Journal of Bioinformatics and Computational Biology
Vol. 16, No. 1 (2018) 1840001 (27 pages)
© World Scientific Publishing Europe Ltd.
DOI: 10.1142/S0219720018400012

ASSA: Fast identification of statistically significant interactions between long RNAs

Ivan Antonov[*,†,**], Andrey Marakhonov[‡,§], Maria Zamkova[¶]
and Yulia Medvedeva[*,†,||]

[*]*Institute of Bioengineering*
Federal Research Center Fundamentals of Biotechnology RAS
Moscow 117312, Russia

[†]*Department of Molecular and Biological Physics &*
Moscow Institute of Physics and Technology
Dolgoprudny, Moscow Region 141701, Russia

[‡]*Laboratory of Functional Analysis of the Genome*
Moscow Institute of Physics and Technology
Dolgoprudny, Moscow Region 141701, Russia

[§]*Federal State Scientific Budgetary Institution*
Research Centre for Medical Genetics, Moscow 115478, Russia

[¶]*Russian N.N. Blokhin Cancer Research Center*
Moscow 115478, Russia

[||]*Vavilov Institute of General Genetics*
RAS, Moscow 119333, Russia
[**]*ivan.antonov@gatech.edu*

Received 24 October 2017
Revised 21 November 2017
Accepted 5 December 2017
Published 26 January 2018

The discovery of thousands of long noncoding RNAs (lncRNAs) in mammals raises a question about their functionality. It has been shown that some of them are involved in post-transcriptional regulation of other RNAs and form inter-molecular duplexes with their targets. Sequence alignment tools have been used for transcriptome-wide prediction of RNA–RNA interactions. However, such approaches have poor prediction accuracy since they ignore RNA's secondary structure. Application of the thermodynamics-based algorithms to long transcripts is not computationally feasible on a large scale. Here, we describe a new computational pipeline ASSA that combines sequence alignment and thermodynamics-based tools for efficient prediction of RNA–RNA interactions between long transcripts. To measure the hybridization strength, the sum energy of all the putative duplexes is computed. The main novelty implemented in ASSA is the ability to quickly estimate the statistical significance of the observed interaction energies. Most of the functional hybridizations between long RNAs were classified as statistically significant. ASSA outperformed 11 other tools in terms of the Area Under the Curve on two out of four test sets. Additionally, our results emphasized a unique property of the *Alu*

[**] Corresponding author.

repeats with respect to the RNA–RNA interactions in the human transcriptome. ASSA is available at https://sourceforge.net/projects/assa/

Keywords: RNA–RNA interactions; natural antisense transcripts (NATs); long noncoding RNAs (lncRNAs); post-transcriptional regulation; hybridization energy, statistical significance.

1. Introduction

Due to the single strand nature of an RNA molecule, its nucleotides are capable of base pairing with the complementary nucleotides. Usually, the hybridization occurs between different regions of the same transcript producing the secondary structure. However, a part of one RNA molecule can bind to a complementary part of another transcript forming inter-molecular duplex. Such RNA–RNA pairing is called *antisense interaction* and the corresponding RNAs are known as natural antisense transcripts or NATs.[1]

Long noncoding RNAs (lncRNAs) are a large and diverse class of transcripts with a length of more than 200 nucleotides that do not encode proteins. The discovery of thousands of lncRNAs expressed in the mammalian cells raises a question about their functionality.[2,3] The fact that the transcription of lncRNAs is regulated,[4] indirectly supports their functionality. Due to the functional diversity,[5] the role and/or the molecular mechanism of only a few hundred lncRNAs have been determined to date. Particularly, it has been shown that some of them function post-transcriptionally via formation of inter-molecular RNA–RNA duplexes.[6–8]

Several experimental methods have recently been developed to identify inter-molecular RNA–RNA duplexes on a large scale (SPLASH,[9] PARIS,[10] LIGR-seq,[11] MARIO,[12] RIA-seq[13]). Nevertheless, due to the limited availability of the experimental data there is still a need for a computational prediction of antisense interactions. Existing thermodynamics-based tools[14–16] compute the free energy of the inter-molecular duplexes (ΔG) to estimate the strength of the RNA–RNA binding. Although these algorithms are effective in working with relatively short RNAs (such as bacterial sRNAs) on a small scale, the computational complexity does not allow to directly use them for genome- or transcriptome-wide searches. To overcome this limitation, a number of large-scale computational studies have utilized sequence alignment tools (such as BLASTn[17] or LASTAL[18]) to predict mammalian NATs.[19–22] However, these approaches do not account for RNA secondary structure which is crucial for the RNA binding.

Here we present a new computational pipeline called "AntiSense Search Approach" (ASSA) that combines sequence alignment and thermodynamics-based tools for efficient prediction of the RNA–RNA interactions between long transcripts. It reduces the running time by fast identification of the putative antisense sites using the sequence alignment tool *LASTAL*. The detected sites along with the flanking sequences form the putative duplexes. The inter-molecular hybridization energy of every duplex is calculated by the thermodynamics-based tool *RNAup* and the *SumEnergy* of all the putative duplexes between two RNAs is computed.

Clearly, the value of the *SumEnergy* depends on several factors including the transcript lengths (longer transcripts produce more putative duplexes) and GC-content (G::C base pairing is stronger than A::T). This makes it difficult to compare RNA–RNA interaction energies computed for transcript pairs with different properties. To tackle this problem, the statistical significance ("Theoretical P-value") of every *SumEnergies* value is estimated in ASSA with respect to the lengths and GC-contents of both the interacting transcripts. For this purpose, we developed a mathematical model that predicts the expected background distribution of *sumenergies* based on the properties of the input sequences. This model ensured that the interaction energies computed for random sequences were not statistically significant while most of the functional hybridizations between mammalian RNAs produced strong P-values. Moreover, sorting predictions by P-value instead of *SumEnergy* improved ASSA accuracy and allowed it to outperform other bioinformatics tools on two test sets containing random sequences.

A similar idea of combining the sequence alignment and thermodynamics-based tools has been used in several recent algorithms.[23,24] However, these approaches do not estimate the statistical significance of the interaction energies and, therefore, may be biased to predict stronger interactions between longer transcripts or RNAs with higher GC content.

2. Methods

2.1. *The ASSA pipeline development*

2.1.1. *Predicting antisense sites by the LAST package*

First, the *lastdb* tool is used to index all the queries by executing the command "`lastdb DB`" and passing all the query sequences via STDIN. Next, a search for antisense sites is performed by submitting the reversed target sequences via STDIN to "`lastal -s 0 -m 9999999 -P 1 -p MATRIX.txt -a 12 -b 6 -e 30 DB`", where `-s 0` is the search strand (0=reverse), `-m 9999999` – the maximum initial matches per query position (the restrictions are removed by using a very large threshold value); `-P 1` – number of parallel threads (the value is updated according to the ASSA launch options); `-a 12` – gap opening penalty; `-b 6` – gap extension cost; `-e 30` – threshold value for alignment scores (can be changed through an ASSA option) and `MATRIX.txt` is the custom substitution matrix adopted from Szczesniak *et al.*[22] (Supplementary Table 1).

In ASSA, we relaxed the gap open/extension penalties suggested by Szczesniak *et al.*[22] from −20/−8 to −12/−6, respectively. With these settings, complementary *Alu*-elements located in different transcripts produced a single local *LASTAL* alignment.

It should be noted that some of the *LASTAL* local alignments overlap either on one of the sequences (the query or the target) or on both of them. The latter cases are resolved by keeping the alignment with the larger score only.

2.1.2. *Comparison of thermodynamics-based tools*

We considered the following 12 thermodynamics-based tools for calculation of the hybridization energies of the putative duplexes produced by ASSA: AccessFold,[15] Bifold,[25] DuplexFold,[25] GUUGle,[26] IRIS,[27] IntaRNA-2,[28] LncTar,[16] RactIP,[29] RIsearch,[30] RNAduplex,[31] RNAPlex,[14] RNAup.[32]

To identify the best tool, we prepared a test set by generating 360 sequence pairs (simulated duplexes) of two types. These sequences were intended to represent two situations with respect to the local secondary structures that might be present at the antisense sites identified by *LASTAL*.

Duplexes of the first type ("true duplexes") did not have strong complementarity to the nearby (flanking) regions on the either side. Consequently, the RNA regions corresponding to an antisense site did not form intra-molecular interactions and were accessible for inter-molecular base pairing. To simulate this situation, we prepared sequence pairs with different lengths and percent complementarity of the antisense sites (all the sequences had GC content = 50%). Each sequence in a pair consisted of a site of length *SL* in the middle and two random flanking sequences of length *FL* each (see Fig. 1). Additionally, the complementarity between the two sites in the simulated duplex was *INTER_CMPL* percent. In this construct, the simulated site in the middle corresponded to a *LASTAL* local alignment. To simulate a variety of possible *LASTAL* hits, different values for *SL*, *FL* and *INTER_CMPL* were used. Sequence pairs with three different combinations of the site/flank lengths (*SL*/*FL* = 10/10, 20/50 or 40/50) were generated (the corresponding total sequence lengths were 30, 120 and 140, respectively). Additionally, three different percents of complementarity were used (*INTER_CMPL* = 100%, 90% or 80%). A total of 20 sequence pairs were simulated for every combination of *SL*,

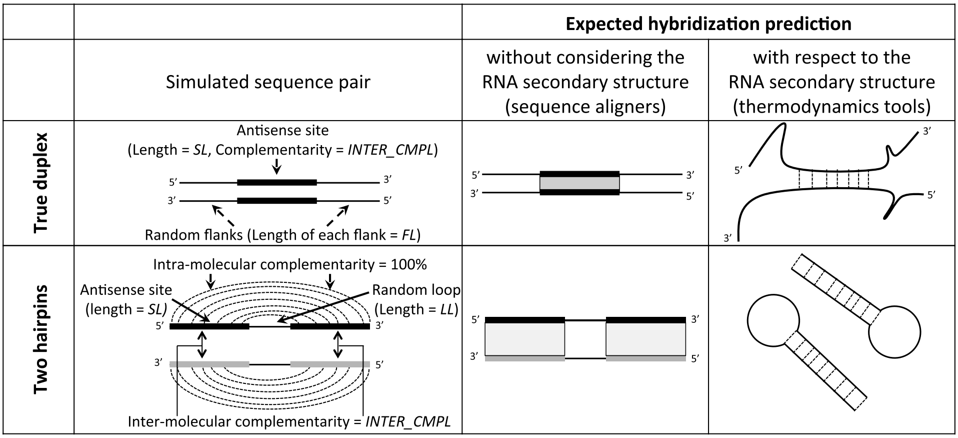

Expected hybridization prediction
Simulated sequence pair
without considering the RNA secondary structure (sequence aligners)
with respect to the RNA secondary structure (thermodynamics tools)
True duplex
Antisense site
(Length = SL, Complementarity = INTER_CMPL)
5′
3′
3′
5′
Random flanks (Length of each flank = FL)
Two hairpins
Intra-molecular complementarity = 100%
Antisense site
(length = SL)
Random loop
(Length = LL)
Inter-molecular complementarity = INTER_CMPL

Fig. 1. Two types of the simulated duplexes and the expected predictions for them produced by a sequence alignment or a thermodynamics-based tool.

FL and *INTER_CMPL* making the total number of "true duplexes" in the test set equal to 180 (= 20 × 3 × 3).

We refer to the second type of the simulated duplexes in the test set as "two hairpins". In addition to some level of inter-molecular complementarity, they also had perfect (100%) intra-molecular complementarity. It should be noted that a loop of length *LL* separated the two complementary regions located on the same sequence (see Fig. 1). We generated nine different variations of the "two hairpins" duplexes by using three combinations of the *SL*/*LL* values (10/5, 20/20 and 40/20 – the corresponding sequence lengths were 25, 60 and 100) and three different complementarity values (*INTER_CMPL* = 80%, 70% or 60%). In total, 180 "two hairpins" duplexes were simulated for the test set by preparing 20 sequence pairs for every combination of the *SL*, *LL* and *INTER_CMPL* values.

On one hand, a sequence alignment tool (such as *LASTAL*) is expected to identify complementarity between sequences of both types. This is why both classes of the simulated duplexes are likely to be among the putative duplexes analyzed by ASSA. On the other hand, due to the presence of 100% intra-molecular complementarity in the "two hairpins" sequences, a thermodynamics-based tool is expected to fold them into two separate RNA molecules. Therefore, a secondary structure aware algorithm is likely to predict strong inter-molecular interaction for the "true duplexes" only (see Fig. 1). So, our goal was to find the thermodynamics-based tool able to distinguish the "true duplexes" from the "two hairpins" types most accurately.

A "good" tool should predict existence of inter-molecular binding for the "true duplexes" (label=1) and no binding for the "two hairpins" (label=0) sequence pairs. However, thermodynamics-based tools do not produce such a binary classification. Instead, they compute inter-molecular hybridization energy (ΔG). Thus, the predicted duplex types were determined by computing tool-specific empirical P-values for all the simulated duplexes.

To do this, 100 "random" duplexes were generated from each simulated duplex by mono-nucleotide shuffling of one of the sequences (the other sequence in the pair did not change). Each tool was applied to the original as well as to 100 "random" duplexes producing 101 interaction energies. The empirical P-value was computed as follows:

$$\text{Empirical } P\text{-value}(\Delta G_{\text{original}} = x | \Delta G_{\text{rand}}, T) = \frac{\text{Num}(\Delta G_{\text{rand}} \leq x | T)}{\text{Num}(\Delta G_{\text{rand}})}, \qquad (1)$$

where x is the inter-molecular hybridization energy ($\Delta G_{\text{original}}$) outputted by the tool T for the original duplex; $\Delta G_{\text{rand}} = \{\Delta G_{\text{rand\_1}}, \ldots, \Delta G_{\text{rand\_}N}\}$ is a set of interaction energies predicted by the tool T for all the corresponding random duplexes; $\text{Num}(\Delta G_{\text{rand}} \leq x|T)$ is the number of random duplexes with the hybridization energy less or equal to x (i.e. the same or stronger interaction is observed for random sequences) and $\text{Num}(\Delta G_{\text{rand}})$ is the total number of random sequences ($N = 100$ in our settings). It should be noted that inter-molecular hybridization energy is a negative value and the smaller it is the stronger the predicted interaction.

Empirical P-value less than 0.05 indicated that a tool predicted the existence of inter-molecular binding (i.e. label $= 1$). On the other hand, a simulated duplex with empirical P-value greater or equal to 0.05 corresponded to the case where no binding was predicted (i.e. label $= 0$). By comparing the predicted labels with the known duplex types ("true duplexes" or "two hairpins"), the numbers of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) predictions were determined. Finally, the Matthews correlation coefficient (MCC) was computed for each tool:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \qquad (2)$$

2.1.3. *Interaction energy calculation by RNAup*

The following command is used in ASSA to compute the hybridization energy for each putative duplex: "`RNAup -o -b -w L`", where `-o` indicates not to produce an output file and report the predicted free energy via STDOUT, `-b` takes into account the probability of unpaired regions in both RNAs and `-w` specifies the maximum length of the region of interaction. Sequence pairs of each putative duplex are concatenated by the '&' character and passed to *RNAup* via STDIN. It should be noted that the value of the `-w` parameter has a great effect on the *RNAup* execution time (the larger the value, the longer the execution time). Our analysis of the 100 human lncRNA–mRNA pairs indicated that the majority of the identified antisense sites (i.e. LASTAL local alignments) were shorter than 50 bp (see Supplementary Fig. 1). Therefore, in ASSA, the value of the `-w` option is defined as $\min(50, \mathit{antisense\_site\_length})$). To speed up the calculation of the hybridization energies for the long (> 50) antisense sites, the *RNAup* is applied to the site sequences only (i.e. flanks are not added) with the `-w` set to 50. The produced energies are then scaled up by multiplying to the ratio "$\mathit{antisense\_site\_length}/50$".

The *RNAup* run is the most time consuming step of the pipeline. Further speed up of this step can be achieved by using the –`num_threads` ASSA option that allows to compute the free energies of all the putative duplexes independently in parallel.

Besides the interaction energies, *RNAup* also outputs the coordinates of the predicted inter-molecular base pairing. Importantly, ASSA only considers hybridization energies from the sequence pairs where the location of the *RNAup* duplex overlaps the original antisense site on both sequences. If the *LASTAL* and *RNAup* predictions do not overlap, the ΔG of the putative duplex is set to 0.

2.1.4. *Estimation of theoretical P-values*

To study the dependence between the properties of the random sequences and the distribution of *SumEnergies*, a training set was prepared. First, groups of random sequences were generated. Each group consisted of 10 sequences with the same length and GC content. Seven different lengths (50, 100, 300, 1000, 2000, 4000 and 8000 nt) and seven different GC contents (30%, 40%, 42%, 50%, 56%, 60% and 70%) were considered. Two groups ("queries" and "targets") were generated for every Length/GC content combination. In total, 49 query and 49 target groups were prepared. It should be noted that short sequences (length $= 50$ and 100 nt) were included in the training set in order to obtain a universal model that can be used to estimate the statistical significance of the interactions between sequences in a wide range of lengths.

Next, ASSA was applied to search queries versus targets. To reduce the number of ASSA runs, the input group pairs were selected so that the queries were never longer than the targets (i.e. $L1 \leq L2$). In total, 1372 ($= 7 \times 7 \times 7 \times (7+1)/2$) ASSA runs were performed and each run produced 100 *SumEnergy* values (10 queries vs 10 targets).

Initially, the RNA–RNA interaction prediction was done with the relatively low *LASTAL* score threshold equal to 30 (or 36 for long sequences with high GC content). To study the influence of this ASSA parameter on the distribution of *SumEnergies*, the produced files were post-processed by increasing the score threshold and re-calculating all the *SumEnergy* values. In total, this allowed to produce 20,532 sets of *SumEnergies* that corresponded to different sequence features as well as a wide range of *LASTAL* score thresholds (from 30 to 105). It should be noted that some random sequence pairs produced *SumEnergies* equal to 0 – e.g. when no putative duplexes were identified because the sequences were too short or the score threshold was too strict.

To select the probability distribution function (PDF) which fits the produced data better, we focused on the 3411 sets where all the 100 *SumEnergies* were negative (i.e. none of them was equal to zero). Three candidate distributions (Gamma, Normal and Log-Normal) were considered. Since the Gamma and Log-Normal PDFs are only defined for values greater than zero, their parameters were identified from the $-1 \times$ *SumEnergy* values using the `fitdistr()` function from the `MASS`[33] R library. The Kolmogorov–Smirnov test P-values were calculated for all the 3411 sets with respect to each of these three PDFs. The Gamma distribution produced the least number of small (< 0.05) KS P-values (for 0.21% of the sets only) and was selected to model the distribution of the *SumEnergies* produced by random sequences.

The Gamma distribution can be parameterized in terms of a shape parameter (α) and a rate parameter (β) – both are positive numbers. To account for the fraction of random sequence pairs that produced $\mathit{SumEnergy} = 0$, we utilized the "hurdle model",[34,35] which in our case is a mixture of Bernoulli and Gamma distributions. The basic idea is that a Bernoulli probability ($P = P(\mathit{SumEnergy} \neq 0)$) governs the binary outcome of whether or not the *SumEnergy* equals to zero. If *SumEnergy* is not zero, the hurdle is crossed, and the Gamma distribution governs the remaining (nonzero) part of the distribution. The 20,532 empirical *SumEnergy* distributions were used to train three linear regressions (see Supplementary Table 2). The obtained models are used in ASSA to predict the expected background distribution based on the features of the input sequences and the *LASTAL* score threshold for Theoretical P-value calculation (see Supplementary text for more details).

2.2. *Transcriptome-wide all-vs-all search and repeat masking*

Information about the genes expressed in the K562 cell line was taken from the FANTOM5[2] database (sample id "CNhs12334.10824-111C5"). All the genes with nonzero expression values were considered. For the alternatively spliced genes, the longest isoform was selected only.

The "all-vs-all" BLASTn[17] search was performed in the antisense mode (`-strand minus`) with the seed length equal to 15 (`-word_size 15`) and without a threshold on the E-value (`-evalue 999999`). Thus, an antisense interaction was recorded between any two transcripts which had at least one perfectly complementary duplex longer or equal to 15 bp. The number of the antisense partners for each RNA was defined as the number of unique transcripts that produced at least one BLASTn local alignment with the query.

We ran the RepeatMasker in a quick mode (the `-qq` option) to mask the human-specific repeats (`-species human`) in all the sequences. Additionally, the `-alu` option was used to restrict masking to the *Alu* repeats only.

2.3. *Test sets to compare the RNA–RNA interaction prediction tools on a large scale*

The performance of different RNA–RNA prediction tools was evaluated based on their ability to detect experimentally identified targets of the lncRNA *TINCR*[13] and mRNA *ACTB*.[9] Among thousands of *TINCR* antisense partners, we randomly selected 100 transcripts with the length between 200 nt and 4000 nt (to reduce execution time of some RNA–RNA prediction tools) and without *Alu*-repeats (in order to focus on the *short-trans* interactions). For the *ACTB*, all the 82 targets identified in the HeLa cells were taken.

To estimate the ability of the computational tools to identify the true targets among a large set of sequences, two types of test sets were prepared. The *TINCR* and *ACTB* test sets of the first type ("mix with human transcripts") consisted of all the selected true targets and a number of randomly selected human transcripts with similar length and GC content. It should be noted that the false targets were selected among the transcripts of the genes expressed in the corresponding cell types. To identify such genes in keratinocytes, (where the *TINCR* pull down has been performed) the reads from the input RNA-seq sample (run ID SRR539976 from the NCBI GEO entry GSM986009) were mapped to the human genome (hg38) and the 7314 NCBI genes with at least 100 mapped reads were identified. In case of *ACTB*, we considered the 1967 highly expressed genes with at least one interaction identified by Aw *et al.*[9] in HeLa cells. Nine "false targets" were selected for every "true target" (see Supplementary Fig. 2) to simulate the assumed situation in cell where a long RNA interacts with a limited number of other RNAs. To prepare the test sets of the second type ("mix with shuffled sequences") every selected true target transcript was used to generate nine di-nucleotide shuffled sequences using the uShuffle tool.[36]

Therefore, each of the two *TINCR* test sets consisted of 1000 sequences with 100 of them being the true targets, while the *ACTB* test sets included 82 true and 738 false targets. Each tool was used to compute the interaction energy between the query transcript (NR_027064 for *TINCR* or NM_001101 for *ACTB*) and sequence from a test set. Predictions of the tools were used to rank all the sequences in each test set and to build a ROC curve by using the ROCR[37] R-package.

3. Results

3.1. *Types of inter-molecular RNA–RNA interactions*

Antisense interactions are usually classified into two groups – *cis* and *trans*. The interactions of the first type occur between the products of the overlapping genes that are transcribed in opposite directions. The resulting RNAs have one or several (due to splicing) relatively long sites with perfect complementarity. All other interactions are classified as *trans* ("not-*cis*") since they are formed between transcripts produced from genes located in different genomic regions.

Analysis of the published cases of the biologically active duplexes formed between long RNAs (i.e. mRNA–mRNA, lncRNA–mRNA or lncRNA–lncRNA) in mammals prompted us to expand the classical "cis-trans" classification of the natural antisense transcripts. It should be noted that some RNAs produced from non-overlapping genes are also able to form long ($>$ 100 bp) highly complementary inter-molecular duplexes. To better discriminate between different classes of *trans*-interactions, we divided the *trans*-category into three sub-categories – *pseudo-cis*, *Alu*-based and *short-trans* interactinos.

The "*pseudo-cis*" sub-type of the *trans* RNA–RNA binding occurs when one of the overlapping genes has an expressed copy (a paralog) at another genomic locus. The gene copy harbors a sequence highly complementary to a part of the other gene in the original overlap and, thus, can form *trans*-antisense duplexes with it. This scenario has been observed in the case of expressed pseudogenes.[38] Moreover, it has been shown that such pseudogene related duplexes can be recognized by RNAi machinery and produce functional siRNAs in mouse oocytes.[39] Inversion of a genomic region during gene duplication is another possible scenario for such NATs formation.[40]

Sequence repeats of several classes occupy a significant portion of the human genome. Up to 350 bp long with relatively high percent identity ($>$ 70%) *Alu*-repeats belong to the short interspersed nuclear elements (SINEs) class. They are frequently present in lncRNAs or in the 3'UTRs of human mRNAs either in the direct or in the reverse-complement orientation. It has been shown that a pair of transcripts with *Alu* repeats in opposite directions are able to interact with each other and trigger Staufen Mediated Decay (SMD)[41,42] and/or regulate mRNA translation.[43]

To evaluate the fraction of the *Alu*-based interactions on a transcriptome scale, "all-vs-all" BLASTn search was performed for the 10,664 genes expressed in the

K562 cell line (see Methods). We observed linear dependance of the number of predicted antisense partners on the query transcript length. Surprisingly, the analyzed RNAs formed three distinct clusters (see Fig. 2(a)). To check whether the origin of these clusters was related to the *Alu* repeats, we applied the RepeatMasker software[44] to the the 10,664 sequences and identified 2212 *Alu*-containing transcripts. Next, all the RNAs were classified into three categories according to the presence and direction of *Alu* repeats that matched well with the three clusters

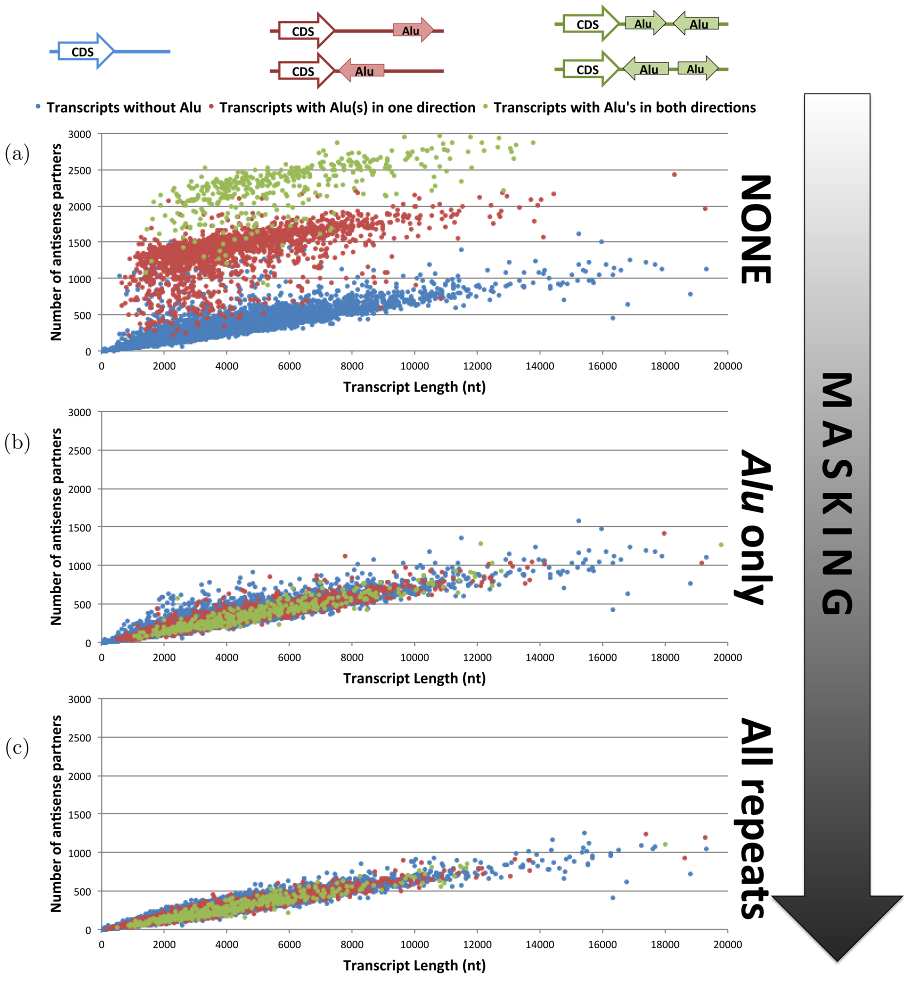


Fig. 2. *Alu*-based hybridization is the main type of the repeat-based antisense interactions between human transcripts. Dependence between the query transcript length and the number of antisense partners is shown for three types of masking: (a) no masking, (b) masking of the *Alu*-repeats only and (c) masking of all the repeats (as identified by the RepeatMasker). Each dot on the graph corresponds to one transcript used as search query. The 8452 transcripts without *Alu* repeats are blue; the 1807 transcripts with all *Alu*(s) in one orientation are red; the 405 transcripts with at least one *Alu* repeat in the direct orientation and at least one *Alu* repeat in the reverse orientation are green.

(see Fig. 2(a)). Indeed, the transcripts with two or more *Alu*-repeats in different directions are able to form inter-molecular duplexes with any other *Alu*-containing transcript. RNAs with *Alu*(s) in one orientation can only hybridize with the transcripts containing complementary repeat sequence. Clearly, transcripts without *Alu*'s have the lowest antisense potential since they can not participate in the *Alu*-based interactions.

To further confirm the observed role of the *Alu* repeats, we masked them in all the 2212 sequences and repeated the "all-vs-all" search (for the 10,664 transcripts). With this modification, all the RNAs followed the same trend of increasing the number of possible antisense partners with the sequence length (Fig. 2(b)). This indicated that the two transcript clusters with the higher number of RNA–RNA interactions (the red and green dots in Fig. 2(a)) appeared due to the presence of *Alu* repeats in the corresponding sequences. Furthermore, we masked repeats of all types (7.8% of the total sequence length) and performed the "all-vs-all" search once again. No significant change was observed this time (compare (B) and (C) in Fig. 2) suggesting that the ability to significantly increase the number of possible antisense targets of a transcript is the unique feature of the *Alu* elements. Thus, we concluded that *Alu*-based base pairing plays the major role in the repeat-associated interactions in the human transcriptome.

It should be noted that a number of large scale computational studies[19–21] have masked repeats of all types in the transcript sequences prior to the antisense search to avoid the prediction of the large number of repeat-based interactions. Our results indicated that the same effect can be achieved by masking the *Alu* repeats only while preserving other potentially informative parts of the sequences. This approach was used in the present study where applicable.

Finally, apparent long sites with high complementarity have not been found in a number of *trans*-antisense interactions.[8,45] This makes it difficult to identify the exact regions of inter-molecular hybridization between the corresponding RNAs. The authors of the corresponding papers have hypothesized that in such cases the observed bindings may be based on several relatively short and not-perfectly complementary duplexes. We refer to this sub-type of *trans* binding as "*short-trans*-interactions".

3.2. *The ASSA algorithm*

Among the four introduced types of inter-molecular RNA–RNA hybridization, three (*cis*, *pseudo-cis* and *Alu*-based) are based on relatively long, highly complementary duplexes. Thus, even sequence alignment algorithms that do not take secondary structure into account can detect interactions of these types.[19,46,47]

However, the *short-trans* category is different. The presence of multiple small duplexes distributed along the sequences makes it difficult to localize the exact interacting regions both experimentally and computationally. In addition, a number of spurious short sites with imperfect complementarity can be predicted by the alignment tools.

On the other hand, thermodynamics algorithms discriminate between RNA regions that are parts of stable secondary structures (i.e. involved in intra-molecular interactions) and the unpaired sites accessible for inter-molecular base pairing. Thus, the computed free energies reflect both the accessibility and the length/complementarity of the specific transcript regions. This is why interaction energy calculation is expected to improve the prediction accuracy of the *short-trans* interactions.[23,24] It is unpractical to apply the traditional thermodynamics tools to the long transcripts (such as mammalian mRNAs or lncRNAs) because the execution time grows quickly with the lengths of the input sequences.[14,15,28] At the same time, recent experimental data[9–13] have indicated that the *short-trans* base-pairing may be the most abundant interaction type in the human transcriptome.

Thus, the two main challenges in predicting *short-trans* hybridizations between long (>200 nt) RNAs on a large scale are (i) the execution time of existing thermodynamics-based algorithms and (ii) the identification of the statistically significant interactions among all the transcript pairs. Here we present a new computational pipeline, called ASSA ("AntiSense Search Approach"), developed in attempt to address both of these problems. It should be noted that ASSA is able to identify interactions of all types, but our main goal was to make a progress in the most challenging direction–prediction of the *short-trans* binding.

Briefly, ASSA performs the following main steps (see Fig. 3) to predict interaction between each pair of input transcripts: (i) identify antisense sites by the local sequence alignment algorithm *LASTAL*[18]; (ii) extract sequence regions corresponding to the predicted antisense sites and compute hybridization energies of the putative duplexes; (iii) compute the RNA–RNA interaction energy by summing the hybridization energies of all the putative duplexes and (iv) estimate the statistical significance ("theoretical *p*-value") of the obtained *SumEnergy*. These steps are discussed in more details below.

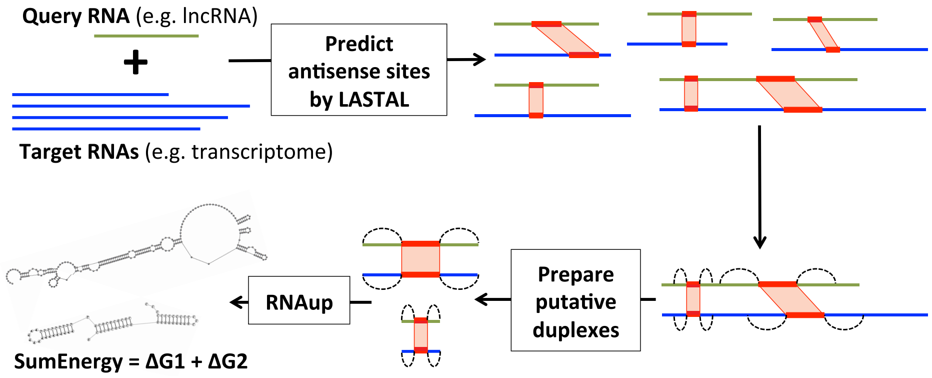

Query RNA (e.g. lncRNA)
+
Target RNAs (e.g. transcriptome)
Predict antisense sites by LASTAL
Prepare putative duplexes
RNAup
SumEnergy = ΔG1 + ΔG2

Fig. 3. The main steps of the ASSA pipeline.

3.2.1. *Predicting antisense sites by LASTAL*

ASSA takes two sets of nucleotide sequences as input (the query and the target transcripts). On the first step, the local sequence alignment tools from the LAST package[18] are used to identify the regions of local complementarity for each sequence pair by searching all the queries versus all the targets (see Methods). In this work, we refer to each produced local alignment as the "*antisense site*". This approach is adopted from another large scale study of the RNA–RNA bindings between human transcripts.[22] The advantage of using *LASTAL* over other aligners (such as BLASTn[17]) for predicting antisense interactions is that it allows to use a custom substitution matrix with appropriate scores for the RNA-specific hybridization rules (i.e. G:C, A:U and G:U base pairs-see Methods).

ASSA antisense sites are selected based on the local alignment scores rather than the alignment length and/or percent complementarity (a similar approach is used in the RIblast algorithm[24]). The threshold on the *LASTAL* alignment scores is one of the pipeline parameters.

On the next ASSA step, a thermodynamics-based algorithm is used to compute the hybridization energies of the "putative duplexes" which are defined as RNA regions consisting of the antisense sites together with the flanking sequences on both sides. Several thermodynamics-based algorithms can be used for this purpose. To choose the tool which suits the ASSA pipeline best, we compared their performances on a simulated dataset.

3.2.2. *Choosing a thermodynamics-based tool to compute hybridization energies*

We considered 12 thermodynamics-based tools as candidates to be used in the ASSA pipeline for calculation of hybridization energies of the putative duplexes. The performances of these algorithms were compared on a test set consisting of 360 short sequence pairs that resembled the putative duplexes in ASSA. Both sequences in a pair had the same length ranging from 25 to 140 nt.

There were two types of simulated duplexes in the test set. The first one ("true duplexes") corresponded to the cases of true inter-molecular hybridizations that were not interfered by the local secondary structures. In these duplexes, the middle part of each sequence represented a putative antisense site (a gapless *LASTAL* local alignment) of a particular length and percent complementarity flanked by the random sequences on both sides. The second type of the simulated duplexes ("two hairpins") represented the situation where the regions of both RNA molecules corresponding to an antisense site interact with the flanking regions to form local secondary structures. Importantly, the percent of complementarity of intra-molecular hybridization was greater than that of the inter-molecular binding (see Methods for details). The test set consisted of 180 simulated duplexes of each type.

Each tool was used to predict the type of every duplex. The prediction accuracy of an algorithm was determined by comparing the predicted and the true duplex types. According to the obtained Matthews correlation coefficients, *RNAup*[32] produced the most accurate labeling of the simulated duplexes (see Supplementary Fig. 3). Thus, this tool was incorporated in the ASSA pipeline.

3.2.3. *Preparing putative duplexes and calculating SumEnergies*

Execution time of the thermodynamics-based algorithms does not allow to directly use them for analysis of long transcripts on a large scale. The interaction energy is efficiently computed in ASSA by applying *RNAup* to the specific sequence chunks ("putative duplexes") rather than the full-length transcripts. A putative duplex is generated by extracting regions of two transcripts that correspond to an antisense site together with the flanking sequences on both sides (see Methods).

Adding flanks to an antisense site allows *RNAup* to compute the inter-molecular hybridization energy with respect to the local RNA secondary structure. The flank length influences both the accuracy and the execution time of the ASSA pipeline. On one hand, longer flanks allow to take more elements of the RNA secondary structure into account. On the other hand, the *RNAup* takes more time to process longer sequences. By default, the length of each flank is equal to the length of the corresponding *LASTAL* alignment. This approach can be considered as a tradeoff between the time and accuracy.

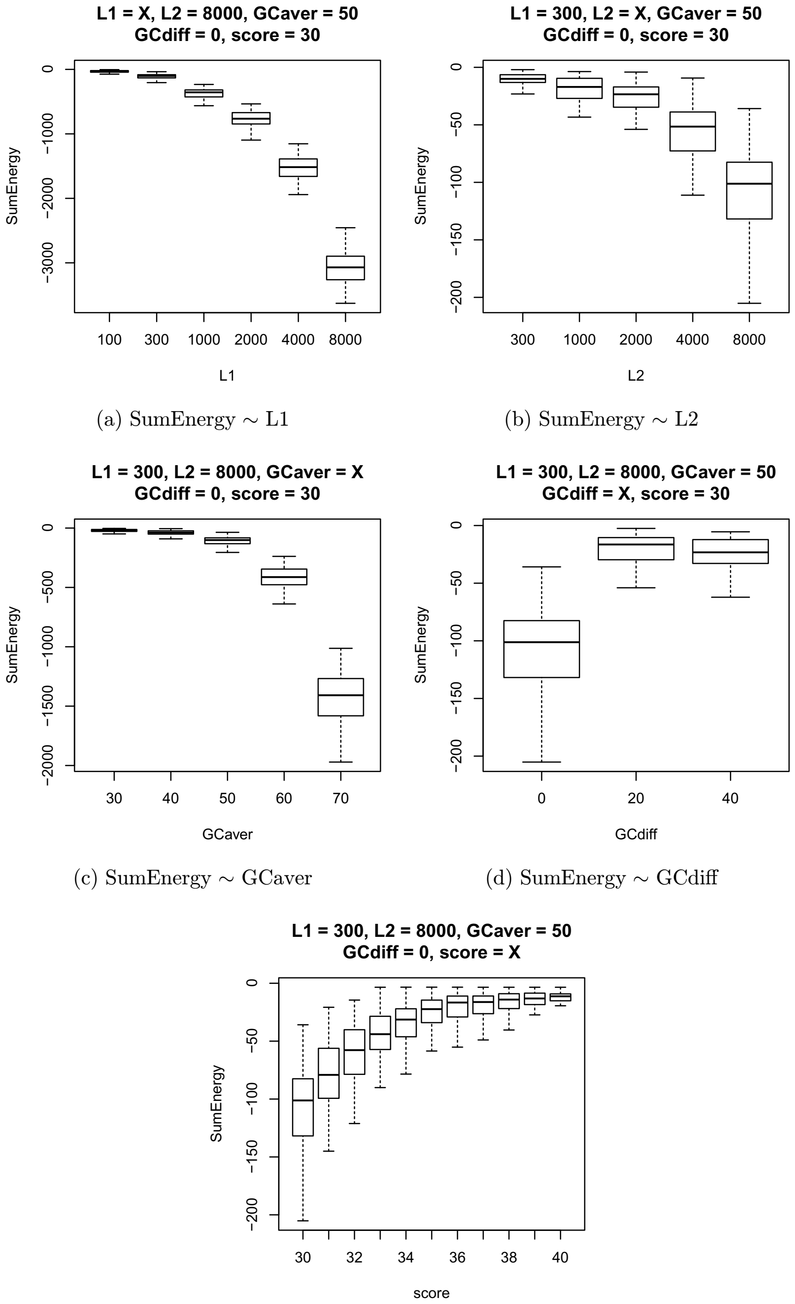
In ASSA, the interaction strength between two transcripts is measured by the sum of the hybridization energies of all the putative duplexes (*SumEnergy*). It has been shown that *SumEnergy* outperforms *MinEnergy* in predicting binding between human transcripts.[23] It should be noted that both the duplex ΔG and the RNA–RNA *SumEnergy* are negative values and the smaller they are the stronger the corresponding hybridization will be.

Notably, any RNA–RNA prediction tool (including ASSA) outputs some value of the interaction energy even for random sequences. We observed that the distribution of the *SumEnergy* values computed by ASSA depended on the features of random sequences (lengths and GC contents) as well as on the *LASTAL* score threshold (see Fig. 4). We were interested in identifying the mammalian transcript pairs with the *SumEnergy* values smaller (stronger interactions) than the ones produced by random sequences with the same lengths and GC contents. Thus, on the next ASSA step, the statistical significance (P-value) of each *SumEnergy* is estimated by comparing the observed value with the distribution produced by the corresponding random sequences.

3.2.4. *Estimating the statistical significance of the SumEnergy*

The ability to quickly estimate the statistical significance ("Theoretical P-value") of the interaction energies taking into account the lengths and GC contents of the corresponding transcripts is the main novelty of the ASSA pipeline.

Any p-value is computed by comparing the *observed value* with the distribution of the values generated by the "null" model (the *background distribution*). In ASSA, the observed value is the *SumEnergy* calculated for a particular transcript pair. Random



(a) SumEnergy ∼ L1

(b) SumEnergy ∼ L2

(c) SumEnergy ∼ GCaver

(d) SumEnergy ∼ GCdiff

(e) SumEnergy ∼ score

Fig. 4. Dependence of the SumEnergy distribution on (a–d) the features of the input sequences (L1, L2, GCaver, GCdiff) and (e) the LASTAL score threshold. Each distribution is based on the SumEnergy values obtained for 100 pairs of random sequences with the parameters indicated in the chart titles. The values of the parameter indicted as "X" are on the *X*-axis of the corresponding chart.

sequences are frequently used in bioinformatics as the "null" model. Thus, a background distribution can be generated by applying ASSA with the same parameters to a number of random sequences generated by shuffling the nucleotides in the original transcripts. The "Empirical P-value" can then be directly calculated from the obtained empirical background distribution (see Eq. (1) in Methods).

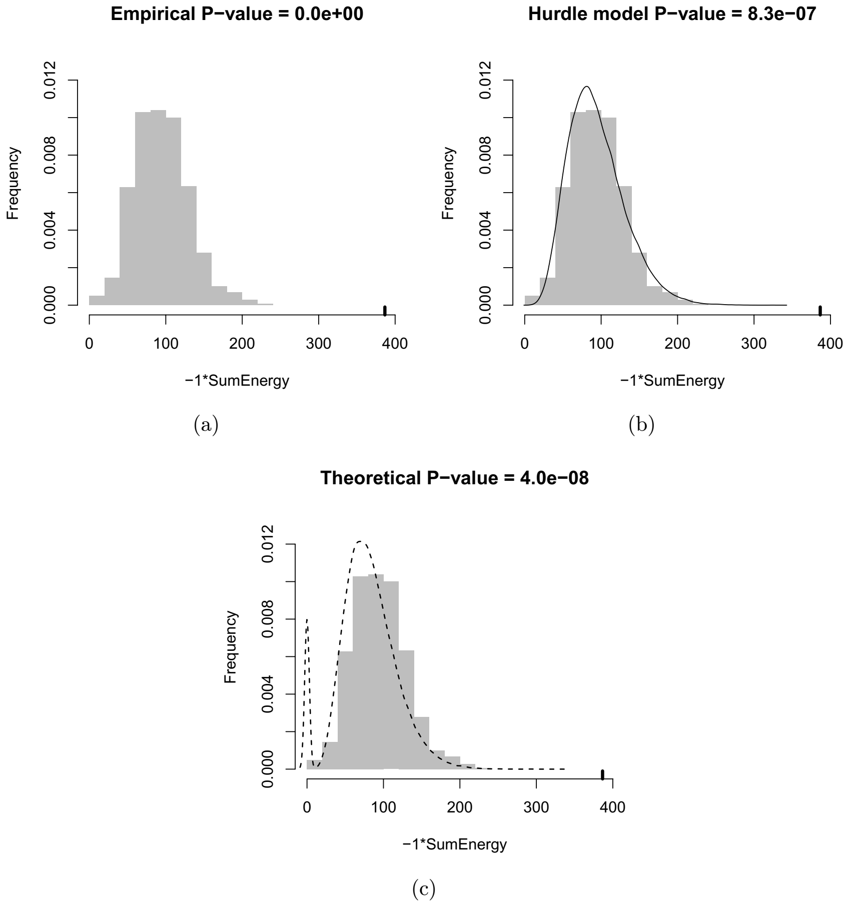
The problem with this approach is that it requires a lot of simulations to estimate small p-values. For example, the *SumEnergy* computed by ASSA for *lncRNA-ATB* and *IL11*[48] was $-386.59\,\text{kcal/mole}$. To estimate the statistical significance of this value each sequence was mono-nucleotide shuffled 30 times and all-versus-all ASSA search was performed producing an empirical background distribution consisting of 900 *SumEnergy* values (Fig. 5(a)). The smallest value observed for random sequences was $-220.56\,\text{kcal/mole}$. Thus, with this number of simulations, the Empirical P-value was equal to zero which meant that the actual P-value was less than $1/900 = 0.0011$ and a larger background dataset was needed to get a better estimate.

A common approach to obtain estimates for small P-values without generating enormous amount of random sequences is to approximate the empirical background distribution with a function and use it to compute the statistical significance of the observed value. The *SumEnergy* distribution produced by ASSA for random sequences can be approximated by a mixture of the Bernoulli and Gamma distributions (the hurdle model[34,35]) — see Methods. We applied this approach to the interaction between *lncRNA-ATB* and *IL11* and obtained the "Empirical hurdle model P-value" $= 8.3 \times 10^{-7}$ (see Fig. 5(b)).

As was mentioned above, the distribution of the *SumEnergy* values depends on several factors – the transcript lengths (longer transcripts produce more putative duplexes) and GC-contents (G::C base-pairing is stronger than the A::T) as well as the *LASTAL* search threshold (Fig. 4). Thus, sequence-specific background distributions should be generated for calculation of the Empirical P-values (with or without the use of the hurdle model). It is computationally challenging to use this measure for estimating statistical significance of the numerous *SumEnergy* values obtained in large-scale searches.

To tackle this problem, we analyzed the dependence between the distribution parameters and the features (length and GC content) of the random sequences as well as the *LASTAL* score threshold. In short, ASSA was applied to a number of random sequences with various lengths and GC-contents. Three parameters of the hurdle model (P, mean and variance) were computed for each distribution of the *SumEnergy* values. The generated dataset was used to train three linear regression models to predict each of the hurdle model parameters based on five features (log ($L1$), log($L2$)), *GCaver*, *GCdiff* and *score*–see Methods for more details). The derived formulas are used to predict the expected background distribution and compute the "Theoretical hurdle model P-value" (or simply "Theoretical P-value") for every transcript pair analyzed by ASSA (Fig. 5(c)).

The goal of the developed mathematical model is to quickly and reliably estimate the statistical significance of an observed *SumEnergy* value. This implies that the



(a)

(b)

(c)

Fig. 5. Three types of P-values computed for the short-trans interaction between the *lncRNA-ATB* (ENST00000493038) and *IL11* (NM_000641.3). With the LASTAL score threshold set to 36, the SumEnergy outputted by ASSA was −386.59 kcal/mole (indicated by the thick bar above each X-axis). Note that distributions of the $-1 \times \mathit{SumEnergy}$ are plotted on the graphs. (a) The "Empirical P-value" was calculated based on the distribution of the SumEnergies obtained for 900 random sequence pairs generated by mono-nucleotide shuffling of the original transcripts. (b) The MLE-estimates of the hurdle model parameters (P = 1.0, shape = 6.467, rate = 0.0686) were obtained from the empirical distribution and the "Empirical hurdle model P-value" was computed. (c) The "Theoretical P-value" was also computed using the hurdle model, but the parameter values (P = 0.94, shape = 6.567, rate = 0.0786) were predicted based on the sequence features (L1 = 2381 nt, L2 = 2446 nt, GCaver = 50.28%, GCdiff = 12%) without using the empirical distribution.

Theoretical and the Empirical P-values should be similar for the same transcript pairs. To check the correspondence between these two measures 10 lncRNAs and 10 mRNAs without *Alu* repeats and with the lengths between 200 and 4000 nt were randomly selected from the human transcriptome. Both the Theoretical and the Empirical hurdle model P-values were computed for all the 100 lncRNA–mRNA pairs using two different *LASTAL* score thresholds. In both cases, the Pearson correlation coefficient between the log's of the P-values was greater than 90% (Supplementary Fig. 4) indicating that the Theoretical P-values can be considered as a reasonable approximation of the Empirical hurdle model P-values.

It should be noted that ASSA output also includes the FDR-corrected Theoretical P-values to account for the database size and multiple testing.

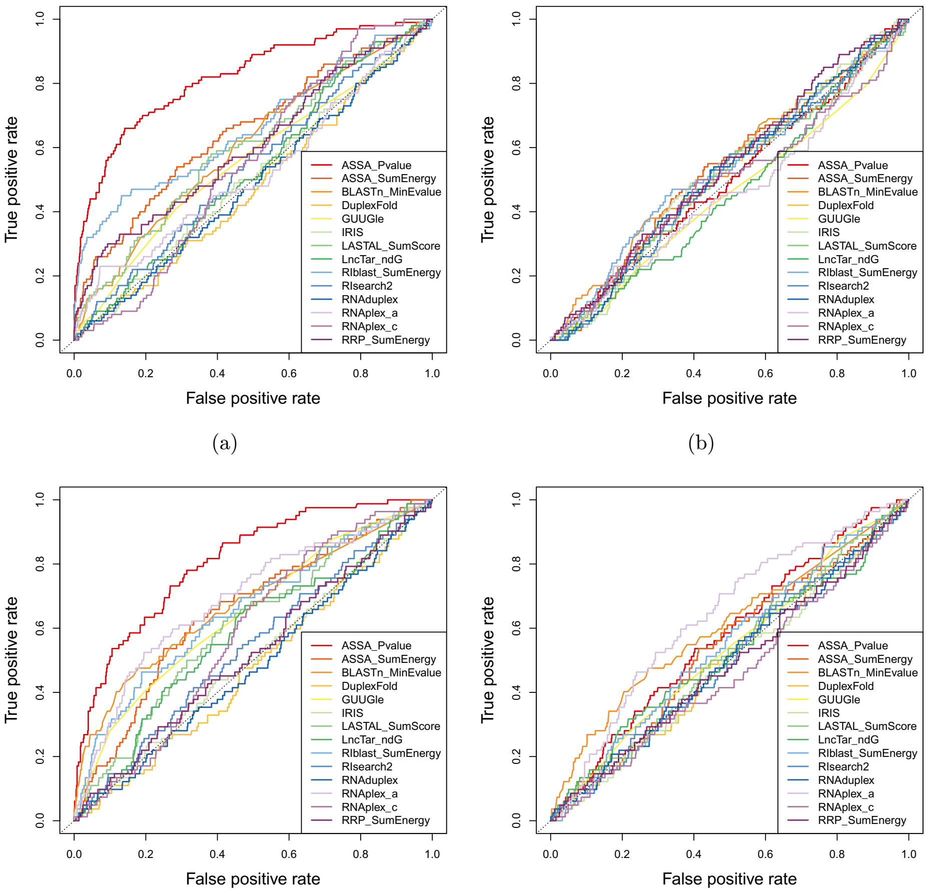
3.2.5. *Choosing the default value for the LASTAL score threshold*

The influence of the $LASTAL$ score threshold on the ASSA performance was evaluated on four validation sets, not overlapping with the test sets. The score values between 30 and 50 were considered. It should be noted that ASSA execution time also depends on the value of this threshold as more putative duplexes are predicted by $LASTAL$ when a weak threshold (a small score) is used. Our analysis demonstrated that the score threshold 36 produced one of the best average AUC values and reduced the ASSA execution time (see Supplementary Fig. 5). Therefore, this value was selected as the default in ASSA and it is used throughout this work (if not stated otherwise).

3.3. *Properties of ASSA P-values computed for random sequences*

One important property of every P-value is that it is uniformly distributed between 0 and 1 when computed for the objects from the "null-model".[49,50] In case of ASSA, the "null-model" is a pair of random sequences. To check whether this property holds for Theoretical P-values, the 10 lncRNAs and 10 mRNAs were mono-nucleotide shuffled and the Theoretical and the Empirical hurdle model P-values were computed for the 100 random sequence pairs. As anticipated at the level of p-value < 0.05, four random sequences pairs out of 100 had Empirical P-value < 0.05. Moreover, the Kolmogorov-Smirnov test for the uniform distribution applied to the 100 Empirical observations produced P-value 0.065 confirming that the distribution of this measure is indeed close to uniform (see Supplementary Fig. 6(a)). On the other hand, the obtained 100 Theoretical P-values were not exactly uniformly distributed (there were 10 sequence pairs with the Theoretical P-value < 0.05 and the corresponding Kolmogorov–Smirnov P-value was 0.032 — see Supplementary Fig. 6(b)). Still, the correlation between the Theoretical and the Empirical hurdle model P-values was 83% (Supplementary Fig. 6(c)). This analysis demonstrated that ASSA have a tendency to predict P-values that are slightly stronger than the corresponding true estimates. This may be explained by the fact that the Maximum Likelihood approach (used for Empirical hurdle models) provides better estimates of the Gamma distribution parameters than the method of moments used in ASSA. Nevertheless, given the high correlation with the Empirical hurdle model P-values (Supplementary Fig. 4 and Supplementary Fig. 6(c)), Theoretical P-values may still be useful to identify the statistically significant RNA–RNA interactions in a large scale search.

As mentioned above, one of the problems with the interaction energy is that it depends on the lengths and GC contents of the input sequences (Fig. 4). Theoretical P-values computed by ASSA not only estimate the statistical significance of the



(a)

(b)

(c)

(d)

Fig. 6. Performance of different RNA–RNA prediction tools on four test sets.

observed *SumEnergies*, but also provide automatic normalization to the features of the input sequences. Indeed, every P-value is computed with respect to the background distribution that takes into account the lengths and the GC contents of both transcripts in a pair. To check whether Theoretical P-values are indeed normalized to the sequence properties, two types of random sequence sets were generated. Each set consisted of 100 random sequence pairs. The four sets of the first type had different lengths (from 1000 to 4000 nt), but the GC content of all the sequences was the same (50%). On the other hand, the sets of the second type had different GC contents (from 35% to 65%), but the length was the same (3000 nt). Theoretical P-values were computed for all sequence pairs. There was almost no dependance between ASSA P-values and the features of the random sequences (see Supplementary Figs. 7(a) and 7(b)). It should be noted that the *SumEnergies* computed by RIblast[24] for the same sequences demonstrated strong dependence on both the length and the GC content (see Supplementary Fig. 7(c) and 7(d)). For example, the median RIblast *SumEnergies* for 3kb sequences with GC contents 35% and 65% were −67 and −694 kcal/mole, respectively (the corresponding ASSA P-values were 0.479 and 0.703).

Overall, the results obtained for random sequences demonstrated that the Theoretical P-values are highly correlated with the Empirical P-values and provide automatic normalization to the sequence properties. The main advantage of this measure is the ability to compute it quickly without generating random sequences. Thus, ASSA P-values can be considered as the first approximation of the statistical significance of the RNA–RNA interaction energy.

3.4. *ASSA predictions for functional NATs*

In order to evaluate ASSA performance on real RNA sequences, we applied it to the 34 mammalian natural antisense transcripts (NATs) collected from the literature – 11 *cis*, 4 *pseudo-cis*, 10 *Alu*-based and nine *short-trans* cases (see Supplementary Table 3). ASSA P-values computed for *cis*, *pseudo-cis* and *Alu*-based interactions were very strong due to the existence of long (>100 bp) duplexes with high percent of complementarity.

By contrast, only five out of nine *short-trans* interactions had Theoretical P-values < 0.05. The inability of ASSA to classify some of the *short-trans* cases as statistically significant could be an artifact of the method of moments used in ASSA to predict the parameters of the background distribution. To find out whether the inter-molecular hybridizations between the corresponding transcript pairs are statistically significant, the same distribution parameters were obtained by the Maximum Likelihood Estimation approach and the Empirical hurdle model P-values were obtained from the *SumEnergies* computed by ASSA or RIblast for the shuffled versions of the same sequences. In all four cases, ASSA Empirical P-values agreed with the Theoretical estimates (the Empirical P-value 0.043 is assumed to be not statistically significant – see Supplementary Figs. 8(a)–8(d). Once again, this result confirmed the good correspondence between the Theoretical and the Empirical ASSA P-values. Interestingly, two out of the four cases were statistically significant according to the interaction energies computed by RIblast (see Supplementary Figs. 8(e)–8(h). This indicates that ASSA may not be sensitive enough to detect some of the *short-trans* interactions. Additional analysis is needed to pinpoint the parts of the pipeline that should be improved to handle such cases.

According to the obtained results, Theoretical P-values computed by ASSA are suitable to predict all types of antisense interactions between long RNAs. However, the identification of the *short-trans* interactions remains the most challenging task.

3.5. *Comparison of ASSA with other tools*

We compared the ability of ASSA to identify *short-trans* RNA–RNA interactions on a large scale with the following tools that were used for this purpose in other studies – BLASTn,[17] DuplexFold,[25] GUUGle,[26] IRIS,[27] LASTAL,[18] LncTar,[16] RIBlast,[24] RIsearch2,[51] RNAduplex,[31] RNAPlex[14] and RRP[23]. To evaluate the ability of the tools to predict the *short-trans* interactions, we used the experimentally identified targets of the lncRNA *TINCR*[13] and mRNA *ACTB*[9]. It has been shown that at least some of the identified transcripts are associated with these RNAs through direct *short-trans* antisense duplexes.

To exclude the possibility of *Alu*-based interactions with *TINCR* (that has an *Alu* repeat), the *Alu*-free transcripts were used for the *TINCR* test sets. The *ACTB* mRNA does not have *Alu* repeats. In total, four test sets were prepared (see Methods). Each tool was used to rank the sequences from every test set according to the predicted hybridization strength with the query (*TINCR* or *ACTB*). The ASSA output sorted by the Theoretical P-values outperformed all the tools in terms of the Area Under the Curve (AUC) and partial AUC (pAUC) on both test sets of the "mix with shuffled sequences" type (see Figs. 6(a) and 6(c), Supplementary Figs. 9 and 10 and Supplementary Table 4). The AUC values produced by ASSA were above 80% which improved the performance of the second best tool by more than 10%.

The performance of all the tools decreased on the two test sets of the "mix with human transcripts" type (see Figs. 6(b) and 6(d). For the *TINCR* and *ACTB* interactions, the most accurate algorithms produced AUC values around 54% (BLASTn and RIblast) and 63.6% (RNAplex-a), respectively.

Our analysis demonstrated that ASSA was able to accurately identify true interactions of long RNAs (*TINCR* or *ACTB*) in the mix of human transcripts with shuffled sequences. At the same time, other tools performed better on the test sets consisting of human transcripts only but the produced accuracies were relatively low. Thus, there is a room for further improvement of the RNA–RNA prediction tools.

4. Discussion

The goal of our work was to improve the accuracy and the speed of prediction of inter-molecular interactions between long transcripts (i.e. lncRNAs and mRNAs). For this purpose, we developed a new computational pipeline ASSA. To speed up the time-consuming traditional thermodymanics tools, we obtained a set of local alignments (by the sequence aligner *LASTAL*) that allowed to restrict the calculation of the interaction energies (by the *RNAup* algorithm) to a limited number of relatively short parts of the input transcripts. The main novelty implemented in ASSA was a mathematical model that allowed to quickly estimate the background distribution and compute the statistical significance (Theoretical P-value) of the observed RNA–RNA hybridization energy. Sorting the predicted interactions by the P-value rather than the *SumEnergy* allowed ASSA to outperform other tools on two out of the four test sets. It should be noted that ASSA is one of the fastest tools that takes RNA secondary structure into account. This makes it a good candidate to perform transcriptome-wide searches.

Roughly speaking, ASSA can be viewed as a three stage algorithm — the *LASTAL* and *RNAup* runs followed by the calculation of *P*-values. The prediction accuracy is improved gradually in the pipeline since every step takes additional piece of information into account. First, the interactions are predicted by *LASTAL* without considering RNA secondary structures and estimating the statistical significance. This step is similar to the approach used by Szczesniak *et al.*[22] and the "*LASTAL* (SumScore)" performance provides an estimate of its prediction accuracy (the average AUC value over the four test sets was 55.7% — see Supplementary Table 4). Next, *RNAup* takes secondary structure into account by computing intermolecular hybridization energy. In terms of ROC statistics, this improves the accuracy by 3.5% (the average AUC value of the "ASSA (SumScore)" approach was 59.3%). Notably, at this step, ASSA is similar to the RIblast algorithm which produces similar average AUC value (60%). Finally, calculation of *P*-values with respect to the sequence features allowed to sort all the predictions in the most accurate way providing an additional 8.4% increase in the accuracy which made the average AUC value produced by ASSA equal to 67.7%.

In our study, we also suggested an improvement to the classical *cis*/*trans* classification of the antisense interactions. Based on the origin of the regions involved in inter-molecular hybridizations, three subtypes of the *trans*-interactions were introduced: *pseudo-cis*, *Alu*-based and *short-trans* interactions. Importantly, we demonstrated that among all types of sequence repeats in the human genome, *Alu* repeats have a striking influence on the ability of a transcript to base pair with other RNAs and, therefore, participate in post-transcriptional gene regulation.[41,42]

Inter-molecular RNA–RNA hybridizations may form yet another layer in the gene regulatory network. It should be noted that the bioinformatics prediction of a RNA–RNA interaction is not sufficient to make conclusions about its functionality or even realization in the cell. There are other factors that should be taken into account, including the cellular localization of the RNAs as well as the presence of specific RNA binding proteins. Probably, these reasons contributed to the poor performance of all the tools on the "mix with human transcripts" test sets. Thus, the search for new biologically active NATs is a more complex task than prediction of antisense partners. Nevertheless, we believe that further improvement of the RNA–RNA interaction detection methods (both computational and experimental) is a necessary step in this direction.

5. Conclusions

- A new computational pipeline ASSA was developed for identification of intermolecular hybridizations between long RNAs.
- ASSA provides statistical significance estimate for every predicted RNA–RNA interaction computed by a custom mathematical model.
- A special role of the *Alu*-based interactions in the human transcriptome was emphasized.

Acknowledgments

We would like to thank Mikhail Skoblov, Alexandra Filatova, Victoriya Serzhanova and Nikolay Zernov for useful discussions about the biological roles of the LNC RNAs; Andrey Leontovich for suggestions regarding aspects of the statistical analysis; Mark Borodovsky for the help in obtaining the RRP tool; Yuriy Shkandybin for the technical support; Slava Batyrev and Barbara Arellano for the help with English grammar.

The classification of NATs and the development of the beta-version of the pipeline was supported by the Dynasty Foundation [Fellowship DP-B-26/14 to I.A.]; the development of the regression model and the estimation of the statistical significance as well as comparison of ASSA with the existing tools was supported by the Russian Science Foundation [Grant No. 14-15-30002 to Y.M.].

References

1. Faghihi MA, Wahlestedt C, Regulatory roles of natural antisense transcripts, *Nat Rev Mol Cell Biol* **10**(9):637–643, 2009. doi: 10.1038/nrm2738, http://www.ncbi.nlm.nih.gov/pubmed/19638999.
2. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassman T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jorgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple C, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JA, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drablos F, Edge AS, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furino M, Furusawa J, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiecki L *et al.*, A promoter-level mammalian expression atlas, *Nature* **507**(7493):462–470, 2014. doi: 10.1038/nature13182, http://www.ncbi.nlm.nih.gov/pubmed/24670764.
3. ENCODE Project Consortium, An integrated encyclopedia of dna elements in the human genome, *Nature* **489**(7414):57–74, 2012. doi: 10.1038/nature11247.
4. Alam T, Medvedeva YA, Jia H, Brown JB, Lipovich L, Bajic VB, Promoter analysis reveals globally differential regulation of human long non-coding rna and protein-coding genes, *PLoS One* **9**(10):e109443, 2014. doi: 10.1371/journal.pone.0109443, http://www.ncbi.nlm.nih.gov/pubmed/25275320.
5. Wang KC, Chang HY, Molecular mechanisms of long noncoding rnas, *Mol Cell* **43**(6):904–914, 2011. doi: 10.1016/j.molcel.2011.08.018, http://www.ncbi.nlm.nih.gov/pubmed/21925379.
6. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent r G, Kenny PJ, Wahlestedt C, Expression of a noncoding rna is elevated in alzheimer's disease and drives rapid feed-forward regulation of beta-secretase, *Nat Med* **14**(7):723–730, 2008. doi: nm1784 [pii] 10.1038/nm1784, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db =PubMed&dopt=Citation&list_uids=18587408.

7. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, Forrest AR, Carninci P, Biffo S, Stupka E, Gustincich S, Long non-coding antisense rna controls uchl1 translation through an embedded sineb2 repeat, *Nature* **491**(7424):454–457, 2012. doi: 10.1038/nature11508, http://www.ncbi.nlm.nih.gov/pubmed/23064229.
8. Xiao L, Rao JN, Cao S, Liu L, Chung HK, Zhang Y, Zhang J, Liu Y, Gorospe M, Wang JY, Long noncoding rna spry4-it1 regulates intestinal epithelial barrier function by modulating the expression levels of tight junction proteins, *Mol Biol Cell* **27**(4):617–626, 2016. doi: 10.1091/mbc.E15-10-0703.
9. Aw JGA, Shen Y, Wilm A, Sun M, Lim XN, Boon KL, Tapsin S, Chan YS, Tan CP, Sim AYL, Zhang T, Susanto TT, Fu Z, Nagarajan N, Wan Y, In vivo mapping of eukaryotic rna interactomes reveals principles of higher-order organization and regulation, *Mol Cell* **62**(4):603–617, 2016. doi: 10.1016/j.molcel.2016.04.028.
10. Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, Davidovich C, Gooding AR, Goodrich KJ, Mattick JS, Mesirov JP, Cech TR, Chang HY, Rna duplex map in living cells reveals higher-order transcriptome structure, *Cell* **165**(5):1267–1279, 2016. doi: 10.1016/j.cell.2016.04.028.
11. Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ, Global mapping of human rna-rna interactions, *Mol Cell* **62**(4):618–626, 2016. doi: 10.1016/j.molcel.2016.04.030.
12. Nguyen TC, Cao X, Yu P, Xiao S, Lu J, Biase FH, Sridhar B, Huang N, Zhang K, Zhong S, Mapping rna-rna interactome and rna structure in vivo by mario, *Nat Commun* **7**:12023, 2016. doi: 10.1038/ncomms12023.
13. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J, Johnston D, Kim GE, Spitale RC, Flynn RA, Zheng GXY, Aiyer S, Raj A, Rinn JL, Chang HY, Khavari PA, Control of somatic tissue differentiation by the long non-coding rna tincr, *Nature* **493**(7431):231–235, 2013. doi: 10.1038/nature11661.
14. Tafer H, Amman F, Eggenhofer F, Stadler PF, Hofacker IL, Fast accessibility-based prediction of rna-rna interactions, *Bioinformatics* **27**(14):1934–1940, 2011. doi: 10.1093/bioinformatics/btr281.
15. DiChiacchio L, Sloma MF, Mathews DH, Accessfold: Predicting rna-rna interactions with consideration for competing self-structure, *Bioinformatics* **32**(7):1033–1039, 2016. doi: 10.1093/bioinformatics/btv682.
16. Li J, Ma W, Zeng P, Wang J, Geng B, Yang J, Cui Q, Lnctar: A tool for predicting the rna targets of long noncoding rnas, *Brief Bioinform* **16**(5):806–812, 2015. doi: 10.1093/bib/bbu048.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool, *J Mol Biol* **215**(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2, http://www.ncbi.nlm.nih.gov/pubmed/2231712.
18. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC, Adaptive seeds tame genomic sequence comparison, *Genome Res* **21**(3):487–493, 2011. doi: 10.1101/gr.113985.110.
19. Li YY, Qin L, Guo ZM, Liu L, Xu H, Hao P, Su J, Shi Y, He WZ, Li YX, In silico discovery of human natural antisense transcripts, *BMC Bioinformatics* **7**:18, 2006. doi: 10.1186/1471-2105-7-18, http://www.ncbi.nlm.nih.gov/pubmed/16409644.
20. Wang P, Yin S, Zhang Z, Xin D, Hu L, Kong X, Hurst LD, Evidence for common short natural trans sense-antisense pairing between transcripts from protein coding genes, *Genome Biol* **9**(12):R169, 2008. doi: 10.1186/gb-2008-9-12-r169, http://www.ncbi.nlm.nih.gov/pubmed/19055728.
21. Li JT, Zhang Y, Kong L, Liu QR, Wei L, Trans-natural antisense transcripts including noncoding rnas in 10 species: Implications for expression regulation, *Nucleic Acids Res*

36(15):4833–4844, 2008. doi: 10.1093/nar/gkn470, http://www.ncbi.nlm.nih.gov/pubmed/18653530.

22. Szcześniak MW, Makałowska I, lncrna-rna interactions across the human transcriptome, *PLoS One* **11**(3):e0150353, 2016. doi: 10.1371/journal.pone.0150353.
23. Terai G, Iwakiri J, Kameda T, Hamada M, Asai K, Comprehensive prediction of lncrna-rna interactions in human transcriptome, *BMC Genomics* **17** Suppl 1:12, 2016. doi: 10.1186/s12864-015-2307-5.
24. Fukunaga T, Hamada M, Riblast: An ultrafast rna-rna interaction prediction system based on a seed-and-extension approach, *Bioinformatics* **33**(17):2666–2674, 2017. doi: 10.1093/bioinformatics/btx287.
25. Reuter JS, Mathews DH, Rnastructure: Software for rna secondary structure prediction and analysis, *BMC Bioinformatics* **11**:129, 2010. doi: 10.1186/1471-2105-11-129.
26. Gerlach W, Giegerich R, Guugle: A utility for fast exact matching under rna complementary rules including g-u base pairing, *Bioinformatics* **22**(6):762–764, 2006. doi: 10.1093/bioinformatics/btk041.
27. Pervouchine DD, Iris: Intermolecular rna interaction search, *Genome Inform* **15**(2): 92–101, 2004.
28. Mann M, Wright PR, Backofen R, Intarna 2.0: Enhanced and customizable prediction of rna-rna interactions, *Nucleic Acids Res*, 2017. doi: 10.1093/nar/gkx279.
29. Kato Y, Sato K, Hamada M, Watanabe Y, Asai K, Akutsu T, Ractip: Fast and accurate prediction of rna-rna interaction using integer programming, *Bioinformatics* **26**(18):i460–i466, 2010. doi: 10.1093/bioinformatics/btq372.
30. Wenzel A, Akbasli E, Gorodkin J, Risearch: fast rna-rna interaction search using a simplified nearest-neighbor energy model, *Bioinformatics* **28**(21):2738–2746, 2012. doi: 10.1093/bioinformatics/bts519.
31. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL, Viennarna package 2.0, *Algorithms Mol Biol* **6**:26, 2011. doi: 10.1186/1748-7188-6-26.
32. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL, Thermodynamics of rna-rna binding, *Bioinformatics* **22**(10):1177–1182, 2006. doi: 10.1093/bioinformatics/btl024.
33. Venables WN, Ripley BD, *Modern Applied Statistics with S*, 4th ed., Springer, New York, 2002, http://www.stats.ox.ac.uk/pub/MASS4, iSBN 0-387-95457-0.
34. Mullahy J, Specification and testing of some modified count data models, *J Econ* **33**(3):341–365, 1986.
35. Hu MC, Pavlicova M, Nunes EV, Zero-inflated and hurdle models of count data with extra zeros: Examples from an hiv-risk reduction intervention trial, *Am J Drug Alcohol Abuse* **37**(5):367–375, 2011. doi: 10.3109/00952990.2011.597280.
36. Jiang M, Anderson J, Gillespie J, Mayne M, ushuffle: A useful tool for shuffling biological sequences while preserving the k-let counts, *BMC Bioinformatics* **9**:192, 2008. doi: 10.1186/1471-2105-9-192.
37. Sing T, Sander O, Beerenwinkel N, Lengauer T, Rocr: Visualizing classifier performance in r, *Bioinformatics* **21**(20):3940–3941, 2005. doi: 10.1093/bioinformatics/bti623.
38. Muro EM, Andrade-Navarro MA, Pseudogenes as an alternative source of natural antisense transcripts, *BMC Evol Biol* **10**:338, 2010. doi: 10.1186/1471-2148-10-338, http://www.ncbi.nlm.nih.gov/pubmed/21047404.
39. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ, Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes, *Nature* **453**(7194):534–538, 2008. doi: 10.1038/nature06904, http://www.ncbi.nlm.nih.gov/pubmed/18404147.

40. Korneev SA, Korneeva EI, Lagarkova MA, Kiselev SL, Critchley G, O'Shea M, Novel noncoding antisense rna transcribed from human anti-nos2a locus is differentially regulated during neuronal differentiation of embryonic stem cells, *RNA* **14**(10):2030–2037, 2008. doi: 10.1261/rna.1084308, http://www.ncbi.nlm.nih.gov/pubmed/18820242.
41. Gong C, Maquat LE, lncrnas transactivate stau1-mediated mrna decay by duplexing with 3' utrs via alu elements, *Nature* **470**(7333):284–288, 2011. doi: 10.1038/nature09701, http://www.ncbi.nlm.nih.gov/pubmed/21307942.
42. Gong C, Tang Y, Maquat LE, mrna-mrna duplexes that autoelicit staufen1-mediated mrna decay, *Nat Struct Mol Biol* **20**(10):1214–1220, 2013. doi: 10.1038/nsmb.2664, http://www.ncbi.nlm.nih.gov/pubmed/24056942.
43. Schein A, Zucchelli S, Kauppinen S, Gustincich S, Carninci P, Identification of antisense long noncoding rnas that function as sineups in human cells, *Sci Rep* **6**:33605, 2016. doi: 10.1038/srep33605.
44. Smit A, Hubley R, Green P, Repeatmasker open-3.0, http://wwwrepeatmaskerorg, 1996, http://www.repeatmasker.org.
45. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M, Lincrna-p21 suppresses target mrna translation, *Mol Cell* **47**(4):648–655, 2012. doi: 10.1016/j.molcel.2012.06.027, http://www.ncbi.nlm.nih.gov/pubmed/22841487.
46. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD, Over 20% of human transcripts might form sense-antisense pairs, *Nucleic Acids Res* **32**(16):4812–4820, 2004. doi: 10.1093/nar/gkh818.
47. Lehner B, Williams G, Campbell RD, Sanderson CM, Antisense transcripts in the human genome, *Trends Genet* **18**(2):63–65, 2002.
48. Yuan JH, Yang F, Wang F, Ma JZ, Guo YJ, Tao QF, Liu F, Pan W, Wang TT, Zhou CC, Wang SB, Wang YZ, Yang Y, Yang N, Zhou WP, Yang GS, Sun SH, A long noncoding rna activated by tgf-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma, *Cancer Cell* **25**(5):666–681, 2014. doi: 10.1016/j.ccr.2014.03.010, http://www.ncbi.nlm.nih.gov/pubmed/24768205.
49. Klammer AA, Park CY, Noble WS, Statistical calibration of the sequest xcorr function, *J Proteome Res* **8**(4):2106–2113, 2009. doi: 10.1021/pr8011107.
50. Murdoch DJ, Tsai YL, Adcock J, P-values are random variables, *Am Stat* **62**(3):242–245, 2008.
51. Alkan F, Wenzel A, Palasca O, Kerpedjiev P, Rudebeck AF, Stadler PF, Hofacker IL, Gorodkin J, Risearch2: Suffix array-based large-scale prediction of rna-rna interactions and sirna off-targets, *Nucleic Acids Res* **45**(8):e60, 2017. doi: 10.1093/nar/gkw1325.

Ivan Antonov received his M.Sc. from Moscow State University, Russia in 2008 and Ph.D. from Georgia Institute of Technology, USA in 2012 (under supervision of Dr. Mark Borodovksy). Currently, he works with Yulia Medvedeva as a postdoctoral research scientist at the Institute of Bioengineering RAS, Moscow, Russia.

His research interests include long noncoding RNAs, RNA–RNA interactions, post-transcriptional and epigenetic gene regulation as well as programmed frameshifting.

Andrey Marakhonov received his M.Sc. from Moscow State University, Russia, in 2006 and Ph.D. from Research Centre for Medical Genetics, Russia, in 2010 (under the supervision of Dr. Mikhail Skoblov). Currently, he works as a Senior Research Scientist at the Research Centre for Medical Genetics, Moscow, as well as at the Moscow Institute of Physics and Technology, Dolgoprudny, Russia. His research interests include molecular pathogenesis of hereditary diseases, functional genomics, long noncoding RNAs, and antisense regulation.

Maria Zamkova graduated from Moscow State University, Russia, in 2006 and received her Ph.D. from Russian Blokhin Cancer Research Center, Russia, in 2013. Currently she works in Blokhin Cancer Research Center. Her scientific interests include immunology, T cells, TCRs, cancer, lncRNAs, RNA–RNA interactions.

Yulia Medvedeva graduated from Lomonosov Moscow State University and got her Ph.D. from GosNIIgenetika, Moscow, Russia. After that, she worked at King Abdullah University of Science and Technology (Jeddah, Saudi Arabia) and Institute of Personalized and Predictive Medicine of Cancer (Barcelona, Spain).

Now she is a research group leader at the Research Center of Biotechnology (Moscow, Russia). Her main interests include computational biology, epigenomics, transcriptomics, regulation of transcription, lncRNA and application of computational methods to medical research.