Published online 11 December 2009

Nucleic Acids Research, 2010, Vol. 38, No. 5 ***e34***
doi:10.1093/nar/gkp1127

SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells

Lorena Pantano[1,2], Xavier Estivill[1,2,3,*] and Eulàlia Martí[1,2,*]

[1]Genetic Causes of Disease Group, Genes and Disease Program, Centre for Genomic Regulation (CRG), [2]Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP) and [3]Experimental and Health Sciences Department, Pompeu Fabra University, Barcelona, Catalonia, Spain

Received June 8, 2009; Revised and accepted November 13, 2009

ABSTRACT

High-throughput sequencing technologies enable direct approaches to catalog and analyze snapshots of the total small RNA content of living cells. Characterization of high-throughput sequencing data requires bioinformatic tools offering a wide perspective of the small RNA transcriptome. Here we present SeqBuster, a highly versatile and reliable web-based toolkit to process and analyze large-scale small RNA datasets. The high flexibility of this tool is illustrated by the multiple choices offered in the pre-analysis for mapping purposes and in the different analysis modules for data manipulation. To overcome the storage capacity limitations of the web-based tool, SeqBuster offers a stand-alone version that permits the annotation against any custom database. SeqBuster integrates multiple analyses modules in a unique platform and constitutes the first bioinformatic tool offering a deep characterization of miRNA variants (isomiRs). The application of SeqBuster to small-RNA datasets of human embryonic stem cells revealed that most miRNAs present different types of isomiRs, some of them being associated to stem cell differentiation. The exhaustive description of the isomiRs provided by SeqBuster could help to identify miRNA-variants that are relevant in physiological and pathological processes. SeqBuster is available at http://estivill_lab.crg.es/seqbuster.

INTRODUCTION

Small silencing RNAs are a family of non-coding RNAs of 20–30 nt in length, associated with members of the Argonaute family of proteins that are effectors of the small RNA-directed silencing. Small non-coding RNAs are involved in the guidance of diverse formats of gene regulation, typically resulting in reduced expression of target genes. The different classes of regulatory RNAs differ in the type of RNA precursor and proteins required for their biogenesis, the constitution of the complexes mediating the regulatory process and the biological functions in which they participate [reviewed in (1) and (2)].

In animals, the small RNA family includes highly abundant and functionally important RNA classes, such as small interfering RNA (siRNA), Piwi interacting RNA (piRNA) and microRNAs (miRNAs). Early examples of siRNA-mediated gene expression regulation included silencing induced by exogenous double stranded RNA (dsRNA) such as that from viruses. However, endo-siRNAs deriving from transposons, heterochromatic sequences, intergenic regions or mRNAs have been recently described in *Drosophila* and mammals, although their biological role remains largely unknown (3,4). piRNAs have thus far been found only in germ cells, repressing the activity of mobile genetic elements (5).

miRNAs are the best-known class of small silencing RNAs. miRNAs are genomically encoded and expressed as long precursor RNAs (pri-miRNAs) that are processed by the RNAses III Drosha and Dicer to 20–24 RNA duplexes. Mature miRNAs use base pairing to guide RNA-induced silencing complexes (RISCs) to the

*To whom correspondence should be addressed. Tel: +3493 316 0201; Email: eulalia.marti@crg.es
Correspondence may also be addressed to Xavier Estivill. Email: xavier.estivill@crg.es

3′UTR of mRNAs with fully or partially complementary sequences. The repression of target mRNA is a common outcome of RISC recruitment and might occur through translational inhibition or mRNA degradation. miRNAs are ubiquitously expressed and are believed to regulate most biological processes in a tissue- and temporal-specific manner, with a potential role in a number of pathological processes, including cancer and neurological disorders (6–8).

The identification of a near complete set of small RNAs in organisms is of fundamental importance to understanding small-RNA-mediated gene regulation. The available second-generation sequencing technologies, including 454/Roche, Illumina/Solexa and SOLID, offer a novel perspective for small RNA characterization, enabling quantitative estimates of expression profiles and the discovery of novel small RNAs by direct observation and validation of the folding potential of flanking genomic sequence (9). One of the distinctive capabilities of direct sequencing is the detection of variation in the mature miRNA sequence. miRNA variability has recently been described using several large scale sequencing strategies in plants (10,11), mouse tissues and human stem cells (12,13) and human brain samples (14). These miRNAs variants have been designated as isomiRs (12). IsomiRs can be the consequence of Drosha and Dicer enzymatic activities during miRNA biogenesis, which cleave the pre-miRNA at variable positions (5′- and 3′-trimming IsomiRs). In addition pri-miRNA post-trancriptional editing as a consequence of adenosine or cytidine desaminase activities results in nucleotide changes at different positions of the mature miRNA (nt-substitution isomiRs) (10–20). Besides, nucleotide additions at the 3′-end of the mature miRNA have been reported as the most common form of miRNA enzymatic modification (3′-addition isomiRs) (11,12). Therefore, deep sequencing provides a more complete view of the miRNA transcriptome in a quantitative and qualitative fashion.

A major problem arising from high-throughput sequencing strategies is the management of huge amounts of data. Illumina in its current sequencing protocols and capacities produces over 7 million reads per sample. The analysis pipelines published to date are focused on general characterization of small RNAs, differential expression analyses between libraries and prediction of new miRNAs (21–24). Here we present SeqBuster, an easy to use web-based toolkit specifically designed to process and analyze large-scale small RNA datasets. SeqBuster offers different types of analyses, including the identification of small RNAs, length and frequency distribution and the comparative expression levels of different small RNA loci between different samples. Notably, SeqBuster is the first web server tool offering several packages capable of deeply characterizing qualitative and quantitative miRNA variability. To demonstrate the pipeline usefulness, Illumina/Solexa deep sequencing data provided by Morin *et al.* (12) have been loaded into SeqBuster and re-analyzed. The results of this analysis suggest that this bioinformatic tool has the potential to uncover small silencing RNA-related mechanisms underlying biological processes.

METHODS

SeqBuster implementation

Raw data processing has been performed in a Java-based GUI (Graphical User Interface) engine as a stand-alone version of SeqBuster available at http://estivill_lab.crg.es/seqbuster/download. SeqBuster web-interface is based on DHTML (Dynamic HTML) and CGI (Common Gateway Interface) architecture. Pre-analysis modules have been developed in Perl language and for the computational modules the R statistical package has been used. The users may upload R/perl-based packages to the server, offering new or modified analyses to the community (SeqBusterDev Center). Data were stored and handled in a MySQL platform. The database is composed of three data classes: general data, sample data and results data. The general data stores all reads that have been pre-analyzed and the different annotations according to the distinct databases used. In subsequent experiments, only new sequences that have not been detected previously will go through the pre-analysis module. Each sequence is assigned with an ID number, which is used as a key to interact with other tables and modules. Sample data contain a table per experiment that stores all the information needed for the posterior analysis (ID sequence, size, frequency, annotation, type of annotation and type of sequence variability). Results data consist in a single table containing the name of every output file resulting from a specific analysis and may be saved permanently by the user.

Sequencing data

Public raw data produced by Illumina deep sequencing of short RNAs in undifferentiated and differentiated human embryonic stem cells (hESC and EB) (12) were downloaded from ftp03.bcgsc.ca/public/hESC.

Pre-analysis of embryonic stem cells sequencing data

The machine used for the pre-analysis was a HP Workstation xw9300 with a Dual Core AMD Opteron(tm) processor 275 2194.15 MHz and 8 Gb of RAM memory. We have used the stand-alone version for the adapter recognition step allowing only three mismatches and no gaps. The minimal size of the adapter recognized was set up to 10 nucleotides. The time process was 30 min for 6 millions reads. After that, sequences were annotated using human pre-miRNA and mature miRNA databases provided by the miRBase (http://microrna.sanger.ac.uk/sequences/) available at the SeqBuster server (15 min required). In addition, using the stand-alone version, the data were also mapped onto mRNA and genome databases (1 and 2.5 h required, respectively) (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). The annotated files were uploaded to the server and stored for subsequent analyses.

For precursor-miRNAs annotation, the following parameters were configured: one mismatch, three nucleotides in the 3′ addition variants and the priority degree equal to 3. For miRNAs and miRNA*annotation, the following parameters were configured: one mismatch,

three nucleotides in the 3′ or 5′ trimming variants, three nucleotides in the 3′ addition variants, a priority degree equal to 1 and 2 for the miRNA and the miRNA* databases, respectively, and the parental database was the precursor miRNA database. These options permitted the annotations of the following alignments: (i) perfect match, where the sequence is completely identical to the reference sequence; (ii) trimming at the 3′-end of the reference miRNA sequence, which is a miRNA variant several nucleotides shorter or longer that matches to the mature or precursor reference sequence, respectively; (iii) trimming at the 5′-end of the sequence, an analogous case focused in the 5′-end of the miRNA; (iv) nucleotide additions at the 3′-end of the sequence and (v) nucleotide substitutions, showing nucleotide modifications with respect to the reference sequence. The parameters for the alignment in mRNA and genome databases allowed as much as one mismatch and up to three nucleotide additions in the 3′-terminus. The priority parameter was equal to 4 and 5 for the mRNA and genome databases, respectively.

IsomiR analysis of embryonic stem cells sequencing data

For deep characterization of miRNA variants, we applied several filters in the different 'IsomiRs analysis' packages. First, the sequences considered in the analysis presented a frequency above 3. Second, 10 was chosen as the 'Contribution Cut-Off' parameter, meaning that every isomiR considered in the analysis contributes in more than 10% to the total number of variants annotated in the same miRNA locus. Third, we applied the *Z*-score option to exclude sequencing errors as the possible cause of the nucleotide changes observed in some variants (25).

Differential expression in embryonic stem cells sequencing data

The sequencing performance in the samples to be compared was evaluated using the 'Sequencing Capacity' package in the 'Basic Analysis' module of SeqBuster with default parameters. In this package, the Willcoxon test was applied to determine statistically significant differences in the frequency distribution between samples. For the expression profiling we used the 'Differential Expression Analysis' module. Two types of sequences were considered in the differential expression analyses: 5′-trimming and nucleotide-substitution isomiRs affecting the seed region of the miRNA and sequences perfectly matching the reference miRNAs. To perform these analyses several options and filters were applied. First, the frequency was normalized to r.p.m. (reads per million). Second, the *Z*-test (26) was applied to show statistical significance in the differential expression. Third, the Hochberg and Benajmini (27) method was applied to correct the *P*-value assigned by the *Z*-test. Fourth, in analyzing the seed region variants, 10 was chosen as the 'Contribution Cut-Off' value (see isomiR analysis for description). Fifth, we applied the *Z*-score option to exclude sequencing errors as the possible cause of the nucleotide changes observed in the isomiR (25). Sixth, 'Reference', '5′ trimming' and 'Nt-substitution' (start position = 2 and end position = 8) options were selected in order to discriminate between sequences with variants affecting the seed region. In the output resulting from the analysis, we only showed sequences with a total count contribution above 50, considered as the sum of the frequencies of the two libraries.

Function enrichment analysis

We used the TargetScan algorithm that predicts biological targets by searching for the presence of conserved 8-mer and 7-mer sites that match the seed region of each isomiR (28). We used the TargetScan custom option (www.targetscan.org) to predict mRNA targets of the seed region isomiRs differently expressed between libraries. The mRNA targets for the corresponding reference miRNAs differently expressed were identified through the SeqBuster 'Target prediction' module using the TargetScan (28) algorithm. Since the cooperative action of multiple miRNAs can be multiplicative and sometimes synergistic (29), mRNAs with more predicted target sites for co-expressed isomiRs or reference-miRNAs should be more drastically affected. Therefore, we considered targets predicted by more than one hESC- or EB-enriched isomiRs or reference miRNAs. Then, ingenuity pathway analysis (IPA) was used for the subsets of genes exclusively targeted by hSCE- or EB-enriched isomiRs and those affected by the corresponding reference miRNAs. The *P*-value associated with a biological process is calculated with the right-tailed Fisher's exact test, considering the number of functions/pathways/lists eligible molecules that participate in that annotation, the total number of knowledge base molecules known to be associated with that function, the total number of functions/pathways/lists eligible molecules and the total number of genes in the reference set (IPA tutorial).

RESULTS

SeqBuster overview

We have developed SeqBuster, a web-based bioinformatic tool offering a custom analysis of deep sequencing data at different levels, with special emphasis on the analysis of miRNA variants or isomiRs. The pipeline for small RNA analysis is available at http://estivill_lab.crg.es/seqbuster (Figure 1) and includes a pre-analysis module, for raw data processing, and an analysis module (tutorials 1–4 provided in the 'Documentation' section of the web-server home page). Pre-analysis consists in the recognition and removal of the adapter and the annotation of the sequences. The several gigabytes of output that are generated after a sequencing experiment are by far too many data to be pre-analyzed using a web server tool. Therefore, to perform different steps of the pre-analysis, SeqBuster includes a java-based, user-friendly stand-alone version available in the 'Download' option of SeqBuster home page. Recognition and removal of the adapter is performed using the stand-alone version (tutorial 1). Annotation of the sequences can be performed with the web server (tutorial 2) that contains the miRNAs and pre-miRNAs databases (http://microrna.sanger.ac.uk/sequences/) or through the stand-alone version using

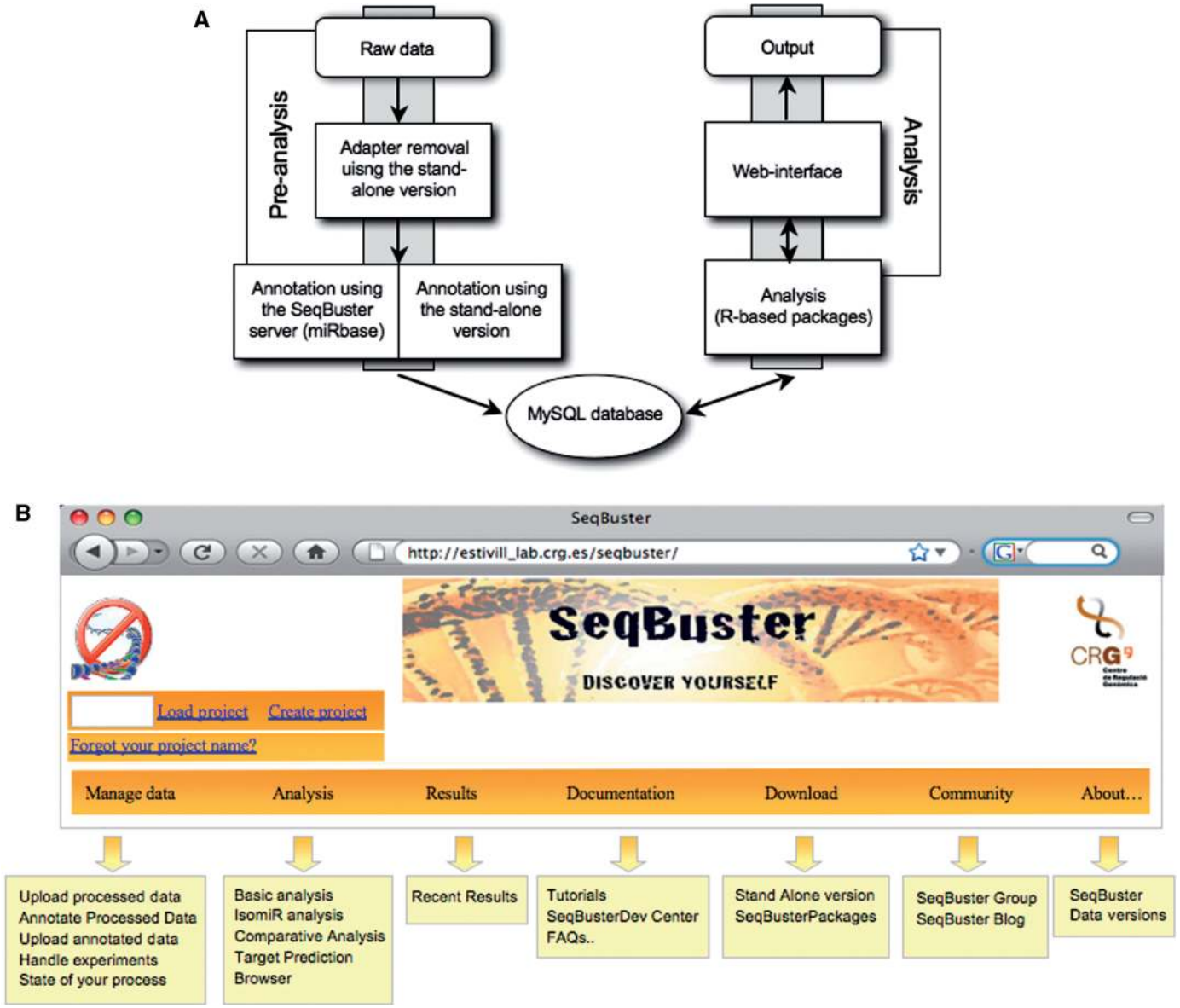

Figure 1. (**A**) Workflow of SeqBuster pipeline showing the architecture and connection of pre-analysis and analysis modules. In the pre-analysis module, raw data are processed for recognition and annotation. Annotation can be performed through the web server that offers the miRNA and miRNA-precursor databases or through a stand-alone version using any custom database. The processed and annotated data are stored in a MySQL database. The web interface permits the analysis of the data using several R-based packages. The output of every analysis is visualized through a Dynamic HTML format and stored in the server or downloaded to the local machine. (**B**) Scheme of the main menu at SeqBuster home page. The different choices offered by each option in the menu are highlighted in light yellow boxes.

any custom database (tutorial 3). The pre-analysis produces a table that is then uploaded to the server in a project for subsequent analysis (tutorial 4). The project may contain as many tables as sequencing experiments.

The analysis modules offered by the web server include (i) a general characterization of small-RNA datasets, (ii) a deep analysis of miRNAs variability (IsomiRs), (iii) differential expression analysis and (iv) target prediction for sets of miRNAs. All tutorials for data pre-analysis and analysis are illustrated with practical examples (the SeqExample project). In addition, other R/perl-based specific analyses are available in the 'Download' section of the SeqBuster web-server home page.

In the following sections, we describe the algorithms SeqBuster applies for pre-analysis and the different options that can be selected by the user. Then we similarly describe the web-based analysis module with special focus on the miRNA variability analysis. Finally, we present the results of applying SeqBuster to human stem cells small-RNA sequencing data (12).

Pre-analysis

For the pre-analysis, we have developed a pipeline tool to parse small RNA sequences from the adapter, collapse the data to uniread set, count the number of reads per unique

sequence, map sequences to different databases and annotate sequences with basic information. The adapter recognition and removal is performed using the SeqBuster stand-alone version, supporting raw data from Illumina/Solexa and FLX454 technologies (see tutorial 1 for a detailed procedure). The algorithm implemented for adapter recognition generates all possible candidate adapters for each sequence, considering that the adapter starts at position 15 to position −10 of the end of the sequence. For instance, for a read of 35 nucleotides, 11 candidate adapters would be generated, starting at 11 different positions (positions 15–25). This strategy ensures that at least the first 10 nucleotides of the adapter are being recognized. The starting and final positions to generate all candidate adapters can be modified according to the user criteria. For each read, these candidate adapters are aligned to the adapter sequence using a modification of the Needleman–Wunsch algorithm (30) that does not allow gaps, therefore increasing the speed of the process. Then the alignment with the best score is selected. In addition, we have established a default value of three mismatches in the alignment process, since these conditions ensure at least 85% of identity in the adapter alignment; however, this threshold can be varied according to the user requirements. For each read, the recognized adapter sequence is removed and a table is generated containing all unique sequences and the corresponding counts. Overall, this strategy resulted in an increased number of recognized sequences compared with the original analysis performed by Morin *et al.* (12) (SeqBuster application on a biological example).

After recognition and removal of the adapter, the sequences need to be annotated against different databases. The individual steps for the computational annotation are as follows: (i) load all the databases against which the sequences are going to be annotated if using the stand-alone version; (ii) chose a database; (iii) map to the chosen database using Mega BLAST; (iv) upload the annotated file to the SeqBuster web server if using the stand-alone version and (v) repeat steps ii—iv with other databases. SeqBuster has implemented the Mega BLAST algorithm of the BLAST repository instead of BLASTN (www.ncbi.nlm.nih.gov/blast/megablast.shtm) because an increased number of sequences were annotated, using the data of Morin *et al.* (12) (Table 1). This is the consequence of the application of two consecutive alignment strategies that consider different reward and penalty parameters for matches and mismatches, respectively (see Mega BLAST at www.ncbi.nlm.nih.gov/BLAST for further information). In one of the procedures, default reward and penalty parameters were considered. In the other method, a reward parameter of 3 and a penalty parameter of −2 were used in order to force the alignment at the beginning of the sequences. In both procedures a word size of seven was used. This resulted in an increased number of sequences being annotated with a mismatch at the beginning and the end of the sequence (Figure 2).

For annotation, SeqBuster differentiates two types of databases: the parental and child databases (tutorials 2 and 3). This permits the identification of sequences in the child database that vary in their extremes but match perfectly with the parental database. For instance, miRNA is a child database and the precursor-miRNA is the corresponding parental database. Using the child and parental databases, the isomiRs resulting from variations in the cleavage positions in the pre-miRNA during miRNA biogenesis are annotated as 5′- or 3′-trimming variants. The number of positions up-stream (5′-trimming) or down-stream (3′-trimming) of the reference sequence to be considered in the trimming variants can be custom set.

Table 1. Benchmarking for the alignment of hESC reads against miRNA data set, using BlastN or megablast

Parameters	BlastN	megablast
word size = 7, penalty = −3, reward = 1	20 790	23 444
word size = 7, penalty = −2, reward = 3	23 644	25 152

The table shows the number of sequences successfully annotated depending on the parameters selected.

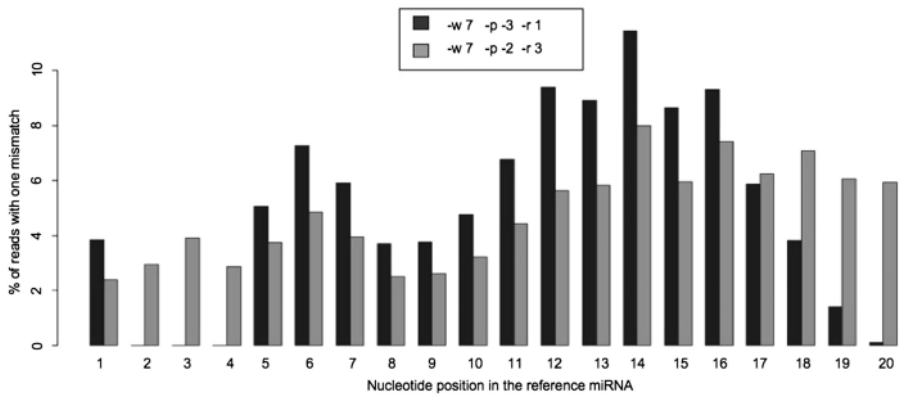

Figure 2. Percentage of reads with a mismatch at different positions of the reference miRNA detected by SeqBuster, considering two different annotation strategies. Penalty and reward parameters of −3 and 1 (black bars) or −2 and 3 (grey bars) were used. In both strategies a word size of 7 was considered.

For annotation using a parental database (for instance precursor-miRNAs or genome) or a child database (for instance miRNA), the parameters offered by SeqBuster web-server or stand-alone version are (i) the 'word size' parameter that is the size of the initial word that must be matched between the database and the query sequence (Blast manual at www.ncbi.nlm.nih.gov/BLAST); (ii) the 'Mismatch' parameter meaning the number of mismatches allowed in the alignment; (iii) the 'Priority' parameter that provides a priority degree to the database in each experiment (for instance a priority of 1 for 'miRNA' and 2 for 'genome' databases means that all sequences annotated as miRNA and genome are going to be labeled as miRNA) and (iv) the 'Addition' parameter that refers to the number of nucleotides that will be considered as a result of the nucleotide addition process that results in an extended sequence in the 3′-terminus.

Analysis module

Once the raw data have been processed and annotated, several analyses can be easily performed through the 'Analysis' option in the SeqBuster web-server home page (tutorial 4). Each analysis module contains several packages, which in turn, hold multiple options and filters both for the analysis and for the visualization of the results. The output is presented in the form of downloadable figures and tables.

Basic analysis: general characterization of small RNA datasets

The basic analysis includes several packages. The 'General information' package offers the analysis of the distribution of the sequences in different lengths or classes. The 'Frequency distribution' package provides an analysis of the frequency distribution of different types of sequences. The 'Adapter quality' package is used to visualize the quality of the adapter (the 10 first nucleotides) in selective sequences. This may provide an idea of the quality of the adapter-attached sequence. The 'Experiment capacity' package explores the sequencing performance or sequencing capacity in multiple sequencing experiments. To this aim, all sequences are ordered by decreasing frequency. The resulting distribution should be similar in experiments that need to be compared in differential expression analysis. The algorithm applies the Willcoxon test to determine statistically significant differences in the frequency distribution between samples. If the distribution is significantly different, several normalization strategies can be applied to make the samples comparable. To this aim, SeqBuster includes a basic equation to scale the frequencies according to the following equation: scaled freq n = (freq n/sum [freq all seqs]) × scale-value. In addition, 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. These extreme values can be removed from the analysis, through the selection of an upper and lower frequency cutoff in the 'discard upper quantile' or 'discard lower quantile options. Finally, different types of metric centers (mean, median, min, max, etc.) may be applied to normalize the frequency distributions. In all the packages several parameters can be specified for the analysis including the sequences length and frequency and the type of database to be considered. In addition, the sequence frequency values can be expressed at different logarithmic scales or represented as a percentage or as absolute values. Finally, for some graphic representations of the analysis output, the user can also choose between pie and bar charts.

IsomiR analysis: characterization of the miRNA variability

One of the more innovative advantages of deep sequencing is the detection of sequence variability. SeqBuster offers the selective analysis of the 5′-trimming, 3′-trimming, nt-substitution and/or 3′-addition isomiRs using different packages. The 'IsomiR distribution' package shows the percentage of miRNAs presenting the different types of variability (or IsomiRs) (Figure 3). In the output analysis, a histogram displays the proportion of miRNAs with different types of isomiRs in all the selected samples. Statistical differences in the abundance of the different types of variability are determined, using the Fisher test.

In thinking about the possible physiological importance of the isomiRs, a reasonable possibility exists that the IsomiRs target new mRNAs; however, since the variant and the reference miRNA sequences are very similar it is likely that both small RNAs compete for a number of target mRNAs. Therefore, to approach the issue of the possible relevance of each variant, the 'IsomiR distribution' output graph illustrates the frequency of the isomiRs with respect to that of reference miRNAs, in a mirror histogram showing a brown color scale. For each isomiR, a ratio is calculated following the equation: fv/(fr + fv) × 100, where fr is the frequency of the reference miRNA sequence and fv is the frequency of the isomiR. Values vary between 100 and 0. The closer the value is to 100 the higher the frequency of the isomiR is with respect to the reference sequence (dark brown). On the contrary, values close to 0 indicate that the frequency of the isomiR is negligible compared with that of the reference sequence (light brown). In addition, in the output analysis, a table appears below the histogram that groups the miRNAs according to the number of variants (one variant in white; two variants in gray; or more than two variants, in black), the type of isomiR and the relative abundance of the isomiR with respect the corresponding 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 'IsomiR by nucleotide position' package shows the percentage of miRNAs with a specific type of variant according to the nucleotide type and position involved (Figure 4A). In the analysis output, a histogram displays the percentage of miRNAs with a selected type of variability, at specific positions. The color pattern in the upper bars indicates the proportion of each type of nucleotide present in the isomiR, at every desired position. To study

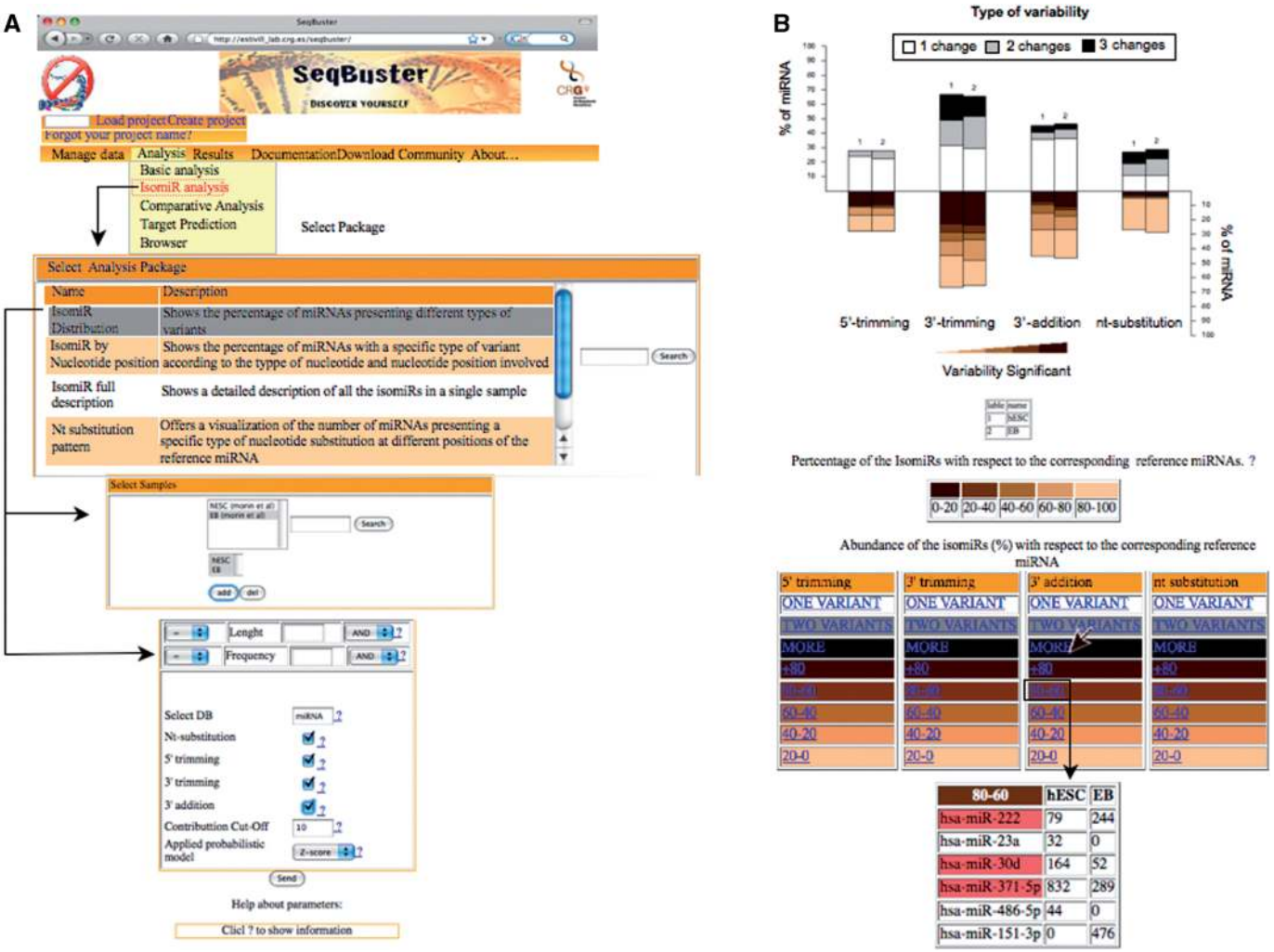

Figure 3. The 'IsomiR distribution' package scheme. (**A**) Within the 'IsomiR analysis' several packages appear in a general menu. After selecting the 'isomiR distribution' package the samples for the analysis should be chosen. Up to four samples can be loaded to the analysis. Different options and parameters may be configured in order to customize the study (tutorial 4). (**B**) In the output analysis, a histogram displays the proportion of miRNAs with different types of isomiRs in all the selected samples. 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 to the corresponding reference miRNA is mirrored in the lower graph in a brown color scale. The five brown color intensities from dark to light indicate the frequency of the isomiR with respect to that of the reference miRNA: 1, $>80\%$; 2, 60–80%; 3, 40–60%; 4, 20–40% and 5, $<20\%$. Below the graph, a table helps to obtain the complete information of the analysis. All the miRNAs contained in the histogram can be listed by clicking on the corresponding link. Those miRNAs highlighted in pink are commonly detected in all the samples examined.

a possible significant enrichment of a nucleotide variant in a certain position (indicated with an asterisk), a *P*-value is assigned using the bootstrapping method (1000 permutations). The nucleotide contribution at every position of the reference-miRNA collection is considered as the population for the permutations. The proportion of the isomiRs with respect to the reference miRNA is presented in a mirror histogram (brown color scale). In the output graph resulting from the analysis of the nt-substitution variants, the upper histogram shows the nucleotides present in the reference-miRNA and the mirror histogram those present in the isomiR. A table below the histogram shows the different nucleotides involved in every position of the variant. By clicking on the nucleotides, a list of the miRNAs and the corresponding frequency appears for every sample.

The 'isomiR full description' package provides a detailed visualization of all the isomiRs (one or several types of variability may be chosen) in a selected sample (Figure 4B). In the output analysis, a table shows a list of the miRNAs presenting the variants specified in the analysis. In this table, the type of nucleotide, the position and the frequency with respect to the reference miRNA are highlighted.

The 'Nt-substitution pattern' package offers a visualization of the number of miRNAs showing a specific type of nucleotide substitution at different positions of the reference miRNA (Figure 4C). This is presented in the form of a table in the analysis output. Nucleotides in the rows correspond to those found in the reference miRNAs and nucleotides in the columns are those found in the isomiRs. The identification of the types of nucleotide substitutions

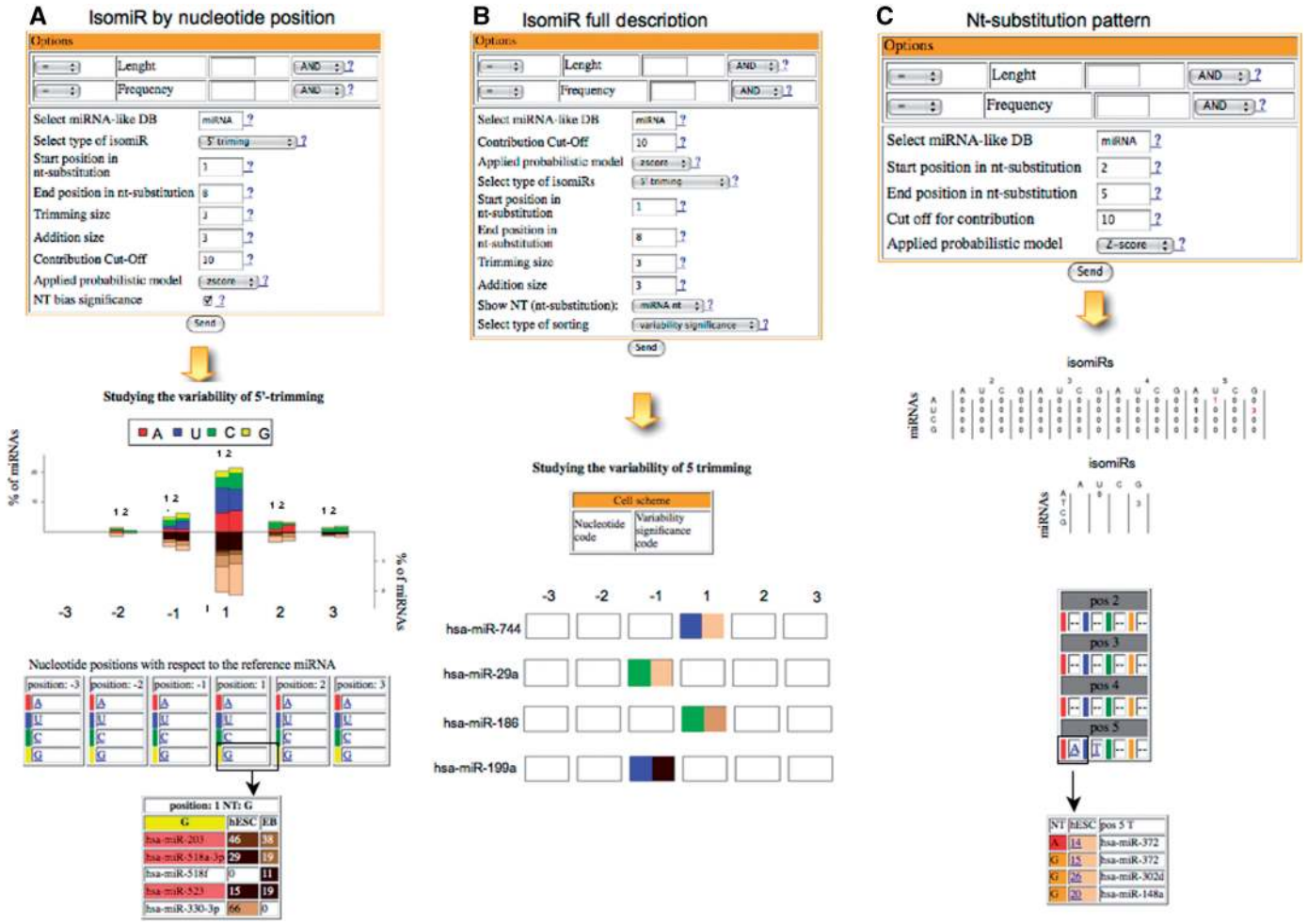

Figure 4. Summary of the main packages available in SeqBuster IsomiR analysis module. For the three packages, the parameters that can be configured (tutorial 4) and the analysis output are shown. (**A**) 'IsomiRs by nucleotide position' package. This example shows 5′-trimming variants involving three positions upstream and three positions downstream of the reference-miRNAs. The output shows the proportion of miRNAs with a trimming variant in a specific position. In the upper bars, the color pattern indicates the proportion of miRNAs showing a specific nucleotide being involved in the trimming variants. The lower bars show the proportion of the isomiRs with respect to the corresponding reference miRNA in a brown color scale, as described in Figure 3. A list of miRNAs involved in each type of variant is displayed when clicking inside the table. (**B**) 'IsomiR full description' package. The output analysis of this example shows a list of some of the miRNAs presenting 5′-trimming variants involving three positions upstream and three positions downstream of the reference-miRNAs. Every cell represents a position. The color pattern on the left half of the cell indicates the type of nucleotide present in the isomiR, and that of the right side the proportion of the isomiR with respect the reference miRNA. (**C**) 'Nt-substitution pattern' package. The output analysis of this example shows a table with the number of miRNAs presenting any of the 12 possible nt-substitution events in positions 2–5 of the reference miRNA. A summary table showing the number of miRNA with any of the 12 possible nt-changes significant changes is also represented. The overall nt-substitution pattern occurring in a statistically significant number of miRNAs is shown in another table. The list of miRNAs for each class of nt-substitution can be retrieved by clicking on the summary table.

affected in a significant number of miRNAs is established with a *P*-value. This *P*-value is assigned using the bootstrapping method above described. To study in the population of modified nucleotides present in the isomiR, each nucleotide is assigned the same probability (1/4) in the statistical method. Finally, the 'Invariable miRNAs' package provides a list of the miRNAs that do not present variability.

For deep characterization of miRNA variants SeqBuster offers several filters in the different packages. Sequences to be considered in the analysis can be filtered according to specific length and/or frequency cutoff. The type of isomiRs and the specific positions and number of nucleotides for characterization can also be specified. In addition SeqBuster permits the selection of isomiRs contributing in more than a specific percentage, to the total number of variants annotated in the same miRNA locus ('Contribution cutoff' parameter). Finally, a probabilistic *Z*-score option can be selected to exclude sequencing errors as the possible cause of the nucleotide changes observed in some variants. This model is based on previous determinations of Illumina sequencing error rates that have been calculated for every position of the sequence and every type of nucleotide change, in genomic libraries (25). The *Z*-score is calculated according to Equation 1 that considers the frequency of the variants

(p_o), the expected frequency error (p_e) and the sum of the frequencies of all the sequences mapping in the same miRNA locus (total).

$$Z = \frac{p_e + p_o}{\sqrt{p_e(1 - p_e)/\text{total}}} \quad \mathbf{1}$$

Since the *Z*-score follows a standard normal distribution, a *P*-value can be established for each variant, being considered as true variants, those presenting a $P < 0.05$. In conditions in which the *Z*-score cannot be applied: $p_e \times \text{total} < 5$ (Equation 1), the *P*-value is calculated as the sum of the binomial distribution tail (Equation 2), where *N* is the number of sequences in a locus, *p* is expected frequency and *v* is the frequency of the observed variant.

$$p = \sum_{n=v}^{N} \text{Binomial}(n, N, p) \quad \mathbf{2}$$

The *P*-value obtained by any of these two strategies is transformed into a *q*-value using the false discovery rate correction. If the *Z*-score option is selected, the variants considered as sequencing errors are not included in subsequent analyses.

Comparative expression analysis

One of the main goals of small-RNA high-throughput sequencing is to characterize the expression across tissues or states and to understand the changes that take place in different physiological and pathological conditions. In sequencing experiments, we can detect and quantify the changes in the relative frequencies of small-RNAs within samples. For differential expression, SeqBuster includes a sample cluster analysis and small-RNA differential expression analysis modules. In both modules, SeqBuster offers two types of packages: (i) a general comparative analysis in which all the sequences or a specific type of database can be chosen and (ii) a package to specifically study the miRNAs, providing the selective analysis of the different types of miRNA-variants. If all sequences mapping in the same miRNA locus are considered collectively, the frequency is expressed as the sum of the frequencies of all sequences (reference miRNAs and variants) mapping onto the same locus. If a type of isomiR is selected, for instance 5′-trimming variants, the output analysis shows for every miRNA locus, the frequency of all the sequences presenting 5′-trimming variants and the frequency of the rest of the sequences collapsed into another group. The frequency of all the sequences mapping in a specific miRNA locus will be distributed in as many groups as different types of variants are selected. Furthermore, different parameters permit to direct the analysis to a selective region of the miRNA. For instance, SeqBuster allows the selective characterization of nucleotide-substitution variants considering only the seed region (nucleotides 2–8 of the reference miRNA). In addition the 5′- and 3′-trimming size as well as the number of added nucleotides in the 3′-addition variants can be chosen. Through the selective specification of the kind of isomiRs, SeqBuster permits the dissection of a complex miRNA expression landscape.

To perform comparative expression analysis, SeqBuster provides several options to express and normalize the frequency of the sequences (see 'Sequencing capacity' in the 'Basic analysis' option of the SeqBuster web-server). As in the isomiR analysis module, the contribution cutoff parameter and the *Z*-score to discard sequencing errors as the possible cause of the nucleotide changes may be applied.

In the hierarchical sample, clustering analysis SeqBuster offers two algorithms based on the overall expression profile of the small-RNAs/miRNAs. One of the algorithms is a standard cluster analysis that uses two basic R functions: (i) *cor* is used to calculate the correlation coefficients used to make the distance matrix and (ii) *hclust* is used to group the samples with the *complete* method (see R tutorials for detailed description). The alternative algorithm is based on the Dirichlet probabilistic model to calculate the overall similarity between the miRNA/small-RNA expression profiles of two samples as described in (21). This analysis uses a Bayesian probabilistic framework that considers frequency distributions in which most small-RNAs occur at low frequency and only a few display high counts.

In the small-RNA differential expression module, different tests can be applied to show statistical significance in the differential expression: the *Z*-test (26), binomial model (31), Bayesian model (32) and Fisher-test (33). Furthermore, SeqBuster offers the possibility to control the false discovery rate by applying the Hochberg and Benjamini method (27) that corrects the *P*-value assigned by the statistical test.

Target prediction

Several programs have been developed to predict miRNA targets. SeqBuster contains an analysis module to explore the miRNA targets using the miRBase (34), PicTar (35) and/or TargetScan (28) algorithms. This analysis requires a file with a list of the miRNAs for which the targets are going to be identified. The user can select the targets predicted by one of the algorithms or only those commonly predicted by two or three algorithms. In addition, genes being targeted by a minimum number of miRNAs can be selected.

Other packages

Additional packages are available in SeqBuster to be run locally in the user machine since the proposed analyses are high time and memory consuming to be processed on the web-server ('SeqBusterPackage' in the 'Download' section in the web server). Furthermore, any user can upload a custom package on this section to share with the community allowing a continue evolving platform dedicated to the small-RNA analysis. At present, the website provides R and Perl-based programs that implement the algorithms for human miRNA prediction, single nucleotide polymorphism (SNP) analysis and transcription factor enrichment analysis (see the corresponding tutorials for a detailed explanation). The 'human

miRNA prediction' package uses the algorithm described in the microPred pipeline (36) that is based on a non-comparative computational method for the effective identification of pre-miRNAs among the hairpin secondary structures predicted from the human genome. A test input file is requested containing all sequences for which the analysis is going to be performed. The package provides an output list of the candidate miRNAs. The 'SNP analysis' package contains genomic coordinates for available SNPs (http://hgdownload.cse.ucsc.edu/downloads.html). The algorithm produces a table showing all the sequences that have a nucleotide substitution event at the same position where a SNP has been detected. Furthermore, allele SNP description can be added to the output table. The 'Transcription factor enrichment' package addresses the problem of comparing and characterizing the promoter regions of miRNAs with similar expression patterns, using the algorithm described by Blanco *et al.* (37). The package requires a test file containing a list of the co-regulated miRNAs and a list with all the expressed miRNAs. To identify a possible significant enrichment of cis-regulatory elements in co-regulated miRNAs, the algorithm assigns a *P*-value using the permutation-based simulation.

SeqBuster test case: human stem cells sequencing data

Using SeqBuster we have computationally analyzed the data of Illumina deep sequencing of short RNAs in undifferentiated (hESC) and differentiated human embryonic (EB) stem cells (12)]. First, we recognized and removed adapters from the reads, and second, we annotated against precursor miRNA and mature miRNA databases downloaded from the miRbase repository. Of the total number of reads, 55% was classified as miRNA and 5% as pre-miRNA in both libraries, improving by 10% the previous annotation procedure (12). Our analysis detected 442 different miRNAs genes, 109 more than the previous analysis (Supplementary Table S1), indicating that our pre-analysis strategy was more efficient for the detection of miRNAs. The counts for the miRNAs detected by SeqBuster that included the reference miRNAs and the corresponding isomiRs correlated in 99% with the counts found in the original analysis. However, our strategy resulted in the recognition of an increased number of different sequences (2-fold) being annotated as miRNAs (most of them were identified as 3′-trimming isomiRs).

In the hSCE and EB libraries the majority of the miRNAs displayed isomiRs and only 17% remained invariable, 19 miRNAs being common between the two libraries (Supplementary Table S2). A total of 3566 different isomiRs were detected, most of them being the result of variations in the cleavage position at the 3′-end of the pri-miRNA (3′-trimming) (Figure S1). In this case, a considerable proportion of miRNAs showed two or more 3′-trimming variants. However, the majority of the 5′-trimming variants (5′-trimming), presented isomiRs with only one change. The majority of the variants resulting from single nucleotide substitutions with respect to the reference sequence (nt-substitution) or 5′-trimming variants showed a low frequency compared to that of the reference miRNAs. However, variants affecting the 3′-terminus of miRNAs, especially the 3′-trimming variants, showed a variable proportion with respect to reference miRNAs.

We analyzed the positions and nucleotides involved in the different types of variants (Supplementary Figure S2). Most 3′-trimming variants involved positions −1 (one nucleotide upstream) and +1 (one nucleotide downstream) of the 3′-end of the reference miRNA, while the majority of the 5′-trimming variants involved the position −1 of the 5′-terminus of the reference miRNA, matching the miRNA precursor. Furthermore, the nucleotide preferentially involved in the most abundant 3′- and 5′-trimming variants was a U, which suggests a preference of the dicing machinery for this nucleotide. In analyzing the 3′-addition variants (data not shown), the vast majority consisted in a single A or U addition in the 3′-end of the mature miRNA, in accordance with previous reports (12,13).

SeqBuster revealed that 50% of the miRNAs detected in each library presented nt-substitution variants (Supplementary Figure S1). Given the putative relevance of these sequences, we analyzed the nt-substitutions at different positions of the mature miRNA (Supplementary Figure S3). In the SeqBuster analysis, we only considered nt-substitutions significantly different from these error rates applying the *Z*-score probabilistic model. In line with previous observations (13), variability of miRNAs was low in positions 1–8 containing the seed region (Supplementary Figure S3A), which agrees with the importance of these sites in selective gene expression regulation. Several nt-substitution variants were identified in both libraries, involving the same type of nucleotide change (Supplementary Table S3). For instance, miR-30a and miR-30d presented nt-substitutions at position 3 of the miRNA involving in all cases a U to G modification. Since this substitution affects the seed region, new targets may be recognized by these isomiRs.

To contrast small RNA expression between biologically comparable samples, similar sequencing efficiencies should be considered. The sample processing for Illumina deep sequencing involves different steps that may influence the sequencing output, including a PCR-amplification of the cDNA obtained by retrotranscription of the small RNAs ligated to specific adaptors. To evaluate the sequencing efficiencies in the samples to be compared we used the 'Sequencing capacity' package of SeqBuster. The analysis revealed an identical frequency distribution in both samples (Supplementary Figure S4), making them comparable for differential expression profiles.

We performed two types of comparative expression analyses using the 'Differential expression analysis' module. In a first approach, we used only sequences annotated as the reference miRNAs. We chose the *Z*-test to show significance in differential expression, since it is the most restrictive test used for evaluation of differently expressed sequences in high-throughput sequencing methodologies (26). A total of 61 upregulated miRNAs (ratio > 1.5; $P < 0.05$; frequency (hESC + EB) > 50) and 39 downregulated miRNAs (ratio < 0.5 and $P < 0.05$;

frequency (hESC + EB) > 50) were detected in the hESC library (Supplementary Table S4). A similar number of deregulated miRNAs was found in the original analysis. Furthermore, 95% of miRNAs showed the same pattern of deregulation (12). When studying the correlation between the expression pattern of all reference miRNAs and the corresponding isomiRs [frequency (hESC + EB) > 2], our analysis showed that the majority (>80%) of up-regulated miRNAs displayed an up-regulation of the corresponding 3′- or 5′-isomiRs. A similar result was obtained when considering the downregulated reference miRNAs (Supplementary Table S5).

In the second approach we considered isomiRs affecting the seed region of the reference miRNA, as the seed region is essential in recognition and expression modulation of target mRNAs. These included 5′-trimming variants that involved dicing at positions −2 to +2 with respect to the mature reference miRNA, and nt-substitution variants involving any nucleotide change in positions 2–8 of the reference miRNA. A total of 21 upregulated isomiRs (ratio ≥ 1.5; $P < 0.05$; frequency (hESC + EB) >50) and 12 downregulated isomiRs [ratio ≤ 0.5; $P < 0.05$; frequency (hESC + EB) >50] were found in the non-differentiated stem cell library (Supplementary Table S6). To gain insights into the biological pathways affected by the isomiRs targets, we compared the biological functions most likely affected by the genes exclusively targeted by the isomiRs with those affected by genes targeted by the corresponding reference-miRNAs. The results highlighted that some of the top significantly enriched biological functions were specific of the targets of hESC- or EB-enriched isomiRs. These results underline the potential contribution of isomiRs in the modulation of new biological pathways.

DISCUSSION

In this work, we present SeqBuster, a web-based bioinformatic tool offering a custom analysis of deep sequencing data at different levels, with special emphasis on the analysis of miRNA variants or isomiRs. While the currently existing pipelines address specific questions such as differential expression or miRNA prediction (21–24), SeqBuster integrates these and other types of analyses in a single user-friendly platform. Furthermore, SeqBuster is the first tool providing an automatized pre-analysis for sequence annotation. Several features point out SeqBuster as a unique tool for the characterization of large-scale sequencing data of small RNAs. First, SeqBuster includes a stand-alone version that permits the annotation against any custom database installed in the local machine independently of the web server. This offers a pre-analysis that is not restricted to the databases stored in the web server, overcoming the limitations in the storage capacities detected in other web-based bioinformatic tools. Second, the R environment, in which the different analysis packages have been developed, permits the incorporation and/or modification of different types of analysis, which may be focused not only on small non-coding RNAs but also on any type of sequence generated in large-scale sequencing strategies. This provides a continuous evolving platform, where future analysis packages may be easily added to the repository. Third, SeqBuster is highly versatile offering a wide range of options both in the pre-analysis for annotation purposes and in the different module analysis for data manipulation. An example of this flexibility is illustrated by the pre-analysis strategy applied to the hSCE and EB raw data. The high number of recognized miRNAs, 109 more than in the previous analysis, corresponds mostly to newly discovered miRNAs. However, 15% of these miRNAs were already known at the time of the original analysis (Supplementary Figure S2B) and were only detected in our study as a consequence of the algorithm used for the adapter recognition/removal and the alignment parameters. For the adapter recognition, we integrated a modified version of the Needleman–Wunsch algorithm that resulted in the adapter detection in a greater number of sequences compared with the original analysis. For the annotation step, we implemented a novel strategy using Mega BLAST instead of BLASTTN to achieve an increased number of annotated sequences. Although the time of the process is shorter using newer algorithms like SOAP (38), the number of sequences annotated is significantly lower compared to more traditional algorithms like BLAST or Mega BLAST (http://www.ncbi.nlm.nih.gov/blast). SeqBuster flexibility is extensive to the analysis modules, where the user can choose different statistical approaches, normalization strategies and the type of visualization of the miRNA variability, therefore providing a deep control of the analysis process. These features allow a wide plasticity that is essential in the highly evolving field of high-throughput sequencing data analysis and the non-coding RNA field.

One of the distinctive packages offered by SeqBuster is the analysis of the variability with respect to the reference mature miRNA that has been recently highlighted in a number of studies (11,12,18,39). The analysis of the small RNAs in hESC and EB revealed different types of variability for the major part of miRNAs, confirming that the miRNA transcriptome is more complex than previously suggested. SeqBuster revealed that most miRNAs displayed 3′ trimming and 3′-addition events that showed a variable proportion with respect to the reference miRNA. This agrees with the concept that 3′-variants are more permissive, having moderate consequences in gene expression regulation, as previously suggested in animal and plants miRNAs (40–42).

The 'IsomiR analysis' packages highlighted significant nucleotide modifications along the mature miRNA. Several lines of evidence argue against RT–PCR and sequencing errors as contributors to sequence discrepancies with respect to the reference miRNA. First, the frequencies of the nucleotide modifications were remarkably higher compared to the estimates attributable to Illumina sequencing errors (25). Second, the positional non-randomness of nucleotide changes along the length of the miRNA seen in both libraries. Third, analogous nucleotide modifications were found in two independent libraries. Finally, nucleotide changes, insertions and

deletions have been reported in previous studies using different sequencing strategies. These nucleotide substitutions have been described in few miRNAs as pri-miR precursor editing changes from A to G attributed in part to A to I deaminations, which lead to a repression in the maturation of the miRNA (11,16–20,43). Other types of nucleotide substitutions have been reported in a meta-analysis of small-sequencing data in plants (10,11) and in the *let* family of miRNAs in different mouse cells lines (13). SeqBuster revealed that the nucleotide substitutions for the majority of the miRNAs were distinct from the classical A to I editing, suggesting that alternative changes are not limited to the *let*-7 family of miRNAs in animals. These modifications may result in alternative base pairing between the variant and the target mRNA, possibly affecting the efficiency of gene regulation, as previously described (13,40).

The 'IsomiR analysis' packages showed a decreased proportion of miRNAs presenting length and sequence heterogeneity at the 5′-end of the miRNA, with the abundance of most of these variants negligible, compared with that of the reference sequence. This suggests that the 5′-terminus of the miRNAs is specially protected from variations, which agrees with the crucial role of Watson–Crick base pairing of the 5′-seed region of the miRNA with the 3′UTR of the mRNA, for gene targeting. In line with this, scarce miRNAs presented nucleotide substitutions at positions 1-11 of the miRNA, containing the 5′-seed (nt 2–8) and the cleavage (nt 10–12) sites (44) that are typically base paired in the miR:mRNA duplex.

The 'miRNA differential expression' package revealed miRNA and isomiR expression modiffications linked to stem cell differentiation processess. Our analysis showed that the expression pattern of the isomiRs correlated with that of the corresponding miRNAs, suggesting that the mechanisms modulating the degree of expression of the isomiRs and the corresponding miRNAs are parallel in most cases. Given the essential role of the seed region for mRNA target recognition and silencing (45), we applied SeqBuster to analyze the differential expression of isomiRs affecting nucelotides 2–8 in hESC and EB. SeqBuster identified 5′ trimming and nt-substitution seed region variants that were differently expressed between hESC and EB. The new putative targets identified by these differently expressed isomiRs highlighted novel enriched biological functions.

Overall, the present analysis strongly suggests a biological function for this sequence plasticity in miRNAs, which may have broad implications in mRNA targeting, stability and/or gene expression regulation mechanism. The exhaustive description of the different types of miRNA variability provided by SeqBuster is extremely useful to uncover tissue-specific isomiR distributions relevant in development, physiology and disease conditions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jerome McDonald for linguistic help, Núria López-Bigas for critical reading of the manuscript, and Oscar González and Judith Flo for informatics support to install the web server for SeqBuster.

FUNDING

Spanish Ministry of Health 'Fondo de Investigaciones Sanitarias (PI081367) and 'Instituto de Salud Carlos III' (CIBERESP); the Sixth Framework Programme of the European Commission through the SIROCCO integrated project LSHG-CT-2006-037900 and the Spanish Ministry of Science and Innovation (SAF2008-00357). E.M. is partially supported by the Spanish Ministry of Health; L.P. is recipient of a fellowship from the Spanish Ministry of Science and Innovation.

Conflict of interest statement. None declared.

REFERENCES

1. Ghildiyal,M. and Zamore,P.D. (2009) Small silencing RNAs: an expanding universe. *Nat. Rev. Genet*., **10**, 94–108.
2. Carthew,R.W. and Sontheimer,E.J. (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.
3. Tam,O.H., Aravin,A.A., Stein,P., Girard,A., Murchison,E.P., Cheloufi,S., Hodges,E., Anger,M., Sachidanandam,R., Schultz,R.M. *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.
4. Watanabe,T., Takeda,A., Tsukiyama,T., Mise,K., Okuno,T., Sasaki,H., Minami,N. and Imai,H. (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev*., **20**, 1732–1743.
5. Aravin,A.A., Sachidanandam,R., Bourc'his,D., Schaefer,C., Pezic,D., Toth,K.F., Bestor,T. and Hannon,G.J. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol. Cell*., **31**, 785–799.
6. Hébert,S.S. and De Strooper,B. (2009) Alterations of the microRNA network cause neurodegenerative disease. *Trends Neurosci*., **32**, 199–206.
7. Spizzo,R., Nicoloso,M.S., Croce,C.M. and Calin,G.A. (2009) SnapShot: MicroRNAs in Cancer. *Cell*, **137**, 586–586.e1.
8. Visone,R. and Croce,C.M. (2009) MiRNAs and cancer. *Am. J. Pathol*., **174**, 1131–1138.
9. Berezikov,E., Cuppen,E. and Plasterk,R.H. (2006) Approaches to microRNA discovery. *Nat. Genet*, **38(Suppl.)**, S2–S7.
10. Lida,K., Jin,H. and Zhu,J.-K. (2009) Bioinformatics analysis suggests base modification of tRNA and miRNA in arabidopsis thaliana. *BMC Genomics*, **10**, 155.
11. Ebhardt,H.A., Tsang,H.H., Dai,D.C., Liu,Y., Bostan,B. and Fahlman,R.P. (2009) Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res*., **37**, 2461–2470.
12. Morin,R.D., O'Connor,M.D., Griffith,M., Kuchenbauer,F., Delaney,A., Prabhu,A.L., Zhao,Y., McDonald,H., Zeng,T., Hirst,M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*., **18**, 610–621.
13. Reid,J.G., Nagaraja,A.K., Lynn,F.C., Drabek,R.B., Muzny,D.M., Shaw,C.A., Weiss,M.K., Naghavi,A.O., Khan,M., Zhu,H. *et al.* (2008) Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5′-seed/cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes. *Genome Res*., **18**, 1571–1581.

14. Kawahara,Y., Megraw,M., Kreider,E., Iizasa,H., Valente,L., Hatzigeorgiou,A.G. and Nishikura,K. (2008) Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res.*, **36**, 5270–5280.
15. Aravin,A. and Tuschl,T. (2005) Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.*, **579**, 5830–5840.
16. Kawahara,Y., Zinshteyn,B., Sethupathy,P., Iizasa,H., Hatzigeorgiou,A.G. and Nishikura,K. (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science*, **315**, 1137–1140.
17. Kawahara,Y., Zinshteyn,B., Chendrimada,T.P., Shiekhattar,R. and Nishikura,K. (2007) RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. *EMBO Rep.*, **8**, 763–769.
18. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell.*, **129**, 1401–1414.
19. Luciano,D.J., Mirsky,H., Vendetti,N.J. and Maas,S. (2004) RNA editing of a miRNA precursor. *RNA*, **10**, 1174–1177.
20. Blow,M.J., Grocock,R.J., van Dongen,S., Enright,A.J., Dicks,E., Futreal,P.A., Wooster,R. and Stratton,M.R. (2006) RNA editing of human microRNAs. *Genome Biol.*, **7**, R27.
21. Berninger,P., Gaidatzis,D., van Nimwegen,E. and Zavolan,M. (2008) Computational analysis of small RNA cloning data. *Methods*, **44**, 13–21.
22. Fahlgren,N., Sullivan,C.M., Kasschau,K.D., Chapman,E.J., Cumbie,J.S., Montgomery,T.A., Gilbert,S.D., Dasenko,M., Backman,T.W., Givan,S.A. *et al.* (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA*, **15**, 992–1002.
23. Hackenberg,M., Sturm,M., Langenberger,D., Falcón-Pérez,J.M. and Aransay,A.M. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, W68–W76.
24. Wang,W.C., Lin,F.M., Chang,W.C., Lin,K.Y., Huang,H.D. and Lin,N.S. (2009) miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinform.*, **10**, 328.
25. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
26. Reinartz,J., Bruyns,E., Lin,J.Z., Burcham,T., Brenner,S., Bowen,B., Kramer,M. and Woychik,R. (2002) Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief. Funct. Genomic Proteomic*, **1**, 95–104.
27. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.
28. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell*, **120**, 15–20.
29. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell.*, **27**, 91–105.
30. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*, **48**, 443–453.
31. Rozowsky,J. (2009) PeakSeq enables systematic scoring of Chip-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
32. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
33. Fisher,R.A. (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc*, **85**, 87–94.
34. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2005) miRanda algorithm: human MicroRNA targets. *PLoS Biol.*, **3**, e264.
35. Krek,A., Grün,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
36. Batuwita,R. and Palade,V. (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995.
37. Blanco,E., Messeguer,X., Smith,T.F. and Guigo,R. (2006) Transcription factor map alignment of promoter regions. *PLoS Comput. Biol.*, **2**, e49.
38. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP:short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
39. Kuchenbauer,F., Morin,R.D., Argiropoulos,B., Petriv,O.I., Griffith,M., Heuser,M., Yung,E., Piper,J., Delaney,A., Prabhu,A.L. *et al.* (2008) In-depth characterization of the microRNA transcriptome in a leukemia progression model. *Genome Res.*, **18**, 1787–1797.
40. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
41. Brodersen,P. and Voinnet,O. (2009) Revisiting the principles of microRNA target recognition and mode of action. *Nat. Rev. Mol. Cell. Biol.*, **10**, 141–148.
42. Lu,L. and Li,J. (2009) A combinatorial approach to determine the context-dependent role in transcriptional and posttranscriptional regulation in Arabidopsis thaliana. *BMC Syst. Biol.*, **3**, 43.
43. Yang,Y., Lv,J., Gui,B., Yin,H., Wu,X., Zhang,Y. and Jin,Y.. (2008) A-to-I RNA editing alters less-conserved residues of highly conserved coding regions: implications for dual functions in evolution. *RNA*, 1516–1525..
44. Ebert,M.S., Neilson,J.R. and Sharp,P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat. Methods*, **4**, 721–726.
45. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.