend in SEQualyzer. Lesser the Gini index, lesser is non-uniformity in coverages. Note that SEQualyzer provides the Lorenz curve for plus channel modification counts as plus channel are stronger determinats of data quality. Lorenz curve gives a transcriptome-wide view of data quality.

8 Sample data and examples

To demonstrate utility of SEQualyzer to quality inspections, we provide partial analysis of three structure-profiling datasets below. All results presented below are derived from analysis using SEQualyzer with default parameters for processing steps.

8.1 Data files

See Table 1

Experiment/ Reference	Features	
1. SHAPE-Seq / Loughrey *et al.*, 2014	Number of transcripts/genes	8
	Replicates	3
	Condition	*in vitro*
	Sequencing	Paired-end
	Lengths of transcripts	71-335 nt
	Probing reagent	SHAPE
	Probed bases	A, C, G, U
	Local coverage information	Yes
	Sequence information	Yes
2. SHAPE-MaP / Siegfried *et al.*, 2014	Number of transcripts/genes	1
	Replicates	1
	Sequencing	Paired-end
	Lengths of transcripts	469 nt
	Probing reagent	SHAPE
	Probed bases	A, C, G, U
	Local coverage information	Yes
	Sequence information	Yes
3. Mod-Seq / Talkish *et al.*, 2014 (Mapped and summarized as counts)	Number of transcripts/genes	7120
	Replicates	2
	Condition	*in vivo*
	Sequencing	Single-end
	Probing reagent	DMS
	Probed bases	A, C
	Local coverage information	No
	Sequence information	No
4. icSHAPE / Spitale *et al.*, 2015 (300 million reads from each channel; mapped and summarized as counts)	Number of transcripts/genes	12729
	Replicates	2
	Condition	*in vivo*
	Sequencing	Single-end
	Probing reagent	SHAPE
	Probed bases	A, C, G, U
	Local coverage information	No
	Sequence information	No

Table 1

8.2 SHAPE-Seq dataset

The SHAPE-Seq dataset [5] contains sequence and local coverage information for all transcripts, has 1545174 reads mapped to 3 replicates of 8 transcripts. The mean SNR values from three replicate comparisons for this dataset range from 1.2 to 5.03, although in pairwise comparisons of replicates, several transcripts have higher mean SNR. Among the three datasets included along with SEQualyzer, the SHAPE-Seq dataset has the most uniform distribution of coverages as shown with the Lorenz curve in Figure 2. Quality of data for specific transcripts can be inspected by choosing the transcript from side panel. We present some comparative insights for cyclic di-GMP of *V. cholera* (labelled 'cdGMP' in dataset) and P4-P6 domain group I ribozyme of *Tetrahymena* (labelled 'tetra.intronI' in dataset). SEQualyzer reveals reactivity profiles (Figure 3) for the three replicates of these transcripts. The error bars are found to be small for both the transcripts. Comparison of the replicates however, becomes simpler if reactivity profiles for three replicates are overlaid in the same plot. Figure 4 reveals that among the two transcripts, the replicates are in better agreement for *tetra.intronI.* This comparison is further illustrated by correlation and pairwise mean SNR summaries as in Figure 5. These illustrations facilitate identification of discordant replicates as they present

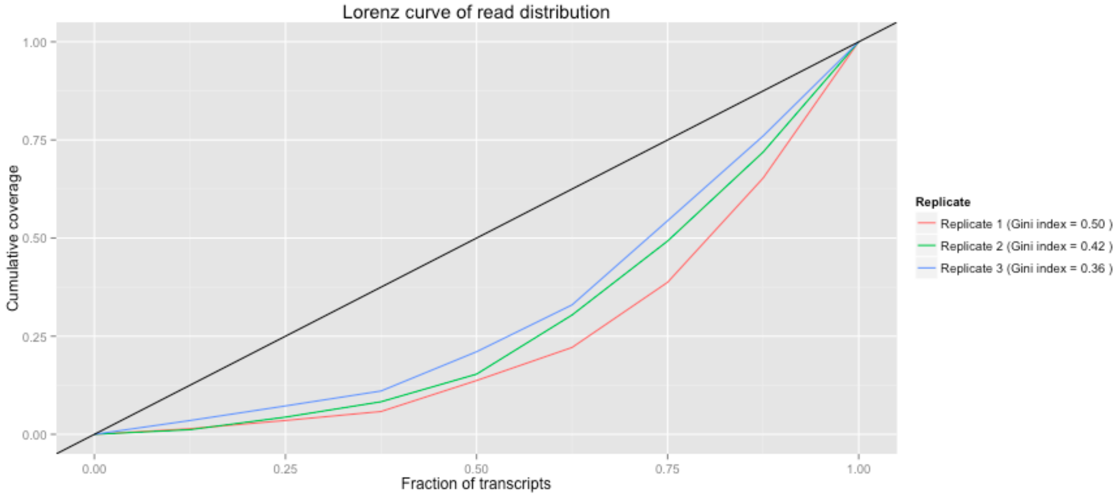


Figure 2: Lorenz curve for distribution of reads among 8 transcripts in SHAPE-Seq dataset. The plot shows that the coverage is most uniform for replicate 3 in this data.

all pairwise comparisons. Figure 5 shows that for *cdGMP*, replicates 1 and 2 are in best agreement, while for *tetra.intronI*, replicates 2 and 3 agree best. Such illustrations and comparisons will be more useful in situations when there were larger number of replicates. Figure 6 uses window SNR to show that all replicates of both transcripts display a loss in quality towards 5' end but this loss is more pronounced for *cdGMP*. The Coverage Quality Indices (Figure 7) show that the coverages are much better for *tetra.intronI* for all categories of reactivities and in each replicate.

8.3 SHAPE-MaP dataset

Included in the sample dataset is SHAPE-MaP data for *O. iheyensis* group II intron RNA [2]. This data features only one replicate but may still be examined for quality of that one replicate. This transcript, 469 residues in length has very high mean local coverage of 168524.7. The data has a high mean SNR of 9.96 and good CQI for high (0.13) and medium (0.39) reactivities while that for low reactivities is 2.17. Results of analysis for this data are shown in Figure 8-11.

8.4 Transcriptome-wide dataset

There are two transcriptome-wide datasets provided with SEQualyzer. One of them is obtained from a Mod-Seq [6] experiment and another from icSHAPE [7]. Mod-Seq dataset has 2661819 reads mapped to 2 replicates of 7120 transcripts. The mean SNR values from two replicate comparisons for this dataset range from 0.7 to 21.6. The Mod-Seq dataset has a highly non-uniform distribution of coverages as shown with the Lorenz curve in Figure 12a. Quality of data for specific transcripts can be inspected by choosing the transcript from side panel. This being a transcriptome-wide dataset, it contains information for 7120 transcripts. Since it is not possible to examine each of these transcripts one-by-one, we make use of filters available in the side panel to select transcripts of interest. In different scenarios, interest may be confined to transcripts with best coverage, or a certain minimum length, or a minimum profiling quality as summarized by mean SNR. Filtering for transcripts with mean SNR $>$ 5 highlights 4 tRNAs, which on closer inspection show signal at only a few sites. Relaxing the mean SNR criterion and adding criterion for mean local coverage, we filter transcripts using two screens - mean SNR $>$ 3 and mean local coverage $>$ 25. SEQualyzer reveals that 2 rRNAs satisfy these criteria. Closer inspection of 18S rRNA reveals unusually high noise at site 1192, which may be warrant some further investigation.

Similar analyses can be performed for icSHAPE data. icSHAPE data has 114608517 reads mapped to 12729 transcripts. It is notable that icSHAPE has more uniformity in coverage of transcripts as can be seen from Figure 12b. Given the huge volume of the dataset, again it is not possible to examine each transcript in depth. Hence, it is desirable to confine study to transcripts of interest. If the filters are set for transcripts with mean local coverage $>$ 25 and mean SNR $>$ 5, 37 transcripts identified as high quality by Choudhary *et al.*, in press are reproduced. Closer inspection reveals that certain transcripts such as ENSMUSG00000005442, ENSMUSG00000047284, etc. give good signal for only the 3' prime end of their length. Additionally, it is interesting to note that the best quality transcripts in Mod-Seq data are all rRNAs while those in icSHAPE data are protein-coding RNAs.

9 Further information

Please check Aviran lab website (`http://bme.ucdavis.edu/aviranlab/sequalyzer`) for updated information on the tool, to provide feedback and/or report bugs.

References

[1] Y. Tang, E. Bouvier, C. K. Kwok, Y. Ding, A. Nekrutenko, P. C. Bevilacqua, and S. M. Assmann. StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics*, 31(16):2668–2675, 2015.

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

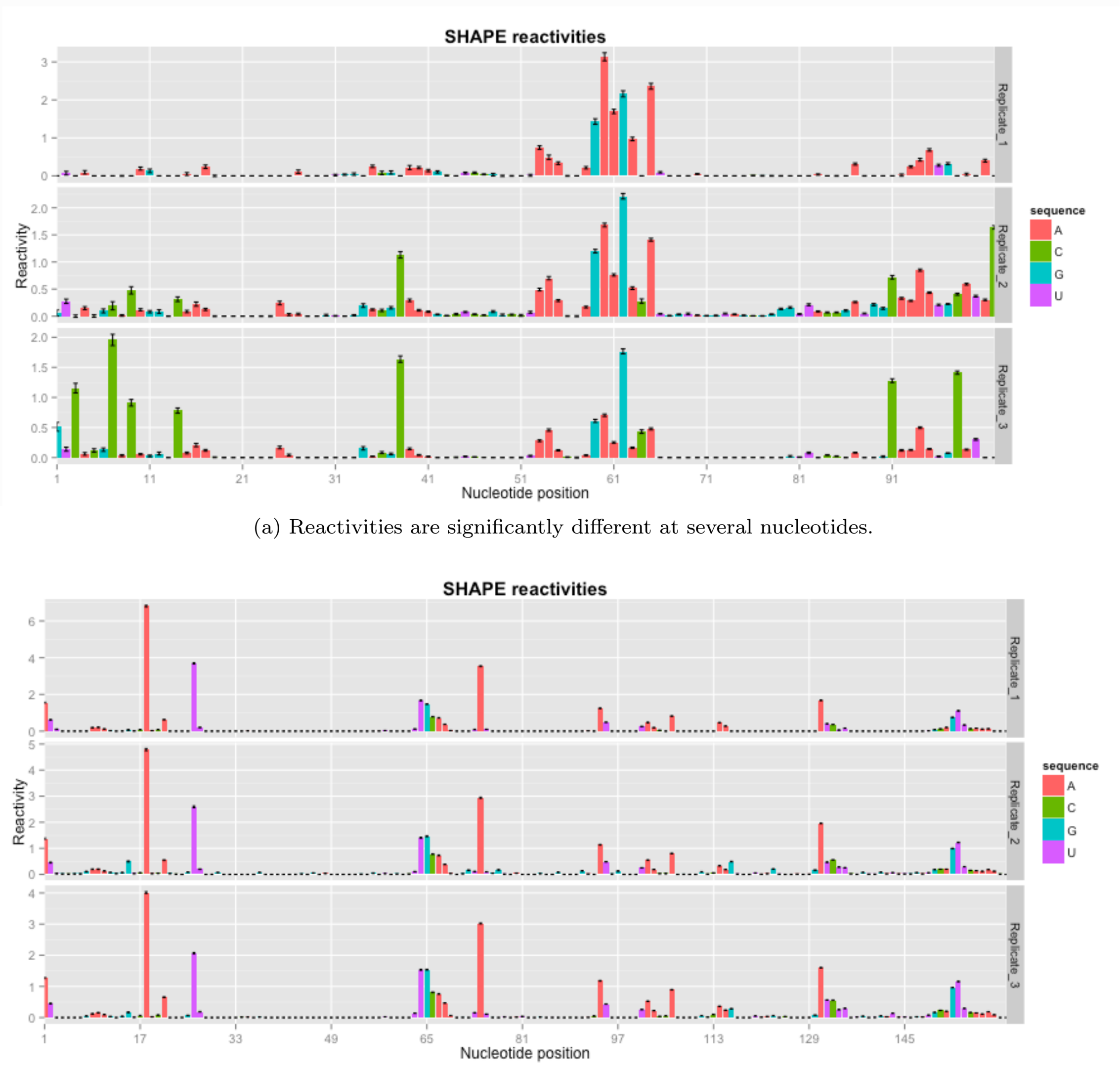
[3] P. Cordero, J. B. Lucks, and R. Das. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics*, 28(22):3006–3008, 2012.

[4] K. Choudhary, N. P. Shih, F. Deng, M. Ledda, B. Li, and S. Aviran. Metrics for rapid quality control in RNA structure probing experiments. *Bioinformatics*, page btw501, 2016.

[5] D. Loughrey, K. E. Watters, A. H. Settle, and J. B. Lucks. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research*, 2014.

[6] J. Talkish, G. May, Y. Lin, J. L. Woolford, and C. J. McManus. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*, 20(5):713–720, 2014.

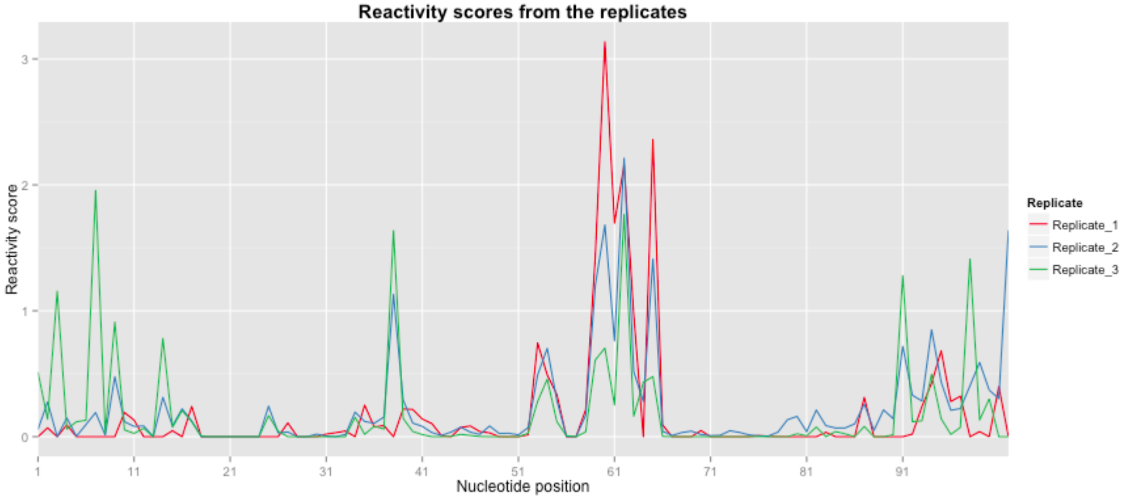
[7] R. C. Spitale, R. A. Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J.-W. Jung, H. Y. Kuchelmeister, P. J. Batista, E. A. Torre, E. T. Kool, et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 2015.



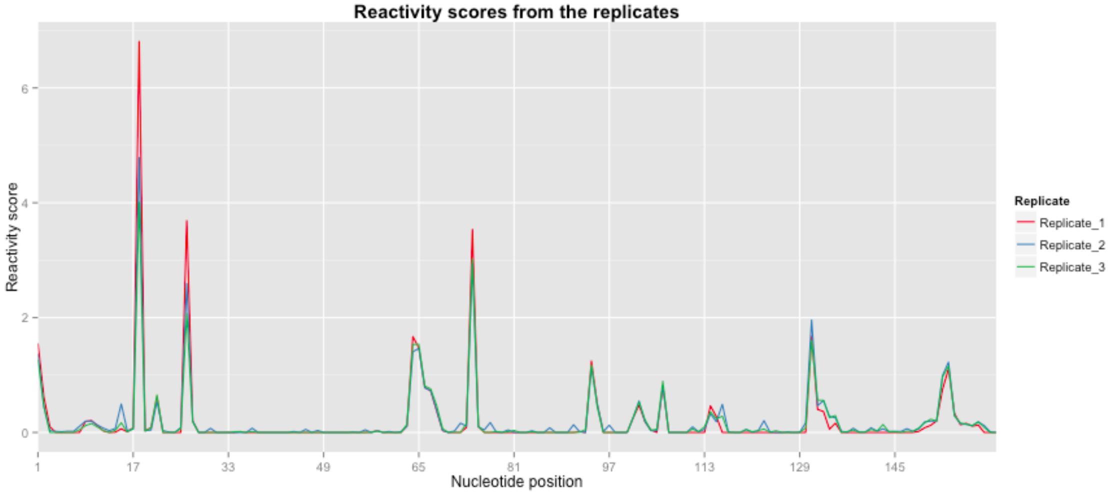
(a) Reactivities are significantly different at several nucleotides.

(b) Reactivities are in strong agreement at all nucleotides.

Figure 3: SHAPE reactivities (normalized) for (a) cyclic di-GMP and (b) P4-P6 domain group I ribozyme.

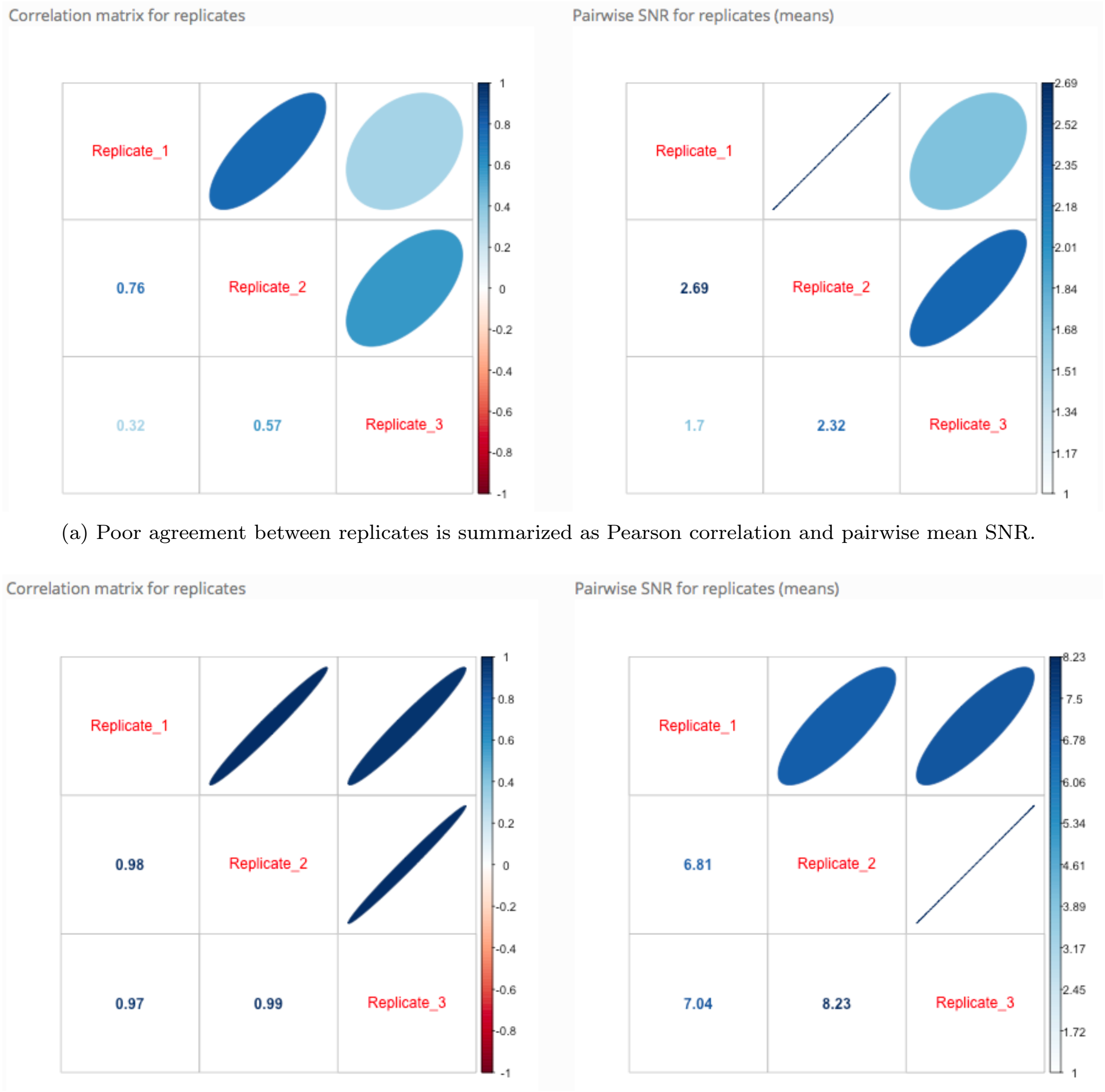


(a) Reactivities are significantly different at several nucleotides.



(b) Reactivities are in strong agreement at all nucleotides.

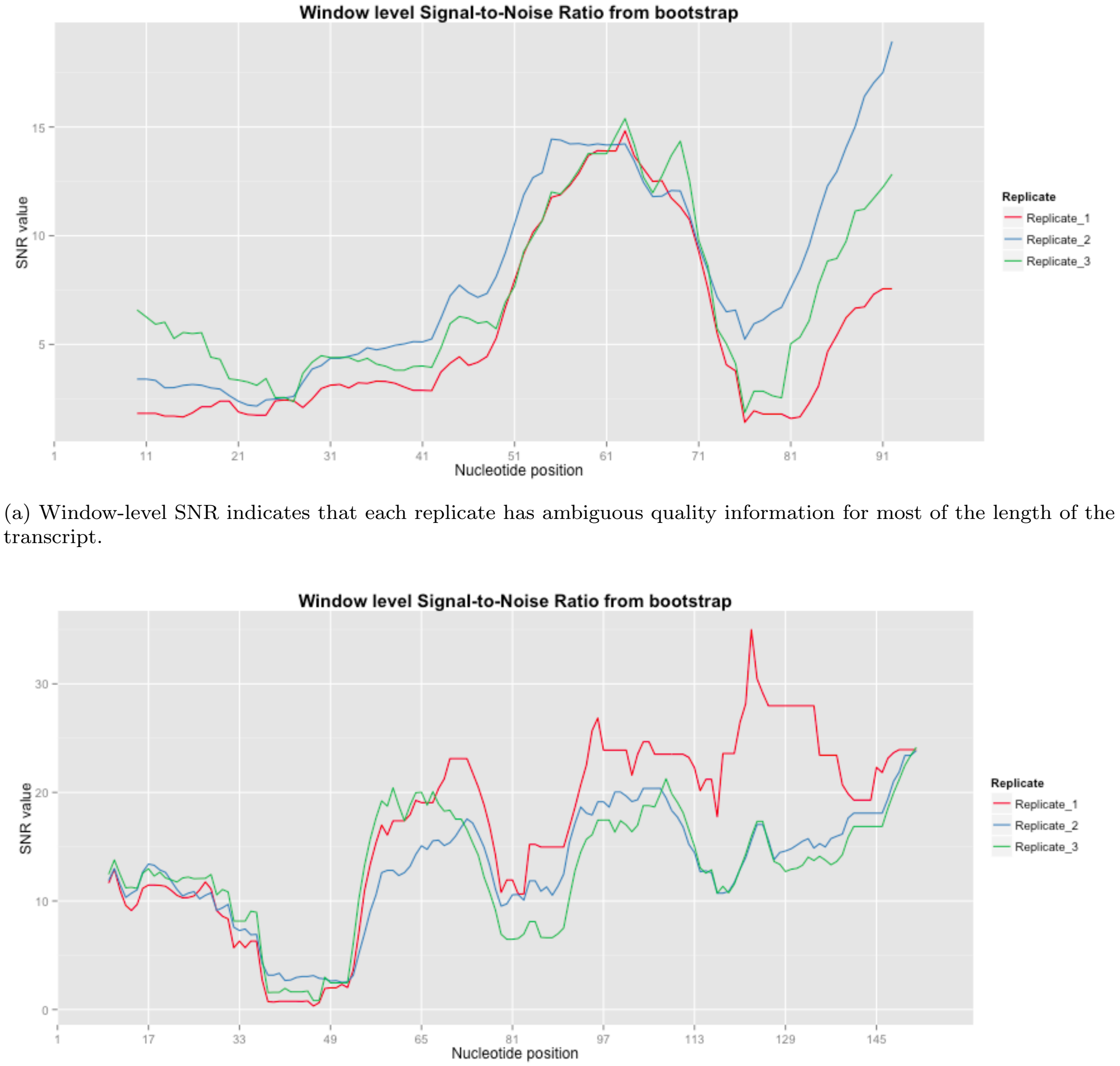
Figure 4: SHAPE reactivities (normalized) for all replicates together in same plot for (a) cyclic di-GMP and (b) P4-P6 domain group I ribozyme.



(a) Poor agreement between replicates is summarized as Pearson correlation and pairwise mean SNR.

(b) Strong agreement between replicates is summarized as Pearson correlation and pairwise mean SNR.

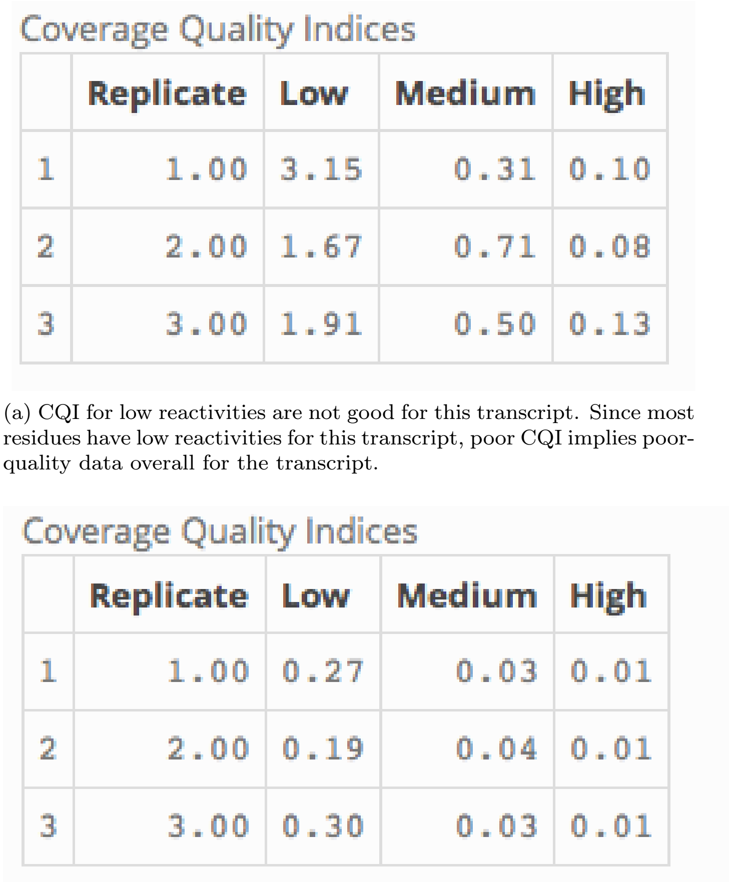
Figure 5: Summarizing replicate comparisons as pairwise mean SNR can help identify discordant replicates. For e.g., in Sub-Figure 5a, replicates 1 and 2 are in better agreement than replicates 1 an 3. Sub-figures summarize replicate agreements for (a) cyclic di-GMP and (b) P4-P6 domain group I ribozyme.



(a) Window-level SNR indicates that each replicate has ambiguous quality information for most of the length of the transcript.

(b) Window-level SNR indicates that each replicate has good quality information for most of the length of the transcript. This provides extra credibility to data for this transcript , since not only are the repicates for this transcript in agreement, they are also each good quality individually.

Figure 6: Window-level SNR with window-size 20 for (a) cyclic di-GMP and (b) P4-P6 domain group I ribozyme.



Coverage Quality Indices

	Replicate	Low	Medium	High
1	1.00	3.15	0.31	0.10
2	2.00	1.67	0.71	0.08
3	3.00	1.91	0.50	0.13

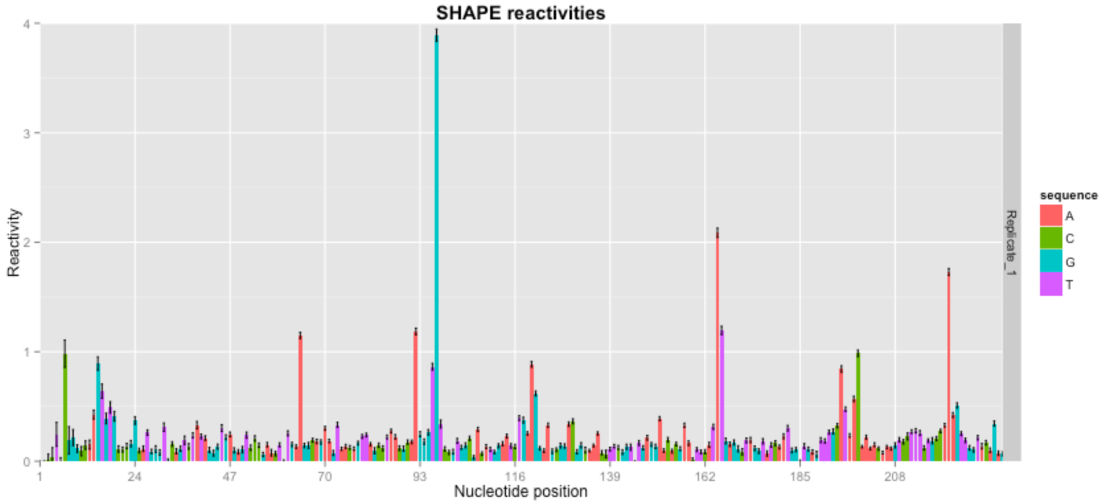
(a) CQI for low reactivities are not good for this transcript. Since most residues have low reactivities for this transcript, poor CQI implies poor-quality data overall for the transcript.

Coverage Quality Indices

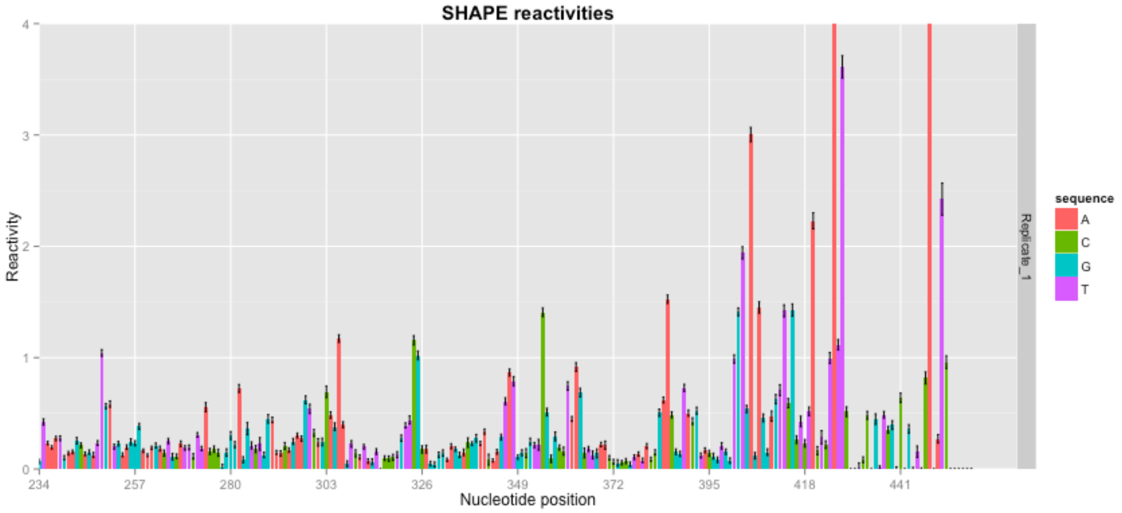
	Replicate	Low	Medium	High
1	1.00	0.27	0.03	0.01
2	2.00	0.19	0.04	0.01
3	3.00	0.30	0.03	0.01

(b) All replicates have good CQI for all reactivity ranges.

Figure 7: CQI indicates the goodness of coverage of (a) cyclic di-GMP and (b) P4-P6 domain group I ribozyme.

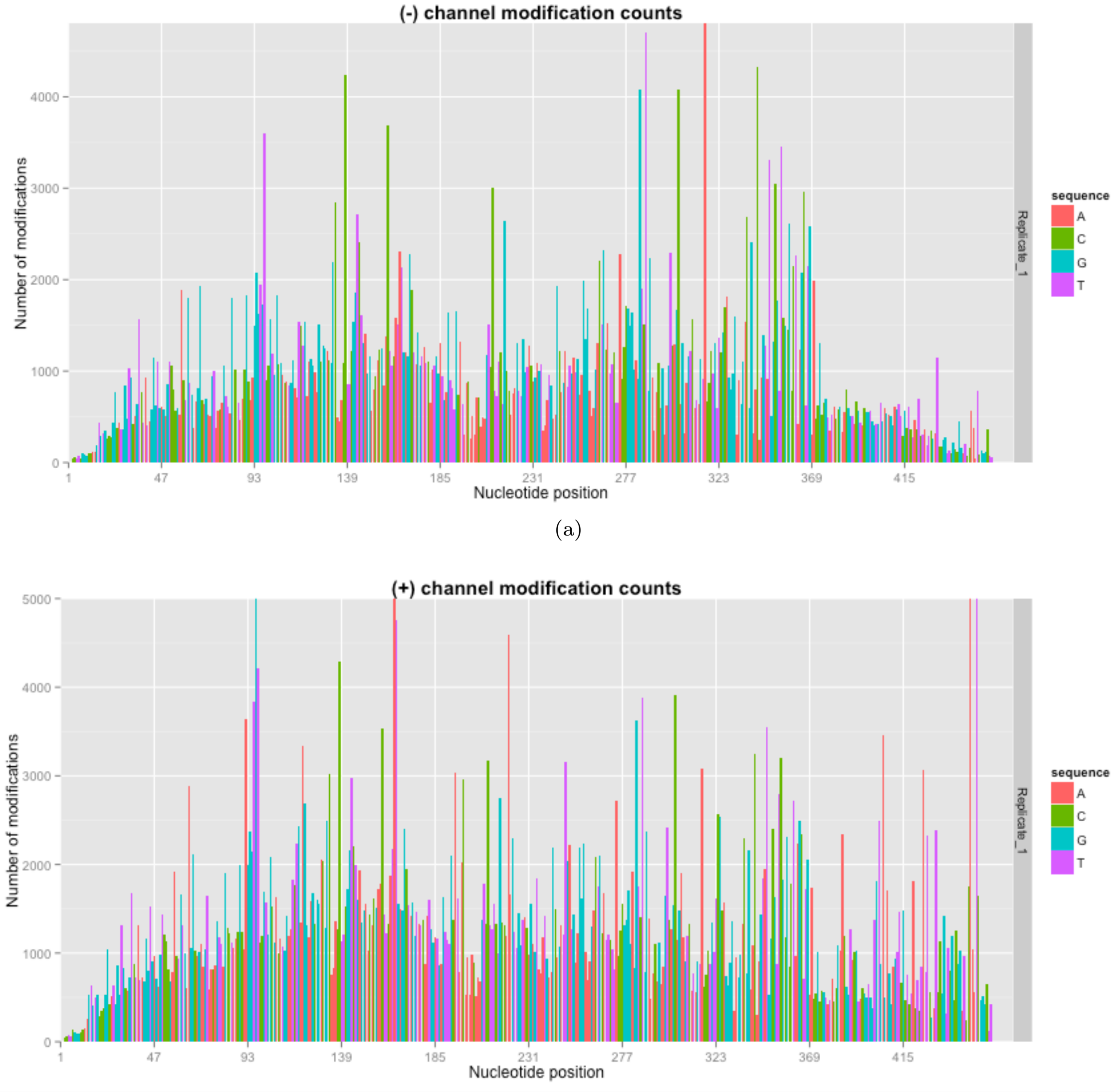


(a)



(b)

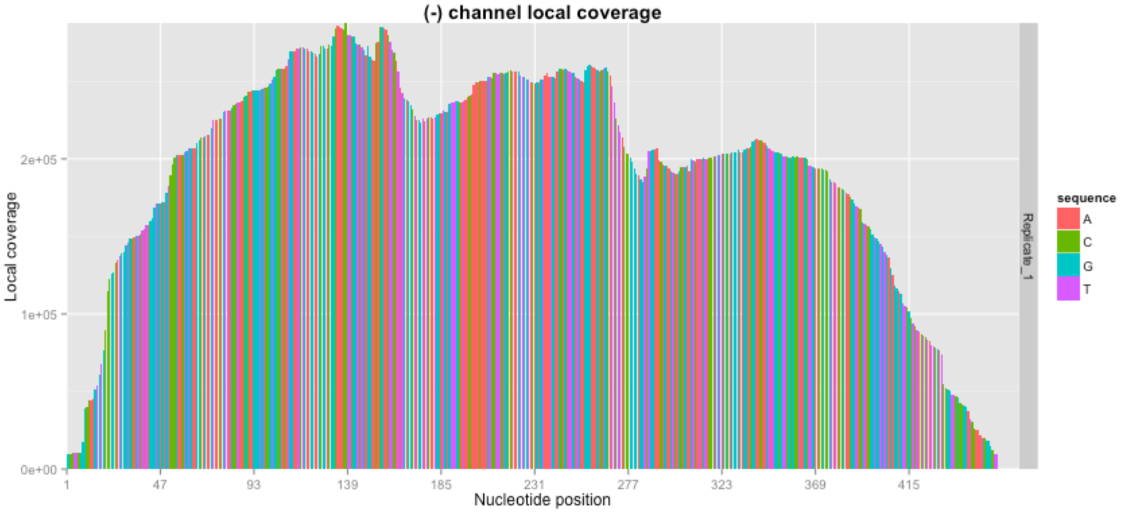
Figure 8: SHAPE reactivities with error bars based on bootsrapping for SHAPE-MaP profiling of group II intron (a) residues 1-234 and (b) residues 234-469.



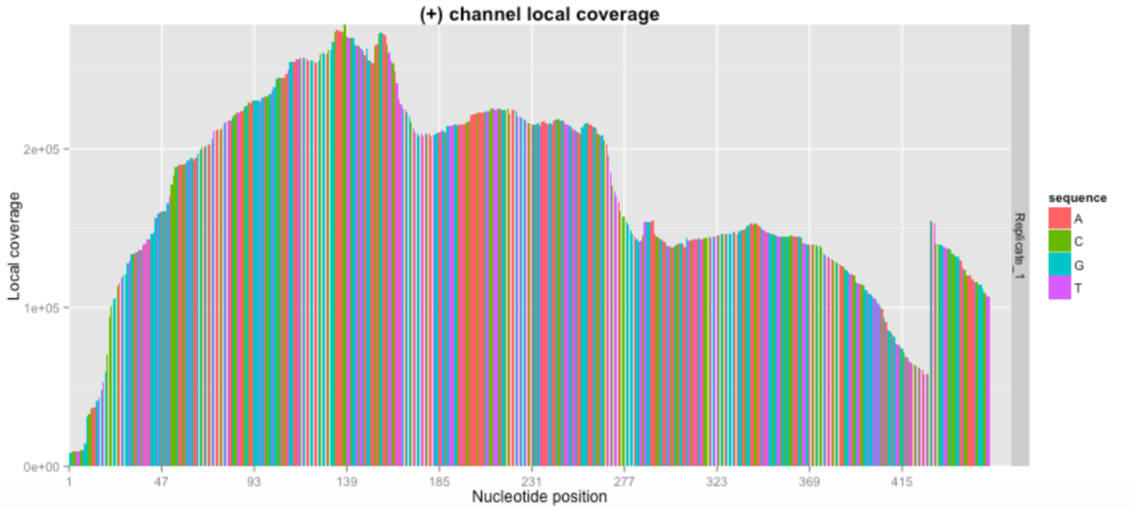
(a)

(b)

Figure 9: SHAPE-MaP modification detection counts for (a) minus channel and (b) plus channel.

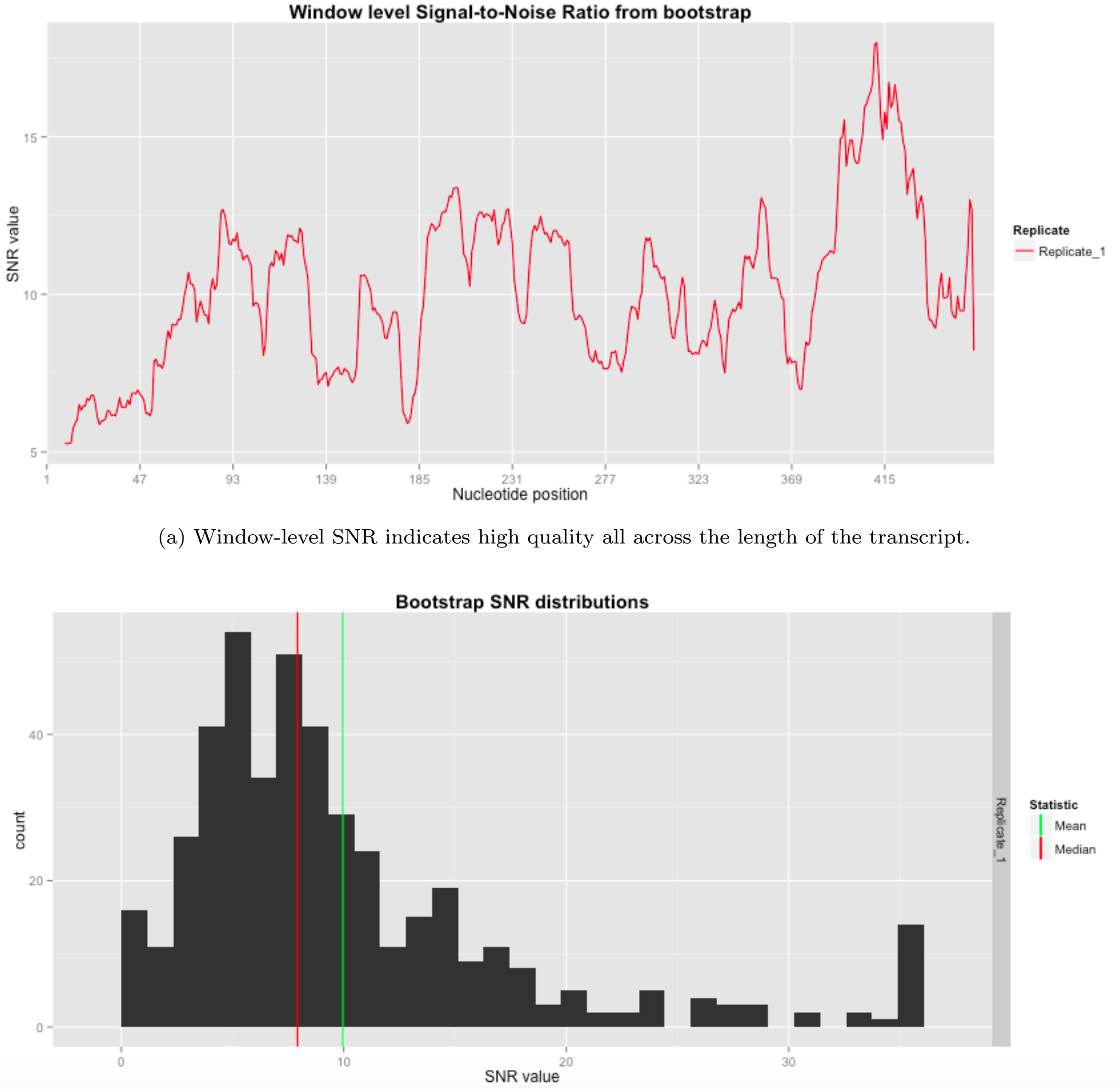


(a)



(b)

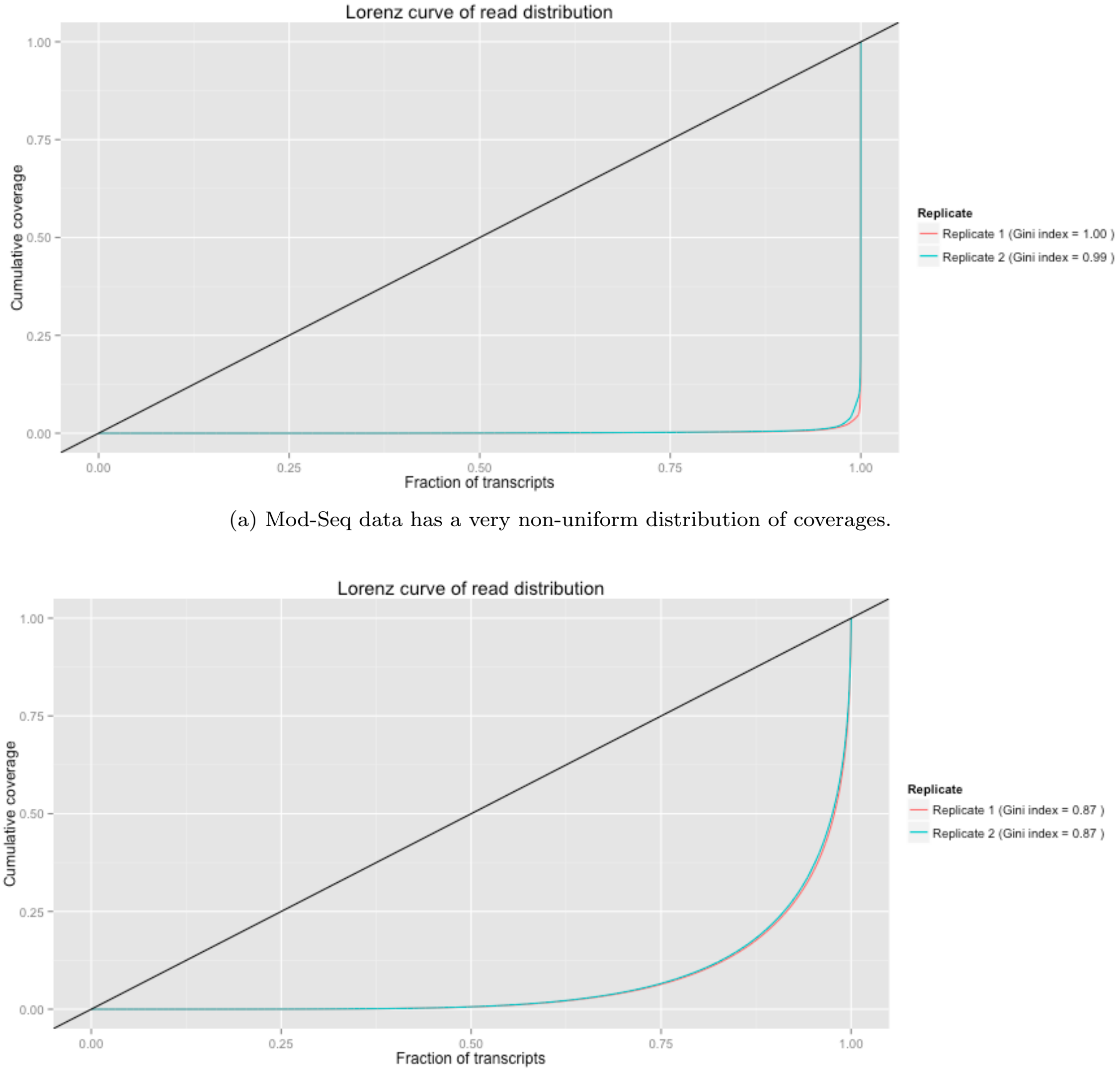
Figure 10: SHAPE-MaP per-residue coverage for (a) minus channel and (b) plus channel. The plots reveal an identical trend for both channels and no 3' bias of coverage, an important feature of SHAPE-MaP profiling method.



(a) Window-level SNR indicates high quality all across the length of the transcript.

(b) Distribution of bootstrap-based SNR reveals that a good majority of residues have high SNR.

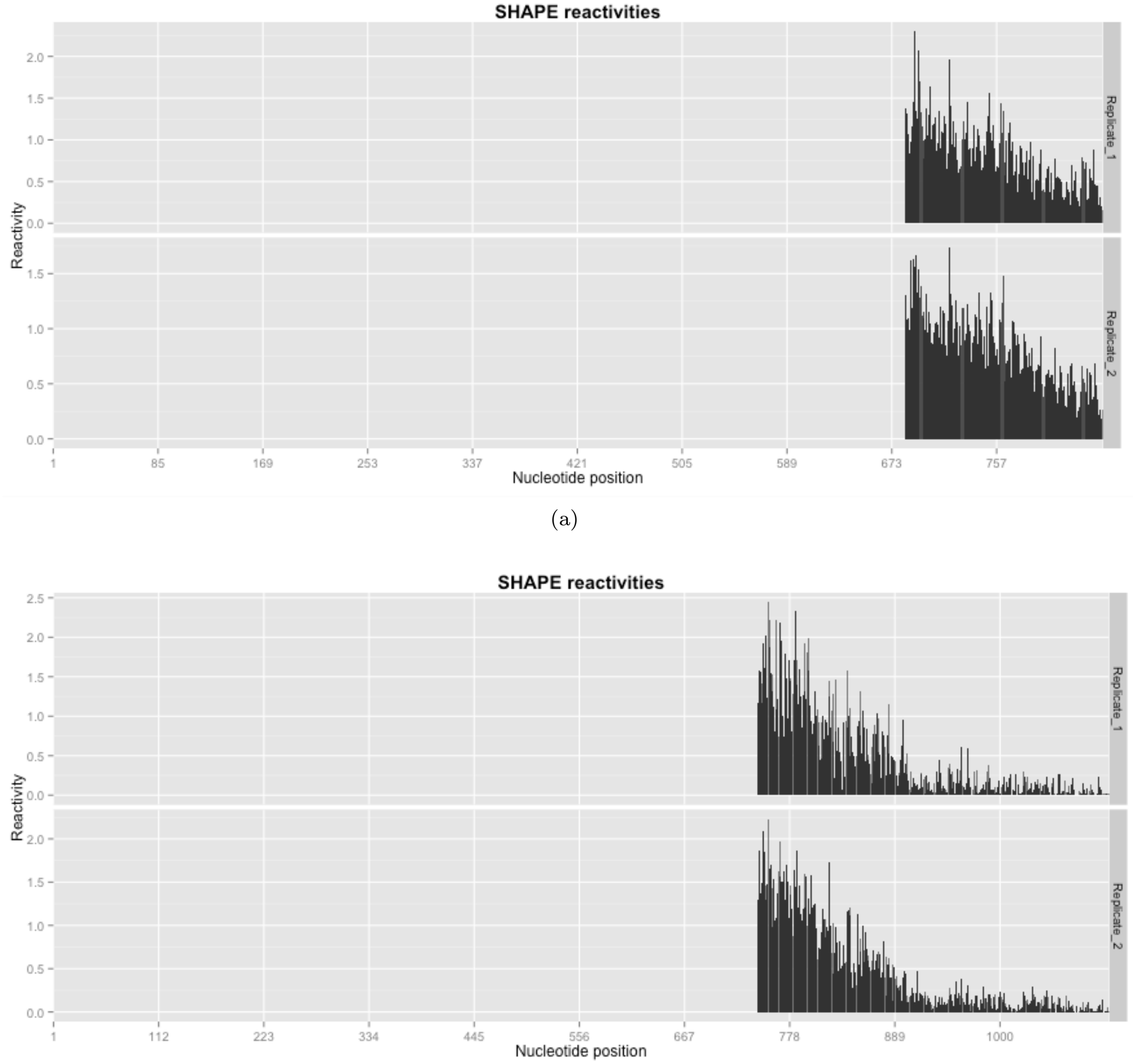
Figure 11: Results from bootstrapping of group II intron data presented as (a) window-level SNR and (b) bootstrap SNR distributoon.



(a) Mod-Seq data has a very non-uniform distribution of coverages.

(b) icSHAPE data has a much better uniformity of coverage than Mod-Seq data.

Figure 12: Lorenz curve for distribution of reads among 8 transcripts in (a) Mod-Seq dataset and (b) icSHAPE dataset.



(a)

(b)

Figure 13: SHAPE reactivities good for only 3' end of (a) ENSMUSG00000005442 and (b) ENSMUSG00000047284.