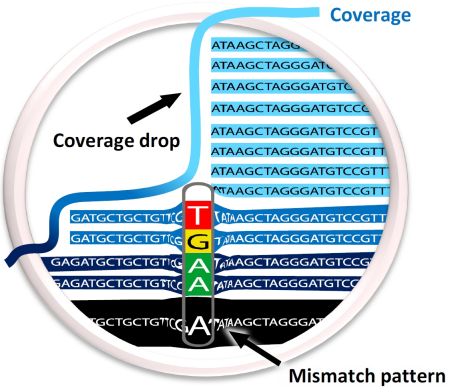
CoverageAnalyzer

Visual Inspection & Automatic Screening of Deep-Seq Data for Detection of RNA Modifications

v. 1.0

02.2016

Manual



released by

Ralf Hauenschild

from
Obernburg am Main

on 1st February 2016

Prof. Dr. Mark Helm

Institute of Pharmacy and Biochemistry

Johannes Gutenberg-University, Mainz

1 Installation

CoverageAnalyzer is available for Windows, Linux and MacOSX 64 bit systems. The software has several dependencies, which are installed during setup. 2 GB of disk space should be available for installation. Reserve additional space for intended analyses.

1. Download CoverageAnalyzer at https://sourceforge.net/projects/coverageanalyzer/files/.
2. Copy to a desired installation folder with full permissions (read/write/execute) and short absolute path (< 30 characters) to avoid installation problems.
3. It is recommended to temporarily disable firewalls/antivirus software or create exceptions.
4. Ensure internet connection and run setup as *administrator*:
   - Windows: Double-click sfx.exe archive to extract and install contents.
   - Linux/MacOSX: Unzip .zip archive and run Setup.sh.
5. Read and accept licence.
6. Wait for download and installation
7. Install Java RE 7+, if prompted.
8. Launch. (Windows: *CoverageAnalyzer.exe*; Linux/MacOSX: *CoverageAnalyzer.sh*)

Troubleshooting:
When setup fails due to accidental launch without administrator rights, kill unterminated processes using Task Manager, such as Miniconda-Setup, vanish install folder and restart setup as administrator. Java RE is a mandatory requirement. When encountering launching problems, make sure Java RE is installed and available via path/environment variable. If Java is not found, try to launch CoverageAnalyzer by '*.../yourjavafolder/java -jar .../yourfolder/CoverageAnalyzer.jar*' or right-click .jar file and specify JRE to open with.

Dependencies:

- Installed by user: Java RE
- Installed automatically: Python (Miniconda), Matplotlib, Numpy, Scipy, Homebrew (MacOSX only), SAMtools

2 Usage

Launch CoverageAnalyzer in any of the following ways:

- Windows
  - Double-click CoverageAnalyzer.exe
  - CMD '*java -jar ...yourpath/CoverageAnalyzer.jar*'
  - Right-click *'...yourpath/CoverageAnalyzer.jar'* and open with JRE
- Linux/MacOSX
  - Terminal: 'sh ...yourpath/CoverageAnalyzer.sh'
  - Right-click and run/execute *'...yourpath/CoverageAnalyzer.sh'*
  - Terminal: '*java -jar ...yourpath/CoverageAnalyzer.jar*'

2.1 Prerequisites & File Formats

Prior to data inspection, the input files for analysis must be deposited in a folder structure such that a mother folder contains an arbitrary number of sample folders named with integer numbers, each equipped with one *SAM* file named ending with '.sam'. Outside the sample number folders, the original FASTA reference of the SAM files must be located named ending with '.fasta'. BAM files can be reconverted to SAM format using SAMtools [3] (see documentation at http://samtools.sourceforge.net/samtools.shtml).

2.2 Profiling & Analysis

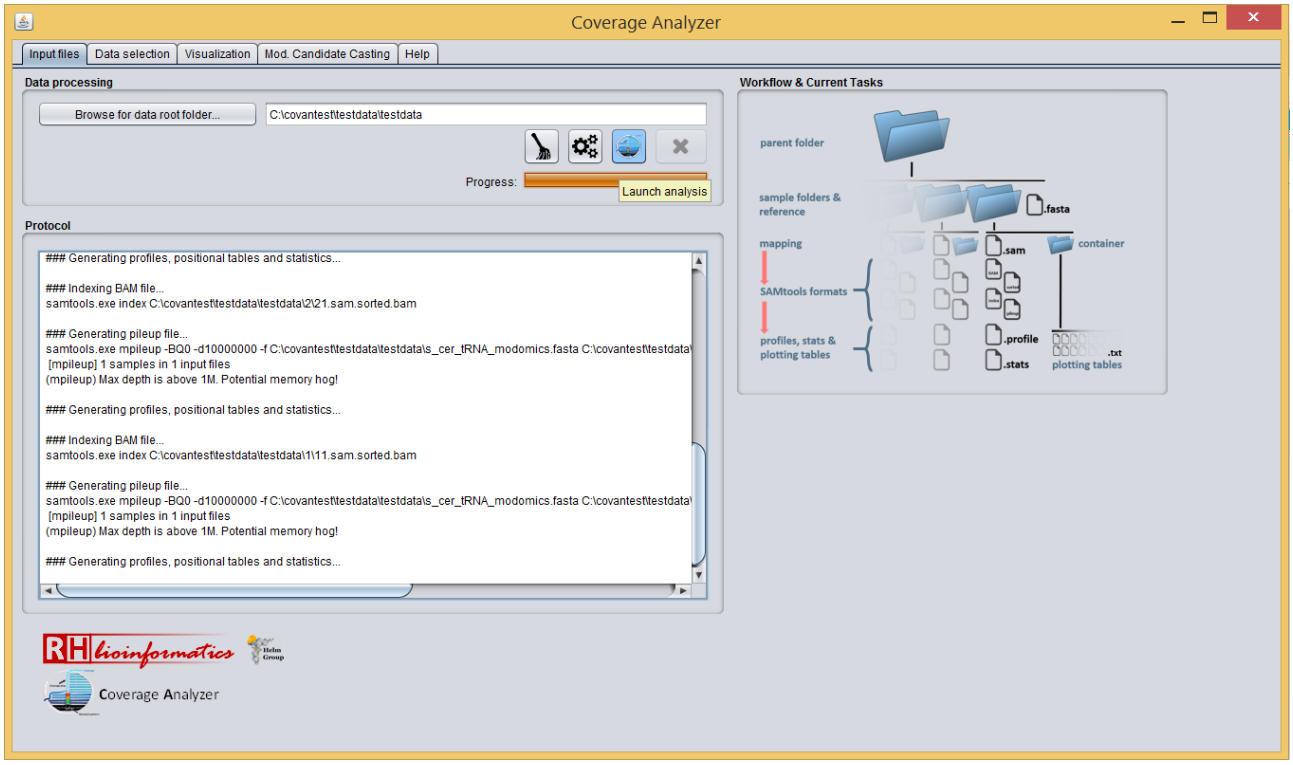


Figure 1: *Input* tab.

CoverageAnalyzer's graphical user interface is fully provided with mouse-hover tool tips. Every analysis starts on the *Input files* tab (Figure 1).

1. Use the browse button Browse for data root folder... to specify the path of the analysis mother folder containing the sample number folders (choose select/open). Alternatively, a path can be pasted into the adjacent text field. The displayed path should be of the scheme '.../path/motherfolder'. The program will remember the last directory you visited.

2. Optionally, you can vanish the mother folder and subfolders from all data generated previously by CoverageAnalyzer using the button.

3. Generation of *Profile* format can be performed using the profiling function . Computation time depends on coverage depth in the *SAM*. The current task running is highlighted in the scheme on the right.

4. Pressing of the abortion button will cancel profiling.

5. When profiling is done, the analysis button is enabled. Press it and proceed on the *Data selection* tab (Figure 2).

2.3 Table Inspection

The *Data selection* tab allows the user to assess the analyzed mapping profiles with respect to multiple parameters. Next to *sample IDs*, *file paths* and *reference segment* names, also *reference lengths*, *sequences*, *maximum coverage* and the total number of *mapped reads* can be chosen as sorting criteria. A more modification specific evaluation can be done *via* sorting by number of significant arrest sites (coverage ≥ 20, arrest rate ≥ 0.1), high mismatch rate m positions (coverage ≥ 20, $m \geq 0.1$) and heterogenous mismatch sites (coverage ≥ 20, 2^{nd}-most frequent mismatch component ≥ 0.1). Additional requirements, exclusions and cutoffs can be set in the *Filters* section also featuring a restore option for the original table.

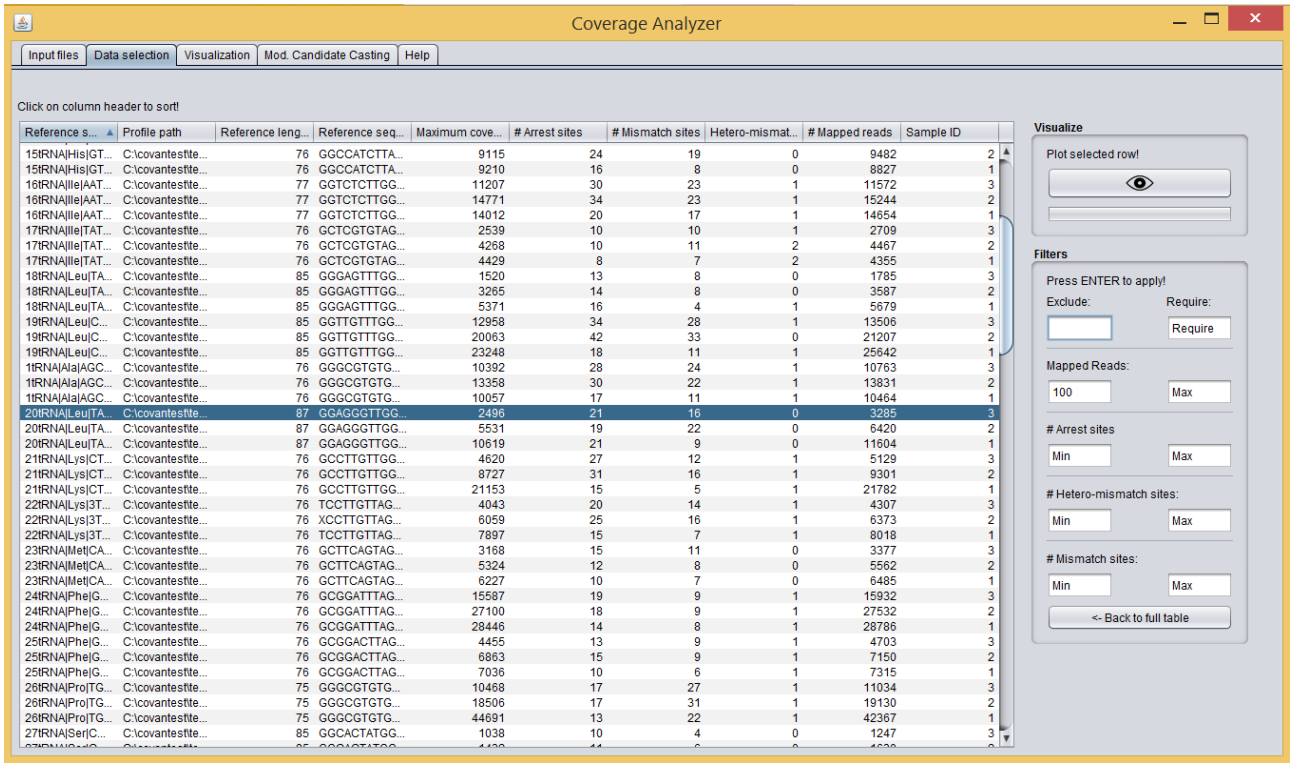


Figure 2: *Data selection* tab.

2.4 Visual Inspection

Click the button in order to launch visualization of a selected alignment. A first plot with standard settings will show up in the *Visualization* tab (Figure 3).

2.4.1 Preferences & Plot Control

The number of plots displayed on one page is changeable to 1 or 2 using the Images/page: 2 spinner. All specifications (parameters, sequence range, plot details) apply to the plot currently selected in plot selection drop-down menu C:\covantest\testdata\testdata\1\positional_tables/201602090356295... . The delete button will remove the currently selected plot and resize the remaining ones according to Images/page: 2 . The button deletes all plots at once. This option is also useful, if erroneous program behavior is observed. Plots can be resized by dragging the main frame of the program window or the right-side divider. Pushing the Apply button or reselecting the plot parameter in the drop-down menu Mismatch Patterns (MP) will redraw the selected graph with all changes in settings are realized. If plots appear in wrong aspect ratio after repeated adding or removal, press in order to refresh all plots (this can take a few seconds depending on the number plots and their contents). Preferences can be specified for reportable bases A G T C, as well as for the visual range r of the environment adaptive parameter Context

Sensitive Arrest rate, CSA, the fold change of a position i's arrest rate A w.r.t. median arrest rate within visual range: $CSA^r(i) = \frac{A_i}{median(A_{i-r},...,A_{i-1},A_{i+1},...A_{i+r})}$. In order to display mismatch patterns only at one single sequence position of interest, enable the *center only* checkbox. If needed, the displayed sequence positions can be shifted back or forth using the *pos. shift* spinner. A toggle button optionally shows and hides the arrest rate curve. The sequence interval of interest can be specified in the start and end spinners or by dragging or shifting the range slider. Note that sequence information is plotted up to 100 bases length only, also depending on image width. Zoom in, using the zoom functionality or the start and end spinners. If longer RNA references are plotted, the software switches to a *reduced-detail mode*, showing coverage only. Display of plot legend, arrest rate curve, reference sequence and grid lines as well as additional PDF image generation on the hard drive can be toggled optionally.

Figure 3: *Visualization* tab. Profiles of two different experimental conditions for the same RNA species were plotted. In addition, a differential plot was generated.

2.4.2 Differential Analysis

Whether for assessment of chemical treatment effects on RT signatures or for comparison of native samples of distinct origins, the differential analysis function is a powerful tool highlighting qualitatively diverging sequence positions. Press button in order to compare the currently displayed references (e.g. Figure 3). The function is active for equal references only. The readouts are absolute differences indicated by the black markers and relative differences indicated by the colors. Red indicates a stronger signal in the lower plot compared to the upper plot and blue a weaker signal in the lower plot, respectively. Fold changes outside of the .1 to 10 are colored the same as the margin values. If the user wants less events highlighted, but is unsure which pvalue to choose, the usage of for instance a top 5 limit (spinner at upper center) is recommended.

2.4.3 Sequence Search

This feature will be activated in future releases.

2.5 Candidate Screening

The *Mod. Candidate Casting* tab (Figure 4) provides a formula editor for generation of screening filters of arbitrary complexity. Samples can be screened in independent or differential mode, specifically or in batch mode (all). Filter modes control whether screening conditions are applied in require-all mode (and-connected conditions), alternative mode (fulfilling one of the conditions is sufficient for the candidate) or mutual exclusive mode (xor-connected conditions). Alternatively, the user can define a custom formular using parentheses and arbitrarily chosen conditions combined by the boolean buttons . The intelligent recognition mechanism helps you to set valid brackets (bracket button) around selected pairs of boolean statements by automatic selection snapping. Formulas can be saved as presets and loaded in later sessions. Filter parameters for a candidate position i include coverage c, 3' coverage, arrest rate A, mismatch rate M, M/A ratio, Context Sensitive Arrest rate CSA for specific visual range r and an auxiliary feature termed diversity. *Diversity* d of a single profile position is parameter for experienced users only. Though this feature can be used for other modifications too, its usage is recommended for m^{1}A screening only. It is represented as a 5-bit binary code, e.g. 10101 with the i^{th} bit set to 1, if the i^{th} of properties
$\left\{\left[(M \geq 0.1 \ \wedge \ M - max(M^G, M^T, M^C) \geq 0.1\right] \ ; \ \left[M \geq 0.2\right] \ ; \ \left[A \geq 0.2\right] \ ; \ \left[CSA \geq 2\right] \ ; \ \left[10 \geq M/A \geq 0.1\right]\right\}$
is fulfilled by the instance, where M is the mismatch rate as sum of mismatch components M^G, M^T, M^C. Correspondingly, $d := \sum_{i=0}^{4} f \cdot 2^i$, where $f = 1$, if condition $5 - i$ is fulfilled and $d \in [0, 31]$ is the decimal representation. Thus, a higher non-m^{1}A diversity score mimics higher similarity to m^{1}A. As a quick-setter of this arbitrary set of conditions, the bit code serves two purposes, providing a rough impression of pronounced signature parameters at first glance, while uprating particularly such m^{1}A candidates meeting the most characteristic features of m^{1}A, *i.e.* heterogeneous mismatch composition as described in [2], by the implied exponential weighting.
Specific base types can included or discarded from screening report. The *ignore* function allows hiding of certain positions specified by comma-separated integers, e.g. known modification sites out of current interest that would increase the number of reported canditates. Casting is started by the search button . Screening results in *Candidate* format can be accessed *via* the folder button . They are located in the sample number folders chosen for screening. *Candidate* files from differential screening are named according to the compared samples.

2.6 Serial Plotting

Positions of interest for a serial plotting can be provided in *Candidates* format. *Zoom* specifies the number of bases left and right of a candidate to be included in the plot. Having selected the *Candidates* file through the *Batch* button . The plotting series is launched *via* clicking the *Serial plot* button and the timecode-labeled plots can be inspected in the result folder (subfolder 'positional_tables' in sample number folder). For every timecode, the reference sequence ID can be obtained from the corresponding '_name' file.

2.7 Export

Positional information on modification candidates and other reported events in the selected plot can be exported *via* the export button . The checkbox Include adjacent positions controls supplementation of reported positions with the adjacent ones. Button allows us to jump to the folder containing the saved tab-separated result text file. Button can be clicked in order to create a PDF file containing all currently displayed plots. For high resolution vector graphics consider the auto-generated PDFs in the plot folder. From these, either JPG snapshots can be generated. Alternatively, vector graphics (EPS recommended) *via* conversion by *e.g.* Inkscape [1] freeware for a flexible graphical post-editing are recommended.

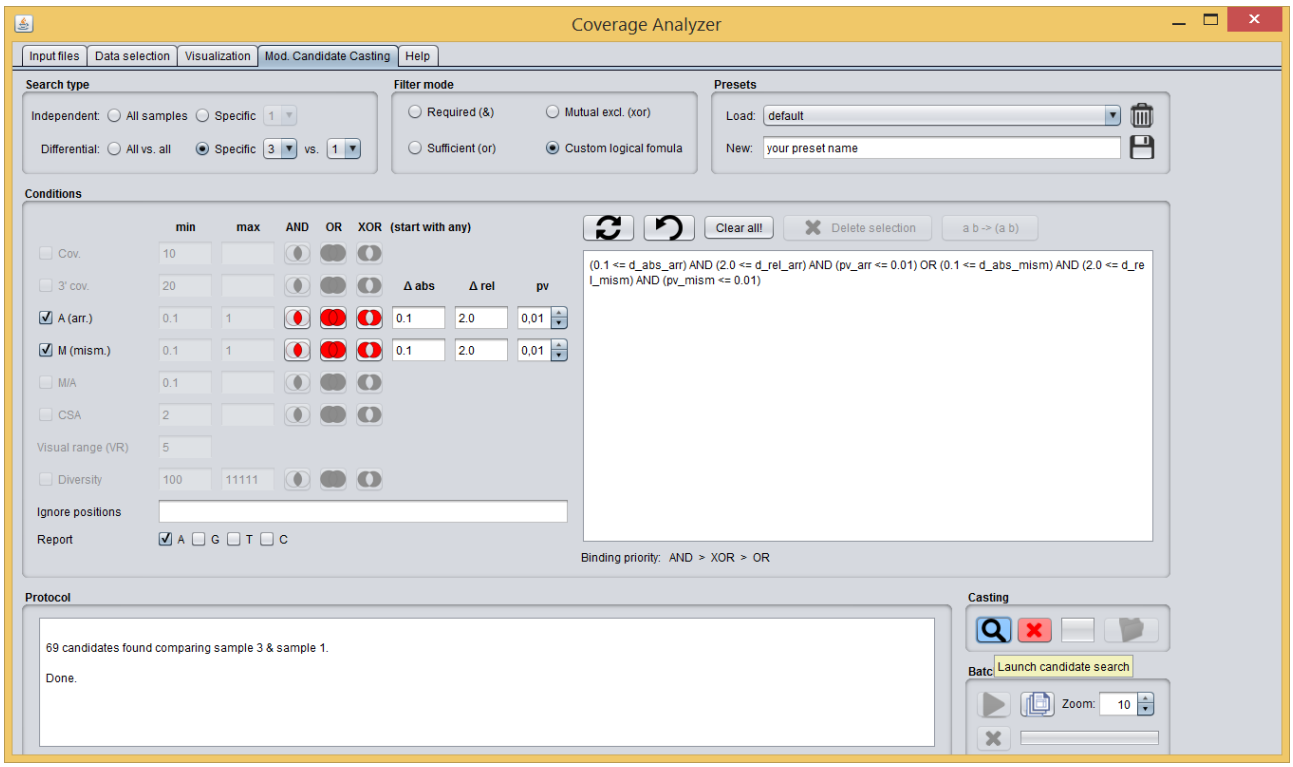


Figure 4: *Mod. candidate casting* tab. A formula for differential analysis of sample 3 vs. sample 1 was generated. By pressing the search button, 69 positions were written to a *Candidates* file on the hard drive.

3 Troubleshooting & Support

Users are encouraged to report any unexpected software behavior to ralf.hauenschild@uni-mainz.de. A common cause of error is invalid formatting of the *Profile* input file. In such case, verification of correct specification and completeness of the tab-separated data should help. In one case, a block of invalid characters turned up at an arbitrary positional line of a reference segment in *Profile*, leading to abortion of the loading process in the *Analysis* step. Reproducibility of this error is not tested yet, but manual correction cured the problem and allowed further study of the data. Other bugs may be resolved by vanishing , followed by restart of the *Analysis* routine (2.2).

4 References

[1] Inkscape Community. Inkscape.

[2] Ralf Hauenschild, Lyudmil Tserovski, Katharina Schmid, Kathrin Thüring, Marie-Luise Winz, Sunny Sharma, Karl-Dieter Entian, Ludivine Wacheul, Denis L. J. Lafontaine, James Anderson, Juan Alfonzo, Andreas Hildebrandt, Andres Jäschke, Yuri Motorin, and Mark Helm. The reverse transcription signature of n-1-methyladenosine in rna-seq is sequence dependent. *Nucleic acids research*, 43(20):9950–9964, 2015.

[3] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.