

Instructions for using VMir

Introduction:

VMir is a program to identify potential miRNA precursors in viral genomes or other nucleotide sequences of up to approx. 2 Mb (Grundhoff and Sullivan, 2011, for a review on virus-encoded microRNAs). Please be aware that, in contrast to other prediction algorithms, the program was designed to identify candidates for further screening by microarray analysis; it therefore deliberately over- (rather than under-) predicts potential pre-miRNAs (for further discussion, see Grundhoff et al., 2006). However, for users who do not wish to perform high throughput validation, the current version of VMir also includes a stringency adjustment tool that allows the user to eliminate many false positive predictions, at the expense of potentially losing some bona fide miRNAs (see the **Adjusting Score Stringency** section of this guide).

To predict pre-miRNA hairpins, the VMir program slides a window of adjustable size over the sequence of interest, advancing each window by a given step size. The secondary structures of RNAs corresponding to each window are predicted using the RNAFold algorithm (Hofacker et al., 1994). Hairpins with a size above a certain threshold are then identified and scored (for further details, see Grundhoff et al., 2006). The hairpins detected in each of the windows are recorded and, after completion of the structural analysis, are classified as Main, Subsidiary or Repeated Hairpins (MHPs, SHPs or RHPs, respectively) to reduce the complexity of the prediction. For this purpose, all recorded hairpins are compared to one another and grouped into local and repeat families. Local families consist of hairpins which form at the same location and which are structurally identical within a core region of a given size, but can differ in flanking regions due to their folding in different sequence windows. The longest hairpin within such a group is designated the MHP while all others are classified as the SHPs of the local family. Repeat families contain hairpins which are of identical sequence and structure (within a core region of given size), but form at different genomic locations. The 5'-proximal hairpin within a repeat group is designated the MHP of the group; all others are classified as RHPs (Individual hairpins, i.e. hairpins which do not belong to a group, are always classified as MHPs). The results of the analysis are written to a file which can be later visualized using a graphical user interface.

The VMir package consists of two individual programs:

- **VMir Analyzer:** Performs the actual analysis of sequence files and generates a results file (*.mir).
- **VMir Viewer:** Reads and filters results files written by VMir Analyzer, visualizes the results in form of a graph, allows visualization of individual hairpin structures/data, exports data to text files. From version 1.5 on, VMir Viewer also contains tools to design and analyze microarrays for high-throughput validation of miRNA predictions.

I. Using VMir Analyzer

Adding Sequence Files: Add sequence files to the list by clicking the "Add File" button (You can choose a single file or multiple files from the "Open File" dialog). The program can read GenBank, FASTA and raw nucleotide / text files and should automatically detect the right format (If you select a raw text file which contains invalid nucleotide characters, the program will remove all characters not conforming to the standard nucleotide code, but will not give any further warning; thus, you should make sure your raw nucleotide file does not contain any additional text). For each added file or group of files, you will be prompted with a dialog asking for the following file/job options (You can return to this options dialog later and make changes by selecting one or more files from the list and clicking the "Edit" button or double-clicking the files in the list.):

Output File: The file to which the results should be saved. This will default to the name of the input/sequence file with the added extension ".mir", saved to the same folder as the input file. You can change the name/location by clicking the "Output File" button (only accessible if a single file was chosen in the "Open File" dialog).

Conformation: the conformation of your input sequence, choose "Linear" or "Circular".

Orientation(s): The orientation(s) in which the program should look for hairpins, choose "Direct", "Reverse" or "Both"

Status: The current status of the file/job, will default to "New" for a newly added file but will change to "Done" once the file is processed. The program will skip all files with a "Done" status. So, if you want to analyze an already processed file again with different settings you will have to reset its status to "New"

Window Size: Define the Window Size (in nucleotides) for the analysis. The program will shift a window of this size over the input sequence. The default value is 500 nucleotides.

Step Size: Define the number of nucleotides by which the sequence window is shifted at each step during the analysis. The default value is 10 nucleotides.

Min./Max.

Hairpin Size: Choose "Any" or a specific value. During analysis, hairpins which are shorter or longer than the given values, respectively, will be rejected. The default value is 50; especially for large sequences this keeps the size of the results file relatively small by ensuring that hairpins which are smaller than bona-fide pre-miRNAs will not be recorded.

Min. Score: Choose "Any" or a specific value. During analysis, hairpins with scores lower than the given values will be rejected.

After you have added files, you can modify settings, delete or move files within the list by selecting the file(s) in the list and then clicking the corresponding buttons underneath the listbox (or right-clicking with the mouse and choosing from the pop-up menu).

Adjusting Global Settings: In addition to the file specific settings, there are a few global settings which apply to all files in the list; you can access/change these settings via the "Edit/Settings.." command in the menu bar. The settings are:

Max. HP Segment to calculate scores:

This is the maximum number of nucleotides which are considered when calculating hairpin scores, counting from the apex position of the hairpin. For scoring purposes, longer hairpins are trimmed down to this size.

Default value: 100 (Note that the default value was 120 in previous versions of VMir Analyzer, but was decreased to 100 in version 2.2 to prevent overpenalization of large bulges at the base of hairpins).

Max. HP Segment to assign SHPs:

Maximum size of the core region (counting from the apex position) which is considered while comparing hairpins for the assignment of local families / Subsidiary hairpins (SHPs, see Introduction).

Default value: 50.

Max. HP Segment to assign RHPs:

Maximum size of the core region (counting from the apex position) which is considered while comparing hairpins for the assignment of repeat families / Repeated hairpins (RHPs, see Introduction).

Default value: 50.

Batch Size:

The program will internally fold each window using the RNAfold algorithm (Hofacker et al., 1994). This number is the number of sequence windows simultaneously submitted to RNAfold (typically between 5 and 50). It is recommended to leave this on automatic; the program will then adjust the number automatically based on window size and machine performance.

Running the analysis: You can start the analysis by hitting the "Go" button in the lower right corner. The program will run through each file with the specified settings and write the results to the indicated output/results file. As the analysis proceeds, the program will provide you with a continuously updated progress report (you can pause or stop the analysis at any time by hitting the "Abort" button). The results files can later be visualized using the VMir Viewer program.

II. Using VMir Viewer

VMir Viewer allows you to view files previously written by VMir Analyzer. VMir Viewer provides information in three panels: i) a chart panel showing a graphic representation of the VMir results in form of a chart, ii) an info panel showing text-based information about the file and individual hairpins, iii) a picture panel showing a graphical representation of individual hairpin structures. You can resize the panels by clicking and dragging the bars which separate the panels. If your computer is slow, you can activate or deactivate the chart and/or picture panels under "Options>Show Chart" and "Options> Show Hairpin Structures" in the main menu bar.

Opening Files: To open a file, use the "File>Open" command on the main menu bar. VMir Viewer will read the file, filter the hairpins according to criteria set in the filter settings dialog (explained below), and display the filtered hairpins on the chart. The chart will show all main / MHP hairpins (as well as optionally also all repeated / RHP hairpins) as representatives of local hairpin families, according to position of the apex position within the sequence (shown on the X axis) and hairpin score (Y axis). MHPs in direct or reverse orientation are shown as solid blue triangles and green diamonds, respectively, while RHPs are shown as open symbols.

The names of the filtered hairpins will also be displayed in the "Filtered Hairpins" drop-down list in the info panel; the list can be sorted by either name or score via choosing the appropriate value from the drop-down list to the right of the hairpin list. A label next to the hairpin list indicates the number of filtered hairpins versus the total number of all individual hairpins contained within the VMir file (Whereas the filtered hairpins list consists of MHPs and (optionally) RHPs, the total number of hairpins includes SHPs as well; therefore the number of filtered hairpins is always smaller than the total number of hairpins, even when no further filters are applied).

You can display information (e.g. parameters of the analysis) about the currently open file by selecting "File>Info.." from the main menu.

Adjusting Filter Settings Filter settings can be changed in the filter settings dialog, accessed either via the "Edit>Filter.." command in the main menu bar or by choosing the "Edit Filter.." command from the pop-up menu after right clicking in the chart. In the filter settings dialog, you can adjust the following values:

Show RHPs: If checked, RHPs will be shown, otherwise all RHPs will be filtered out.

Orientation(s): Choose "Any" to display all hairpins, or "Direct" or "Reverse" to show only those in specific orientations.

Hairpin Score: Choose "Any" or specify a value; hairpins with scores lower than the given value will be filtered out. 115 is a good starting value, as the vast majority of bona fide pre-miRNAs reach values > 115.

Hairpin Size: Choose "Any" or specify a value; hairpins which are shorter or longer than the given value will be filtered out. We typically use 50 and 220 as lower and upper bounds, respectively.

Window Count: Choose "Any" or specify a value, The Window Count indicates in how many of the shifted sequence windows the hairpin or any of its

subsidiary hairpins were detected during the initial analysis. Higher numbers give higher confidence that the hairpin represents a stable structure within the specific sequence context (most bona-fide pre-miRNAs hairpins fold in 30 or more windows). We typically use values between 10 and 35; 15 is a good starting value.

As an example, Figure 1A shows the primary output from an analysis of the Epstein-Barr virus (EBV) genome, and Figure 1B shows the filtered output after selecting a minimal window count of 35 and a hairpin length between 50 and 220 nucleotides. The filter settings are saved when you close a file and exit VMir viewer; they will be restored when you re-start VMir or open another file.

Adjusting Score Stringency Although VMir was originally designed as a low-stringency *ab initio* prediction method which is optimally used in combination with high-throughput validation methods such as microarrays (see the Designing/Analyzing Microarrays sections below), starting with VMir Viewer version 1.6 the program also allows the user to increase the stringency of the predictions by imposing score penalties upon hairpins which have features that are unusual for *bona fide* pre-miRNAs (Grundhoff, 2011). The score stringency tool is based upon a statistical analysis of structural features of over 6500 mammalian and viral pre-miRNA hairpins listed in the release 14 of the miRNA registry (Griffiths-Jones, 2004). Hairpins which deviate significantly from this reference set in one or more categories (e.g., the number, size and symmetry of internal bulges, or the overall percentage of paired vs. unpaired nucleotides in the hairpin stem) are assigned a penalty that is subtracted from the hairpin score. In contrast to the original scoring system implemented by VMir Analyzer, this penalty system is much more restrictive and will eliminate false positive hairpins more efficiently. However, as it is based on an averaged and thus "ideal" feature set, many authentic pre-miRNAs will also be penalized and it therefore may be necessary to lower the score threshold.

To use the stringency adjustment feature, select "Adjust Score Stringency" from the "Edit" menu. A window with a slider control will open; move the slider from left to right to increase the stringency from its lowest (0%) to the highest possible value (100%). The effects of adjusting the stringency to 50% and 100% of its maximum value on the prediction of EBV pre-miRNAs can be seen in Figures 1C and -D. If one was to perform northern blotting confirmation of the highest scoring candidates conditions where no stringency adjustment is used (Figure 1B), the first two blots would identify authentic miRNAs, but the third would already represent a false positive prediction. In contrast, 9 and 12 bona fide miRNAs would be identified before encountering the first false positive when adjusting the stringency as shown in Figures 1B and 1C, respectively. Using a score cutoff value of 115 under conditions of maximum stringency (Figure 1C), however, would also eliminate several authentic pre-miRNAs. The score threshold in the filter settings (see previous section) should therefore be lowered to 50 when the score stringency adjustment is set to 100%.

Note that, in contrast to the filter settings, the stringency parameter is not saved and will default to 0 each time you open a VMir file.

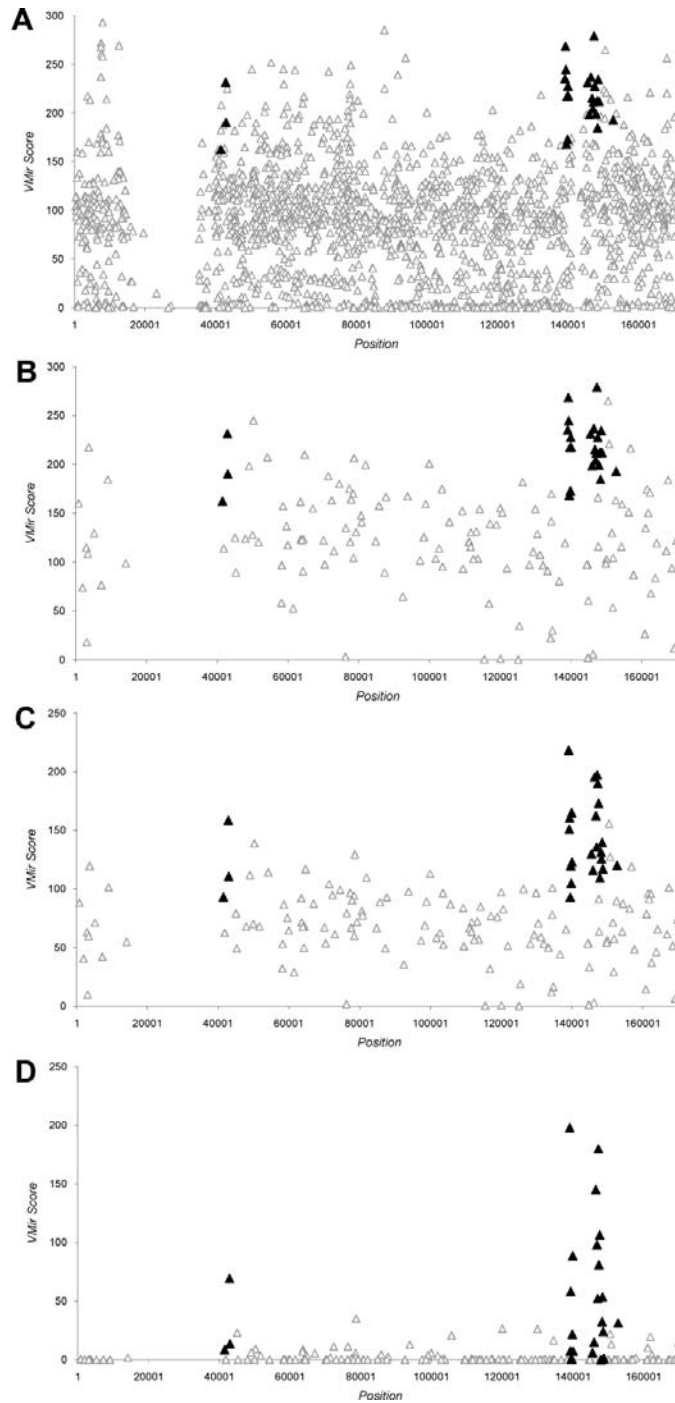


Figure 1: Adjustment of Filter Settings and Score Stringency (Modified after Grundhoff, 2010)

Depicted is the output from the prediction of pre-miRNAs encoded by Epstein-Barr virus (EBV, GenBank accession NC_007605) under different filtering and stringency conditions. Primary analysis was performed as described in Grundhoff et al, 2006. Only hairpins in direct orientation are shown. Solid black triangles indicate the 25 authentic (i.e. experimentally confirmed) pre-miRNA stem loops encoded by EBV. Open grey triangles represent false-positives. **A** shows all predicted hairpins, without any filtering or stringency adjustment. In **B**, the stem loop predictions are filtered to those which fold in 35 or more windows, and which are between 50 and 220 nt in size. In **C** and **D**, the stringency adjustment of the scoring algorithm was additionally set to 50 and 100%, respectively, of its maximum value.

Navigation within the Chart: The position of the hairpins within the sequence is shown on the X axis. The complete sequence is represented by a black bar underneath the chart; a grey rectangle indicates the location of the segment which is currently displayed in the chart within the sequence.

You can zoom the chart window in three ways: i) by right-clicking in the chart window and selecting "Zoom in" or "Zoom out" from the pop-up menu; this will zoom in on the position of the mouse cursor, ii) by left-clicking in the chart and dragging open a window while keeping the button pressed; the window will be indicated by a blue rectangle and once the mouse button is released the chart will be zoomed to that window, iii) if a hairpin is selected, scrolling with the mouse wheel will zoom in or out on that hairpin. In each case, if you keep the control/ctrl key pressed while selecting the zoom operation, only the X axis will be zoomed. If zoomed in, the chart window can be panned left, right, up or down by clicking in the appropriate area of the chart panel, outside of the plot; the mouse cursor will change to an arrow indicating the panning direction. Selecting "Reset Zoom" from the pop-up menu after right clicking in the chart will reset the X as well as Y axis.

The upper and lower bounds of the Y and X Axis can be set in two ways: either double-clicking on the axis or selecting "Set Axis>Set X Axis" or "Set Axis>Set Y Axis" will bring up a menu in which you can define the minimum and maximum values for that axis.

You can also copy a picture of the chart (as a bitmap) to the clipboard by right-clicking on the chart and selecting "Copy Picture" from the pop-up menu.

Searching for Sequences: To find a specific sequence, click the "Find Sequence.." Button or select "Find Sequence" from the Tools menu. After pasting the sequence in the text field of the search dialog, click the "Find Matches" button. The list box at the bottom of the dialog will show the location and orientation of all matches. If you scroll through the matches and "Preview" is checked, the selected match will be highlighted in the VMir chart (matches in direct orientation will be shown in orange, those in reverse orientation in green). If you click the "OK" button (or double-click a given match) the selected region will stay highlighted. To remove the selection, right click in the chart, then select "Clear Selection" from the context menu.

Viewing Individual Hairpins: You can select individual hairpins by either left-clicking on them in the chart or selecting them from the "Filtered Hairpins" drop-down list in the info menu (You can search for hairpins by name using the "Find Hairpin.." Button in the info panel).

Once you select a hairpin, the info panel will display the following hairpin information under "Selected Hairpin" (You can also either copy all of this information to the clipboard or save it to a text file by clicking on the corresponding buttons in the upper right corner of the "Selected Hairpin" info section):

- Hairpin *Name*, *Orientation* (direct or reverse), *Size*, *Location* (start and end position of the hairpin)
- Hairpin *Score* and *Rank* of the hairpin within the filtered hairpins

- *SHPs* and *RHPs*: Indicates the numbers, if any, of subsidiary or repeated hairpins of the current hairpin.
- Absolute and relative Window Count: *Wind. Cnt. Abs.* and *Rel.*: These values indicate in how many of the shifted sequence windows either the currently selected main/MHP hairpin only (*Abs.*) or the MHP as well as any of its subsidiary/SHP hairpins (*Rel.*) were detected during the initial analysis.

In addition, text-based sequence and structural information will be displayed in the bottom part of the info panel:

- The DNA sequence of the region spanned by the selected hairpin and, underneath that sequence, structural information for the selected hairpin as well as all subsidiary hairpins of the local family in form of dots (unpaired nucleotides), open brackets (paired nucleotides belonging to the 5p arm) or closed brackets (paired nucleotides belonging to the 3p arm). The number of windows in which each of the hairpins was detected during the initial analysis is indicated by the numbers in brackets after the hairpin names.
- A text-based structure map of the selected hairpin.

When a hairpin is selected, the picture panel will also show a graphical representation of the hairpin structure (If your computer is slow, you can either deactivate the panel under Tools> Show Hairpin Structures or switch of automatic drawing of the hairpins under Tools> AutoDraw Structures; in the latter case, the program will not automatically draw a hairpin once it is selected but only if you double-click in the picture panel). You can copy the picture as a bitmap to the clipboard by right-clicking on the picture.

By default, the program will draw the complete hairpin. However, you can display only a portion of the hairpin by checking "Trim Size" at the bottom of the picture panel; you will be prompted to select a size value (this defaults to the *Max. HP Segment to calculate score* size used during the initial analysis, i.e. the segment of the hairpins that was considered for the calculation of the VMir scores). Longer hairpins will be trimmed down (starting from the apex) to the given value. The value can be changed by clicking on the label showing the size value.

Exporting Data: You can export the data for all filtered hairpins by choosing the "File> Export.." command from the main menu bar. You will be prompted for the name of a text file to which the following data will be saved in a TAB delimited format (which you can open in, e.g., Excel):

- A header providing information about the analysis file created by VMir Analyzer and the parameters of the analysis, as well as the current filter settings of VMir Viewer.
- A block containing the following data for all filtered hairpins:
Hairpin Name and Orientation, Start Position of the hairpin, Position of the hairpin Apex (=center position of the terminal loop) , hairpin Size, hairpin Score and Rank, number of *SHPs* and *RHPs*, absolute and relative window count (*Wind.Cnt.Abs./Wind.Cnt.Rel.*) of the hairpin.

Designing Microarrays: To design a microarray based on a given VMir analysis, open the results file in VMir Viewer, then select "Microarrays > Microarray Design.." from the "Tools" menu. The software will design probe and control oligos as follows (see figure 2 for a depiction of the general design outline): Two specific oligo probes will be generated for each hairpin, one for the left and one for the right arm, covering the positions of putative 5p- and 3p-miRNAs. To account for inaccuracies in the structure prediction, the default size of these oligos is significantly larger than the typical size of a mature miRNA (30 instead of 22 nt). The oligos are designed such that they end or start (for 5p- or 3p-oligos, respectively) at the junction of hairpin stem and terminal loop. In addition to the specific oligos, three types of control oligos can be generated: scrambled, mismatched and flanking controls. Scrambled oligos are generated by random shuffling of 5p or 3p oligo sequences. Scrambled oligos will have the same GC content as the specific oligo, but they will not control for cross-hybridization due to sequence similarity. Therefore, the software will only design mismatched and flanking control oligos by default. Mismatched oligos are generated by introducing mismatches into the specific miRNA sequences. The frequency with which these mismatches are introduced can be adjusted by the user. Flanking oligos are picked adjacent to the specific oligos.

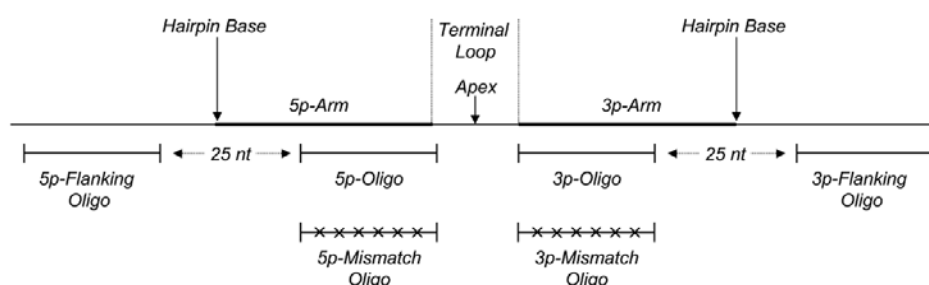


Figure 2: Outline of microarray probe and control oligo design

Shown is a linear representation of a miRNA hairpin. Sequences comprising the 5p- and 3p-arms of the stem are shown as thick lines whereas the terminal loop and regions which flank the hairpin are shown as thin lines. The position of specific oligonucleotides (5p- and 3p-oligos) as well as default controls (5p-/3p-mismatch and flanking oligos) is indicated.

While mismatched (or scrambled) oligos control for cross-hybridization, flanking oligos are intended to account for unspecific signals owing to, for example, small degradation products of transcripts which encompass the hairpin site. Such products will not be enriched in the region encoding the putative miRNA and therefore also should hybridize to the flanking control. The design dialog allows the user to adjust the following settings:

miRNA Probes

- *Oligo Size:*

Determines the size of all oligos (specific miRNA probes as well as control oligos). The length of additional linkers attached to the 5'- and/or 3'ends of the oligos (see below) does not count towards this size limit.

- *Measure from Terminal Loop / Hairpin Apex:*

Determines whether the software picks 5p- and 3p-oligos such that they end (5p-oligo) and start (3p-oligo) at the positions where the terminal loop opens or closes (default), or end and start at the hairpin's apex position.

- *Offset 5p/3p-oligo:*

Positive offset values will shift the oligo towards the 3'-end of the hairpin (thus 5p-oligos will move towards the terminal loop whereas 3p-oligos will move away from it). To account for inaccuracies in terminal loop prediction and ensure complete miRNA coverage, by default the 5p-oligo is offset by 2 nt whereas no offset is used for the 3p oligo. Given that 5p-miRNAs usually terminate at the stem/loop junction whereas 3p-miRNAs start 2 nt into the stem, this corresponds in both cases to oligos which are extended by 2 nt towards the terminal loop,

Scrambled Control Oligos

- *Generate '#' scrambled Oligos*

Determines how many scrambled oligos will be generated by random shuffling of the specific 5p and 3p oligos.

Mismatch Control Oligos

- *Introduce Mismatch every '#' nt:*

Defines the frequency with which mismatches are introduced into the specific oligo sequence (the respective nucleotide will be replaced by its complementary nucleotide). Counting starts from the position which is proximal to the terminal loop. The default is one mismatch introduced every 5 nucleotides.

Flanking Control Oligos

- *Generate '#' flanking oligos each in 5p/3p position:*

Defines the number of oligos which are picked adjacent to the 5p and/or 3p oligos. If more than one oligo is selected, the oligos will be arranged in a tiled fashion with no overlap between adjacent oligos. In cases where a flanking control oligo overlaps with specific oligos from a neighboring hairpin, the control oligo is shifted further away from the hairpin until the overlap is eliminated. By default, VMir picks one flanking oligo each adjacent to the specific 5'- and 3'-oligos.

- *Distance: '#' nt from miRNA Oligo/Hairpin Base:*

Determines the distance at which the first flanking oligo is picked, and whether the distance is measured from the location of the specific miRNA oligos or the base of the hairpin. By default, flanking oligos are picked in a distance of 25 nt from the miRNA oligos (Increasing the distance or calculating the distance from the base of the hairpin would be useful if the size selection procedure used to produce the hybridization material does not efficiently exclude pre-miRNAs, as these precursors might overlap with and hybridize to the flanking oligos).

Attach Linkers

- 5'-Linker / 3'-Linker:

The specified linker sequences will be attached to each of the specific and control oligos in the indicated position.

In addition to the above options, the number of identical replicates generated for each of the specific or control oligos can be adjusted by the user. A counter in the top/left corner of the dialog informs about the total number of oligos which will be generated when the current settings are used.

Once you hit the "Generate Oligos" button, the program will generate a tab-delimited text file with the names and sequences of the individual oligos. The oligos are named by a string consisting first of an identifier for the hairpin (e.g., "MD1540") followed by a string identifying the type of oligo (e.g., "_5p", "_5pSc", "_5pMm" or "_5pFI" for specific, scrambled, mismatched or flanking 5p-oligos, respectively) followed by a number identifying different oligos of the same type, followed by an indicator for replicates (if any). For example, "MD1540_5pFI_1_R2" indicates the second replicate of the first (i.e. proximal) flanking oligo next to the 5p-position of hairpin MD1540.

Analyzing Microarrays: Owing to the design of the array (i.e., individual control spots for every specific oligo), it is not feasible to analyze the arrays using commercial array analysis software. Therefore, VMir incorporates a tool which reads the results generated by the primary array scanning/analysis software, then matches the oligos to the original hairpins and integrates the results. VMir is natively able to read result files generated by Genepix (*.gpr files), but it accepts any tab-delimited text file. Nearly every array analysis software will allow you to export the scanned data in such a format (consult your manual for specific instructions). In order to read a file, the data for individual oligos have to be arranged in rows, with the specific data items for each oligo arranged in tab-separated columns. VMir minimally requires two columns: one which contains the oligo names, and one which contains the data representing the array signals (e.g., the signal from one or two fluorescent channels, or a ratio of two channels calculated by the primary array analysis software).

To open the tool, select "Microarrays > Process Microarray Data..." from the "Tools" menu. After selecting the file containing the array data, you will see the import dialog shown in Figure 3A. From this dialog, you will have to identify the header row which precedes the actual data (or the first data row if there is no header row; make sure to

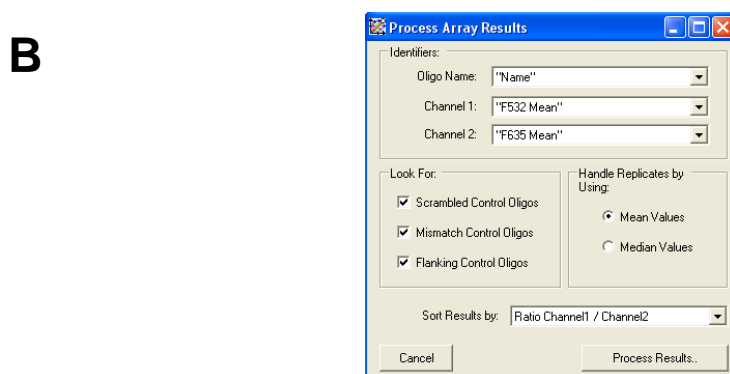
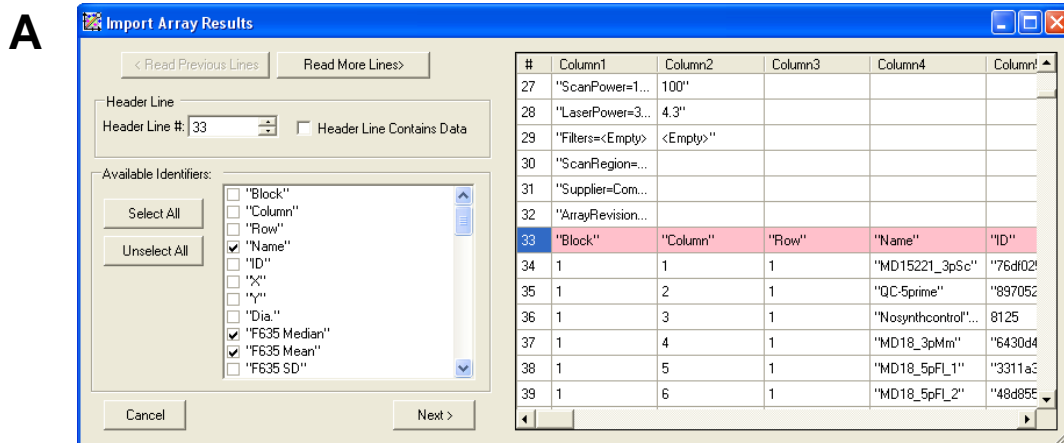


Figure 3: Screenshots of microarray import and processing dialogs

The data import dialog shown in (A) defines which data columns are imported and analyzed according to the settings defined in the data processing dialog (B).

check "Header Line Contains Data" in such a case) and identify the columns which contain the oligo name and fluorescence data. VMir reads the file in portions of 500 lines. The file contents are shown in a spreadsheet in the right half of the dialog. If necessary, click the "Read Previous Lines" or "Read Next Lines" buttons to navigate through the file. VMir will attempt to identify the header line automatically. If the automatic identification fails, you will have to identify the header manually by entering the appropriate line number in the field in the top left corner of the dialog, or by double-clicking on the appropriate line within the spreadsheet.

Once you have selected the header line, the individual columns will be listed in the "Available Identifiers" field in the lower left part of the dialog. Every column which is checked in this field will be available as a data column in the next dialog. If the result file is in the Genepix format (*.gpr), the software will automatically select the columns which contain the oligo name and fluorescence data. The software will also recognize the columns containing Flag and Auto-Flag values, and automatically filter out all flagged spots. However, if you are using software other than Genepix for the primary array analysis, you will have to ensure that the result file only contains non-flagged spots.

After clicking the "Next" button, you will see the "Process Array Results" dialog shown in Figure 3B. In this dialog, you have the following choices:

Identifiers:

From the contingent of column identifiers selected in the previous dialog, choose the columns in which VMir should look for the oligo name and the data for the first and second fluorescence channels. Select "None" for the second channel if you are performing a one color experiment, or if the first data column already contains ratio values.

Look for:

By default, the program searches for all types of control oligos (scrambled, mismatched and flanking), but you can manually select or de-select individual control types. De-selected controls will not be used in the subsequent analysis.

Handle Replicates by Using:

If your array contains multiple replicates of a given oligo, you can select whether VMir should handle them by calculating mean or median average values (default is mean). The same method will also be used to calculate average values for multiple different scrambled or flanking oligos.

Sort Results by:

Here you can select whether the results should be sorted according to the values of a single channel or whether VMir should calculate ratio values.

After you click the "Process Results" button, you will be prompted to provide a name for an output file. The program then reads the file containing the array data and identifies the oligos based on the naming scheme described under "Designing Microarrays". After matching the oligos to the individual hairpins, the program will calculate average values for each set of specific probes and control oligos, using the method chosen in the data processing dialog (mean or median average calculation). Average values will be calculated first for replicates of identical oligos, using the method chosen in the data processing dialog (mean or median averaging). Depending on the settings chosen during the microarray design, there might be also multiple different scrambled or flanking oligos. If so, the same averaging method will be used to calculate representative values for these categories. Scrambled oligos are considered sequence specific controls for 5p- and 3p-probes, thus separate 5p- and 3p-correction values will be calculated. In contrast, for flanking controls a single average correction value will be calculated from all oligos, regardless of whether they were originally chosen adjacent to the 5p- or the 3-p arm of the hairpin. For each of the 5p- and 3p-probes, the control type with the highest value is then selected and its value is subtracted from that of the specific oligo to calculate a corrected value. However, to ensure that the correction procedure does not produce exaggerated or negative ratios (and to prevent division by zero problems), only the value fractions which are above a minimal background hybridization value are subjected to correction. This background hybridization value is calculated individually for each of the fluorescence channels and represents the median value calculated from all control oligos across the array (if you should use an array without any controls, this value will be set to 1). For example, if the background hybridization value in one of the channels is 300, and the values for a specific probe is 600 whereas the maximum value from the controls is 400, the specific probe value would be corrected by the fraction of the control value which is above 300 (i.e. 100) to 500. Assuming the probe signal would be 350, it

```

VMir Hairpin Array Analysis
Source File: rLCV_HPArray.gpr
Header Line #: 33
Header Line Contains Data: False
OligoName ID: Name
Channel1 ID: F635 Mean
Channel2 ID: F532 Mean
Score By: Channel1/Channel2
Calc. Average Method: Mean
Use Scrambled Controls: True
Use Mismatched Controls: True
Use Flanking Controls: True
No. of Hairpins: 3848
No. of Oligos: 45832
No. of Spots: 183328
No. of Flagged Spots: 86
Avg. Backgrnd. Channel1: 305
Avg. Backgrnd. Channel2: 303

Hairpins:

Name      Score    Score 5p    Score 3p    Probe Corrected 5p    Ch1    Ch2    3p    Ch1    Ch2    Probe Average 5p    Ch1    Ch2    3p    Ch1    Ch2    Controls Average 5p    Mismatch    Ch1    Ch2    3p    Mismatch    Ch1    Ch2    5p    Scrambled    Ch1    Ch2    3p    Scrambled    Ch1    Ch2    Flanking    Ch1    Ch2
MD1432    50.762    50.762    0.98    15545.8    306.25    297    303    15553.5    315.75    297    312.5    312.75    312.5    290    313.75    295.125    292.5    299.125    296.625    305    308.688
MD1428    46.973    38.646    46.973    11468.3    296.75    14608.8    311    11478.8    296.75    14610.5    321.5    315.5    306.75    306.75    313.5    299    298.25    305.125    301.75    282    297
MD1441    44.41    44.41    0.963    12712.5    286.25    305    316.75    12726.5    286.25    329.5    316.75    319    313    277.5    276.5    308    309.375    330    300.625    297.688    301.25
MD1434    38.141    1.095    38.141    323.625    295.5    11556.8    303    326.75    295.5    11556.8    303.5    266.25    295.25    300    305.75    308.125    303.375    292.5    296.875    297.875    306.938
.....
MR1820    1.153    1.153    1.01    349.5    303    349.5    303.75    291.5    288.5    291.5    288.5    287    302.5    304    309.375    321.25    304.25    294.25    315.125    285.438    297.875
MD986     1.153    0.958    1.153    290.25    303    290.25    311.75    334.75    290.25    334.75    290.25    283.5    312    312.625    306.75    295.75    283    296.875    298    290.375    310.375

Spots:

Name      Channel1    Channel2    Flags
MD1_5p_R1    248    289    0
MD1_5p_R2    293    289    0
MD1_3p_R1    353    324    0
MD1_3p_R2    322    288    0
MD1_5pSc_1_R1    355    334    0
MD1_5pSc_1_R2    289    295    1
MD1_5pSc_2_R1    312    342    0
MD1_5pSc_2_R2    342    332    0
.....

```

Figure 4: Layout of the output file

The file is in a standard tab-delimited format and contains the following elements:

- a header which shows settings and statistics for the current analysis, including background hybridization values for each of the channels
- a block which lists all hairpins in descending order according to their score. The columns contain the following data: *Name*: Hairpin name, *Score*: the score assigned to the hairpin, *Score 5p* and *Score 3p*: background corrected scores for 5p and 3p probes, *Probe corrected: 5p* and *3p*: corrected values in each channel for 5p- and 3p-probes, *Probe Average: 5p* and *3p*: average values in each channel for 5p- and 3p-probes before correction, *Controls Average: 5p* and *3p-Mismatch*, *5p* and *3p-Scrambled*, *Flanking*: Average values for each of the control types.
- a list of all hairpin-specific spots on the array (sorted by name) together with the original values in each of the channels and a value which indicates whether the given spot was flagged (0 means no flag, whereas 1 means the spot was flagged)

would be corrected down only to the background value of 300, but not any further. No correction at all will be performed if the probe signal is below the background hybridization value of the given channel.

Following value correction, ratios are calculated and awarded to the 5p- and 3p-probes as a score (if ratio calculation is not desired, the score simply represents the corrected value from the selected channel). The hairpin is given the score of either its 5p- or 3p-probe, whichever is greater. The results of the analysis are subsequently written to a file which lists the hairpins, sorted according to their score, along with the corrected and original values of the specific 5p- and 3p oligos, as well as the averaged values of each of the control types (see figure 4 for an example of the output). Following the hairpin data, the file also lists all individual spots along with their fluorescence and flag values. The file is in a standard, tab delimited text format which can be opened, annotated, edited and re-sorted in other applications such as Excel. Based on these results, the top scoring candidates can then be chosen for final confirmation by northern blotting.

REFERENCES

Hofacker I. L., Fontana W., Stadler P. F., Bonhoeffer S., Tacker P., Schuster P. (1994). Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f Chemie* 125, 167-188.

Griffiths-Jones S. (2004). The microRNA Registry. *Nucleic Acids Res* 32 (Database issue):D109-11.

Grundhoff A, Sullivan CS. (2011). Virus-encoded microRNAs. *Virology*. 411(2):325-43. Epub 2011 Jan 31.

Grundhoff, A. (2011). Computational prediction of viral miRNAs. *Methods Mol Biol.*;721:143-52.

Grundhoff A., Sullivan C. S., Ganem D. (2006). A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA* 12(5):733-50. Epub 2006 Mar 15.

Sullivan C.S., Grundhoff A. (2007). Identification of viral miRNAs. *Methods Enzymol* ;427:3-23.

Walz N, Christalla T, Tessmer U, Grundhoff A. (2010). A global analysis of evolutionary conservation among known and predicted gammaherpesvirus microRNAs. *J Virol*. 84(2):716-28. Epub 2009 Nov 4.

.