Research Article Open Access

Guillermo Barturen, Antonio Rueda, Maarten Hamberg, Angel Alganza, Ricardo Lebron, Michalis Kotsyfakis, Bu-Jun Shi, Danijela Koppers-Lalic*, Michael Hackenberg*

sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments

Abstract: MicroRNAs and other small RNAs are known to play important functions in gene regulation. Over the last years, it became also apparent that many virus genomes encode microRNAs and that those strongly interact with the host transcriptome. Important functions include the evasion of the immune response and the regulation of the switch to lytic infection. Since the advent of deep sequencing protocols for small RNAs, expression profiles can be routinely determined. However, currently the tools developed for the data analysis of small RNA deep sequencing experiments are limited to the analysis of only one species at a time. In order to facilitate the analysis of experimental setups that include genetic material from several species, we developed sRNAbench. It maintains the main features implemented in its predecessor program, miRanalyzer, and includes new functionalities such as full isomiR support including statistical test on differential frequency, improved prediction of novel microRNAs, extended summary files and data visualization support. Both a standalone program and a webserver are available at: http://bioinfo5.ugr.es/sRNAbench/.

***Corresponding author: Danijela Koppers-Lalic and Michael Hackenberg**: E-mail: d.koppers@vumc.nl, mlhack@gmail.com
Guillermo Barturen, Maarten Hamberg, Angel Alganza, Ricardo Lebron, Michael Hackenberg: Department of Genetics, University of Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain
Guillermo Barturen, Maarten Hamberg, Angel Alganza, Ricardo Lebron, Michael Hackenberg: Bioinformatics Group, Biomedical Research Center (CIBM), PTS, Avda. del Conocimiento s/n, 18100-Granada, Spain
Antonio Rueda: Genomics and Bioinformatics Platform of Andalusia (GBPA), Edificio INSUR, Calle Albert Einstein, 41092-Sevilla, Spain.
Michalis Kotsyfakis: Biology Centre, Academy of Sciences of Czech Republic, Branisovska 31, 37005 Budweis, Czech Republic
Bu-Jun Shi: Australian Centre for Plant Functional Genomics, the University of Adelaide, South, Australia 5064, Australia
Danijela Koppers-Lalic: Department of Pathology, Cancer Center Amsterdam, VU University Medical Center, 1007MB Amsterdam, the Netherlands

Keywords: microRNA, small RNA, isomiRs, expression profiling, multi-species experiment, webserver

DOI 10.2478/mngs-2014-0001
Received May 26, 2014 accepted August 11, 2014

1 Introduction

The availability of high-throughput sequencing (HTS) technologies plays now a pivotal role in the expression profiling of known small RNAs and the discovery of novel classes of non-coding RNA [1]. Over the last decade, in both plants and animals a notable number of novel small RNA classes have been described [2, 3]. It became apparent that the biogenesis of microRNAs and other small RNA classes, their contribution to gene regulation mechanisms and the cellular control of their expression are more complex than previously envisioned. Therefore, apart from microRNA profiling and detection, the analysis of other small RNAs became an important task. In plants, small interfering RNAs (siRNA), trans-acting RNAs (ta-RNA) and heterochromatic siRNAs have been shown to regulate gene expression and being responsible for the deposition of repressive chromatin marks (DNA methylation and histone marks) [4]. In animals, apart from microRNAs and piRNAs, other putatively important processed small RNA fragments do exist, including yRNA, tRNA and snoRNA fragments [5, 6]. Furthermore, microRNAs are reproducibly diversified at the sequence level in both animals and plants, including post-transcriptional modifications such as adenylation [7–9].

Many tools have been developed to analyze HTS small RNA data covering many crucial aspects, such as quality control, expression profiling, prediction of novel microRNAs, snoRNAs, tRNAs, piRNAs or ta-siRNAs, identification of siRNAs or piRNAs clusters and isomiR quantification. In order of appearance: small RNA toolkit

[10], miRDeep [11], miRanalyzer [12], SeqBuster [13], NiBLS [14], miRTRAP [15], MIReNA [16], miRanalyzer [17], DARIO [18], UEA sRNA workbench [19], segmentSeq [20], CPSS [21], ncPRO-seq [22], miRDeep2 [23], ShortStack [24], isomiRex [25], mirDeep* [26], mirTools 2.0 [27] among others. These tools detect and profile small RNAs either by mapping the sequencing reads first to the genome (using a genome annotation on a second step) or by direct mapping to reference sequence libraries in FASTA format.

Apart from expression profiling and the detection of differential expression between two experimental groups, many other important questions arose over the last years in small RNA research. For example, Host-Microbe interactions have been studied at a small RNA level [28] and it has been shown that viral microRNAs play important roles in the infection process and the maintenance of a latent infection [29, 30]. Furthermore, small RNAs have been detected in most studied bodily fluids opening the door for biomarker applications like the detection of parasite presence in released microvesicles called exosomes [31, 32]. In order to analyze this kind of experimental assays, tools are needed that can treat with datasets obtained from more than one species. However, none of the existing tools have been conceived for multi-species experiments. As a consequence, workarounds need to be performed, such as pooling all species into a single index, which is virtually impossible for users with less bioinformatics background.

We generated sRNAbench, a user friendly tool for the analysis of small RNA high-throughput sequencing data, especially suited to detect and profile small RNAs in multi-species assays.

2 Methods

sRNAbench is based on the well-established miRanalyzer tool [12], extending its scope incorporating novel features like multi-species support, genome and library mapping approaches, improved prediction of novel microRNAs and full isomiR support, which also includes the detection of enzymatically added nucleotides to other small RNA species. Figure 1 sketches the workflow of sRNAbench that can be roughly divided into pre-processing, mapping, profiling/detection and several output options.

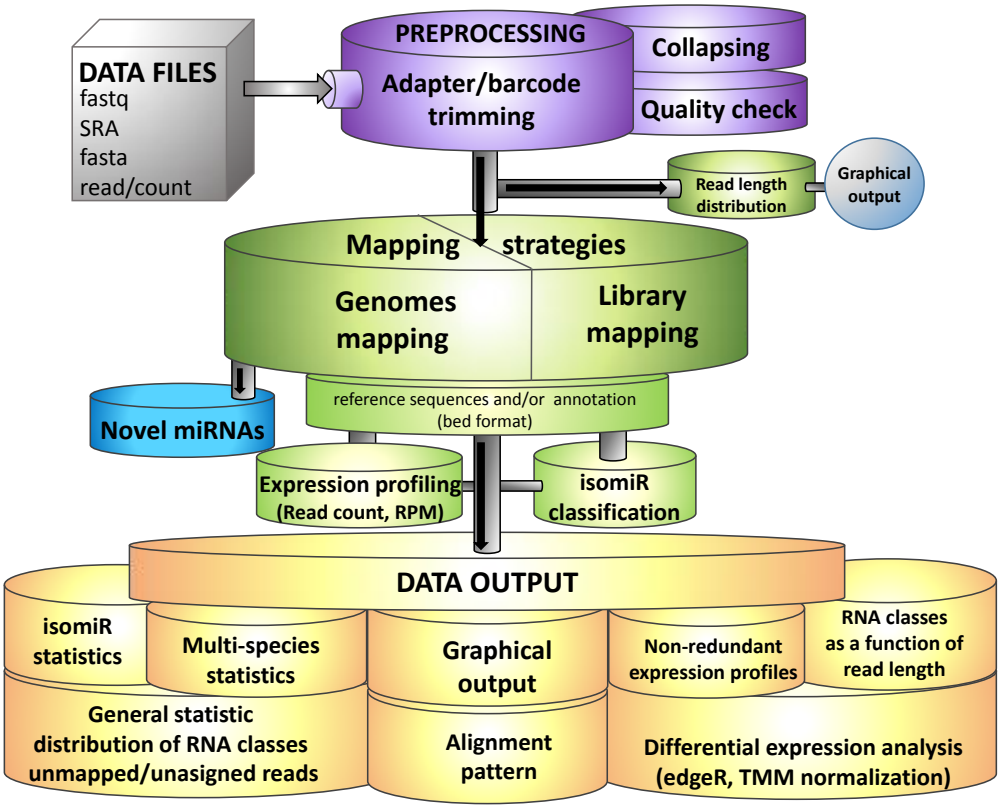

Figure 1: Schematic overview on sRNAbench features.

2.1 Input and pre-processing

sRNAbench accepts several input formats including fastq, sra (needs SRA toolkit), read/count and FASTA. All formats can be GNU zip compressed and both nucleotide and color space input is accepted. Furthermore, barcode and adapter sequences can be detected and trimmed. As a novelty, sRNAbench can automatically detect the adapter sequence in the following way: i) a subset of reads (by default 100,000) is aligned with Bowtie seed option [33] to the genome, ii) the putative adapter sequence is defined as the sequence that starts at the first mismatch, iii) all putative adapter sequences are grouped and the most frequent sequence is determined as the adapter sequence. Finally, sometimes certain classes of small RNAs need to be filtered out prior to the expression profiling steps (for example rRNA fragments). sRNAbench first maps all reads to the user defined 'filter libraries', removing all mapped reads from the analysis. These libraries should be given in FASTA format or as bowtie indexes.

2.2 General mapping strategies and expression profiling

The microRNA sequences from miRBase [34] are generally used to profile the expression of known microRNAs. However, microRNAs represent less than 0.01% of a typical mammal genome. Therefore, especially if mismatches are allowed, a non-microRNA derived read can give a best hit to a microRNA reference sequence, although it might map even better to another non-microRNA position in the genome. For example, we observed that reads from HY3 scRNA pseudogenes (545 copies in the hg19 assembly of the human genome) can be mapped under lax settings to miR-4301 (see Figure 2 for an illustration). Therefore, we implemented both a genome mapping and a library mapping mode (like performed by miRanalyzer). Briefly, the genome mapping mode first aligns all reads to the reference genome, extracting the coordinates in internal BED format. The coordinates are then compared to a user-provided annotation. In general, a read is assigned to the reference RNA if its coordinates lie completely within the chromosome coordinates of the reference RNA. If a good genome assembly is available, the genome method is recommended.

Currently, the users can provide the reference annotations in BED, GTF/GFF, FASTA format and as bowtie indexes. A reference library given in FASTA format will be mapped first to the genome in order to obtain the corresponding chromosome coordinates. Therefore, only un-spliced reference sequences can be provided in FASTA format, while spliced protein coding genes should be given either in BED format or as bowtie indexes. Both mapping strategies contain a certain number of parameters like the maximum number of mismatches or the seed length that can be modified by the user (please see the sRNAbench manual for more details). Furthermore, different measures of the expression level are generated like the read count (total number of reads assigned to the reference RNA), adjusted read count (read count normalized by the number of times that the read maps to the library or the genome) and normalized RPM (reads per million). The output files are explained in detail within the manual.

2.3 microRNA expression profiling

MicroRNA expression profiling includes some particularities when compared to the general expression profiling that was explained in the last section. First, for known microRNAs, both the pre-microRNA and

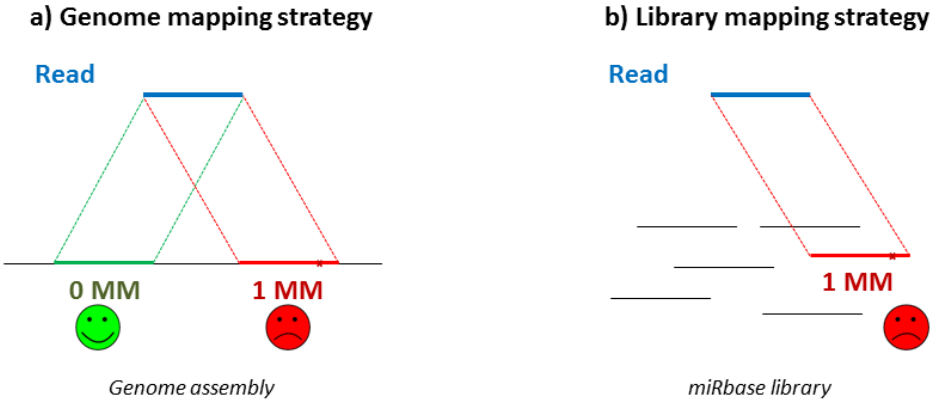

Figure 2: Schematic comparison between genome mapping and library mapping. The left part shows that the read is correctly assigned (the mapping with fewer mismatches) in the genome mapping approach. However, if the green genome region is not represented in the library (right part), then the read would be incorrectly assigned to another reference sequence (red region).

mature sequence can be profiled. sRNAbench labels the corresponding expression profiling files with 'hairpin' and 'mature' respectively. In the genome mapping mode, sRNAbench quantifies additionally the mappings to the antisense strand while in library mapping mode only the forward strand is considered. This leads to a maximum of 4 output files: 'mature_sense.grouped', 'mature_antisense.grouped', 'hairpin_sense.grouped' and 'hairpin_antisense.grouped' (please see the manual for a description of the file formats). These files accept the reference sequences (by default miRBase) "as they are", performing the expression profiling by direct mapping to the sequences or via the corresponding genome annotations. However, the used microRNA database might contain false positives and therefore after the profiling step, sRNAbench carries out a consistency check of the provided microRNA sequences. First, the secondary structure of the pre-microRNA sequences are determined using RNAfold [35], and second the corresponding mature sequences are localized in the secondary structure, assigning as well the corresponding reads. Pre-microRNA reference sequences are reported in the 'hairpin_suspicious.txt' file if: i) no stem-loop structure was detected (with the default RNAfold parameters), ii) one of the detected mature microRNAs folds back onto itself in the secondary structure (for example it overlaps the loop) or iii) the corresponding passenger strand of the annotated guide mature microRNA is not fully contained within the pre-microRNA sequence (this can happen if only one arm is annotated in the database).

Furthermore, if only one arm is annotated in the used database and the corresponding opposite arm is represented by a read in the sample, this previously unannotated mature microRNA will be reported in the output ('hairpin_NovelStar.txt').

Finally, it is known that microRNAs suffer diversification at a sequence level and therefore not only the canonical sequence can be found in the sample (see also 'isomiR/isoRNA detection and classification'). This raises the question how the 'expression level' of a given mature microRNA should be assessed: only by its canonical form or summing the canonical form and all its length and sequence variants. By default, sRNAbench assign all reads to a given mature microRNA that map within a window around the canonical form. By default, this window is defined as: [Canonical $miRNA_{start}$ – 3nt ; Canonical $miRNA_{end}$ + 5 nt]. All previously mentioned files contain expression value estimates (read count and normalized RPM – Reads Per Million) obtained in this way. However, sRNAbench writes out another file including only those microRNAs that passed the consistency check mentioned before ('miRBase_main.txt'). This file contains additionally the read count of the canonical sequence, in case the user wants to perform the expression profiling based on the canonical sequence only.

2.4 Hierarchical mapping

In both the genome and library mapping mode a hierarchy can be established. By default, first known microRNAs are profiled and the mapped reads are removed. After this step, the remaining reads are mapped successively to all other reference libraries provided by the user. By default, the mapped reads are removed after each analyzed library. However, the user can also choose to not remove those reads, allowing the assignment of the reads to different libraries. For examples, this option might be interesting as a way to detect the origin of putative siRNAs. Reads can be mapped first to the antisense strand of protein coding genes and afterwards to other libraries like tRNA, repetitive DNA, snRNA, snoRNA etc... For those reads that map to both the antisense strand of protein coding genes and another RNA class, the origin of these putative siRNAs can be established.

2.5 Ambiguous mapping treatment

A general problem in all high-throughput sequencing experiments is how to deal with ambiguously mapping reads. Ambiguous mapping can arise if i) several alternative transcripts of one gene are given in the reference library or ii) the read maps with the same quality to different genome loci. The first point is of particular importance in mRNA-seq experiments in order to infer the correct expression values of the different isoforms of a gene [36, 37]. For small RNA sequencing, the problem of different transcripts from the same locus virtually does not exist; however, several microRNAs have more than one gene in the genome. In order to address this problem, sRNAbench generates two different output files: i) a multiple assignment file – each read counts for all loci to which it maps and ii) a single assignment file – each read is assigned only to one locus. The single assignment expression files are generated starting from the multiple assignments (for example mature_sense.grouped):

- The multiple assignment file is ordered by the read count in a descending way.
- The most frequent RNA maintains its expression value and all reads that map to it are removed.
- From the second most frequent RNA to the last, in each step the remaining reads that have been assigned previously to this RNA are summed and removed

afterwards. In this way, each read is assigned only once – out of the RNAs to which it maps with the same quality, it is assigned to the most frequent one (based on the read count).

Figure 3 illustrates the result of this procedure. The mature microRNA hsa-miR-92a-3p can be obtained from two genes, located on chromosomes 13 and X. In this example, the read counts are 19620 and 19248 respectively, while the corresponding single assignment read counts are 19620 and 35. This means that all ambiguously mapping reads are assigned to the locus on chromosome 13 leaving a single assignment read count of 35 for the mature microRNA located on the X chromosome. Note that these 35 reads do map to this locus in an unambiguous way, i.e. they map with higher quality to the X chromosome than to chromosome 13. In this way, the number of uniquely mapped reads can be calculated. The locus on the X chromosome had a total read count of 19248, out of which 35 mapped exclusively to this locus. Therefore we can infer that 19248-35=19213 reads mapped to both loci. This allows us finally to conclude that 19620-19213 = 407 reads map exclusively to the locus on chromosome 13. Note that these numbers could mean different things: i) the microRNA is transcribed from chromosome 13 and the 35 exclusively mapping reads from the X chromosome are obtained due to sequencing errors or ii) the microRNA is transcribed from both loci, but at much higher levels from chromosome 13. In either case, the single assignment file might serve in some cases as starting point for further investigation. In this concrete example it shows that the locus on chromosome 13 is very likely the important one.

The generation of the non-redundant files is performed in a strict order. First, the 'sense' files have preference over the 'antisense' files. This implies that a read that maps to both the sense and the antisense strand of a given library will be always assigned to the sense strand. Therefore, the single assignment antisense expression files count only 'true' antisense reads.

In the same line, the single assignment files of mature microRNAs have preference over the pre-microRNA files ('hairpin'). This implies that the single assignment pre-microRNA expression files only contain those reads that mapped to the precursor sequence but not to the corresponding mature microRNAs. Those reads can correspond to unannotated mature sequences (mostly for those microRNAs that only have one arm annotated), loop sequences, degradation products or atypical cleavage products.

The methods used to generate single assignment files depend on the order in the multiple assignment file. This order might change between replicates and conditions. Therefore those files should not been used for differential expression analysis.

2.6 isomiR/isoRNA detection and classification

Most mature microRNAs do not only exist in their canonical form in the cell, but a high number of different sequence variants, i.e. isomiRs are stably generated [9]. Sequence variants include 5' and 3' trimming and extension, non-templated additions (enzymatically addition of a nucleotide to the 3' end, i.e. adenylation, uridylation).

Non-templated additions frequently do not match the genome or the pre-microRNA sequence and would therefore cause mismatches in the alignment. In order to not discard those cases, the Bowtie seed alignment option [33] is used. This option scores only the first L nucleotides (L=20 by default) and therefore does not take into account the mismatches at the 3' end of the read caused by the post-transcriptionally added nucleotides.

Name	unique reads	read count	read count (mult. map. adj.)	RPM (lib)	RPM (total)	Chromosome string
Multiple Assignment						
hsa-miR-92a-3p	51	19620	10013.5	17408.2	15158.76	chr13:92003615-92003636 (+)
hsa-miR-92a-3p	46	19248	9641.5	17078.14	14871.35	chrX:133303574-133303595 (-)
name	unique reads	read count (SA)	read count (MA)	RPM (lib)	RPM (total)	chromosomeString
Single Assignment						
hsa-miR-92a-3p	51	19620	19620	17408.2	15158.76	chr13:92003615-92003636 (+)
hsa-miR-92a-3p	2	35	19248	31.05	27.04	chrX:133303574-133303595 (-)

Figure 3: Comparison between multiple and single read assignment.

In order to detect isomiRs, sRNAbench, i) maps the reads to the genome or pre-microRNA sequences using the Bowtie seed option [33], ii) determines the coordinates of the mature microRNAs , iii) clusters all reads that map within a window of the canonical mature microRNA sequence (see 'microRNA expression profiling'), iv) applies a hierarchical classification schema which is described below.

Frequently, a single read can have more than one modification compared to the canonical microRNA sequence. For example, it can be both 5' trimmed and adenylated. This fact opens two possibilities: 1) work with a redundant classification (allowing that a read can belong to more than one isomiR class) or 2) apply a hierarchical classification schema. sRNAbench works exclusively with a non-redundant classification schema which is summarized in Figure 4.

Briefly, all reads assigned to a given mature microRNA are checked if they belong to one of the following classes:

1. The read is identical to the canonical sequence (usually the miRBase entry).
2. The read starts and ends at the same position as the canonical sequence in the pre-microRNA, but shows sequence variation (most likely due to sequencing errors, but RNA editing events and SNPs might exist as well).
3. The read has non-templated additions (added A, T(U), C or G), i.e. nucleotides at the 3' end that do not match to the reference (template). By default, sRNAbench starts at position 18 detecting the longest run of A's, G's, C's or T's that do not match to the template.
4. The read starts or ends at the same position as the canonical version. For this case we can distinguish 4 groups:
   (a) 3' trimmed read: the read starts at the same position as the canonical sequence (same 5' end) but it is shorter than the canonical sequence.
   (b) 3' extended read: the read starts at the same position as the canonical sequence (same 5' end), but it is longer than the canonical sequence.
   (c) 5' trimmed read: the read ends at the same position as the canonical sequence (same 3' end), but it is shorter than the canonical sequence.
   (d) 5' extended read: the read ends at the same position as the canonical sequence (same 3' end), but it is longer than the canonical sequence.
5. The read coincides neither in 5' nor in 3' with the canonical sequence (multiple length variant).

Note that this classification gives preference to non-templated additions compared to other variants. This is because some NTAs might be biologically meaningful, i.e. at least some microRNAs are stabilized by monoadenylation [38]. Finally, sRNAbench does not only detect isomiRs, but non-templated additions can be detected for any small RNA species (isoRNAs).

2.7 Prediction of novel microRNAs

In comparison to miRanalyzer, we improved the prediction of novel microRNAs lowering the number of false positive predictions. The implemented method is based on both

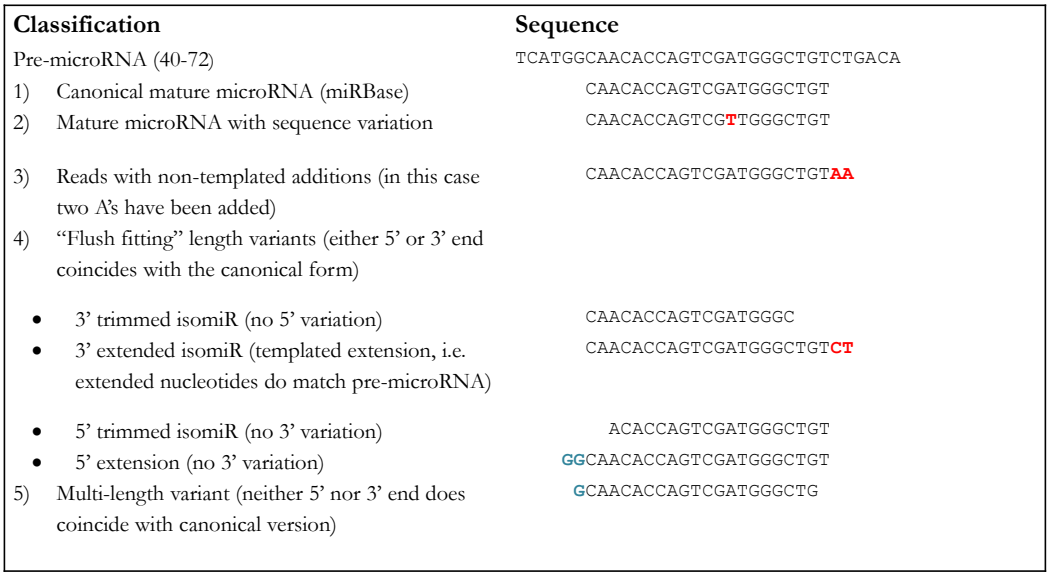

Figure 4: The hierarchical isomiR classification applied by sRNAbench.

structural and biogenesis features and has been used recently to predict novel microRNAs in plants [39]. It is conceptually based on different previous work [10, 40–42]. Briefly, the method works as follows:

1. The reads are mapped to the genome sequence.
2. Reads that map to nearly identical positions in the genome are clustered into 'read clusters' in the following way:
   - The reads are sorted by read count (read frequency).
   - The most frequent read is assigned to the first read cluster (the coordinates of the read cluster are given by the coordinates of the most frequent read).
   - For all other reads, sRNAbench checks if the read lies within a window defined by ClusterStart – 3 nt and ClusterEnd + 5 nt on the same strand (flanks are added in order to assign all isomiRs to the same read cluster just like performed in the microRNA expression profiling).
   - If the read belongs to an existing cluster, the associated read information (sequence and the read count) is added to the cluster.
   - If the read does not belong to an existing cluster, a new cluster is opened.
3. After clustering all reads, read clusters at distances of less than 180 nt for plants or less than 60 nt for animals are extracted. For most *bona fide* miRNAs there should be two read clusters corresponding to the two mature microRNAs processed from the pre-microRNA sequence.
4. Next, the genomic sequence spanned by the two read clusters is extracted (adding 5 nt to both ends) and the secondary structure and alignment pattern of the derived pre-microRNA is analyzed. Candidates are retained for which:
   - The secondary structure shows a stem-loop structure.
   - The reads map to the stem of the pre-microRNA.
   - The read clusters form a Drosha/Dicer (DCL) 2 nt overhang (allowing 1 nt error) in the secondary structure.
   - The read cluster sequence (the putative mature microRNA sequence) does not fold back onto itself (i.e. not spanning the loop region).
   - All calculated features are above the thresholds (See figure 5).

```
hsa-mir-181a-1
TGAGTTTTGAGGTTGCTTCAGTGAACATTCAACGCTGTCGGTGAGTTTGGAATTAAAATCAAAACCATCGACCGTTGATTGTACCCTATGGCTAACCATCATCTACTCCA
.((((..((((((((((..((.(.(((.(((((((..(((((((.(((((.((.......)).)))))))))))))))))).))).).))..))).)))))...)))))..
                      AACATTCAACGCTGTCGGTGAG                                                                  804
                      AACATTCAACGCTGTCGGTGAGT                                                                 797
                     GAACATTCAACGCTGTCGGTGAGTT                                                                607
                                                          ACCATCGACCGTTGATTGTACC                              574
                      AACATTCAACGCTGTCGGTGA                                                                   564
                      AACATTCAACGCTGTCGGTGAGTTT                                                               521
                      AACATTCAACGCTGTCGGTGG                                                                   488
                        CATTCAACGCTGTCGGTGAGG                                                                 315
                      AACATTCAACGCTGTCGGTGAA                                                                  306
                      AACATTCAACGCTGTCGGTGGG                                                                  284
                                                          ACCATCGACCGTTGATTGTACT                              217
                       ACATTCAACGCTGTCGGGGAG                                                                  143
                      AACATTCAACGCTGTCGGTGT                                                                   106
                     GAACATTCAACGCTGTCGGTGAG                                                                   71
                                                          ACCATCGACCGTTGATTGTA                                 55
                                    GAGTTTGGAATTAAAATCAAAAC                                                    47
```

Feature	Feature description	Example value and thresholds
Within cluster ratio	Measures the fraction of mapped (to the pre-microRNA sequence) reads that belong to the 5p and 3p strand (guide and passenger strand) – all the NON-blue reads in the example alignment above.	5852/ 5899 = **0.992** **Plant threshold: 0.521** **Animal threshold: 0.908**
5' Fluctuation	The fraction of reads that start at the same position as the canonical sequence (or most frequent read). Red reads in the example above.	**Plant threshold: 0.12** **Animal threshold: 0.17**
Most frequent to all ratio	The ratio between the read count of the most frequent read divided by the total number of reads. Read count of the green read divided by the read count sum of all reads mapping to the same region	804 / (804 + 797 + 607 + 564 + 521 + 488 + 315 + 306 + 284 + 143 + 106 + 71) = **0.161** **Plant threshold: 0.106** **Animal threshold: 0.52**
Minimum number of hairpin bindings		19 for animal and plant
Minimum number of mature bindings		17 for animal and plant
Most Frequent Read	The minimum read count of the most frequent read	10 for animal and plant
Length interval		Minimum length 19; maximum length 23
Minimum reads	The minimum number of reads in a cluster (minimum number of isomiRs)	3 for animal and plant

Figure 5: Features used in the prediction of novel microRNAs. The thresholds have been determined using the same training set as described before [17]. They represent the percentile 5 (P5) of the distributions obtained for known microRNAs.

The method described above can detect those novel microRNAs for which both arms are represented by reads in the sample. Additionally, in order to detect those novel microRNAs for which only one arm can be detected, we use the same candidate generating method as used in miRanalyzer [17]. Those candidates are then scored in the same way as described above.

2.8 Differential expression

Apart from the sRNAbench main program, a differential expression module was developed. The module generates an expression matrix and uses edgeR [43] to infer differential expression. Note that in order to use the differential expression module, all samples must be individually analyzed by the main sRNAbench module first. Additionally, statistically significant differences in the isomiR patterns are established by defining first the isomiR ratio as the number of reads that belong to a given isomiR type divided by the total number of reads mapped to a given microRNA (canonical read count plus all isomiRs). Significant differences in the isomiR ratios between two conditions are then accessed by means of a standard t-test. Note that by using edgeR, sRNAbench applies implicitly TMM normalization in the detection of differentially expressed small RNAs, which was reported to be among the most stable methods [44, 45]. The RPM values calculated by sRNAbench should not be used for differential expression analysis but 'only' as an easily interpretable measure of the expression level.

3 Results

sRNAbench is available as both standalone application and webserver. The webserver implements the most important parameters and the full miRBase database but only a limited number of genome assemblies. This means that the microRNA expression can be profiled for all species available in miRBase, but novel microRNAs can only be predicted for the species with a genome assembly. However, we offer the possibility to add genome assemblies to the webserver on demand.

In order to show the usefulness of the developed tool, we processed by means of the webserver a publically available data set of B-cell lymphoma origin (Gottwein et al., 2011). BC-1 cell lines contain two virus types, human herpesvirus type 8 (Kaposi Sarcoma herpes virus, KSHV) and human herpesvirus type 4 (Epstein-Barr virus, EBV), and is therefore ideal to show the strength of sRNAbench. Both viruses belong to a Gammaherpesvirus family and each one encodes viral microRNAs [46, 47].

The results can be permanently seen under this link:http://bioinfo5.ugr.es/sRNAbench/sRNAbench.php?launched=true&id=99670255.

Figure 6 shows some of the graphical outputs generated by the sRNAbench webserver (these graphics can be generated with the standalone version as well). One of the generated multi-species output files contains the relative amount of microRNAs derived from the different species. sRNAbench detects that 56.6%, 27.9% and 15.5% of all detected microRNAs are of human, KSHV and EBV origin, respectively. Furthermore, when analyzing the provided isomiR summary file for microRNAs with higher read counts than 100, it can be observed that virus microRNAs show on average more 3' length variants (40.5% vs. 29.9%) but less adenylation events than host microRNAs (4.2% vs. 6.8%). Such data might be important to decipher the interaction between host and virus especially regarding microRNA stability and turnover (Libri et al., 2013). Furthermore, we note that sRNAbench found a novel microRNA in the EBV genome over-lapping the position of a longer non-coding gene.

To show the prediction capacity of sRNAbench, we take advantage of the fact that the barley genome became recently available [48] predicting novel microRNAs on previously published data [49, 50]. In Golden Promise cultivar, sRNAbench predicts 40 novel microRNAs (http://bioinfo5.ugr.es/sRNAbench/sRNAbench.php?showDetails=novel&id=51906132) while in Pallas 39 novel microRNAs are predicted (http://bioinfo5.ugr.es/sRNAbench_dev/sRNAbench.php?showDetails=novel&id=39001951). Figure 7 shows a screenshot of a novel microRNA detected in the 'Golden Promise' cultivar. The microRNA name starts with 'X' indicating that this microRNA has not been annotated in any other species before. If a novel microRNA can be aligned to a known microRNA from another species, the corresponding name is assigned. Furthermore, this microRNA has a relatively high expression value (over 2% of all mapped reads) which might indicate its importance. This in turn shows that sRNAbench is capable to detect putatively important microRNAs that have been missed so far.

Finally, to run the program locally, a small database needs to be set up. In order to facilitate this process, we generated several small helper tools which are available at: http://bioinfo5.ugr.es/sRNAbench/sRNAbenchParser.php. These tools allow the user to retrieve automatically genome annotations and reference sequence libraries in sRNAbench format from databases such as Reference Sequence Database (RefSeq) and Ensembl [51, 52]

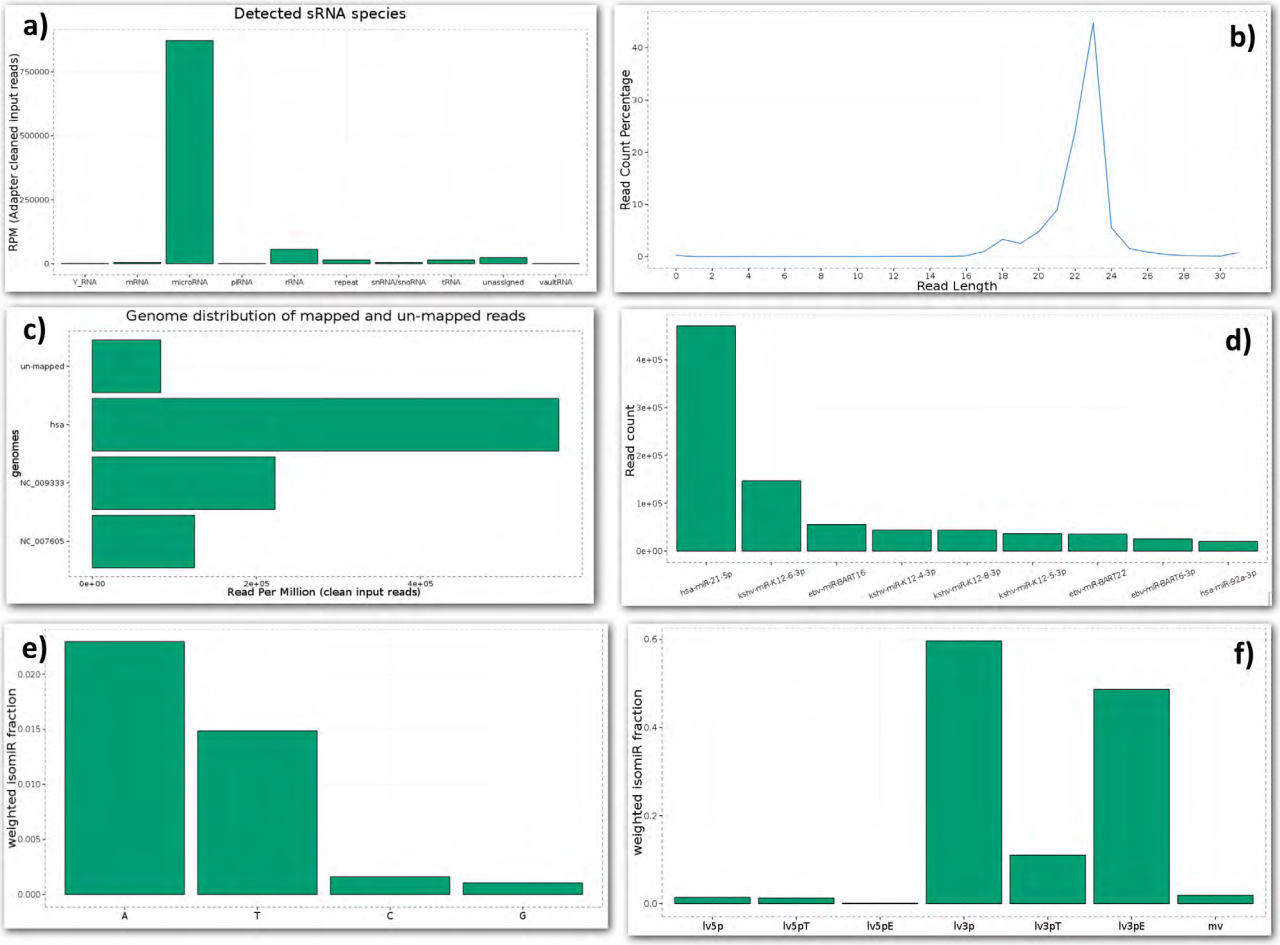

Figure 6: Some of the graphics generated by the sRNAbench webserver. a) the relative frequency of different RNA classes: in this sample microRNAs are by far the most frequent class of small RNAs, b) the read length distribution: the dominant peak at 22 nt corresponds to microRNAs while a very small peak can be seen at 18 nt which might be attributable to tRNA fragments, c) the read distribution over the different genomes, d) the top 10 most frequent microRNAs, e) the relative frequency of non-templated additions: adenylation and uridylatin is clearly more frequent than added G's or C's and f) the relative frequencies of the different length variants: this graphic confirms that 3' length variants are far more frequent than 5' length variants.

```
ATGAAAACTTGTAGAGGTGATTTGGTGATCACCGAATCTCTTGTATTCGGTGCACACCAAATCACTTCCACATGCTTTC  105,661
..((((...(((.(((((((((((((((..(((((((((.....)))))))))..))))))))))))))).)))...)))))
     AACTTGTAGAGGTGATTTGGT                                                       100,908
     AACTTGTAGAGGTGATTTGG                                                        2,684
        TGTAGAGGTGATTTGGTGATC                                                    782
      ACTTGTAGAGGTGATTTGGT                                                       375
     AACTTGTAGAGGTGATT                                                           161
     AACTTGTAGAGGTGATTTG                                                         144
         TGTAGAGGTGATTTGGTGAT                                                    113
    AAACTTGTAGAGGTGATTTGG                                                        112
         TGTAGAGGTGATTTGGTGA                                                     100
     AACTTGTAGAGGTGATTT                                                          66
     AACTTGTAGAGGTGAT                                                            62
     AACTTGTAGAGGTGATTTGGTG                                                      45
     AACTTGTAGAGGTGA                                                             14
                                          ACACCAAATCACTTCCACATG                  13
      ACTTGTAGAGGTGATTTGG                                                        11
         TGTAGAGGTGATTTGGTGATCA                                                  10
 TGAAAACTTGTAGAGGTGATT                                                           9
       CTTGTAGAGGTGATTTGGT                                                       8
        TTGTAGAGGTGATTTGGT                                                       7
     AACTTGTAGAGGTGATTTGGTGA                                                     6
                                          CAAATCACTTCCACATGCTTT                  5
   AAAACTTGTAGAGGTGATTTGGT                                                       5
   AAAACTTGTAGAGGTGATTTGG                                                        5
    AAACTTGTAGAGGTGATTTGGT                                                       4
                                          CAAATCACTTCCACATGCTT                   3
                                          CAAATCACTTCCACATGCT                    3
                                        CACACCAAATCACTTCCACAT                    3
      ACTTGTAGAGGTGATTTGGTG                                                      3
```

Figure 7: Novel microRNA predicted in barley cultivar 'Golden Promise'. The guide strand is shown in red and the passenger strand in green.

4 Discussion

sRNAbench is a user-friendly tool for the expression profiling of small RNAs from high-throughput experiments. The existence of a webserver implementation makes it easy to use for users with less bioinformatics background, while the standalone version allows accessing the full parameter space and more customized analysis steps.

Acknowledgements: Funding: Funding was provided by the Spanish Government (FIS2012-36282) and Basque country 'AE' grant (GB).

Conflict of Interest: none declared.

References

[1] Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu A-L, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA: Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* 2008, 18:610–21.

[2] Finnegan EJ, Matzke MA: The small RNA world. *J Cell Sci* 2003, 116(Pt 23):4689–93.

[3] Grosshans H, Filipowicz W: Molecular biology: the expanding world of small RNAs. *Nature* 2008, 451:414–6.

[4] Axtell MJ: Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol* 2013, 64:137–59.

[5] Tuck AC, Tollervey D: RNA in pieces. *Trends Genet* 2011, 27:422–32.

[6] Hall AE, Turnbull C, Dalmay T: Y RNAs: recent developments. *Biomol Concepts* 2013, 4:103–110.

[7] Neilsen CT, Goodall GJ, Bracken CP: IsomiRs--the overlooked repertoire in the dynamic microRNAome. *Trends Genet* 2012, 28:544–9.

[8] Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, Krishnan K, Gardiner B, Wang X, Nones K, Steen JA, Matigian NA, Wood DL, Kassahn KS, Waddell N, Shepherd J, Lee C, Ichikawa J, McKernan K, Bramlett K, Kuersten S, Grimmond SM: MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol* 2011, 12:R126.

[9] Burroughs AM, Ando Y, de Hoon MJL, Tomaru Y, Nishibu T, Ukekawa R, Funakoshi T, Kurokawa T, Suzuki H, Hayashizaki Y, Daub CO: A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. *Genome Res* 2010, 20:1398–410.

[10] Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, Moulton V: A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 2008, 24:2252–3.

[11] Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 2008, 26:407–15.

[12] Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM: miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* 2009, 37(Web Server issue):W68–76.

[13] Pantano L, Estivill X, Martí E: SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res* 2010, 38:e34.

[14] MacLean D, Moulton V, Studholme DJ: Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *BMC Bioinformatics* 2010, 11:93.

[15] Hendrix D, Levine M, Shi W: miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol* 2010, 11:R39.

[16] Mathelier A, Carbone A: MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 2010, 26:2226–34.

[17] Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM: miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res* 2011, 39(Web Server issue):W132–8.

[18] Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S: DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* 2011, 39(Web Server issue):W112–7.

[19] Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V: The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 2012, 28:2059–61.

[20] Hardcastle TJ, Kelly KA, Baulcombe DC: Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics* 2012, 28:457–63.

[21] Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, Cooke HJ, Xue Y, Shi Q: CPSS: a computational platform for the analysis of small RNA deep sequencing data. *Bioinformatics* 2012, 28:1925–7.

[22] Chen C-J, Servant N, Toedling J, Sarazin A, Marchais A, Duvernois-Berthet E, Cognat V, Colot V, Voinnet O, Heard E, Ciaudo C, Barillot E: ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics* 2012, 28:3147–9.

[23] Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N: miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012, 40:37–52.

[24] Axtell MJ: ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 2013, 19:740–51.

[25] Sablok G, Milev I, Minkov G, Minkov I, Varotto C, Yahubyan G, Baev V: isomiRex: web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. *FEBS Lett* 2013, 587:2629–34.

[26] An J, Lai J, Lehman ML, Nelson CC: miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res* 2013, 41:727–37.

[27] Wu J, Liu Q, Wang X, Zheng J, Wang T, You M, Sheng Sun Z, Shi Q: mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol* 2013, 10:1087–92.

[28] Katiyar-Agarwal S, Jin H: Role of small RNAs in host-microbe interactions. *Annu Rev Phytopathol* 2010, 48:225–46.
[29] Skalsky RL, Cullen BR: Viruses, microRNAs, and host interactions. *Annu Rev Microbiol* 2010, 64:123–41.
[30] Libri V, Miesen P, van Rij RP, Buck AH: Regulation of microRNA biogenesis and turnover by animals and their viruses. *Cell Mol Life Sci* 2013, 70:3525–44.
[31] Bernal D, Trelis M, Montaner S, Cantalapiedra F, Galiano A, Hackenberg M, Marcilla A: Surface analysis of Dicrocoelium dendriticum. The molecular characterization of exosomes reveals the presence of miRNAs. *J Proteomics* 2014.
[32] Hoy AM, Lundie RJ, Ivens A, Quintana JF, Nausch N, Forster T, Jones F, Kabatereine NB, Dunne DW, Mutapi F, Macdonald AS, Buck AH: Parasite-derived microRNAs in host serum as novel biomarkers of helminth infection. *PLoS Negl Trop Dis* 2014, 8:e2701.
[33] Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10:R25.
[34] Kozomara A, Griffiths-Jones S: miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014, 42(Database issue):D68–73.
[35] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL: ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011, 6:26.
[36] Taub M, Lipson D, Speed TP: Methods for Allocating Ambiguous Short-reads. *Commun Inf Syst* 2010, 10:69–82.
[37] Roberts A, Pachter L: Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* 2013, 10:71–3.
[38] D'Ambrogio A, Gu W, Udagawa T, Mello CC, Richter JD: Specific miRNA stabilization by Gld2-catalyzed monoadenylation. *Cell Rep* 2012, 2:1537–45.
[39] Hackenberg M, Shi B-J, Gustafson P, Langridge P: Characterization of phosphorus-regulated miR399 and miR827 and their isomirs in barley under phosphorus-sufficient and phosphorus-deficient conditions. *BMC Plant Biol* 2013, 13:214.
[40] Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, Griffiths-Jones S, Jacobsen SE, Mallory AC, Martienssen RA, Poethig RS, Qi Y, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu J-K: Criteria for annotation of plant MicroRNAs. *Plant Cell* 2008, 20:3186–90.
[41] Meng F, Hackenberg M, Li Z, Yan J, Chen T: Discovery of novel microRNAs in rat kidney using next generation sequencing and microarray validation. *PLoS One* 2012, 7:e34394.
[42] Kozomara A, Griffiths-Jones S: miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 2011, 39(Database issue):D152–7.
[43] Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26:139–40.
[44] Maza E, Frasse P, Senin P, Bouzayen M, Zouine M: Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Commun Integr Biol* 2013, 6:e25849.
[45] Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013, 14:671–83.
[46] Pfeffer S, Zavolan M, Grässer FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T: Identification of virus-encoded microRNAs. *Science* 2004, 304:734–6.
[47] Cai X, Lu S, Zhang Z, Gonzalez CM, Damania B, Cullen BR: Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Natl Acad Sci U S A* 2005, 102:5570–5.
[48] Mayer KFX, Waugh R, Brown JWS, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N: A physical, genetic and functional sequence assembly of the barley genome. *Nature* 2012, 491:711–6.
[49] Hackenberg M, Shi B-J, Gustafson P, Langridge P: A transgenic transcription factor (TaDREB3) in barley affects the expression of microRNAs and other small non-coding RNAs. *PLoS One* 2012, 7:e42030.
[50] Hackenberg M, Huang P-J, Huang C-Y, Shi B-J, Gustafson P, Langridge P: A comprehensive expression profile of microRNAs and other classes of non-coding small RNAs in barley under phosphorous-deficient and -sufficient conditions. *DNA Res* 2013, 20:109–25.
[51] Pruitt KD, Tatusova T, Brown GR, Maglott DR: NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012, 40(Database issue):D130–5.
[52] Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón