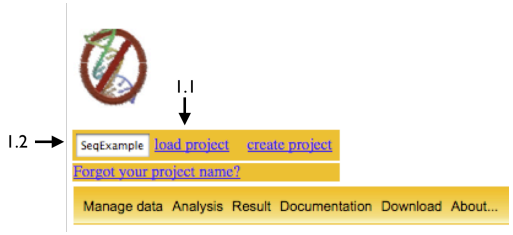
Tutorial 4: Web Server Analysis Module

Once the raw data have been processed and annotated (see Tutorials 1, 2 and 3), three main types of analysis may be performed through the SeqBuster web server: 1) a general characterization of small-RNA datasets, 2) deep characterization of miRNAs variability (IsomiRs), and 3) differential expression analysis. In the present tutorial, a practical example is shown in pre-analyzed data resulting from deep sequencing of undifferentiated (hESC) and differentiated human embryonic stem cells (EB) short-RNAs (Morin et al., 2008, Genome Res. 18(4):610-21).

1. Load/Create a Project:

The project contains the pre-analyzed sequencing data. In the present example the project has been named "SeqExample" (see 5.1. section of Tutorial 1). Notice that the project name is case sensitive.

1.1. Introduce "SeqExample" in the textbox at the upper left part of the web page
1.2. Click on the "load project" link.



2. Basic analysis: "General Information" package

This package offers the analysis of the distribution of the sequences in different lengths or classes

2.1. Go to the "Analysis" option in the menu and choose "Basic analysis"
2.2. Go to the "General Information" package
  2.2.1. In the "Select Samples" panel, choose a sample, for instance "hESC (Morin et al 2008, Genome Res.18(4):610-21)"
  2.2.2. Press the "add" button. Repeat 2.2.1 and 2.2.2 to add the "EB (Morin et al 2008)" sample
2.3. In present example, several parameters are chosen in the "Options" panel to analyze the length distribution of the miRNAs. Question marks in all the analysis options describe the characteristics of the parameter
  2.3.1. "Length" and "Frequency" filters the sequences to be considered in the analysis according to specific cutoff value of size and frequency, respectively. In the present example, all sizes and frequencies are being considered, since no cutoff is established.
  2.3.2. The "Locus" textbox is used to analyze all sequences mapped in a specific locus. In the present example all loci will be analyzed (empty textbox)

2.3.3. In the “Type of DB” text box “miRNA” has been introduced in the present example. Otherwise, all annotated databases would be considered in the analysis.

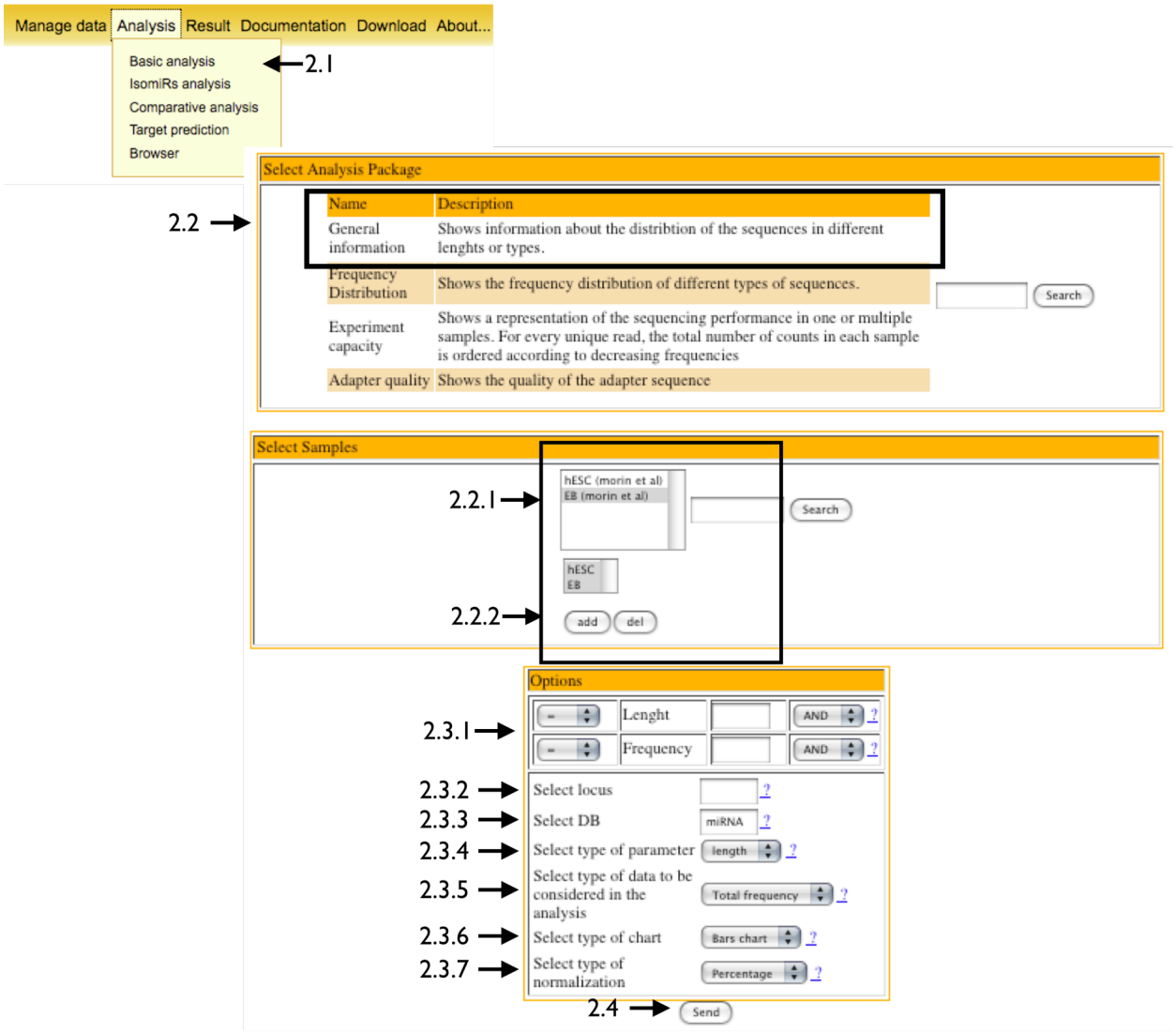
2.3.4. In the “Type of distribution” select box the “length” option has been chosen in the present example in order to represent the length distribution.

2.3.5. In the “Select the type of parameter” select box “Total frequency” has been chosen in the present example

2.3.6. In the “ Select type of chart” select box the “Bars chart” option has been chosen in the present example

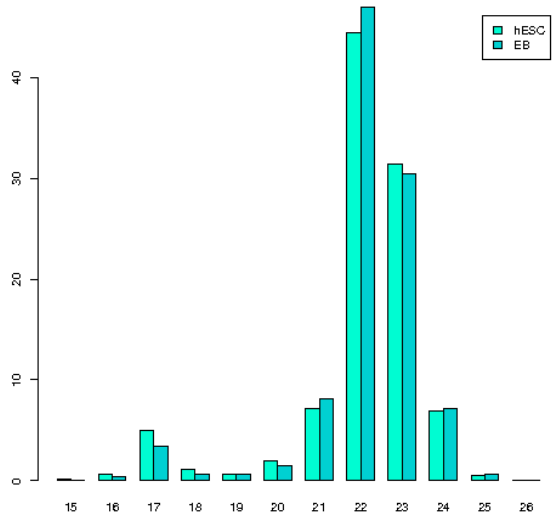
2.3.7. In the “Select type of normalization” select box the “Percentage” has been chosen in the present example

2.4. Press the “Send” button.



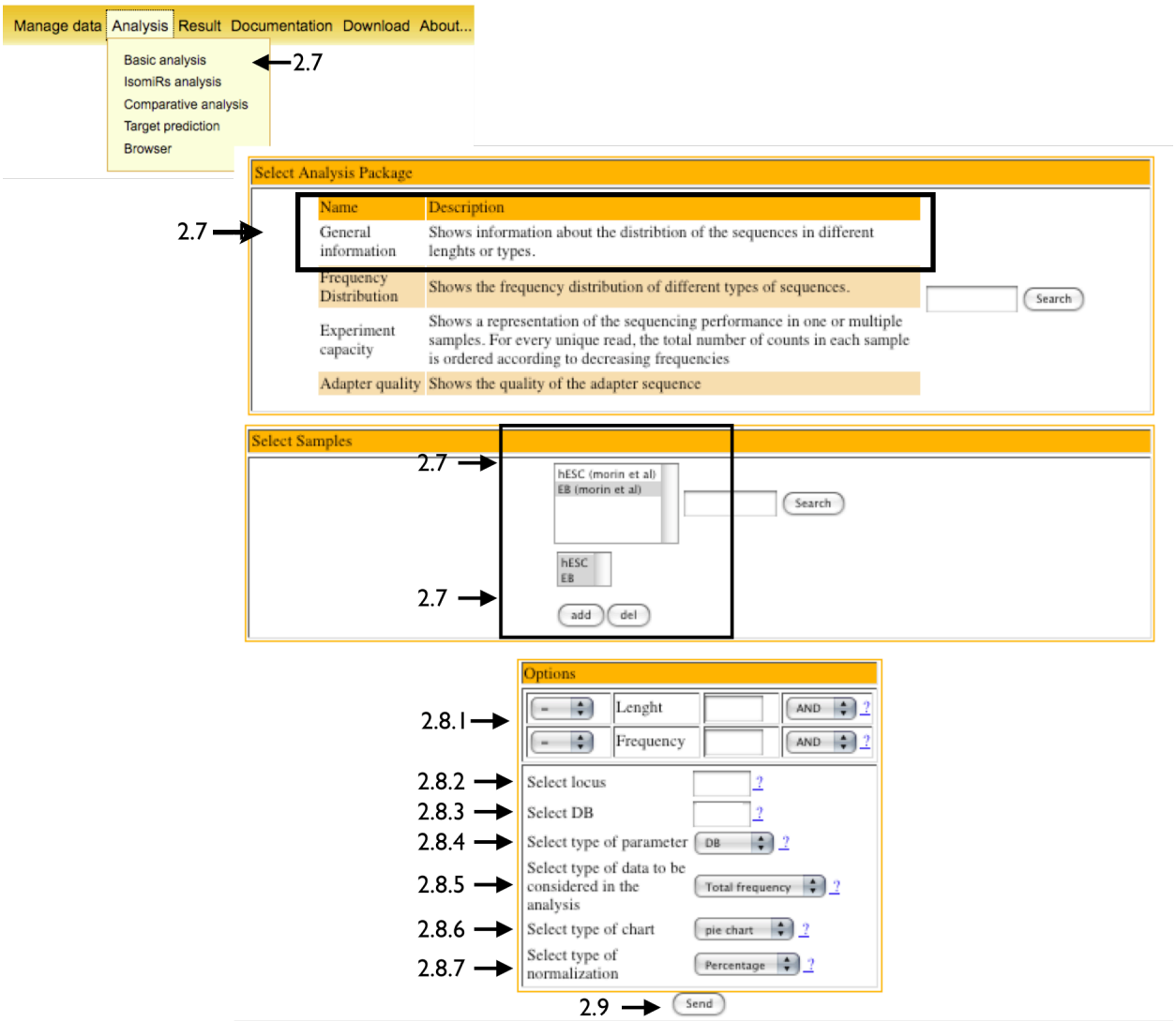
2.5. In the analysis output the following histogram appears, showing the percentage of sequences (in this case miRNAs) showing the indicated nucleotide length (x-axe):

Figure 1



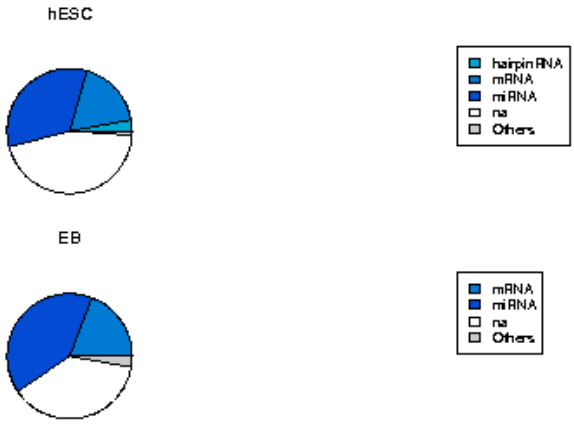
2.6. In the next example the percentage of the different types of sequences in the hESC and EB samples is analyzed.
2.7. Proceed as in the steps 2.1. and 2.2.
2.8. In the “Options” panel, several parameters have been chosen to analyze the percentage of the different types of sequences
  2.8.1. No cutoff for “Length” and “Frequency” have been established (textbox empty); therefore all sizes and frequencies are considered.
  2.8.2. The “Locus” option has not been specified (empty textbox), therefore all loci are considered
  2.8.3. In the “Type of DB” select box the “DB” database has been chosen in the present example which means that all available databases will be considered
  2.8.4. In the “Type of distribution” select box the “DB” option has been chosen in the present example in order to studying the different type of sequence in both libraries.
  2.8.5. In the “Select the type of parameter” select box “Total frequency” option has been chosen
  2.8.6. In the “ Select type of chart” select box the “Pie chart” option has been chosen
  2.8.7. In the “Select type of normalization” select box the “Percentage” has been chosen in the present example

2.9. Press the “Send” button



2.10. In the analysis output the following pie chart appears, showing the percentage of different types of small RNAs in hESC (right pie chart) and EB (left pie chart):

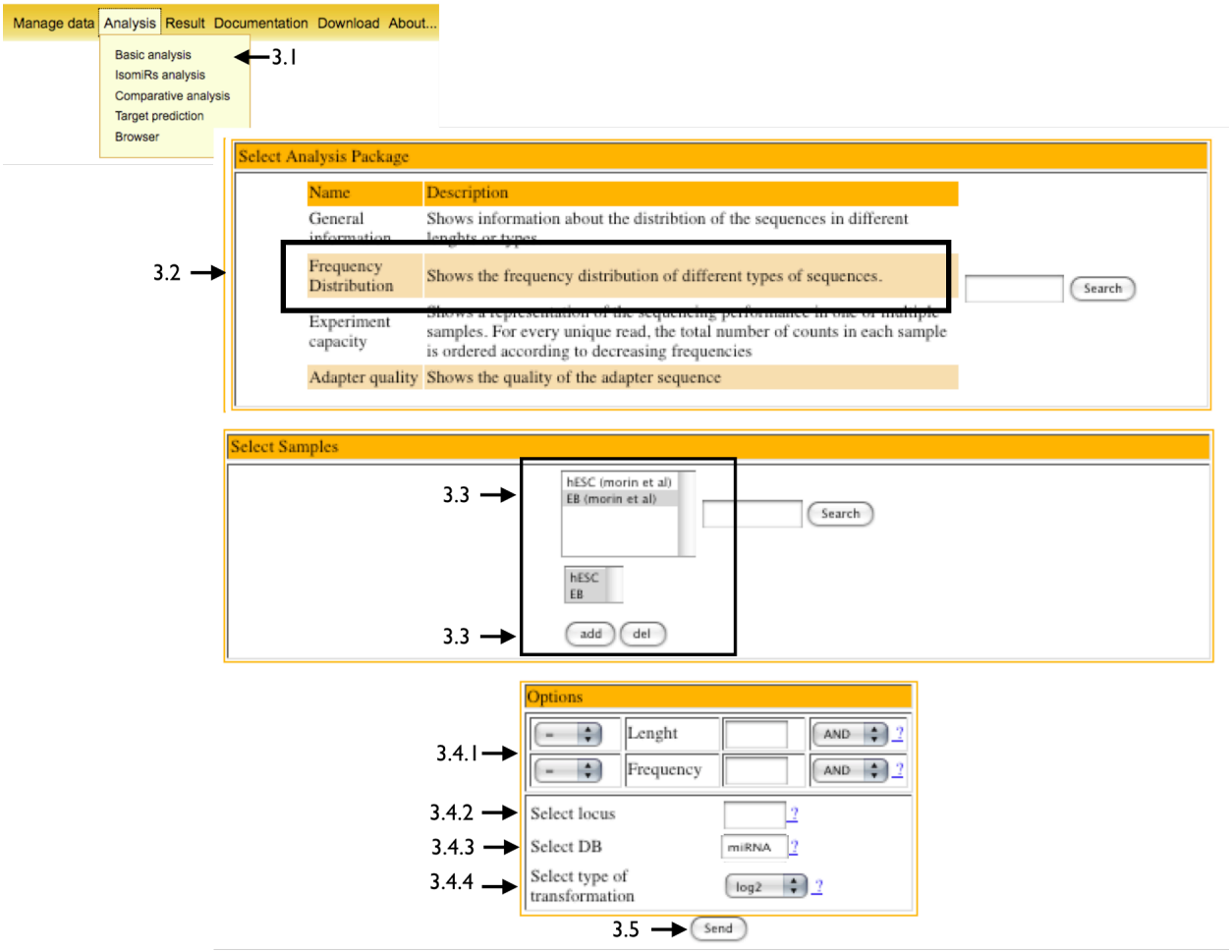
Figure 2



3. Basic analysis: "Frequency distribution" package

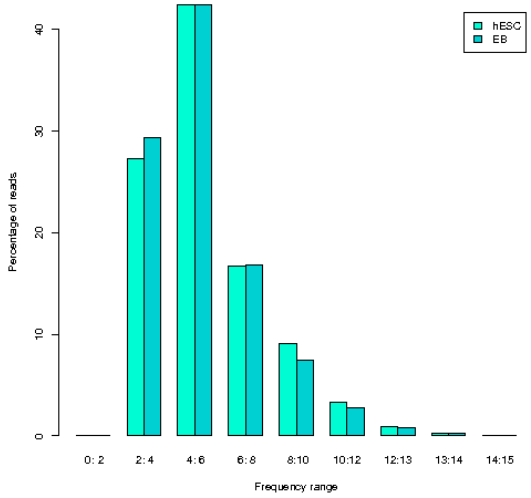
This package provides an analysis of the frequency distribution of different types of sequences. In the next example the frequency distribution of all sequences mapped as miRNA is analyzed

3.1. Go to the "Analysis" option in the menu and choose "Basic analysis"
3.2. Select "Frequency distribution" in the "Select Analysis Package" panel
3.3. Proceed as in steps 2.2.1. and 2.2.2 to select the samples
3.4. In the "Options" panel, several parameters may be chosen to analyze the frequency distribution of miRNAs
  3.4.1. No cut-off for "Length" and "Frequency" has been established (textbox empty). In the present example, all sizes and frequencies are being considered
  3.4.2. The "Locus" option has not been specified (empty textbox) mining that all miRNA loci are considered in the analysis
  3.4.3. In the "Type of DB" select box the "miRNA" database has been chosen
  3.4.4. In the "Select type of transformation" select box "log2" has been chosen, that represents the logarithm in base 2 of the frequency of the sequences
3.5. Press the "Send" button



3.6. In the analysis output the following histogram appears, showing the percentage of miRNAs (y-axe) in hESC (blue bars) and EB (green bars) with a range of frequencies (logarithmic transformed, x-axe):

Figure 3



4. Basic analysis: "Experiment capacity" package

This package explores the sequencing performance or sequencing capacity in different samples to know whether or not they are comparable in differential expression analysis (see section 12 and 13 of the present tutorial). To this aim, all sequences are ordered by decreasing frequency in each sample. If the resulting graph is similar between samples, these are comparable for differential expression analysis. If the sequencing capacity is different, several parameters may be modified to make possible a differential expression analysis

4.1. Go to the "Analysis" option in the menu and choose "Basic analysis"

4.2. Select "Experiment Capacity" in the "Select Analysis Package" panel

4.3. Proceed as in steps 2.2.1. and 2.2.2. to select the samples

4.4. In the "Options" panel, several parameters may be chosen to analyze the sequencing performance.

4.4.1. No cutoff for "Frequency" has been established (textbox empty). In the present example, sequences with all ranges of frequencies are being considered.

4.4.2. In the "Scale" textbox 1.000.000 is chosen. This means that the frequency for every sequence is scaled according to the following equation: Scaled freq n= (freq n/ sum[freq all seqs]) x 1.000.000

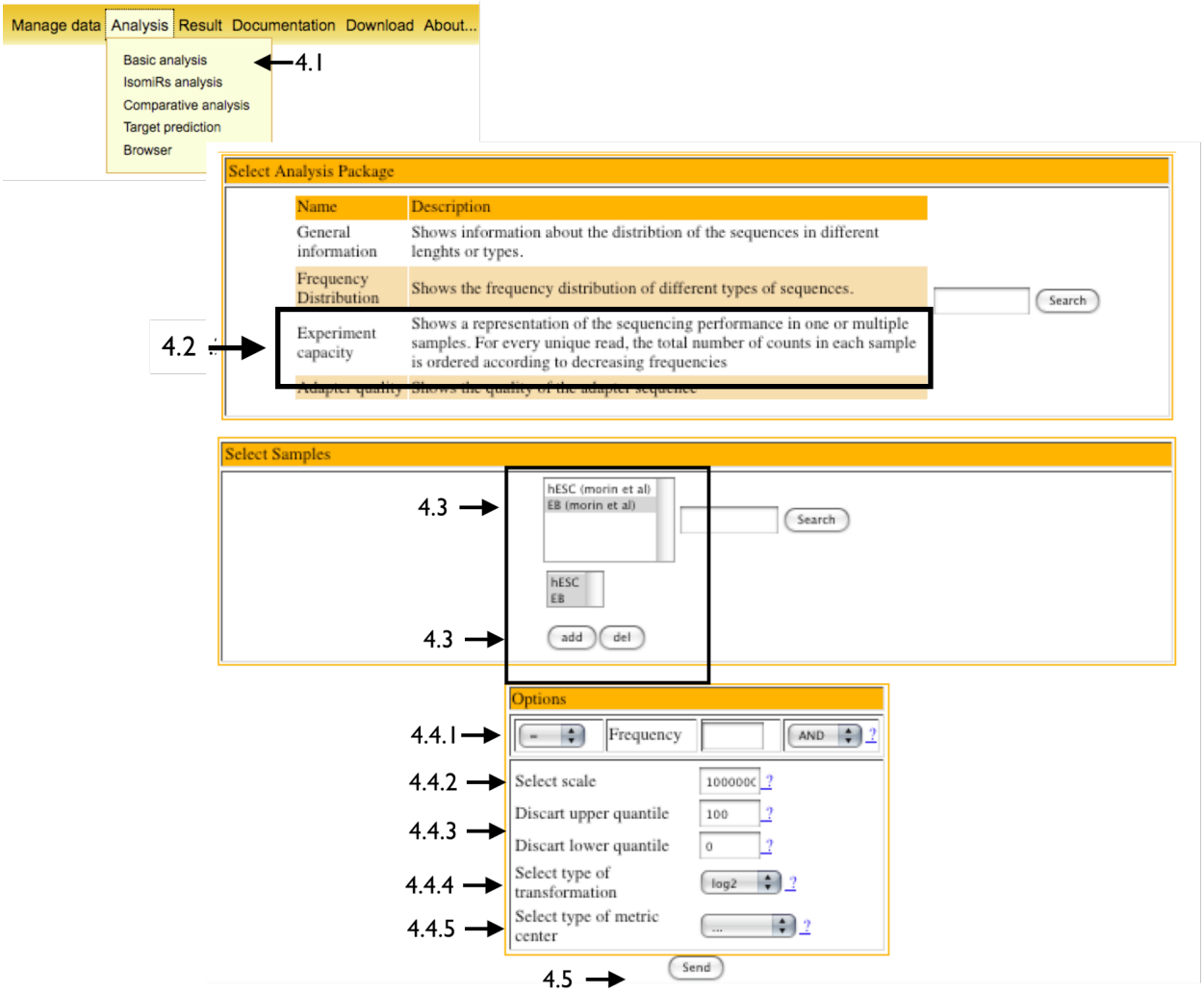
4.4.3. In the "Discard upper quantile" and "Discard lower quantile" textbox certain sequence populations may be not included in the analysis. The quantile (in this case is a percentage or percentile) value represents a specific proportion of sequences showing a certain frequency. For instance, a percentile of "y = x" means that a "y %" of the sequences present a frequency "≤ x". These quantile values depend on the

frequency distribution of the sequences in the sample. The upper and lower quantile values provide an upper and lower frequency cutoff to select a population of sequences in the analysis. It may happen that two samples show a different sequencing capacity due to extreme values; for instance few sequences in one of the samples presenting an extreme number of counts or many sequences showing scarce counts. If these extreme values (percentiles) are removed from the analysis, the sequencing capacity may be comparable between samples. In the present example all sequences are considered

4.4.4. In the “Select type of transformation” select box “log2” has been chosen, that represents the logarithm (base 2) of the frequency of the sequences

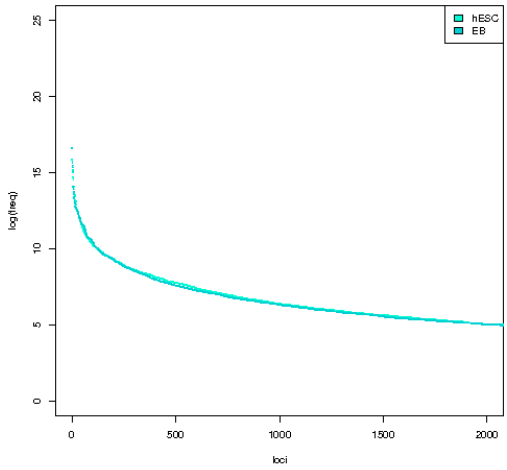
4.4.5. In the “Select type of metric center” the type of statistical parameter to center the frequency distribution may be chosen. If sequencing capacity is not equivalent between samples, different types of metric centers may be applied to normalize the frequency distributions between samples, making them comparable for differential expression. In the present example, no metric center parameter is selected, since the sequencing capacity between samples is equivalent.

4.5. Press “send” button



4.6. In the analysis output the following graph appears, showing for every single sequence (x-axe), the associated frequency (logarithmic transformed, y-axe) in hESC (black) and EB (grey) samples:

Figure 4



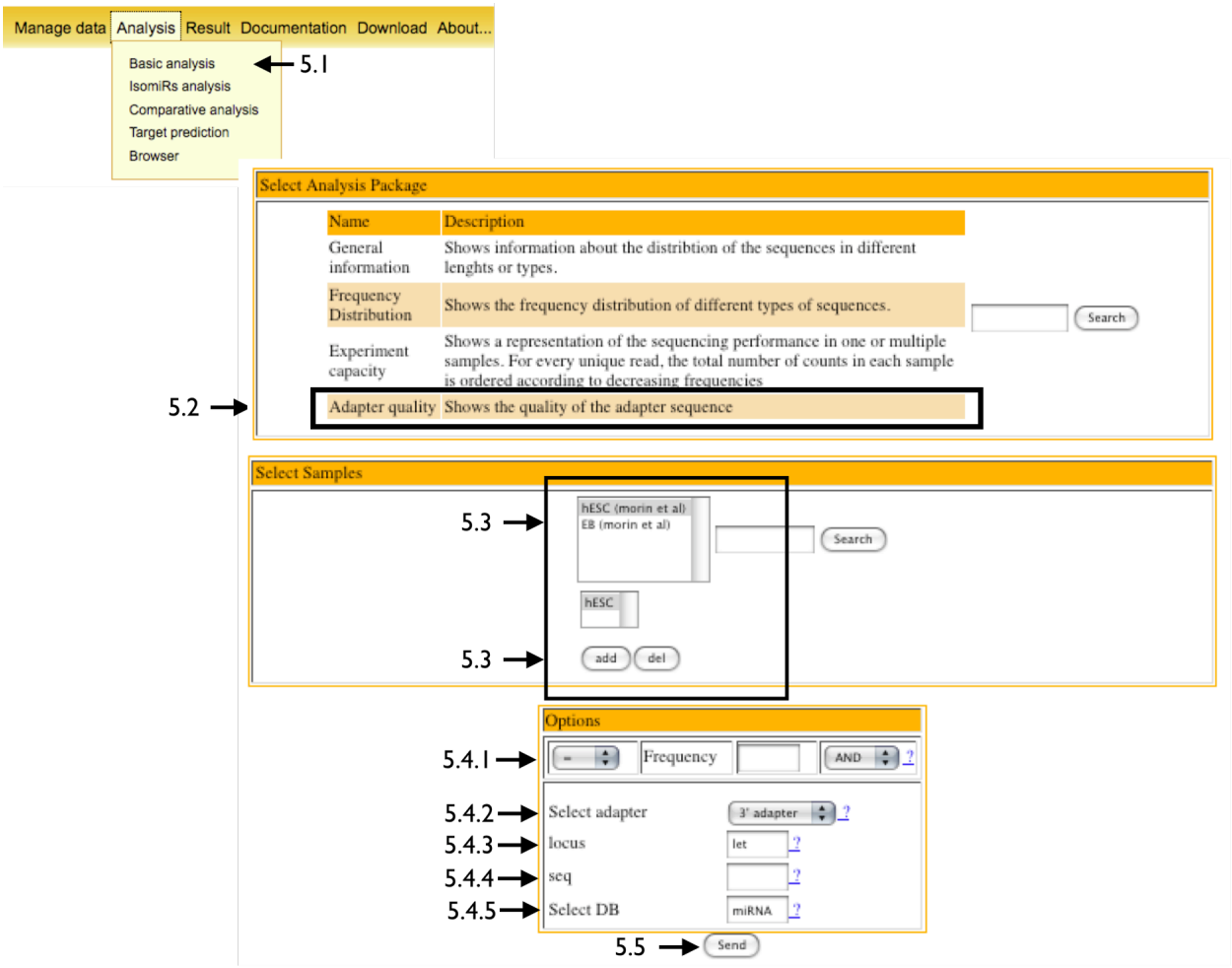
5. Basic analysis "Adapter quality" package

This package is designed to visualize the quality of the adapter in selective sequences. This is an approach to the quality of the sequences

5.1. Go to the "Analysis" option in the menu and choose "Basic analysis"
5.2. Select "Adapter quality" in the "Select Analysis Package" panel
5.3. Proceed as in steps 2.2.1. and 2.2.2. to select the samples
5.4. In the "Options" panel, several parameters may be chosen to analyze the sequencing performance.
  5.4.1. No cutoff for "Frequency" has been established (textbox empty). In the present example, sequences with all ranges of frequencies are being considered.
  5.4.2. In the "Select adapter" chose "3'adapter" if sequencing has been performed using Illumina/Solexa technology, and "3' adapter" or "5' adapter" if 454 sequencing technology has been used. In the present example "3' adapter" is selected
  5.4.3. The "locus" textbox is used to analyze all sequences mapped in a specific locus. In the present example "let" is written to analyze the adapter sequence in all the sequences mapping onto the let family of miRNAs
  5.4.4. The "seq" textbox is used to analyze the adapter of a specific sequence. In the present example all sequences are considered (empty textbox)
  5.4.5. In the "Select DB" textbox a specific database may be chosen for the analysis, according to the name of the databases used in the pre-

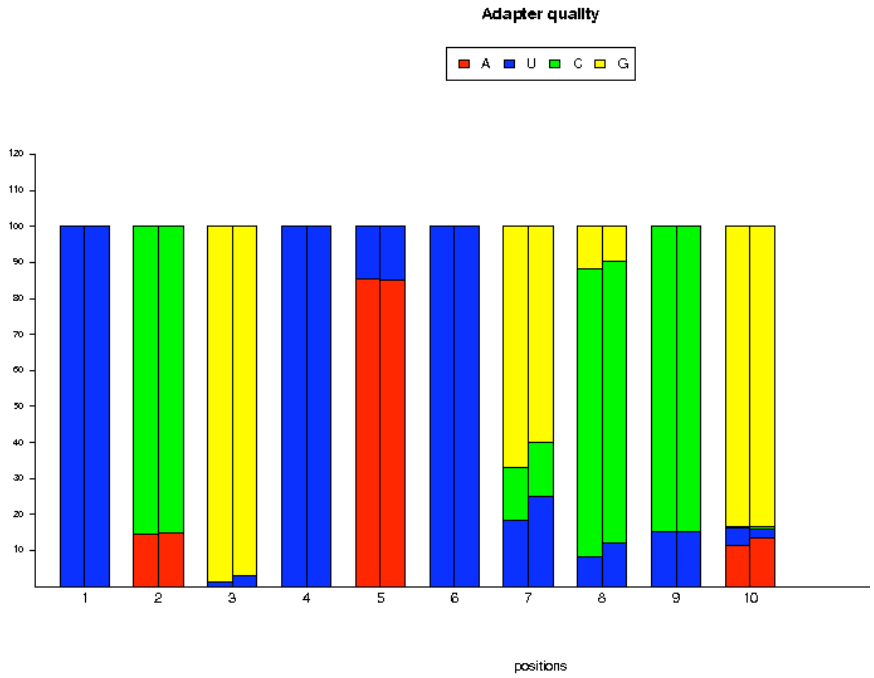
analysis (see Tutorials 1, 2 and 3). In the present example “miRNA” is chosen.

5.5. Press “send” button



5.6. In the analysis output the following histogram appears, showing the nucleotide pattern in the first 10 nucleotides of the 3’-adapter sequence for all let miRNA sequences, in hESC (left bars) and EB (right bars) samples. The x-axe shows the 10 first positions of the adapter sequence. The y-axe shows the percentage of each nucleotide in all the reads:

Figure 5



6. IsomiRs analysis: "IsomiRs distribution" package

This package analyzes the percentage of miRNAs presenting different types of variability (or IsomiRs). miRNA variability can be the consequence of Drosha and Dicer enzymatic activities during miRNA biogenesis(5'-trimming and 3'-trimming variants); nucleotide additions at the 3'-end of the mature miRNA (3'-addition variants); and modifications of the mature miRNA nucleotides at different positions (nt-substitution variants)

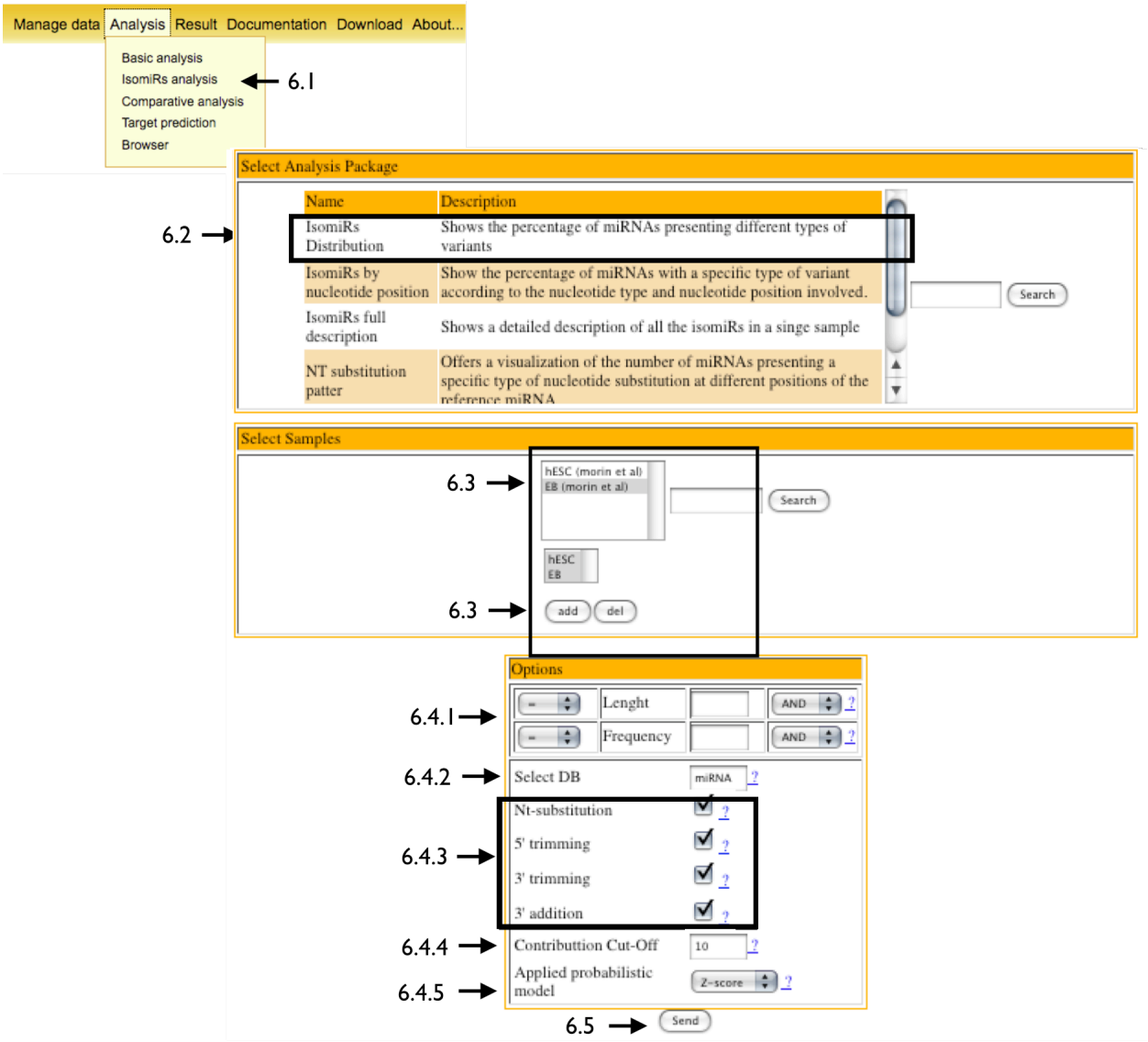
6.1. Go to the "Analysis" option in the menu and choose "IsomiRs analysis"
6.2. Select "IsomiRs distribution" in the "Select Analysis Package" panel
6.3. Proceed as in steps 2.1. and 2.2. to select the samples
6.4. In the "Options" panel, several parameters may be chosen:
    6.4.1. No cutoff for "Length" and "Frequency" have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis
    6.4.2. In the "Select DB", chose all sequences mapped onto the miRNA database. Enter the name of the miRNA database (miRNA if annotated using the web server or a custom name if annotated using the stand alone version). In the present example "miRNA" is chosen.

6.4.3. The type of variant to be considered in the analysis is selected in “Nt-substitution”, “5’-trimming”, “3’-trimming” and/or “3’-addition” options. In the present example all types of variants are selected

6.4.4. Select the “Contribution cut-off” option. This value is used to filter variants contributing in more than a specific percentage to the total of sequences annotated in the same miRNA locus (the reference miRNA and all the corresponding variants). In the present example, the analysis will only consider variants contributing in more than a 10 % .

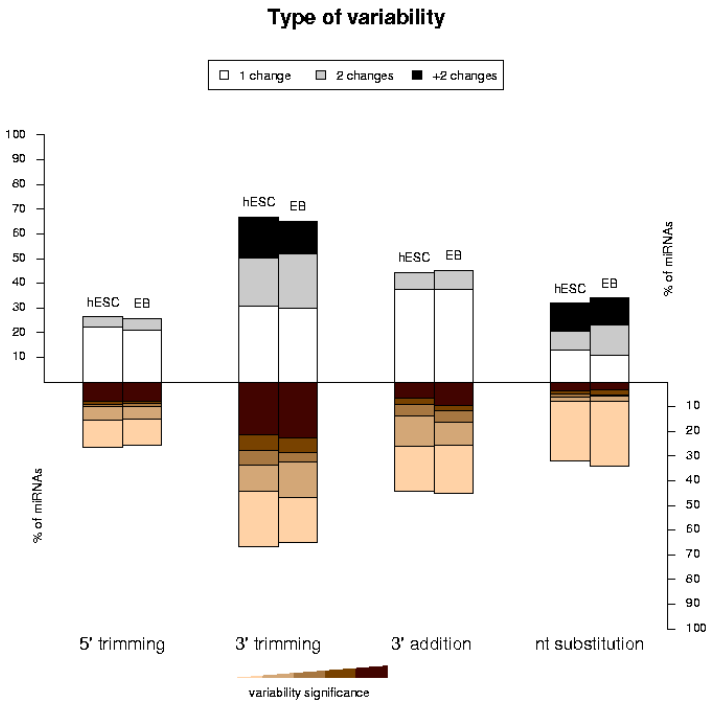
6.4.5. Select the “Applied probabilistic model” option. The “Z-score” option excludes sequencing errors as the possible cause of the nucleotide changes observed in some variants according to (Dohm et al., 2008, Nucleic Acids Res. 36(16):e105). In the present example the “Z-score” option is selected, mining that the variants that are considered as sequencing errors are not going to be included in the analysis

6.5. Press “Send“ button



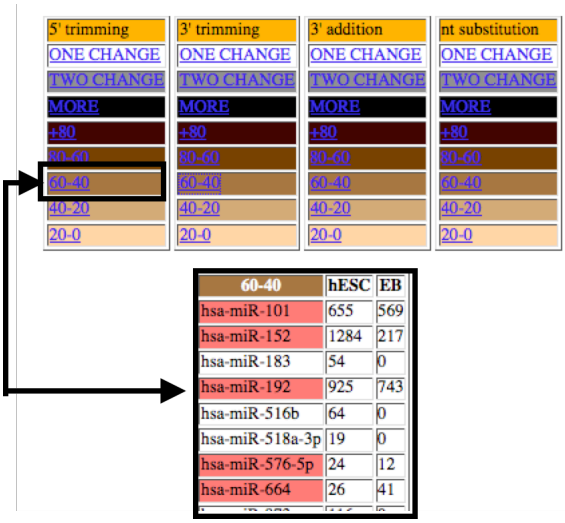
6.6. In the output analysis the following histogram appears displaying the proportion of miRNAs with different types of isomiRs in hESC (left bars) and EB (right bars) small RNA libraries. For every type of variability, the upper part of the graph shows the proportion of miRNAs presenting one (white), two (grey) or more than two (black) isomiRs. The abundance of the isomiR with respect the corresponding reference miRNA is mirrored in the lower graph in a brown scale. The 5 borwn color intensities from dark to light indicate: 1, isomiR frequency >80% with respect to the reference miRNA; 2, isomiR frequency is 60%-80% of that of the reference miRNA; 3, isomiR frequency is 40%-60% of that of the reference miRNA; 4 isomiR frequency is 20%-40% of that of the reference miRNA; and 5, isomiR frequency <20% of the reference miRNA.

Figure 6



In the output analysis, a table appears below the histogram. This table groups the miRNAs according to the number of variants (1 variant in white; 2 variants in grey; or more than 2 variants, in blak), the type of isomiR and the relaive abundance of the isomiR with respect the correpondent reference miRNA (brown color scale). All the miRNA appear listed when clicking inside any part of the table, showing the frequencies of the variant in each sample. The miRNAs highlighted in pink present the same type of isomiR in both samples:

Figure 7



7. IsomiRs analysis: "IsomiRs by nucleotide position" package

This package analyzes the percentage of miRNAs presenting a specific type of variability or IsomiRs, according to the nucleotide type and nucleotide position

7.1. Go to the "Analysis" option in the menu and choose "IsomiRs analysis"
7.2. Select "IsomiRs distribution" in the "Select Analysis Package" panel
7.3. Proceed as in steps 2.2.1. and 2.2.2. to select the samples
7.4. In the "Options" panel, several parameters may be chosen:
- 7.4.1. No cutoff for "Length" and "Frequency" have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis
- 7.4.2. In the "Select type of DB" Select the type of database to be considered in the analysis. In the present example "miRNAs" is chosen.
- 7.4.3. In the "Select type of isomiR" chose the class of miRNA variability to consider in the analysis, in the present example "5'-trimming" is selected.
- 7.4.4. The "Start position in nt-substitution" and "End position in nt-substitution" parameters indicate the positions in the reference miRNAs that are considered in the nt-substitution variants. In the present example these options will be skipped by SeqBuster since the analysis is focused on 5'-trimming isomiRs
- 7.4.5. The "Trimming size" value indicates the number of nucleotides upstream and downstream of the reference miRNA first position (in 5'-

trimming variants, the reference 5'-start site). In the present example 2 positions upstream and downstream are considered.

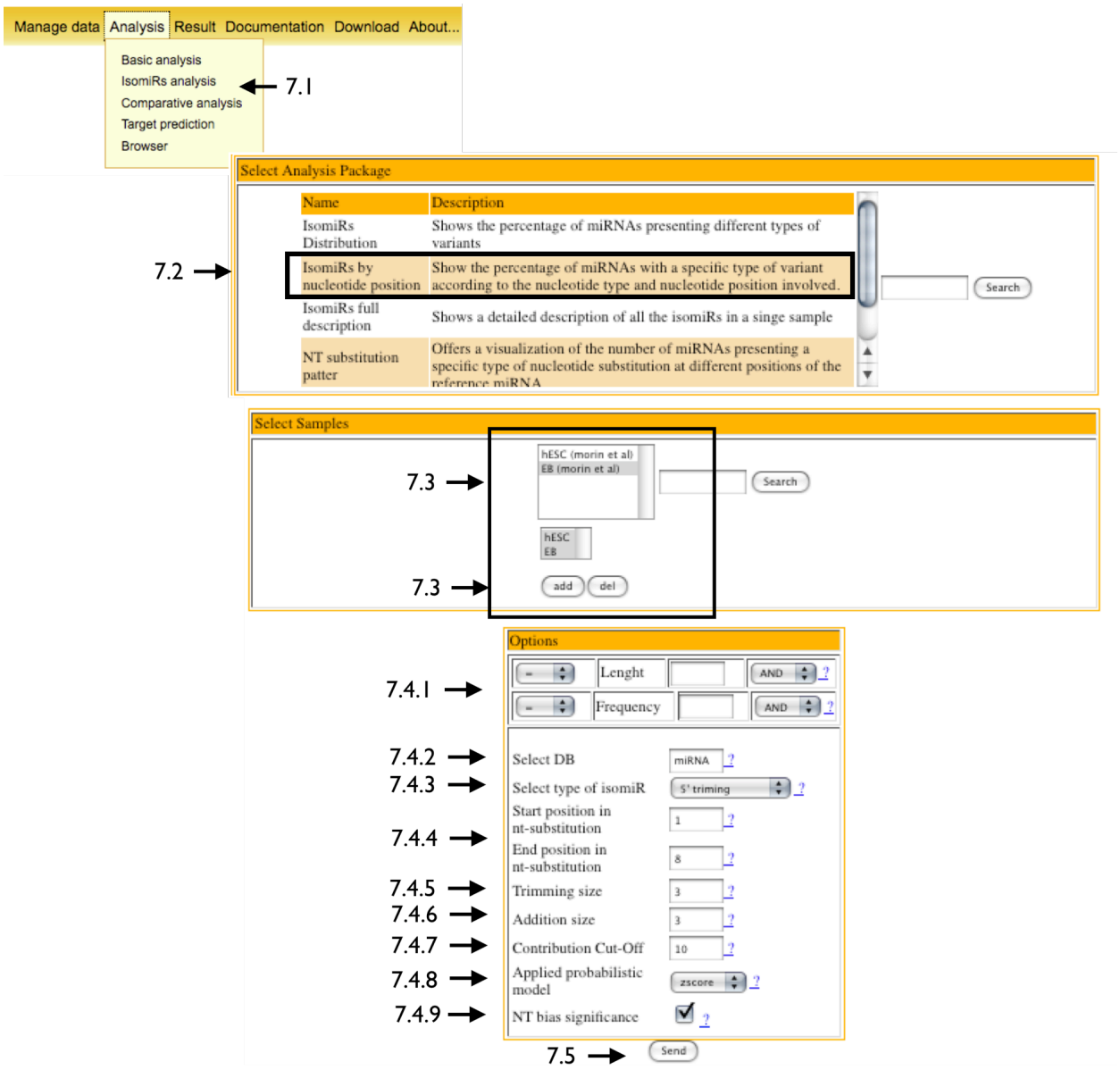
7.4.6. The "Addition size" value indicates the number of nucleotides to be considered in the 3' addition variants. In the present example, this option will be skipped by SeqBuster since the analysis is focused on 5'-trimming isomiRs.

7.4.7. Select the "Contribution cut-off" option (see 6.4.3. section). In the present example, the analysis considers variants contributing in more than 10 %

7.4.8. Select the "Applied probabilistic model" option (see 6.4.5. section). In the present example the "Z-score" option is selected, mining that the variants considered as sequencing errors are not going to be included in the analysis

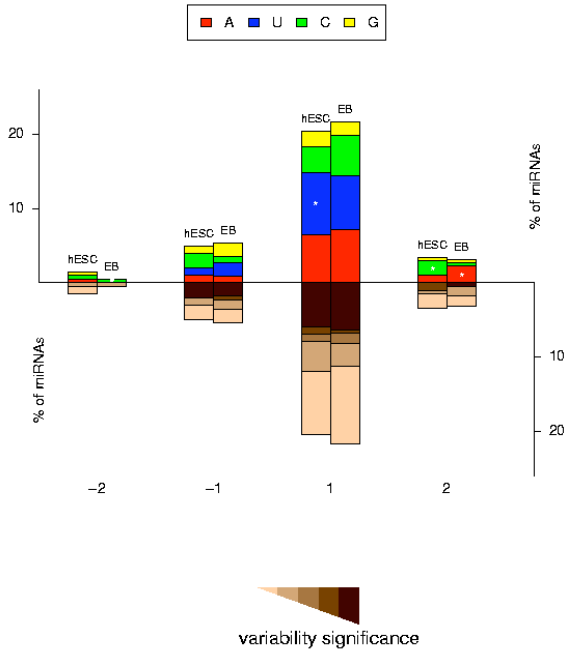
7.4.9. Select the "Nucleotide bias significance" to highlight a possible bias for a specific nucleotide being involved in the trimming variants

7.5. Press the ¨Send¨ button



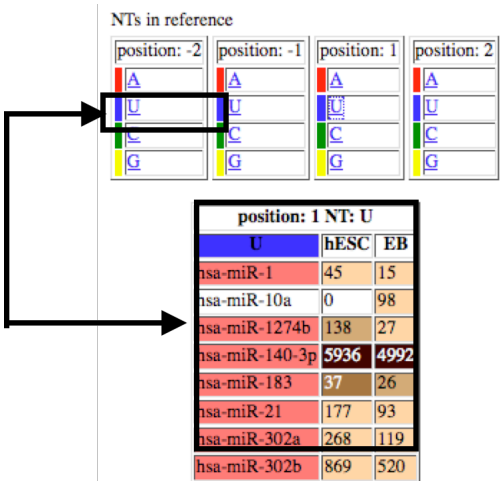
7.6. In the analysis output the following histogram appears displaying the percentage of miRNAs with 5'-trimming variants in hESC (left bars) and EB (right bars) samples, according to the nucleotide position involved in the trimming variant. Pre-miRNA slicing occurring at different positions upstream or downstream of the reference miRNA is indicated by -2 to -1 and +1 to +2, respectively. In the upper bars the color pattern indicates the nucleotides involved in the trimming variants at every position: A (red), U (blue), C (green) and yellow (G). Significant bias toward a nucleotide is shown with a white asterisk. In every histogram the lower bars show the proportion of the isomiR with respect the reference miRNA at different positions as described in Figure 6.

Figure 8



In the output analysis, a table appears below the histogram, showing the different nucleotides involved in every position of the variant. By cliking on the nucleotides, a list of the miRNAs involved and the corresponding frequency appears for every sample. The miRNAs highlighted in pink present the same type of isomiRs in both samples:

Figure 9



7.7. In the next example, the “nt-substitution” variants are analyzed, using the same “IsomiRs by position” package in the “Select Analysis Pakage” panel. In the “Options” panel, several parameters may be chosen:

7.7.1. No cutoff for “Length” and “Frequency” have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis

7.7.2. In the “Select type of DB” Select the type of database to be considered in the analysis. In the present example “miRNAs” is chosen.

7.7.3. In the “Select type of isomiR” chose the class of miRNA variability to consider in the analysis. In the present example “nt-substitution” is selected.

7.7.4. The “Start position in nt-substitution” and “End position in nt-substitution” parameters indicate the positions in the reference miRNAs that are considered in the nt-substitution variants. In the present example positions 9 to 16 of the miRNA are considered.

7.7.5. The “Trimming size” value (see 7.4.5. section) is skipped by SeqBuster in the present example.

7.7.6. The “Addition size” value (see 7.4.6. section) section is skipped by SeqBuster since the analysis is focused on 5’-trimming isomiRs.

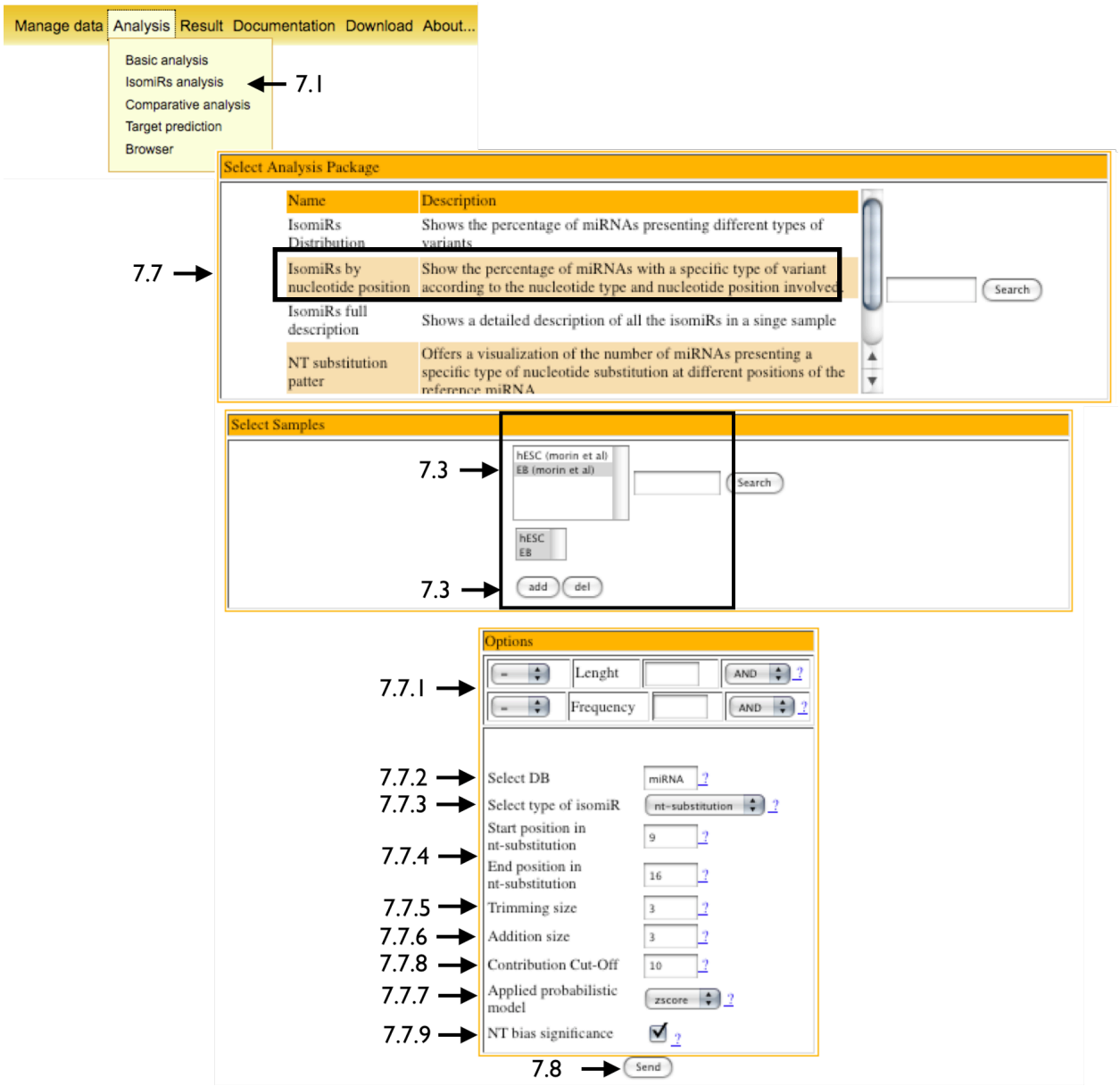
7.7.7. Select the “Contribution cut-off” option (see 6.4.3 section). In the present example, the analysis will only consider variants contributing in more than 10 %

7.7.8. Select the “Applied probabilistic model” option (see 6.4.5. section). In the present example the “Z-score” option is selected, mining that the

variants considered as sequencing errors are not going to be included in the analysis

7.7.9. Select the “Nucleotide bias significance” to highlight a possible bias for a specific nucleotide of the reference miRNA being substituted

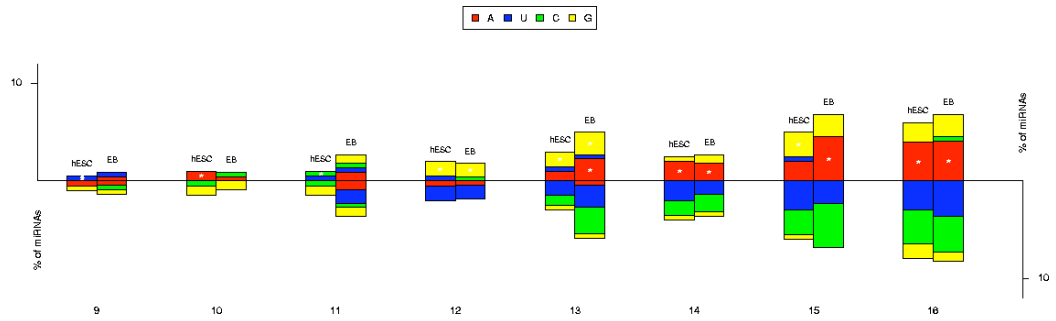
7.8. Press “Send” button



7.9. In the analysis output the following histogram appears, showing the percentage of miRNAs with nucleotide modifications in hESC (left bars) and EB (right bars), at several positions of the miRNA. The upper bars show the

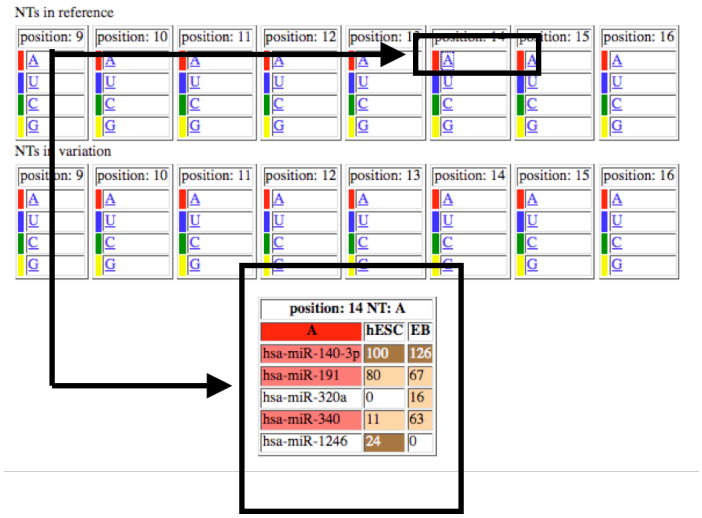
type of nucleotide in the reference miRNA that presents a modification, A (red), U (blue), C (green) and yellow (G). The type of nucleotides present in the isomiRs is mirrored in the lower bars. Significant bias toward a specific nucleotide in the reference miRNA is labeled with a white asterisk:

Figure 10



In the output analysis, two tables appear below the histogram, showing the different nucleotides in every position of the reference miRNA (upper table) and the corresponding isomiR (lower table). By cliking on the nucleotides, a list appears showing the miRNAs involved and the corresponding frequencies, for both samples. The miRNAs highlighted in pink present the same type of isomiR in both samples:

Figure 11



7.10. In the next example, the “3’-addition” variants are analyzed, using the same “IsomiRs by position” option in the “Select Analysis Pakage” panel. In the “Options” panel, several parameters may be chosen:

7.10.1. No cutoff for “Length” and “Frequency” have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis

7.10.2. In the “Select type of DB” Select the type of database to be considered in the analysis. In the present example “miRNAs” is chosen.

7.10.3. In the “Select type of isomiR” chose the class of miRNA variability to consider in the analysis, in the present example “3’-addition” is selected.

7.10.4. The “Start position in nt-substitution” and “End position in nt-substitution” parameters (see 7.4.4. section) are skipped by SeqBuster since the analysis is focused on 3’-addition isomiRs

7.10.5. The “Trimming size” value (see 7.4.5. section) is skipped by SeqBuster in the present example.

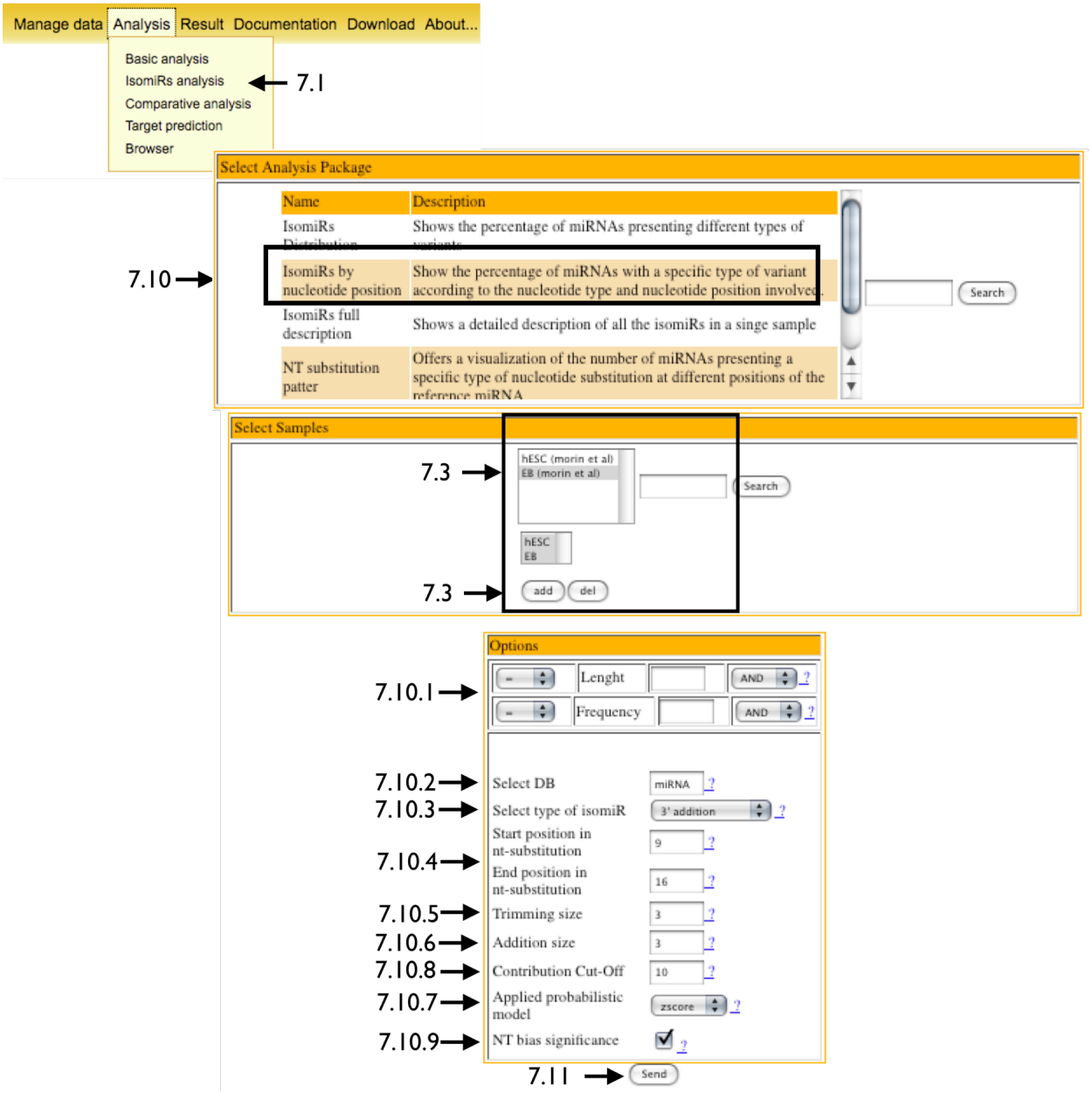
7.10.6. The selected “Addition size” value (see 7.4.6. section) is 3, meaning that the variants considered in the analysis present up to 3 nucleotides added at the 3’-end of the reference miRNA

7.10.7. Select the “Contribution cut-off” option (see 7.4.7 section). In the present example, the analysis will only consider variants contributing in more than 10 %

7.10.8. Select the “Applied probabilistic model” option. (see 6.4.5. section) In the present example the “Z-score” option is selected, mining that the variants considered as sequencing errors are not going to be included in the analysis

7.10.9. Select the “Nucleotide bias significance” to highlight a possible bias for a specific nucleotide of the reference miRNA being substituted

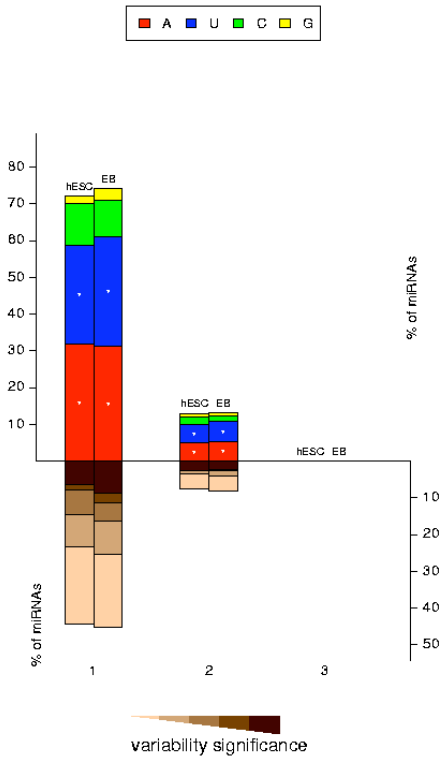
7.11. Press the “send” button



7.12. In the analysis output the following histogram appears, showing the percentage of miRNAs presenting an addition in the 3’-terminus of the reference miRNA. Up to 3 nucleotide extensions are considered. The

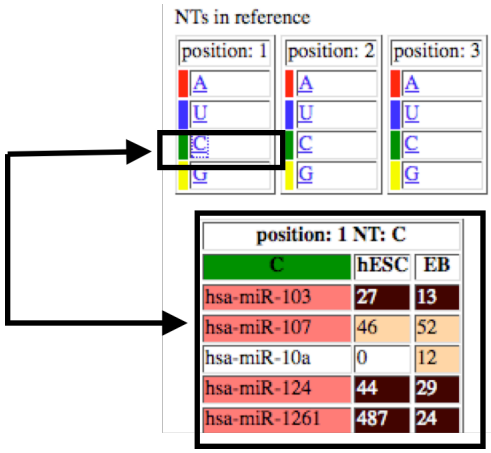
upper bars show the type of nucleotide involved in the addition at every position. The abundance of the 3'-addition isomiRs with respect to the corresponding reference miRNA is mirrored, as shown in Figure 6:

Figure 12



In the output analysis, a table appears below the histogram, showing the different nucleotides involved in every position of the variant. By cliking on the nucleotides, a list of the miRNAs involved and the corresponding frequency appears for every sample. The miRNAs highlighted in pink present the same type of isomiR in both samples:

Figure 13



8. IsomiRs analysis: "IsomiRs full description" package

This package provides a detailed visualization of all the isomiRs in a single sample

8.1. Go to the "Analysis" option in the menu and choose "IsomiRs analysis"
8.2. Select "IsomiRs full description" in the "Select Analysis Package" panel
8.3. Proceed as in steps 2.2.1. and 2.2.2. to select one sample
8.4. In the "Options" panel, several parameters may be chosen:
  8.4.1. No cutoff for "Length" and "Frequency" have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis
  8.4.2. In the "Select type of DB" Select the type of database to be considered in the analysis. In the present example "miRNAs" is chosen
  8.4.3. Select the "Contribution cut-off" option (see 7.4.7. section). In the present example, the analysis will only consider variants contributing in more than 10 %
  8.4.4. Select the "Applied probabilistic model" option (see 6.4.5. section). In the present example the "Z-score" option is selected, mining that

the variants considered as sequencing errors are not going to be included in the analysis

8.4.5. In the “Select type of isomiR” chose the class of miRNA variability to consider in the analysis, in the present example “5’-trimming” is selected.

8.4.6. The “Start position in nt-substitution” and “End position in nt-substitution” parameters (see 7.4.4. section). In the present example this option is skipped by SeqBuster since the analysis is focused on 5’-trimming isomiRs

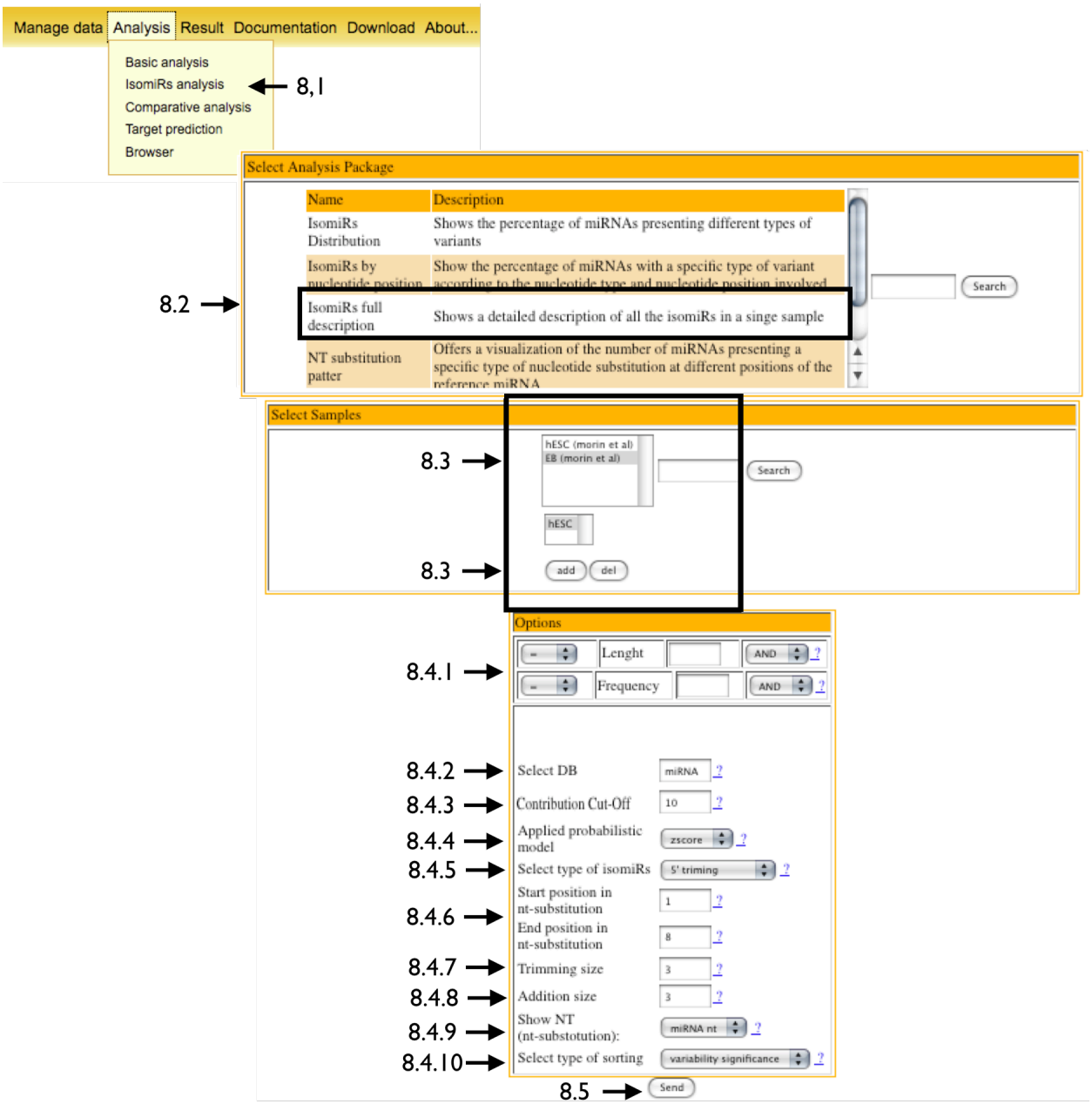
8.4.7. The “Trimming size” value indicates the number of nucleotides upstream and downstream of the reference miRNA first position (see 7.4.5. section). In the present example 3 positions upstream and downstream of the reference miRNA 5’-end is considered.

8.4.8. The “Addition size” value (see 7.4.6. section). In the present example, this option is skipped by SeqBuster since the analysis is focused on 5’-trimming isomiRs.

8.4.9. In the “Show NT (nt-substitution)”, select “reference nt” or the “changed nt” to show the nucleotides present in the reference miRNA or in the isomiR, respectively. In the present example, “changed nt” is chosen

8.4.10. In the “Select type of sorting” options, choose “variability significance” or “frequency” to rank the miRNAs according to the abundance of the isomiR with respect the corresponding reference miRNA (see 6.6. section) or the total frequency, respectively. In the present example, “variability significance” is chosen

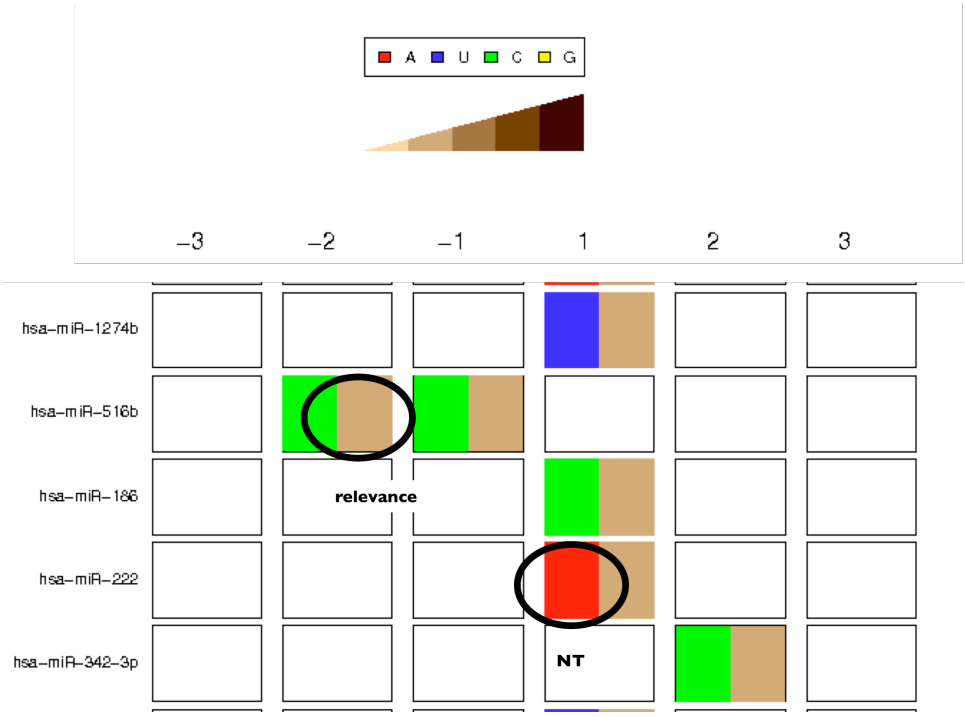
8.5. Press the “Send” button



8.6. In the output analysis a table appears showing a list of the miRNAs presenting the 5' –trimming variants specified in the analysis. The different rows show every miRNA presenting a variant. Every column represents the positions considered in the analysis. In the present example, columns 1-3 represent the upstream trimming positions -3 to -1, and columns 4-6 represent trimming positions +1 to +3, with respect the starting position in every reference miRNA. For each position, the cells involved in the trimming variants are divided in two colors. Color in the left side of the cell refers to the nucleotide that is affected in the isomiR and the color in the

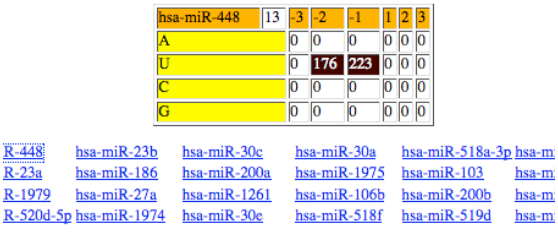
right side of the cell refers to the significance of the isomiR frequency with respect that of the reference miRNA using the same brown color scale used in previous graphs (see Figure 6 and section 6.6.). For instance, in the Figure 14 Hsa-miR-222 presents a variant involving an A in position +1, showing a frequency of 20-40% respect to the reference Has-miR-222

Figure 14



In the output analysis, a list of the miRNAs with isomiRs appears below the table. By clicking on the names, a table showing the nucleotides at each position involved in the 5' trimming variants, and the corresponding number of counts is shown.

Figure 15



9. IsomiRs analysis: "Nucleotide substitution pattern" package

This package offers a visualization of the number of miRNAs that present a specific type of nucleotide substitution at different positions of the reference miRNAs

9.1. Go to the "Analysis" option in the menu and choose "IsomiRs analysis"

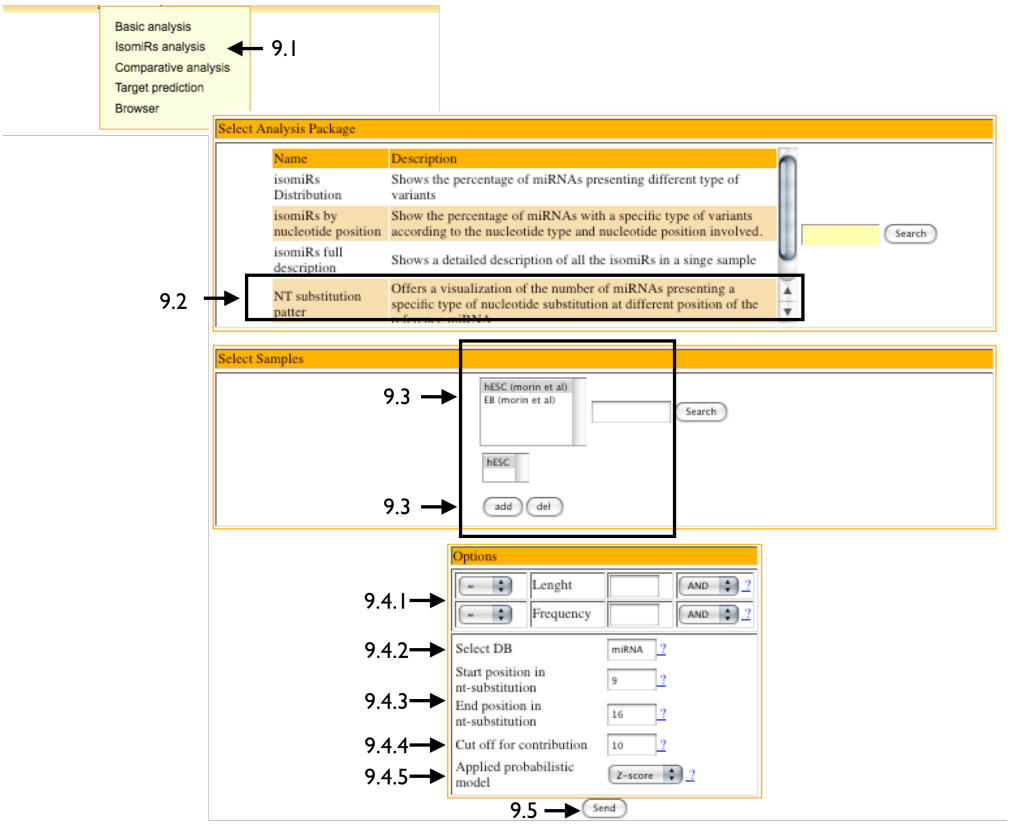
9.2. Select "Nucleotide-substitution pattern analysis" in the "Select Analysis Package" panel

9.3. Proceed as in steps 2.2.1. and 2.2.2. to select one sample

9.4. In the "Options" panel, several parameters may be chosen:

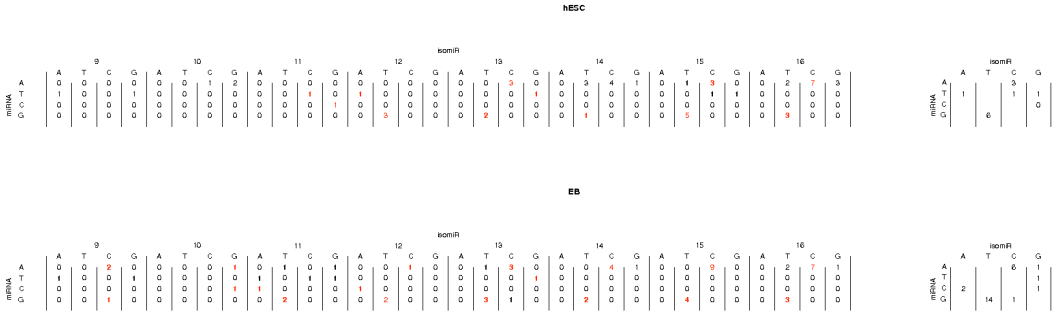
- 9.4.1. No cutoff for "Length" and "Frequency" have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis
- 9.4.2. In the "Select type of DB" Select the type of database to be considered in the analysis. In the present example "miRNAs" is chosen.
- 9.4.3. The "Start position in nt-substitution" and "End position in nt-substitution" (see 7.4.4. section) are selected. In the present example positions 9 to 16 of the miRNA, are considered
- 9.4.4. Select the "Contribution cut-off" option (see 7.4.7. section). In the present example, the analysis will only consider variants contributing in more than 10 %
- 9.4.5. Select the "Applied probabilistic model" option (see 6.4.5. section). In the present example the "Z-score" option is selected, mining that the variants considered as sequencing errors are not going to be included in the analysis

9.5. Press the "Send" button



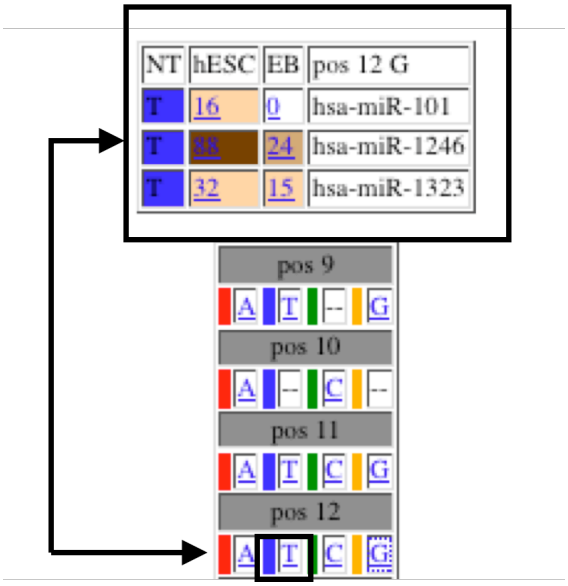
9.6. In the analysis output a table appears showing the number of miRNAs presenting a specific type of nucleotide modification at every position. The overall nucleotide substitution pattern is shown in a single table at the right side. Nucleotides in the rows correspond to those found in the reference miRNAs and nucleotides in the columns are those found in the isomiRs. Within the table, numbers in bold-black indicate a significant bias in the nucleotide of the reference miRNA that presents modifications; numbers in red indicate a significant bias in the nucleotide of the isomiR; and numbers in bold-red indicate a significant bias in the nucleotide of both in the reference miRNA and the isomiR.

Figure 16



In the output analysis, a table appears below showing the nucleotides at every position of the reference miRNA. By clicking on the nucleotide, a list of all the miRNAs presenting variants and the corresponding pattern of the nucleotide change, with the count number for hESC and EB samples is shown. The cells displaying the frequency of the variants are colored according to the frequency of the variant with respect to that of the reference miRNA (brown color scale as explained in Figure 6)

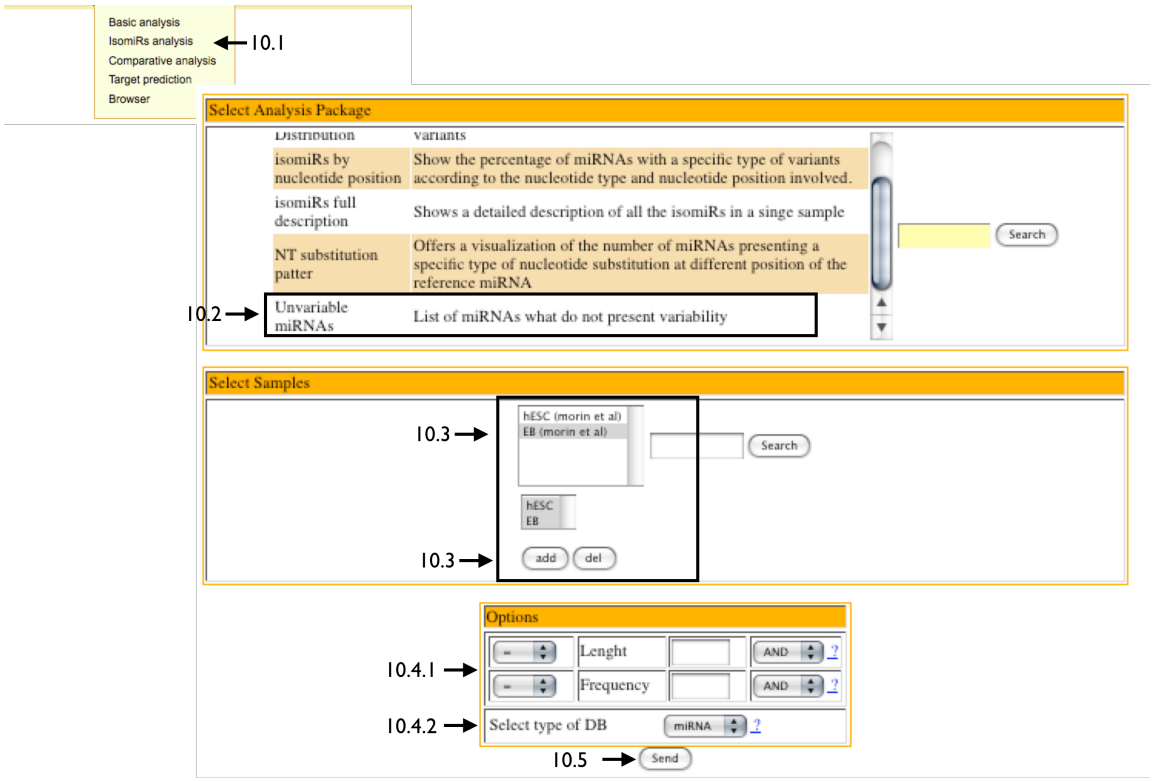
Figure 17



10. IsomiRs analysis: "Unvariable miRNAs" package

This package provides a list of the miRNAs that do not present variability

10.1. Go to the "Analysis" option in the menu and choose "IsomiRs analysis"
10.2. Select "Unvariable miRNAs" in the "Select Analysis Package" panel
10.3. Proceed as in steps 2.2.1. and 2.2.2. to select one sample
10.4. In the "Options" panel, several parameters may be chosen:
    10.4.1. No cutoff for "Length" and "Frequency" have been established (textbox empty); therefore all sizes and frequencies are considered in the analysis
    10.4.2. In the "Select type of DB" Select the type of database to be considered in the analysis. In the present example "miRNAs" is chosen.
10.5. Press the "Send" button



10.6. In the output analysis a table appears showing the invariable miRNAs with the corresponding counts in hESC and EB libraries Invariable miRNAs. Some miRNAs appear only in hESC, and the number of counts is displayed as “-1” in EB, meaning that in EB these miRNAs present variants. A similar observation is detected for invariant miRNAs present only in EB. The miRNA lacking isomiRs in both libraries are highlighted in pink.

Figure 18

name	hESC	EB
hsa-let-7d	4	-1
hsa-miR-1243	2	-1
hsa-miR-1250	5	-1
hsa-miR-1252	4	-1
hsa-miR-1253	2	-1
hsa-miR-1255b	3	5
hsa-miR-1256	2	2
hsa-miR-1257	3	-1
hsa-miR-1258	2	-1
hsa-miR-1263	3	-1
hsa-miR-1265	2	-1
hsa-miR-1267	2	2
hsa-miR-1277	69	33
hsa-miR-1284	6	3
hsa-miR-1288	2	-1
hsa-miR-1294	11	-1
hsa-miR-1295	6	4
hsa-miR-1296	4	-1
hsa-miR-1299	7	2
hsa-miR-149	9	-1

11. Comparative analysis package

Two main analyses can be performed using this package to compare sequencing experiments: 1. The cluster analysis, that sorts through several sequencing experiments and groups them into clusters, and 2. Differential expression analysis showing significant differences in the number of counts of all sequences, between two samples or two groups of samples. In the present example the analysis is focused on differential expression, since more than two samples are required for cluster analysis.

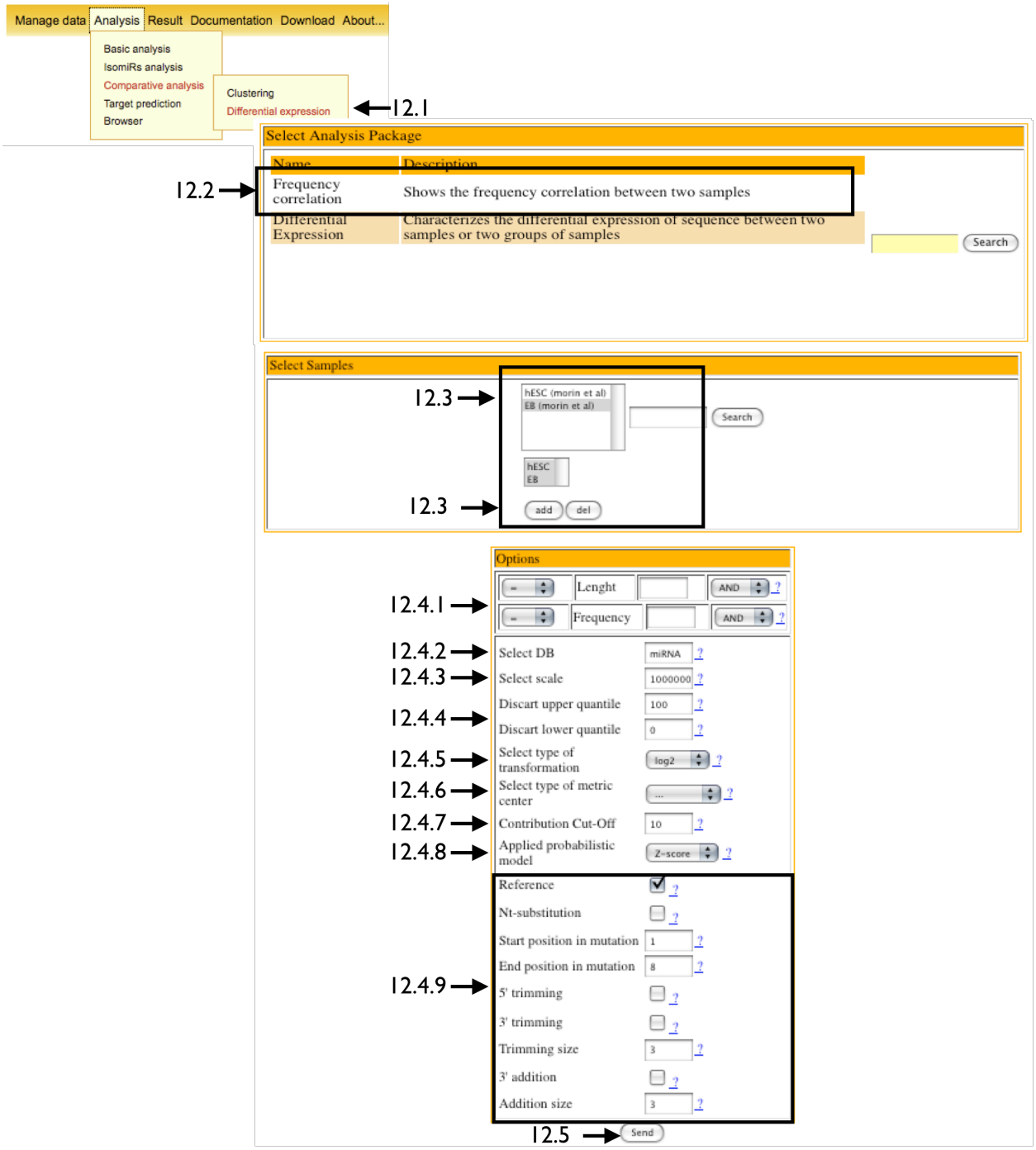
12. Differential expression analysis: "Frequency correlation" package

This package characterizes the frequency correlation between two samples. Several parameters in the "Options" panel (see 12.4. section) may be modified according to the output of the "Secuencing capacity" basic analysis package (see section 4).

12.1. Go to the "Analysis" option in the menu and choose "Comparative analysis"
12.2. Select "Frequency correlation" in the "Select Analysis Package" panel
12.3. Proceed as in steps 2.2.1. and 2.2.2. to select the samples
12.4. In the "Options" panel, several parameters may be chosen.:
- 12.4.1. "Length" and "Frequency" filters the sequences to be considered in the analysis according to specific cutoff value of size and frequency, respectively. In the present example, all sizes and frequencies are being considered, since no cutoff is established
- 12.4.2. In the "Select DB" text box "miRNA" has been introduced in the present example to analyse all sequences (reference and variants) annotated as miRNAs
- 12.4.3. In the "Scale" textbox 1.000.000 is chosen (see 4.4.2. section)
- 12.4.4. In the "Discard upper quantile" and "Discard lower quantile" textbox certain sequence populations may be not included in the analysis (see 4.4.3. section). In the present example all sequences are considered
- 12.4.5. In the "Select type of transformation" select box "log2" has been chosen, that represents the logarithm of the frequency of the sequences
- 12.4.6. In the "Select type of metric center" the type of statistical parameter to center the frequency distribution may be chosen (see 4.4.5. section)
- 12.4.7. Select the "Contribution cut-off" value (see 7.4.7 section). In the present example a value of 10 is selected
- 12.4.8. Select the "Applied probabilistic model" option (see 6.4.5. section). In the present example the "Z-score" option is selected, mining that the variants that were considered as sequencing errors were not included in the analysis

12.4.9. In the options below, select the type of reference miRNA or isomiR that is going to be included in the frequency correlation analysis. If “nt substitution” variants are selected, specify the “Start position in nt-substitution” and “End position in nt-substitution” (see 7.4.4. section). If trimming variants are selected chose a “Trimming size” value (see 7.4.5. section). If 3’-addition variants are selected a “Addition size” value should be chosen. In the present example, only the reference miRNAs are selected for the analysis

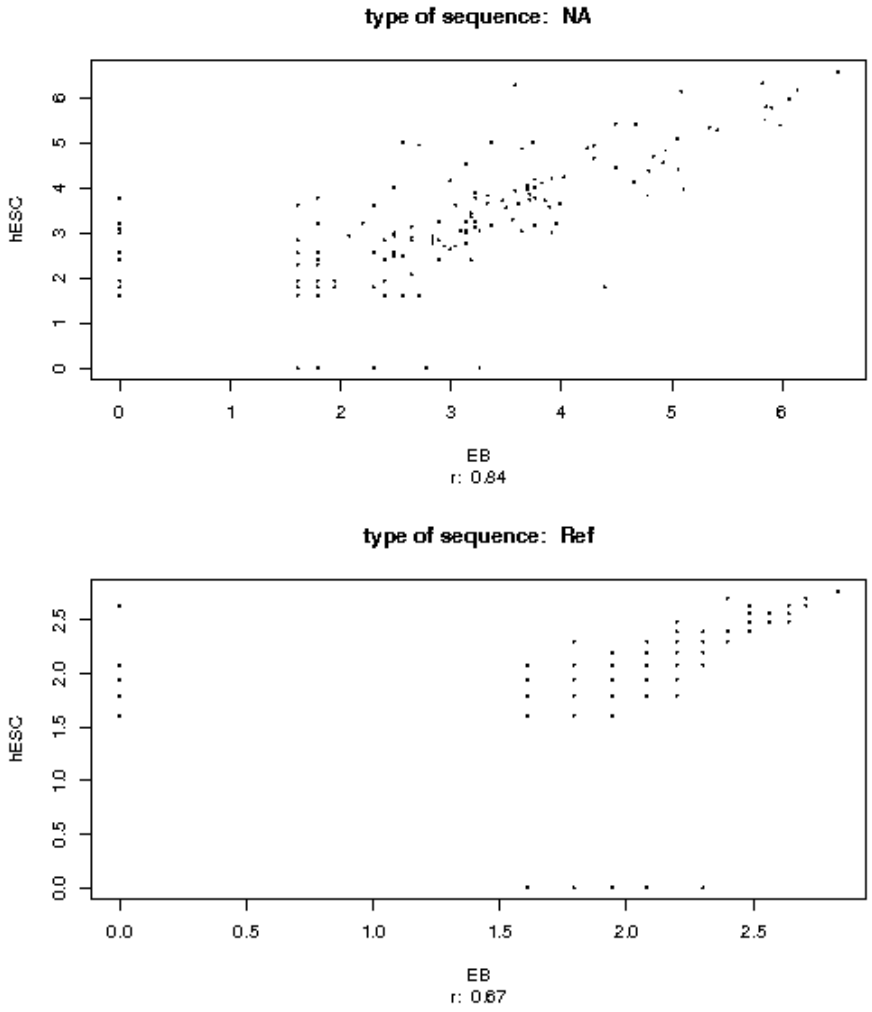
12.5. Press “Send” button



12.6. In the output analysis a graph appears showing the correlation between the two types of sequences frequencies (log2 normalized) in hESC (y axe) and EB (x axe). The label “Ref” indicates that the correlation is between sequences

mapping perfectly onto miRNAs. The label “NA” refers to all the different isomiRs mapping onto the miRNA loci. The correlation value appears below.

Figure 18



13. Differential expression analysis: “Differential expression” package

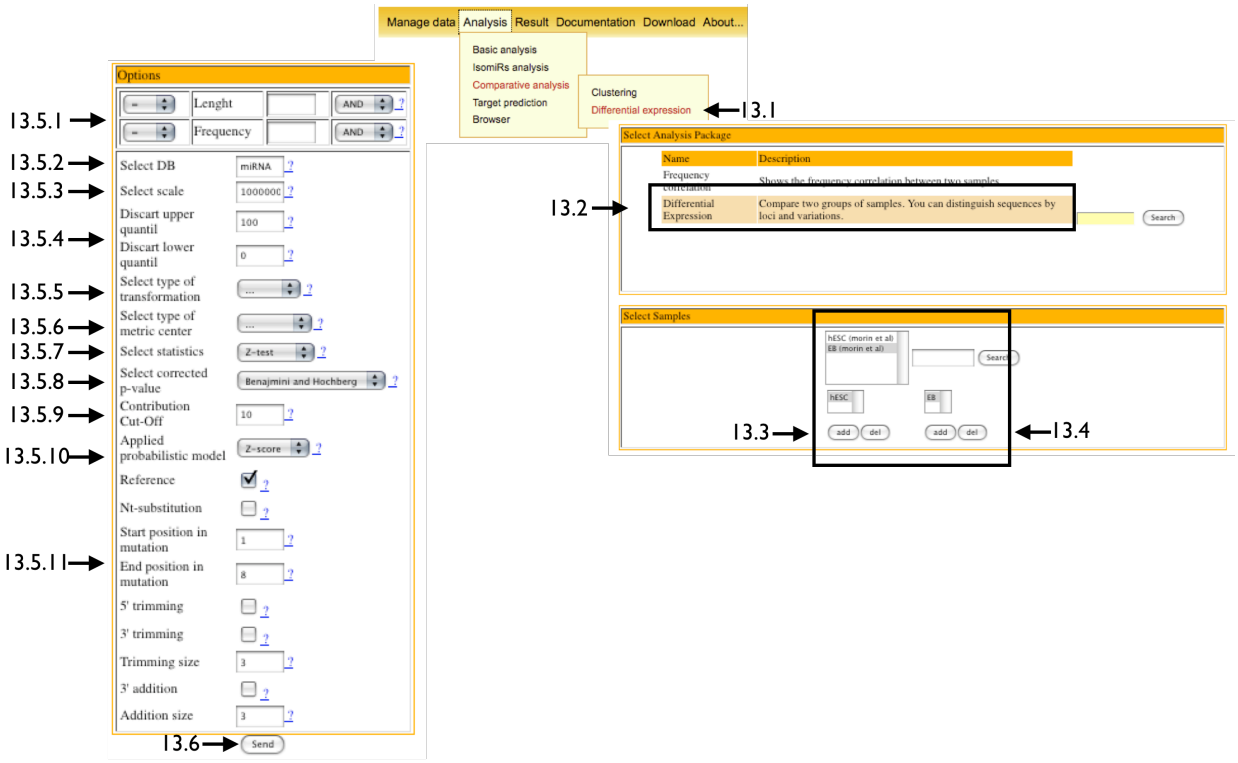
This package characterizes the differential expression of sequences between two samples or two groups of samples. Several parameters in the “Options” panel (see

13.5. section) may be modified according to the output of the “Secuencing capacity” basic analysis package (see section 4)

13.1. Go to the “Analysis” option in the menu and choose “Comparative analysis”
13.2. Select “Differential expression” in the “Select Analysis Package” panel
13.3. Select the “hESC” sample and press left “Add” button
13.4. Select the “EB” sample and press right “Add” button
13.5. In the “Options” panel, several parameters may be chosen:
    13.5.1. “Length” and “Frequency” filters the sequences to be considered in the analysis according to specific cutoff value of size and frequency, respectively. In the present example, all sizes and frequencies are being considered, since no cutoff is established.
    13.5.2. In the “Type of DB” text box “miRNA” has been introduced in the present example to analyse all sequences (reference and variants) annotated as miRNAs
    13.5.3. In the “Scale” textbox 1.000.000 is chosen (see 4.4.2. section)
    13.5.4. In the “Discard upper quantile” and “Discard lower quantile” textbox certain sequence populations may be not included in the analysis (see 4.4.3. section). In the present example all sequences are considered
    13.5.5. In the “Select type of transformation” select box “log2” has been chosen, that represents the logarithm of the frequency of the sequences
    13.5.6. In the “Select type of metric center” the type of statistical parameter to center the frequency distribution may be chosen (see 4.4.5. section)
    13.5.7. In the “Select statistics” select box, choose the type of statistical test to determine differences in the frequency of all sequences between samples. In the present example the Z-test is chosen
    13.5.8. The “Select corrected p-value” select box offers the possibility to correct the p-value applying the Benjamini and Hochberg method on the p-value assigned by the statistical test used to determine differential expression. In the present example, this option has been selected
    13.5.9. Select the “Contribution cut-off” value (see 7.4.7 section). In the present example a value of 10 is selected
    13.5.10. Select the “Applied probabilistic model” option (see 6.4.5. section). In the present example the “Z-score” option is selected, mining that the variants that were considered as sequencing errors were not included in the analysis
    13.5.11. In the options below, select the type of reference miRNA or isomiR that is going to be included in the differential expression analysis. If “nt substitution” variants are selected, specify the “Start position in nt-substitution” and “End position in nt-substitution” (see

7.4.4. section). If trimming variants are selected chose a “Trimming size” value (see 7.4.5. section). If 3’-addition variants are selected a “3’-addition size” value should be chosen. In the present example, only the reference miRNAs are selected for the analysis.

13.6. Press “Send” button



13.7. In the output analysis a table appears showing a list of miRNAs in the “locus” column. In the “Type” column, the reference miRNA is named “Ref” and the rest of sequences mapped in the same locus that contain all the variants are named “NA”. The miRNA “Ref” and the miRNA “NA” types of sequences and counts are shown, by clicking on the miRNA name in the “locus” column. Additional columns show the normalized frequency for hESC (G1) and EB (G2) small RNAs. The ratio of the frequencies between hESC and EB samples is shown in the column “ratio”. The p-value is shown in the last column.

Figure 19

locus	type	norm G1	norm G2	ratio	pvalue
hsa-let-7a	NA	4568.00	795.00	5.75	0.00
hsa-let-7a	Ref	18381.00	4754.00	3.87	0.00
hsa-let-7b	NA	69.00	58.00	1.19	0.58
hsa-let-7b	Ref	33.00	0.00	9999.00	0.00
hsa-let-7c	NA	12810.00	1776.00	7.21	0.00
hsa-let-7c	Ref	336.00	0.00	9999.00	0.00
hsa-let-7d	Ref	45.00	0.00	9999.00	0.00
hsa-let-7e	NA	252.00	749.00	0.34	0.00
hsa-let-7e	Ref	590.00	2228.00	0.26	0.00
hsa-let-7f	NA	1668.00	790.00	2.11	0.00
hsa-let-7f	Ref	2590.00	2157.00	1.20	0.00
hsa-let-7g	NA	66.00	0.00	9999.00	0.00
hsa-let-7g	Ref	363.00	152.00	2.39	0.00
hsa-let-7i	NA	62.00	0.00	9999.00	0.00
hsa-let-7i	Ref	71.00	0.00	9999.00	0.00

14. Target prediction analysis: "Get predicted targets" package

This package allows miRNA target prediction using the most common target prediction algorithms: Targetscan, miRBase and Pictar

14.1. Go to the "Analysis" option in the menu and choose "Target Prediction"

14.2. Select "Common Prediced Targets" in the "Target Prediction" panel

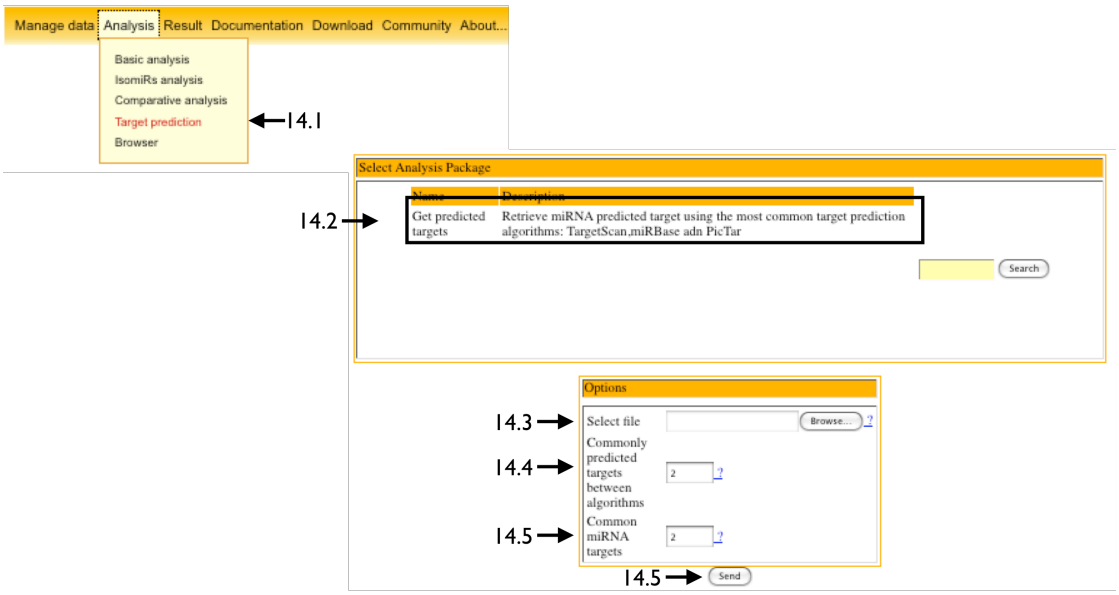
14.3. In the "Options" panel, several parameters may be chosen:

14.3.1. Upload a "txt" file containing a list of miRNAs in the the "Select file" textbox using the "Browse" button

14.3.2. In the "Commonly predicted targets" textbox chose a minimum number of algorithms predicting the same mRNA targets. Default is 2

14.3.3. The “Common miRNA targets” textbox choose a minimum number of miRNAs in the provided list, targeting the same mRNA

14.4. Press “Send” button



15. Browser

This package provides a list of all the sequences in a sample

15.1. Go to the “Analysis” option in the menu and choose “Browse”

15.2. In the “Chose table” select the sample to be considered in the analysis, in the present example “hESC” is chosen

15.3. A message appears indicating “You are Browsing into hESC”

15.4. In the “Select Parameters” panel, different options may be chosen:

15.4.1. In the “Sequence” textbox write the sequence that is going to be analyzed, indicating with the select box if the sequences should be exactly as the written sequence (LIKE), or contain the indicated sequence (CONTAIN), or exclude it (NOT CONTAIN). In the present example, no specific sequences are specified (empty box)

15.4.2. “Frequency” and “Length” ranges may be chosen. In the present example all frequencies and lengths are considered

15.4.3. In the “Locus” textbox a specific locus may be selected for the analysis. Chose “LIKE” “CONTAIN” “NOT CONTAIN” with the select box as described in 15.4.2. section. In the present example all loci are selected (empty box)

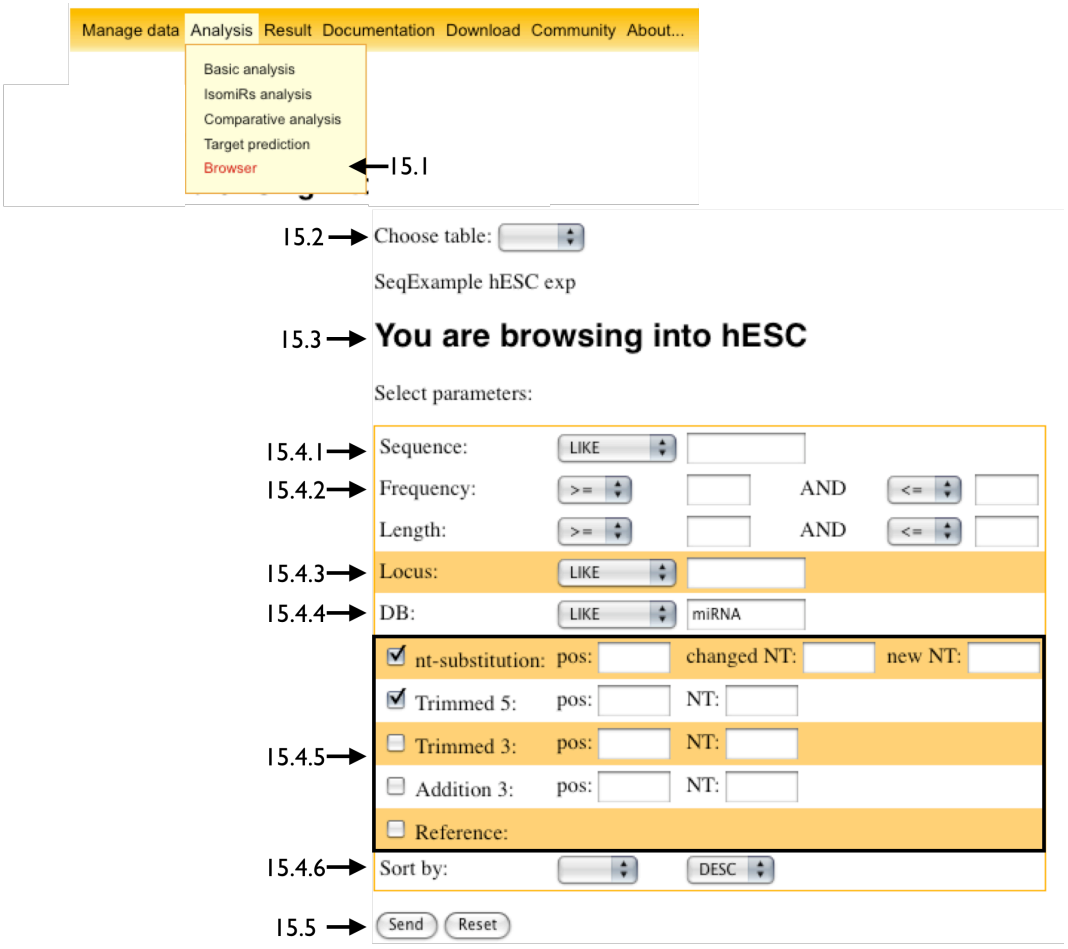
15.4.4. In the “DB” textbox chose one of the databases used for annotation purposes (see Tutorials 2 and 3). Chose “LIKE” “CONTAIN” “NOT CONTAIN” with the select box as described in

15.4.2. section. In the present example “miRNA” database is selected

15.4.5. In the options below select the type of miRNAs (reference or isomiRs) to be included in the list. For each of the variants the position and the type of nucleotide involved may be selected. In the present example “nt-substitution” and “5’-trimming” variants are considered

15.4.6. In the “Sort by” options for the output list may be sorted by frequency (freq) length (len), chromosome (chr) or database (DB) in the left select box, and whether data in the output list are showed in increasing (ASC) or decreasing (DES) order is selected in the right selectbox

15.5. Press the “Send” button



15.6. In the output analysis a table appears showing a list of the sequences (seq), with the corresponding frequences (freq), the length (len), locus where it has been maped (loc), start (start) and end (end) with respect the

reference miRNA and the selected data base (DB). A description of the different variants is shown in the last four columns. 5'-trimming and 3'-trimming variants are identified with a "t" if the trimming occurs downstream or upstream of the reference miRNA, respectivelly; the nucleotide or nucleotides appearing after the "t" indicate those nucleotides present in the reference miRNA and absent in the trimming variant. For instance "tU" means a trimming variant lacking a "U" that was present in the 3'-end or 5'-end of the reference miRNA (1). 5'-trimming and 3'-trimming variants identified with a "q" indicate that the trimming occurs upstream or downstream of the reference miRNA, respectively. For instance qG (2) indicates a trimming variant presenting a additional "G" at the 5'-end or the 3'-end of the reference miRNA. 3'-Addition variants are always identified with a "q"; for instance a 3'-addition variant with a "qAUC" label (3) indicates an isomiR with an extension of 3 nucleotides (AUC) at the 3'-end of the reference miRNA. The "nt-substitution" variants are described by a number followed by 2 nucleotides; for instance "9GU" label (4) indicates a variant affecting the position 9 of the miRNA that presents a "G" to "U" modification.

Figure 20

ID	seq	freq	len	locus	start	end	strand	DB	Trimmed 5	Trimmed 3	Addition 3	Mutation
174	AAGUGCUUCCAUGUUUUAUUAG	23	22	hsa-miR-302b	2	23	+	miRNA	tU	0	0	20UG
612	ACCACAGGGUAGAACCCCGGAC	101	22	hsa-miR-140-3p	2	21	+	miRNA	tU	qAC	0	18CA
1120	UGGAUAAGGCAUUGGCAUC	15	19	hsa-miR-1261	2	17	+	miRNA	tA	tUU	qAUC	12AU
1519	ACCACAGGGUAGGACCACGGAC	13	22	hsa-miR-140-3p	2	21	+	miRNA	tU	qAC	0	14GA
2003	GUACAGUACUGUUAUAACUGAA	16	22	hsa-miR-101	1	21	+	miRNA	qG	0	0	12UG
2109	AAGUGCUUCCAUGUUUGCGUGU	11	22	hsa-miR-302d	2	23	+	miRNA	tU	0	0	19CA
2280	AUGGAUUUUUGUAGCAGGG	13	19	hsa-miR-1246	2	19	+	miRNA	tA	0	qG	13UG
2535	UCCCACCGCUGCCACCA	31	17	hsa-miR-1260	2	18	+	miRNA	tA	0	0	9GU
2676	GUACAGUACUGUGAUAACUGCA	49	22	hsa-miR-101	1	21	+	miRNA	qG	0	0	20CA
2853	AAGUGCUUCCAUGUUUUAGUGU	12	22	hsa-miR-302d	2	23	+	miRNA	tU	0	0	18UG
2915	GUACAGUACUGUGUUAACUGAA	12	22	hsa-miR-101	1	21	+	miRNA	qG	0	0	13UA
2944	ACCACAGGGUAGCACCACGGAC	13	22	hsa-miR-140-3p	2	21	+	miRNA	tU	qAC	0	14CA