The discovery potential of RNA processing profiles

Amadís Pagès[1,2], Ivan Dotu[1,3], Roderic Guigó[1,2] and Eduardo Eyras[1,4,]*

[1]Universitat Pompeu Fabra (UPF), E08003 Barcelona, Spain.
[2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, E08003 Barcelona, Spain.
[3]IMIM - Hospital del Mar Medical Research Institute. E08003 Barcelona, Spain.
[4]Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain

*correspondence to: eduardo.eyras@upf.edu

Abstract

Small non-coding RNAs are highly abundant molecules that regulate essential cellular processes and are classified according to sequence and structure. Here we argue that read profiles from size-selected RNA sequencing, by capturing the post-transcriptional processing specific to each RNA family, provide functional information independently of sequence and structure. SeRPeNT is a computational method that exploits reproducibility across replicates and uses dynamic time-warping and density-based clustering algorithms to identify, characterize and compare small non-coding RNAs, by harnessing the power of read profiles. SeRPeNT is applied to: a) generate an extended human annotation with 671 new RNAs from known classes and 131 from new potential classes, b) show pervasive differential processing between cell compartments and c) predict new molecules with miRNA-like behaviour from snoRNA, tRNA and long non-coding RNA precursors, dependent on the miRNA biogenesis pathway. SeRPeNT facilitates the fast and accurate discovery and characterization of small non-coding RNAs at unprecedented scale.

Introduction

Small non-coding RNAs (sRNAs) are highly abundant functional transcription products that regulate essential cellular processes, from splicing or protein synthesis to the catalysis of post-transcriptional modifications or gene expression regulation[1]. Major classes include micro-RNAs (miRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and transfer RNAs (tRNAs). Developments of high-throughput approaches have facilitated their characterization in terms of sequence and structure[2,3,4] and have led to the discovery of new molecules in diverse physiological and pathological contexts. However, the function of many of them remains unknown[5,6] and their characterization may be essential to understand multiple cellular processes in health and disease.

Sequence and structure are traditionally used to identify and characterize small non-coding RNAs[7,8]. Although sequence is a direct product of the sequencing technology, structure determination is still of limited accuracy and requires of specialized protocols[3,4,9]. On the other hand, extensive processing is a general characteristic of non-coding RNAs[10,11,12]. The best characterized cases are miRNAs, which are processed from precursors and preferentially express one arm over the other depending on the cellular conditions[13,14]. Furthermore, snoRNAs and tRNAs can be processed into smaller RNAs, whose function is often independent of their precursor[10,15,16,17,18].

These findings suggest that a new path to systematically characterize RNA molecules emerges through the genome-wide analysis of their sequencing read profiles. Here we argue that sequencing profiles can be used to directly characterize the function of small non-coding RNAs, in the same way that sequence and structure have been used in the past. We report here on SeRPeNT (Small Rna ProfiliNg Toolkit), a fast and memory efficient software for the discovery and characterization of known and novel classes of small non-coding RNAs exploiting their processing pattern from small RNA sequencing experiments.

Results

Fast and accurate discovery of small non-coding RNAs

Using multiple size-selected (<200nt) small RNA sequencing (sRNA-seq) experiments mapped to a genome reference, SeRPeNT enables the discovery and characterization of known and novel small non-coding RNAs (sRNAs) through three operations: *profiler*, *annotator* and *diffproc*, which can be used independently or together in a pipeline (**Fig. 1**). Initially, sRNA read profiles are calculated from the mapped sRNA-seq reads, and filtered according to the reproducibility between replicates, and to the length and expression constraints given as input (**Fig. 1a** and Online Methods). Pairwise distances between profiles are calculated as a normalized cross-correlation of their alignment calculated using a time-warping algorithm (**Fig. 1b** and Online Methods). Profiles are clustered into families according to pairwise distances using an improved density-based clustering algorithm (**Fig. 1b** and Online Methods). Novel profiles are annotated using the class label from known profiles in the same cluster by majority voting (**Fig. 1c** and **Online Methods**). Additionally, SeRPeNT allows the identification of differential processing of sRNAs between two conditions that are independent of expression changes.This is calculated for each sRNA from the pairwise distance distributions with sister sRNAs from the same cluster in either condition. Profiles are considered as differentially processed according to the fold-change and significance of the change (**Fig. 1d** and Online Methods).

To assess the accuracy of SeRPeNT, we performed a comparison against two other methods that predict known small non-coding RNA families from sRNA-seq data, BlockClust[19] and DARIO[20]. We evaluated the accuracy to detect known miRNAs, tRNAs and C/D-box snoRNAs from the Gencode annotation[21], using cross-fold validation (**Supplementary Fig. 1** and

Supplementary Methods). Using the same data[8] for the three methods, SeRPeNT shows overall higher precision and a dramatic improvement of the recall in all tested sRNA families. In particular, miRNAs show the best accuracy (precision=0.99, recall=0.99), followed by tRNAs (precision=0.95, recall=0.96). For snoRNAs of C/D-box type, which proved hard to predict by the other methods (recall=0.39 and 0.52 by BlockClust and DARIO, respectively), SeRPeNT achieves 0.70 recall (**Supplementary Table 1**). Notably, SeRPeNT analysis took ~3 minutes and less than 200Mb of RAM in a single core AMD Opteron 64 with 4Gb of memory. In contrast, the same analysis for BlockClust, which includes the execution of Blockbuster[22], took ~15 minutes and used nearly 30Gb of memory.

We also assessed the accuracy of SeRPeNT differential processing operation *diffproc* by analyzing the differential expression of miRNA arms and arm-switching events in miRNAs between normal and tumor liver tissues[23]. From the 49 miRNAs tested[23], 41 passed our filters of reproducibility and clustered with other sRNAs. Imposing a significance threshold of p-value < 0.01 and a fold-change of at least 2.5 (**Supplementary Fig. 2**), SeRPeNT identified as differentially processed 10 out of 24 miRNAs described to exhibit different 5'-arm to 3'-arm expression ratio[23], including 4 out of 5 arm-switching events (**Supplementary Fig. 3**). Moreover, only 1 out of the remaining 17 miRNAs that did not exhibit difference in 5'-arm to 3'-arm expression ratio was identified as differentially processed by SeRPeNT. We further compared SeRPeNT against RPA[24], a recent method for differential processing analysis, using data from 9 cell lines[25]. SeRPeNT detects many more differentially processed events, with a moderate overlap with RPA predictions (**Supplementary Fig. 4**). Notably, SeRPeNT took 2 hours in a single core AMD Opteron 64 with 4Gb of memory, whereas RPA took about 10 hours in a cluster of 32 cores each having 8 Gb of RAM.

An extended annotation of small non-coding RNAs in human

We decided to exploit SeRPeNT speed and accuracy to produce an extended annotation of small non-coding RNAs in human. We applied SeRPeNT *profiler and annotator* to sRNA-seq data from 9 cell lines[25] (**Supplementary Table 2**). There is a higher proportion of known RNAs compared to novel sRNAs, with an increase of novel sRNAs in samples sequenced at a higher depth: A549, IMR90, MCF-7 and SK-N-SH (**Fig. 2a**). We further measured the accuracy of SeRPeNT with these datasets and found an overall high accuracy consistently across all cell lines (**Supplementary Table 3**).

We annotated new sRNAs with SeRPeNT and obtained a total of 4,673 non-unique sRNAs across all tested cell lines that are not in Gencode (**Supplementary Table 4**). We were able to assign a label to 2,140 of them. From the remaining 2,533 unlabeled sRNAs, 323 formed 92 clusters with three or more unlabeled profiles per cluster, suggesting possible new classes of non-coding RNAs with a coherent processing pattern. We called these *clustered uncharacterized* RNAs (cuRNAs) and kept them for further study. Interestingly, some known and predicted sRNAs with the same class labels are grouped into different clusters, indicating subfamilies. For instance, SeRPeNT separates C/D-box and H/ACA-box snoRNAs according to their processing profiles (clusters 1 and 2 in **Fig. 2b**), and separates miRNAs into subtypes according to their different arm-processing patterns (clusters 5 and 11 in **Fig. 2b**). Thus SeRPeNT identifies functional families and subfamilies of non-coding RNAs in a scalable and robust way, independently of the granularity of the available annotation.

We established the consistency of the sRNAs across the multiple experiments using an entropy measure of the label assignment across cell lines and removed low mappability regions (**Supplementary Methods**), producing a total of 929 unique novel sRNAs, 787 from the major classes (79 miRNAs, 475 snoRNAs, 82 snRNAs and 151 tRNAs) plus 142 cuRNAs, the majority of them being expressed in only one cell line (**Supplementary Fig. 5**). These, together with the sRNAs annotated in Gencode, conform an extended annotation of the catalogue of small non-coding RNAs in the human genome reference available in (**Supplementary Data 1**) and in GTF format in (**Supplementary Data 2**). Direct access to a UCSC track is available from http://regulatorygenomics.upf.edu/sessions/rgs005/serpent.html.

From the 79 newly predicted miRNAs, 37 were validated as miRNA precursors using FOMmiR[26] (**Supplementary Data 3**). To further characterize these miRNAs, we searched for sequence and secondary structure similarities in Rfam using Infernal[27,28], with threshold e-value < 0.01 (**Supplementary Methods**). We found that 23 of them had a hit to a known miRNA family (**Supplementary Data 1**). Repeating these analyses for the other new sRNAs we found 47 snoRNA and 15 tRNAs with a hit to an Rfam family, from which 3 snoRNAs and 4 tRNAs had the hit to a family of the same class predicted by SeRPeNT (**Supplementary Data 1**). The rest of predicted sRNAs did not have a hit in Rfam. We compared our extended annotation predicted sRNAs with DASHR[6], the most recently published database of human small human non-coding RNAs, and with a compendium of human miRNAs from a recent study using multiple samples[29]. We found that 802 out of the 929 predicted sRNAs (51 miRNAs, 430 snoRNAs, 69 snRNAs, 121 tRNAs and 131 cuRNAs) were not present in those sRNA catalogues. In particular, 4 predicted miRNAs had a hit to an Rfam miRNA family, were validated with FOMmiR and were not present in previous catalogues[6,29] (**Fig. 2c**).

Pervasive differential processing of non-coding RNAs between cell compartments

To further characterize the extended sRNA annotation defined above, we studied their differential processing between four different cell compartments: chromatin, nucleoplasm, nucleolus and cytosol for the cell line K562 using replicated data[25] (**Supplementary Table 2**). The majority of sRNAs from the extended annotation showed expression in one or more cell compartments: 599 in chromatin, 763 in cytosol, 554 in nucleolus and 651 in nucleoplasm. The majority of sRNAs in cytosol are tRNAs (45%), followed by miRNAs (15%). Although tRNAs are enriched in the cytosol (Fisher's one-sided test p-value < 0.001), they are abundant in all four cell compartments (**Supplementary Table 5**). This is compatible with tRNA biogenesis, which comprises early processing in the nucleolus and later processing in the nucleoplasm before export to the cytoplasm[30]. In contrast, miRNA clusters appear almost exclusively in the cytosol (Fisher's one-sided test p-value < 0.001) and are coherently grouped into large clusters (**Fig.**

3a) (**Supplementary Table 5**). The nucleolus is enriched in snoRNAs accounting for 38% of the found profiles (Fisher's one-sided test p-value < 0.01). Interestingly, snoRNAs are also enriched in the chromatin compartment (Fisher's one-sided test p-value <0.001) accounting for 23% of the sRNAs found there, suggesting new candidates for their recognized role on establishing open chromatin domains[31]. Finally, snRNAs and cuRNAs appear at low frequency in most compartments (**Supplementary Table 5**). We applied SeRPeNT *diffproc* operation for each pair of compartments, using fold-change ≥ 2.5 and p-value < 0.01. A large proportion of snoRNAs show differential processing between the nucleus and nucleolus, where they exert their function, and the rest of cellular compartments (**Fig. 3b**). Only 4 of the cuRNAs identified show expression in at least two compartments, nucleolus and cytosol, and 3 of them show differential processing. Overall, tRNAs show the largest proportion of differentially processed profiles between the cytosol and the different nuclear compartments (**Fig. 3b** and **Supplementary Data 4**). Many of these tRNAs show a more prominent processing in the cytosol from the 30-35nt part of their 3' part (**Fig. 3c** and **Supplementary Fig. 6**), also called tRNA halves[32,33].

SeRPeNT uncovers new RNAs with potential miRNA-like function

The analysis of the compartments showed that some clusters at the cytosol group together snoRNAs and miRNAs, suggesting similar processing patterns. Additionally, SeRPeNT analysis on individual cell lines identified a cluster that groups together snoRNA SCARNA15 (ACA45) with 2 miRNAs in NHEK**,** and a cluster that groups snoRNA SCARNA3 with several miRNAs and a tRNA in A549 (**Supplementary Table 6**), agreeing with a previous study showing that these snoRNAs can function as miRNAs[15]. SeRPeNT clusters in cell lines provide additional evidence of 6 other snoRNAs that group with miRNAs: SNORD116, SNORA57, SNORD14C, SNORD26, SNORD60 and SNORA3 (**Supplementary Table 6**). Interestingly, we also found 7 clusters with a majority of miRNAs that included annotated tRNAs: tRNA-Ile-GAT, tRNA-Glu-GAA, tRNA-Gly-CCC, tRNA-Ala-AGC and tRNA-Leu-AAG. In particular, tRNA-Ile-GAT-1-1 clusters with miRNAs in 3 different cell lines: MCF-7, A549 and SK-N-SH, suggesting new tRNAs with miRNA-like function[10,34]. These results support the notion that sRNA read-profiles facilitate the direct identification of functional similarities without the need to analyze sequence or structure.

To search for new cases of miRNA-like non-coding RNAs in the extended annotation, we tested their potential association with components of the canonical miRNA biogenesis pathway, using sRNA-seq data from controls and individual knockouts of *DICER1*, *DROSHA* and *XPO5*[35] (Supplementary Methods). We validated the dependence of a number of known and predicted miRNAs (**Fig. 4a**) (**Supplementary Figs. 7 and 8**) and recovered the previously described dependence of ACA45 and SCARNA3 with *DICER1*[15]. Additionally, we found 18 sRNAs predicted as snoRNAs with similar behaviour upon *DICER1* knockout (**Fig. 4b**). Interestingly, 14 out of 20 *DICER1*-dependent snoRNAs did not show dependence on *DROSHA* (**Supplementary Fig. 7b**), including ACA45 and SCARNA3, in agreement with previous findings[15,35] (**Supplementary Data 4**). We also found a strong dependence on *DICER1* for 128 tRNAs, 82 of which changed expression in the direction opposite to most miRNAs, suggesting that they may be repressed by DICER (**Fig. 4c**). Further, 4 cuRNAs showed similar results to miRNAs, suggesting some association with the miRNA biogenesis machinery (**Supplementary Fig. 9a-c**) (**Supplementary Data 4**). Although they were not confirmed as miRNA precursors using FOMmiR, 2 of these miRNA-like cuRNAs overlap with the protein-coding genes *SEC24C* and *DHFR* (**Supplementary Fig. 9d**).

Certain long non-coding RNAs (lncRNAs) are known to act as precursors of miRNAs[36,37] and tRNAs[38]. We thus analyzed whether the new sRNAs could originate from lncRNAs. We found that 8 miRNAs, 16 snoRNAs, 7 tRNAs and 4 cuRNAs overlap annotated lncRNAs (**Supplementary Data 4**). These lncRNAs include *MALAT1*, which overlaps with 2 miRNAs, 2 tRNAs and 1 cuRNA. Interestingly, 3 of the miRNAs predicted and validated by FOMmiR were found on lncRNAs MIR100HG, CTD-23C24-1, and RP11-141B14.1. From these, the new miRNA in RP11-141B14.1 is not present in recent miRNA catalogues (**Figs. 4d and 4e)**. As the processing from lncRNAs is a recognized biogenesis mechanism for certain small non-coding RNAs, these results provide further support for the relevance of the newly predicted sRNAs in our extended annotation.

Discussion

SeRPeNT provides a fast and accurate method to identify known and novel non-coding RNAs exploiting read profiles from stranded size-selected RNA sequencing data. SeRPeNT does not depend on the annotation granularity and avoids many drawbacks inherent to sequence and secondary structure based methods, which may be affected by post-transcriptional modifications or limited by the reliability of structure determination. Here we have shown that read profiles, by capturing the post-transcriptional processing that is specific to each sRNA family, provide functional information independently of sequence or structure. In particular, a number of known snoRNAs and tRNAs clustered with miRNAs according to their profiles. Beyond the known cases, we detected new candidates of this dual behaviour. It remains to be determined whether these new sRNAs can indeed function as miRNAs and associate with AGO2[39]. It is possible that they compete with more abundant miRNAs to be loaded on the RNA-induced silencing complex, hence they might become more prominent in specific cellular conditions. Incidentally, many sRNAs increase expression when this is measured from the sequencing of AGO2-associated reads in *DICER1* knocked-down cells (data not shown), suggesting a repression by *DICER1*[40] or an association to alternative biogenesis pathways[35].

We have generated an extended annotation for human that includes hundreds of previously unannotated sRNAs from known classes. These included new miRNAs, which we validated comparing to known families, confirming the structure of the precursor, and by measuring their expression dependence with the miRNA biogenesis machinery. We further observed the frequent differential processing of sRNAs across cell compartments, especially for tRNAs. As differential processing of tRNAs has been observed in association to disease[41,42,43], the observed patterns may be indicative of relevant cellular processes that are worth investigating further.

We also detected 131 new sRNAs that could not be labeled, which we named clustered uncharacterized RNAs (cuRNAs), and which are not present in current sRNA catalogues, hence could correspond to novel sRNA species. Although cuRNAs did not show frequent differential processing across cell compartments, they showed dependencies with the miRNA processing machinery and overlap with lncRNAs; suggesting some form of biogenesis. The role of lncRNAs as possible general precursors of multiple types of sRNAs in fact suggests new possible ways to classify lncRNAs beyond the current proposed frameworks[44]. A subset of lncRNAs may act as precursors of a wide variety of small non-coding RNAs, including those from known families.

We envision a wide variety of future applications of SeRPeNT, including the fast identification and differential processing of non-coding RNAs from size-selected RNA-sequencing from tumor biopsies, circulating tumor cells, or exosomes, as well as the rapid discovery and characterization of non-coding RNAs families in multiple organisms. SeRPeNT differential processing operation can also be very powerful at, for instance, discovering RNAs that are differentially processed in tumor cells, thus generating biomarkers and potential drug targets. In summary, SeRPeNT provides a fast, easy to use and memory efficient software for the discovery and characterization of known and novel classes of non-coding RNAs.

Methods

SeRPeNT is written in C. The source code and a binary for Linux are available at https://bitbucket.org/regulatorygenomicsupf/serpent under the MIT license. Makefiles to reproduce the analyses described in this manuscript are available from the same site.

Profile building from aligned short RNA-Seq reads

The tool *profiler* uses as input one or more small RNA sequencing replicates in BAM format. Consensus read-contigs are built by pooling all the reads that overlap on a genomic region and that are at a distance smaller than a user-defined threshold. Each contig is scored per individual replicate by counting the number of reads mapped within its boundaries and reproducibility is measured across all the biological replicates. For all analyses of reproducibility in this paper we used npIDR[45] with cut-off of 0.01. SeRPeNT allows using also SERE[46] for reproducibility. SERE (simple error ratio estimate) compares the observed variation in the raw number of reads of a contig to an expected value, accounting for the impact of variation in read depth across replicates, whereas npIDR (non-parametric irreproducible discovery rate) determines the reproducibility of a contig in one or more replicates with similar sequencing depths. Contigs that do not pass the user-defined cutoff of reproducibility are discarded from further analysis. For each of the remaining contigs, a profile is built by counting the number of reads per nucleotide in the genomic region delimited by the contig boundaries. Each sRNA is defined as a genomic region and a vector of raw read counts, or heights, of length equal to the number of nucleotides spanned by this genomic region. Profiles are additionally trimmed at the 3'-end positions, when heights were either below 5 reads or below 10% of the highest position, but not when having more than 20 reads. Only profiles of lengths between 50 and 200 nucleotides, and of minimum height 100 in pooled replicates, were considered. All these parameters can be configured on SeRPeNT command line. The consistency of sRNA profiles across multiple experiments was determined by calculating a normalized entropy of the different labels for the same sRNA locus across experiments (**Supplementary Methods**).

sRNA profile clustering

The *annotator* tool assigns a distance between each possible pair of profiles resulting from the previous step. This distance is computed with a novel algorithm (**Supplementary Fig. 10**) based on dynamic time-warping[47,48] that accomplishes the task of finding the optimal alignment between two profiles by placing the height of each profile along the axes of a grid, representing alignments as paths through the grid cells, and finding the path with maximum normalized cross-correlation score across it. Given a pair of profiles of the same length $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_n)$, where a_i and b_i are the heights of nucleotide i in profile A and B, respectively, we define the cross-correlation score between A and B as:

$$A \bullet B = \sum_{i=1}^{n} a_i \cdot b_i \quad (1)$$

and the normalized cross-correlation score as:

$$r_{A,B} = \frac{A \bullet B}{\sqrt{(A \bullet A)(B \bullet B)}} \quad (2)$$

This optimal alignment maximizes its normalized cross-correlation score between the two profiles. Given two profiles $S = (s_1, \ldots, s_n)$ and $Q = (q_1, \ldots, q_m)$ of length n and m nucleotides respectively, each position (i, j) in the dynamic matrix D will store a vector of three values $D(i,j) = (x, y, z)$ such that they maximize the value $x/\sqrt{y \cdot z}$ in formula (2) amongst all the possible partial alignments between S_i and Q_j, where $S_i = (s_1, \ldots, s_i)$ and $Q_j = (q_1, \ldots, q_j)$ are the profiles spanning the first i and j nucleotides of the profiles S and Q. The dynamic programming equation is then defined as:

$$\mathrm{D(i,j) = (x, y, z)\ among}\ \begin{array}{l} \mathrm{D(i-1, j) + (s_i \cdot \phi_j,\ s_i \cdot s_i, \phi \cdot \phi)} \\ \mathrm{D(i-1, j-1) + (s_i \cdot q_j, s_i \cdot s_i, q_j \cdot q_j)} \\ \mathrm{D(i, j-1) + (\phi \cdot q_j, \phi \cdot \phi, q_j \cdot q_j)} \end{array}\ \text{that maximizes } x/\sqrt{y \cdot z} \quad (3)$$

where ϕ represents a negative Gaussian white noise function used to penalize an expansion or contraction in the alignment. When applied to a profile S, $\phi(S)$ returns a negative value taken from a random uniform distribution with mean and standard deviation defined by S. Once all the pairwise distances are calculated, profiles are clustered using a modified version of a density-based clustering algorithm[49] (**Supplementary Fig. 11**). For each sRNA profile i we define the local density ρ_i as the number of sRNA profiles that are close to profile ρ_i. We use an exponential kernel as described before[50], such that the local density ρ_i is defined as:

$$\rho_i = \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \quad (4)$$

where d_{ij} is the distance between profiles *i* and *j,* and d_c is a distance cutoff. The clustering algorithm is based on the assumption that cluster centers with high density are surrounded by neighbors of lower local density and lie at large distance from other profiles of high local density. At each iteration of the algorithm, the distance cutoff d_c is calculated[51] and the profile with the highest density is identified. This profile and all the profiles that are closer than d_c, are assigned to the same cluster. We introduced a novel step in the algorithm by which all the clustered profiles are removed before the next iteration step. The remaining unassigned profiles are then assigned to a different cluster, and so on. The algorithm stops when the calculated d_c is higher than a user-defined threshold. Although this modified version of the density-based clustering is slower than the original version, it shows slightly better accuracy in the cross-fold validation. Our software allows skipping the calculation of the threshold as an option, at the expense of a lower accuracy.

Profile annotation

The *annotator* tool performs the sRNA profile annotation. Every detected profile that overlaps an annotated short non-coding RNA is marked as known and labeled with the corresponding class label (e.g. H/ACA snoRNA). The overlap amount required between the sRNA profile and the annotated RNA is user-defined. Profiles that do not overlap with any annotation or do not satisfy the overlapping requirements are marked as unknown. For each cluster with two or more profiles, the different labels from all the known profiles are counted, and all the unknown profiles within the cluster are labeled by majority vote with the most abundant label. In case of a tie, the label of the closest profile is assigned. All the remaining profiles are denoted as unlabeled.

Differential processing analysis

The *diffproc* tool assesses if a profile P_a in a particular condition *A* shows a different processing pattern P_b in another condition *B*. Given A pair of profiles P_a and P_b from conditions A and B, respectively, such that their reference coordinates overlap as described above, are compared as follows. GIven K_a the cluster in condition *A* that contains the profile P_a and K_b the cluster in condition *B* that contains the profile P_b, *diffproc* calculates all the pairwise distances D_{ab} between P_a and all the profiles in K_b, and the pairwise distances D_b between profiles in K_b (**Fig. 1**). These two distance distributions are then compared using a one-sided Mann-Whitney *U* test and a fold-change is calculated as the ratio of the medians between both distributions. The same method is applied to profile P_b and cluster K_a. P_a and P_b are then reported as differentially processed if both tests are significant according to the p-value and fold-change cutoffs defined by the user. When there are not enough cases to perform a Mann-Whitney U test, only the fold-change is taken into account.

Figures

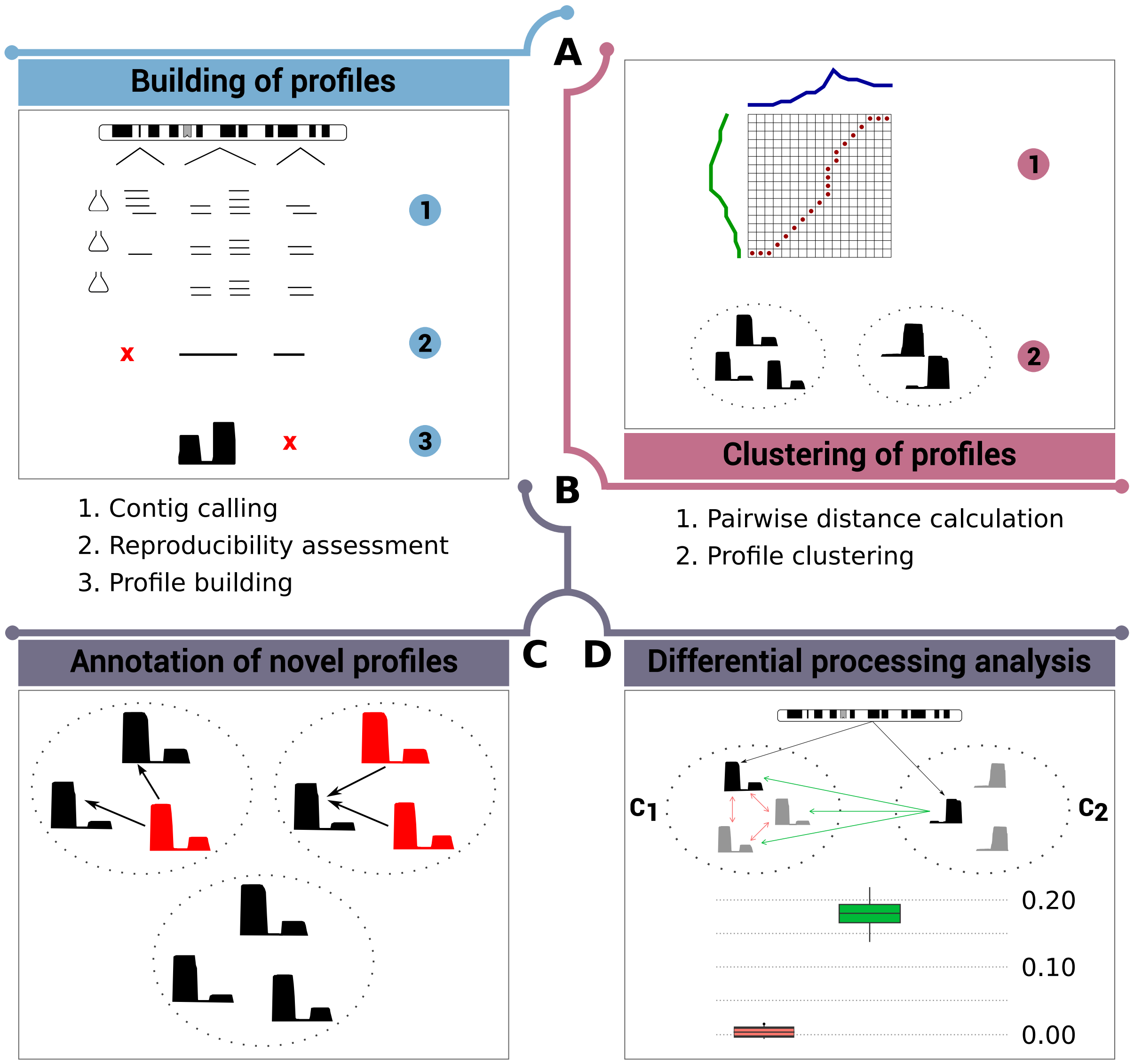

Figure 1. Overview of SeRPeNT. Overview of the operations performed by the SeRPeNT tools: (**a**) Building of profiles from short RNA-Seq mapped reads using reproducibility across replicates. A profile is a collection of reads overlapping over a given genomic locus and can be regarded as a vector where each component contains the number of reads at each nucleotide of that locus. (**b**) Density-based clustering of profiles based on pairwise distances calculated with a dynamic time-warping algorithm. (**c**) Annotation of novel profiles using majority vote in clusters. (**d**) Differential processing calculation. The distribution of distances between a profile and its clusters sisters in one condition cluster (C_1) and across conditions (C_2) are compared (panel below). Differential processing is determined in terms of a Mann-Whitney U test and a fold-enrichment (**Supplementary Fig. 2**).

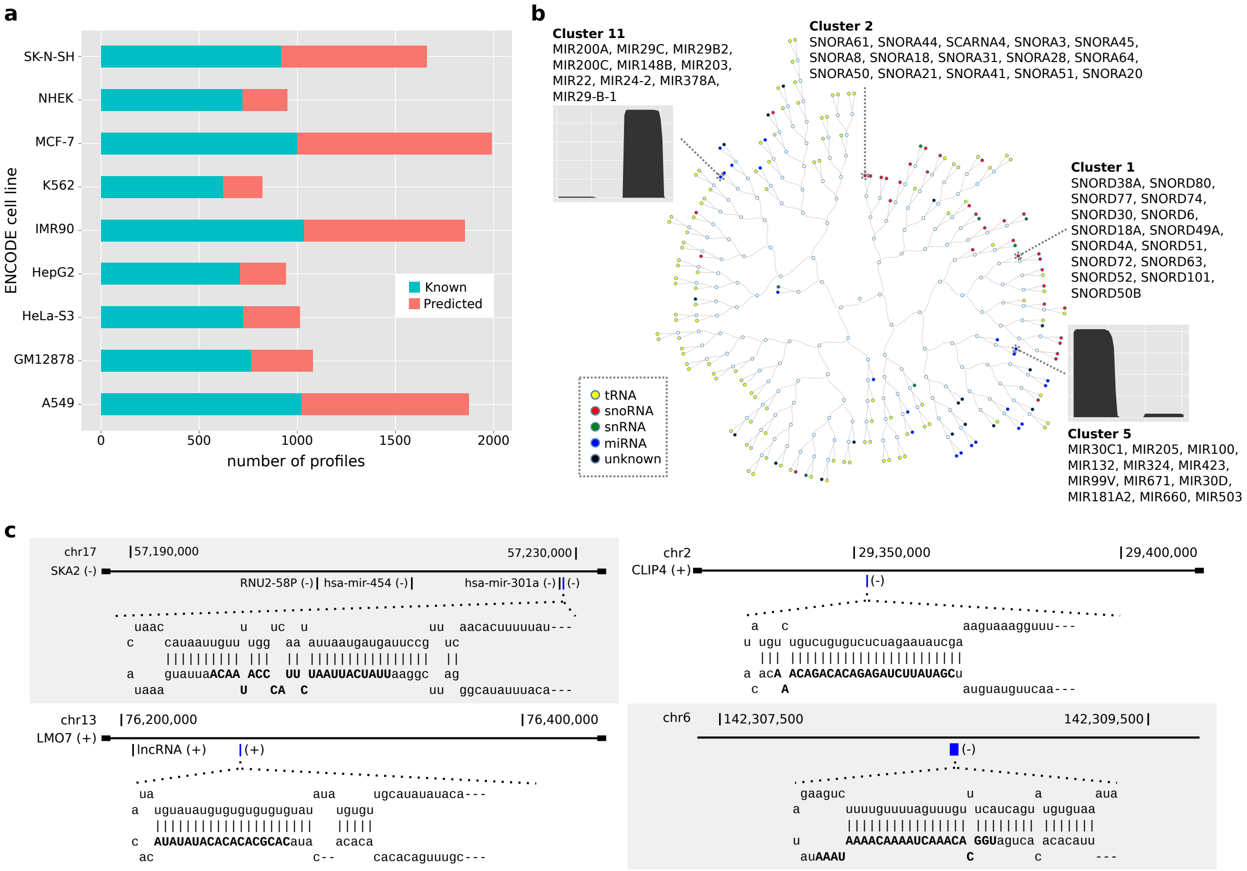

Figure 2. **Extended annotation derived from ENCODE cell lines**. (**a**) Number of known and novel sRNAs across 9 ENCODE cell line dataset. (**b**) Hierarchical clustering representation of the clusters obtained for the NHEK cell line. Distance between clusters is calculated by averaging all the distances between profiles from both clusters. Colored circles represent clusters of sRNAs at the leaves of the tree labeled by class. Empty circles represent internal nodes of the tree. The read profiles in clusters 5 and 111 are for one of its members, for which we plot the number of reads per nucleotide in the sRNA. (**c**) Genomic loci and graphical representation of the hairpins for the four novel microRNA. Predicted mature microRNAs are highlighted in blue in the gene loci SKA2 (MCF-7), LMO7 (SK-N-SH), CLIP4 (A549) and one intergenic region (K562).

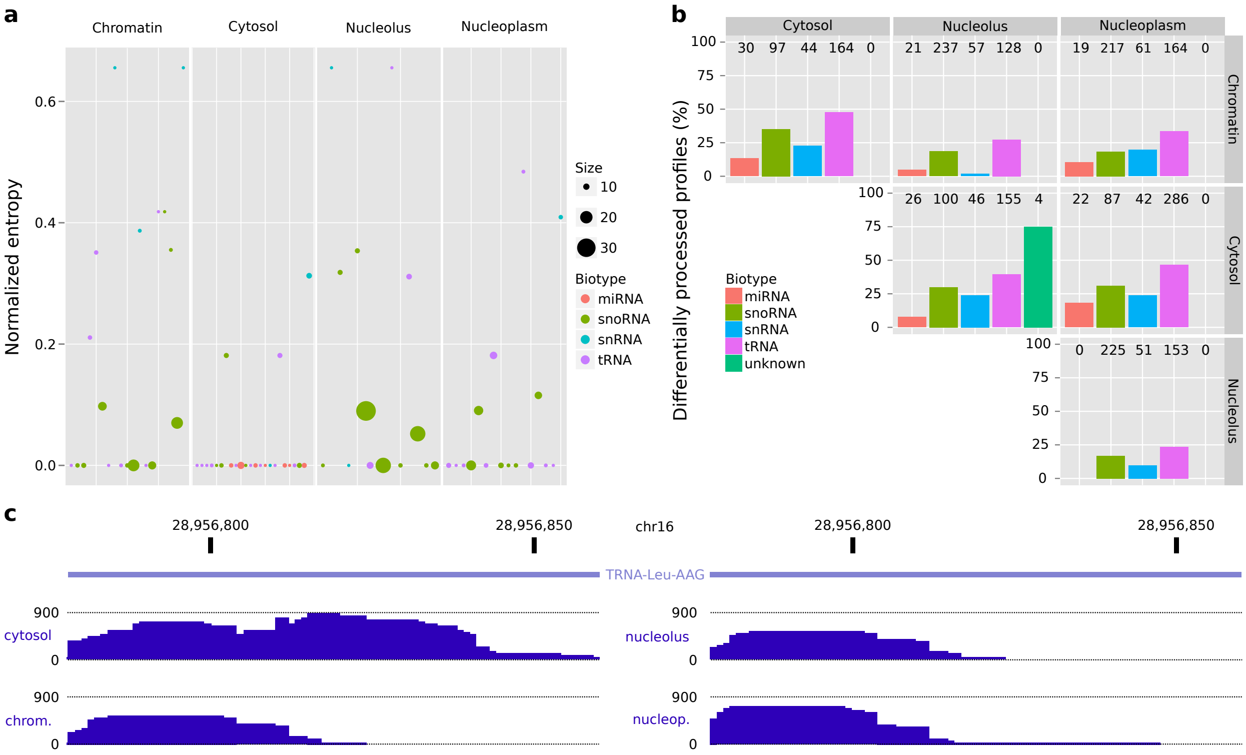

Figure 3. **Differential processing across ENCODE cell compartments**. (**a**) Representation of clusters containing 5 or more sRNAs across all four ENCODE cell compartments. The size of the points represents the number of sRNAs from the extended annotation contained in the cluster. The normalized entropy (y axis) represents the purity of a cluster (**Supplementary Methods**), the lower the entropy, the higher the purity of the cluster. (**b**) Proportion of profiles from the extended annotation that are differentially processed between cellular compartments separated by non-coding RNA family (y axis). Numbers in the top of the bars represent the total number of profiles detected in both compartments. (**c**) Representation of the read profiles for the tRNA-Leu-AAG transfer RNA showing abundant processing of the 3'-half in the cytosol compared to the chromatin compartment. The plot represents the number of reads per nucleotide in the same scale for each compartment.

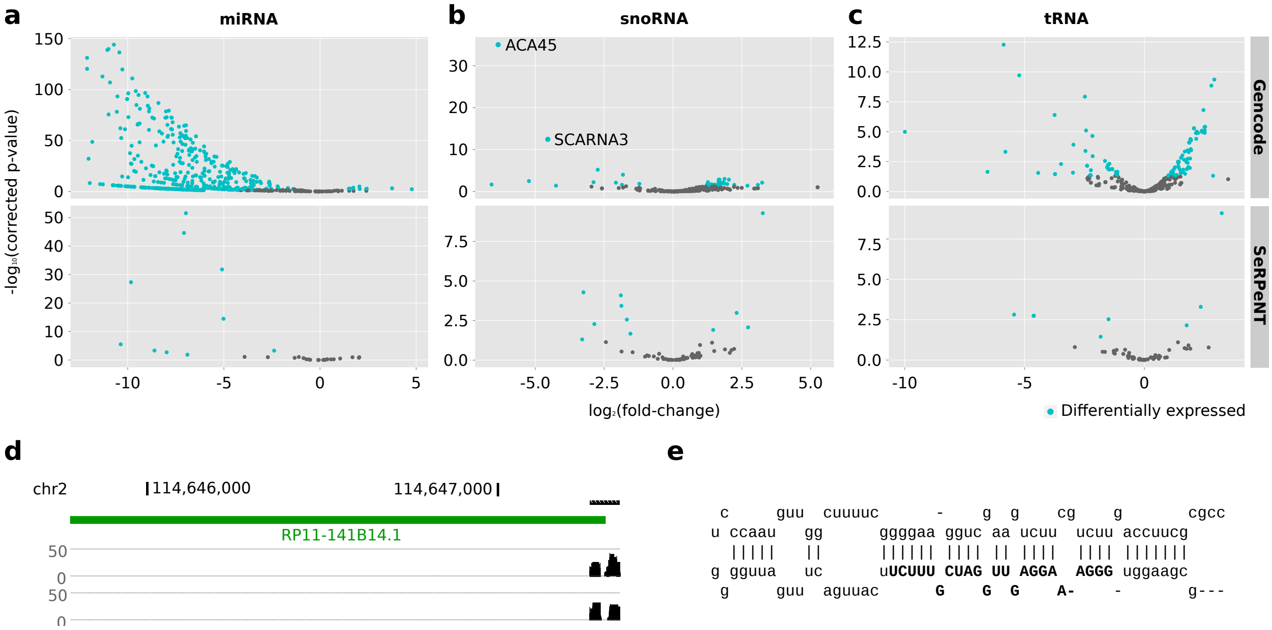

Figure 4. **Detection of miRNA-like sRNAs**. Differentially expressed sRNAs (blue) from extended annotation in the comparison between *DICER1* knockout and control experiments in human HCT116 cell lines for (**a**) miRNAs, (**b**) snoRNAs and (**c**) tRNAs. The analyses for the knockout of *DROSHA* and *XPO5* are available as Supplementary Figures. (**d**) Representation of a novel miRNA detected by SeRPeNT (depicted as a read profile) whose precursor is the lncRNA RP11-141B14.1 (depicted as a green line). Profiles for both replicates are included. (**e**) Secondary structure prediction of the miRNA precursor by FOMmiR.

Acknowledgements

We are thankful to D. Gautheret for useful discussions. EE and AP were supported by the MINECO and FEDER (BIO2014-52566-R), Consolider RNAREG (CSD2009-00080), by AGAUR (SGR2014-1121) and by the Sandra Ibarra Foundation for Cancer (FSI2013). RG and AP were supported from MINECO (BIO2011-26205), National Human Genome Research Institute of the National Institutes of Health (U54HG007004) and MINECO Centro de Excelencia Severo Ochoa 2013-2017 (SEV-2012-0208).

References

1. Morris, K. V. and Mattick, J. S. The rise of regulatory RNA. *Nature Reviews Genetics* **15**, 423-437 (2014).
2. Jha, A. et al. A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Research* **43**, 8713-8724 (2015).
3. Ge, P. and Zhang, S. Computational analysis of RNA structures with chemical probing data. *Methods* **79**, 60-66 (2015).
4. Ding, Y. et al. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nature protocols* **10**, 1050-1066 (2015).
5. Vickers, K. et al. Mining diverse small RNA species in the deep transcriptome. *Trends in biochemical sciences* **40**, 4-7 (2015).
6. Leung, Y. et al. DASHR: database of small human noncoding RNAs. *Nucleic Acids Research* **44** (Database issue), D216-D222 (2016).
7. Ritchie, W. et al. RNA stem–loops: To be or not to be cleaved by RNAse III. *Rna* **13**, 457-462 (2007).
8. Friedländer, M. et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* **40**, 37-52 (2012).
9. Siegfried, N. et al. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods* **11**, 959-965 (2014).
10. Lee, Y. S. et al. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development* **23**, 2639-2649 (2009).
11. Macias, S. et al. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nature structural & molecular biology* **19**, 760-766 (2012).
12. Chen, C.J. and Heard, E. Small RNAs derived from structural non-coding RNAs. *Methods* **63**, 76-84 (2013).
13. Griffiths-Jones et al. MicroRNA evolution by arm switching. *EMBO Reports* **12**, 171-177 (2011).
14. Ha, M., Kim, V.N. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology* **15**, 509-524 (2014).
15. Ender, C. et al. A human snoRNA with microRNA-like functions. *Molecular Cell* **32**, 519-528 (2008).

16. Kishore, S. et al. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Human molecular genetics* **19**, 1153-1164 (2010).
17. Brameier, M. et al. Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic acids research* **39**, 675-686 (2011).
18. Li, Z. et al. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic acids research* **40**, 6787-6799 (2012).
19. Videm, P. et al. BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-Seq profiles. *Bioinformatics* **30**, i274-i282 (2014).
20. Fasold, M. et al. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research* **39** (Web server issue), W112-W117 (2011).
21. Harrow, J. et al.GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760-1774 (2012).
22. Langenberger, D. et al. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics* **25**, 2298-2301 (2009).
23. Li et al. MicroRNA 3' end nucleotide modification patterns and arm selection preference in liver tissues. *BMC Systems Biology* **6**, S14 (2012).
24. Pundhir, S. and Gorodkin, J. Differential and coherent processing patterns from small RNAs. *Scientific reports* **5** (2015).
25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
26. Shen, W. et al. MicroRNA Prediction Using a Fixed-Order Markov Model Based on the Secondary Structure Pattern. *PLOS One* **DOI:10.1371/journal.pone.0048236** (2012).
27. Nawrocki, E. et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research* **43** (Database Issue), D130-D137 (2014).
28. Nawrocki, E. and Eddy, S. Infernal 1.1: 100-fold faster homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
29. Friedländer, M. et al. Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biology* **15**, R57 (2014).
30. Kirchner, S. and Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nature Reviews Genetics* **16**, 98-112 (2015).
31. Schubert, T. et al. Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. *Molecular Cell* **48**, 434-444 (2009).
32. Cole, C. et al. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15**, 2147-2160 (2009).
33. Telonis, A.G. et al. Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget* **22**, 24797-24822 (2015).
34. Venkatesh, Thejaswini, Padmanaban S Suresh, and Rie Tsutsumi. "tRFs: miRNAs in disguise." *Gene* (2015).
35. Kim, Y. et al. Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis. *Proceedings of the National Academy of Sciences* (2016).

36. Bevilacqua, V. et al. Identification of linc-NeD125, a novel long non coding RNA that hosts miR-125b-1 and negatively controls proliferation of human neuroblastoma cells. *RNA biology* **12**, 1323-1337 (2015).
37. Ballarino, M. et al. Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. *Molecular and cellular biology* **35**, 728-736 (2015).
38. Wilusz, J. et al. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135**, 919-932 (2008).
39. Kishore, S. et al. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biology* **14**, R45 (2013).
40. Rybak-Wolf, A. et al. A variety of dicer substrates in human and C. elegans. *Cell* **159**, 1153-1167 (2014).
41. Honda et al. Sex-hormone dependant tRNA halves enhance cell proliferation in breast and prostate cancer. *PNAS* **29**, E3816-25 (2015).
42. Chen, Q. et al. Sperm tRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science* **351**, 397-400 (2016).
43. Sharma, U. et al. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* **351**, 391-396 (2016).
44. Laurent, G. et al.. The Landscape of long noncoding RNA classification. *Trends in Genetics* **31**, 239-251 (2015).
45. Dobin, A. et al. STAR : ultrafast universal RNA-Seq aligner. *Bioinformatics*, **29**, 15-21 (2013).
46. Schulze, S. et al. SERE: Single-parameter quality control and sample comparison for RNA-Seq. *BMC Genomics* **13**, 524 (2012).
47. Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* **26**, 43-49 (1979).
48. Kruskal, J. and Liberman, M. The symmetric time-warping problem: from continuous to discrete. In Sankoff, D. (eds.), *Time Warps, String Edits, and Macromolecules: The theory and Practice of Sequence Comparison*. CSLI Publications, Stanford, pp. 125-161.
49. Rodriguez, A. and Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492-1496 (2014).
50. Cheng, Y. et al. Mean shift, mode seeking and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**, 790-799 (1995).
51. Wang, S. et al. Comment on “Clustering by fast and fins of density peaks”. arXiv:1501.04267v1 (2015).