# A method for RNA structure prediction shows evidence for structure in lncRNAs

Riccardo Delli Ponti[1,2], Alexandros Armaos[1,2], Stefanie Marti[1,2]

and Gian Gaetano Tartaglia[1,2,3]

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain

[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluís Companys, 08010 Barcelona, Spain

* To whom correspondence should be addressed.
; Email: gian.tartaglia@crg.eu; Tel: +34 93 316 01 16; Fax: +34 93 396 99 83

## Abstract

To compare the secondary structure of RNA molecules we developed the *CROSSalign* method. *CROSSalign* is based on the combination of the Computational Recognition Of Secondary Structure (CROSS) algorithm to predict the RNA secondary structure at single-nucleotide resolution using sequence information only and the Dynamic Time Warping (DTW) method to align profiles of different lengths. We applied *CROSSalign* to investigate the structural conservation of long non-coding RNAs such as *XIST* and *HOTAIR* as well as ssRNA viruses including *HIV*. The algorithm is able to find homologues between thousands of possible matches identifying the exact regions of similarity between profiles of different length. *CROSSalign* is freely available at the webpage http://service.tartaglialab.com//new_submission/crossalign.

### Keywords

Non-coding RNA, Secondary Structure, Structural Alignments

## Introduction

Sequence similarity is often considered the key feature to investigate evolutionary conservation of coding transcripts [1]. Yet, knowledge of secondary structure provides important insights into the biological function of RNAs by allowing the study of physical properties, such as for instance molecular interactions [2]. In most cases, information about the RNA folding complements sequence analysis [3] and is useful to understand mechanisms of action: microRNA precursors, for example, are processed by DGCR8 if folded in specific hairpin loop structures [4]. Similarly, the architecture of ribosomal RNAs evolves in a self-contained way through conservation of stem loops present in ancient species [5,6], indicating distinct requirements for structural elements.

Long non-coding RNAs (lncRNAs) are regarded as a mystery in terms of sequence and structural conservation [7]. The vast majority of lncRNAs evolve under little or no selective constraints, undergo almost no purifying selection, are poorly expressed and without easily identifiable orthologues [7,8]. The average sequence homology of evolutionarily conserved lncRNAs is only 20% between human and mouse and drops to 5% between human and fish [7]. Thus, primary structure does not provide relevant information on lncRNA conservation and secondary structure should be used for better characterization. In addition to lncRNAs, the transcriptomes of single-stranded RNA (ssRNA) viruses retain their fold even if sequences mutate rapidly [9], which indicates that secondary structure investigation can reveal important properties.

To study the structural conservation of RNA molecules, we developed the *CROSSalign* method. *CROSSalign*, available at our webpages

http://service.tartaglialab.com//new_submission/crossalign, is based on the combination of two methods: 1) Computational Recognition Of Secondary Structure (CROSS), which is an algorithm trained on experimental data to predict RNA secondary profiles without sequence length restrictions and at single-nucleotide resolution [10]; 2) the Dynamic Time Warping (DTW) algorithm to assess the similarity of two profiles of different lengths [11]. DTW flexibility allows managing profiles of different length without having to sacrifice computational time.

We applied *CROSSalign* on lncRNAs of different species as well as ssRNA viruses. *CROSSalign* is able to find structural homologues among millions of possible matches identifying structural domains with great accuracy.

## Results

To test the performances and functionality of CROSS combined with DTW (**Supplementary Figures 1** and **2**), we selected a dataset of 22 structures for which crystallographic (exact base pairing from between nucleotides) and Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE; chemical probing of flexible regions used to asses double and single-stranded state of nucleotide) data are available [12]. Using DTW, we calculated structural distances between all possible couples in the dataset considering crystallographic (dots and parentheses were transformed in binary code) data as well as 1) SHAPE (Area Under the ROC Curve AUC of 0.76, Positive Predictive Value PPV of 0.76 with crystallographic data) and 2) CROSS profiles (AUC 0.72, PPV 0.74 with crystallographic data, see also http://service.tartaglialab.com/static_files/algorithms/cross/documentation.html#5).

Structural distances with CROSS show higher correlations (Pearson's correlation of 0.91) than SHAPE (**Figure 1A, 1B**, correlation of 0.50). Moreover, CROSS shows better performances than algorithms such as *RNAstructure* [13,14] and *RNAfold* [12] (respective correlations 0.71 and 0.47 with crystals; **Supplementary Figures 3** and **4**).

Sequence similarity (computed with EMBOSS; see **Material and Methods**) show comparable correlations with structural distances calculated with either CROSS or crystallographic profiles (respectively: 0.80, 0.83, 0.38 with crystallographic, CROSS and SHAPE data). While CROSS and crystallographic profiles identify specific clusters of RNA molecules with low sequence identity and high structural similarity (colored in red, orange and green according to difference confidence thresholds; **Figures 1C, 1D, 1E**), SHAPE data cannot be used to identify the structures.

***XIST***

We used *CROSSalign* to study structural similarities of *XIST* domain RepA in 10 different species [15]. Our analysis reveals that primates cluster close to human (**Supplementary Figure 5A**) while other species are more distant (**Supplementary Figures 5B** and **6**). By contrast, calculating sequence similarity with respect to human *XIST* (computed with EMBOSS; see **Material and Methods**), we could not identify a specific cluster for primates (**Supplementary Figure 5**). Thus, our results indicate that secondary structure shows a higher degree of conservation than sequence.

We then selected RepA of orangutan and searched for structural similarities in all human intergenic lncRNAs (lincRNAs 8176 sequences; ENSEMBLE 82). *XIST* was ranked as the best significant match in the pool (structural distance 0.01; p-value < $10^{-6}$) and RepA was correctly identified (predicted coordinates: 328-764; 95% overlap with the query region; **Figure 2A**; **Supplementary Table 1**). Same results were observed for baboon RepA (best result: 0.032; 86% overlap with the query region) and lemur RepA (best result: 0.075; p-value; 97% overlap with the query region), suggesting a strong structural conservation within primates (**Figure 2B**). By contrast, human and mouse RepA showed larger distance in terms of both structural and sequence similarity, which is in agreement with previous studies on lncRNA conservation [16].

We used *CROSSalign* to search human RepA in all mouse lncRNAs and identified *XIST* as the 5$^{th}$ best hit (structural distance 0.085; p-value < $10^{-6}$; **Figure 3A)**. In this case, the position of RepA was not correctly assigned (coordinates: 10306-10698; 0% overlap) but the best match falls in the regulatory region of exon 7 and the structural relation between RepA and exon 7 has been reported [17]. Importantly, the correct coordinates of human RepA within mouse *XIST* rank second in our analysis (structural distance 0.086; p-value < $10^{-6}$), while the best match is a miRNA-containing gene *Mirg* (ENSMUSG00000097391) and the two secondary structure profiles show a strong correlation of 0.92 (**Figure 3B**). Interestingly, even if little information is available on *Mirg*, it is be prevalently expressed in the embryo [18]. This result unveils a previously unreported relationship between *XIST* and *Mirg*, in which structural and functional homologies can be linked. Intriguingly, also the second best result, *Rian* (ENSMUSG00000097451), is expressed in embryo, while no information is available on the third and fourth hits (ENSMUSG00000107391 and ENSMUSG00000085312).

We note that the five matches here are not listed in the top 20 hits obtained by analysis of sequence similarity (<34%).

Our results suggest that RepA secondary structure is conserved among primates, and diverges between human and mouse. However, analyzing the information contained in structured nucleotides (i.e., nucleotides with CROSS score < 0 are set to 0) we could identify *XIST* as the 1$^{st}$ best hit of human RepA in all mouse lncRNAs (structural distance 0.034; p-value < $10^{-6}$; **Figure 3C**). This result indicates that double-stranded regions are more conserved than single-stranded regions. In addition, we note that the sequence identity ranks *XIST* as the 14$^{th}$ hit of human RepA in all mouse lncRNAs, thus showing significantly lower ability to identify structural homologues.

***HOTAIR***

We selected the D2 domain of *HOTAIR* (predicted by CROSS to be highly structured; **Supplementary Figure 7**) to measure its conservation in 10 species [15] using *CROSSalign* (**Supplementary Figure 8A**). As for *XIST*, structural distance analysis indicates that primates cluster close to human and other species are more distant (**Supplementary Figure 8B)**.

Orangutan D2 was searched in all human lncRNAs and *HOTAIR* was identified as the best match (structural distance 0.032; p-value < $10^{-6}$) with overlapping coordinates (nucleotides: 666-1191; 78% overlap with the query region; **Figure 4A**). Searching mouse D2 in all human lncRNAs, *HOTAIR* was found as the best (0.092; p-value < $10^{-4}$) and matching position (nucleotides: 284-788; 57% overlap; **Figure 4B**). These results

suggest that D2 secondary structure is not only conserved in primates but also in mouse.

To further investigate *HOTAIR* secondary structure, we selected the D4 (**Supplementary Figure 9)** domain that is predicted by CROSS to be poorly structured (**Supplementary Figure 7**).

Orangutan D4 is the best results among all human lncRNA (structural distance 0.023; p-value < $10^{-6}$) and correctly matching in sequence position (predicted coordinates: 1650-2291; overlap of 79%; **Figure 5A**). By contrast, mouse D4 shows poor ranking ($1849^{th}$) in all human lncRNAs (structural distance 0.104; p-value = 0.061), indicating little structural homology between human and mouse (**Figure 5B; Supplementary Figure 9**).

***HIV***

HIV is one of the most studied ssRNA viruses with a complex secondary structure [19] that is accurately predicted by CROSS [10] (see also http://service.tartaglialab.com/static_files/algorithms/cross/documentation.html#4 ).

We divided HIV in 10 non-overlapping regions of ~1000 nucleotides and searched each of them against a database of ssRNA viruses having as host human (292 cases, downloaded from NCBI; **Supplementary Table 2**) to identify structurally similar domains. We found that coronavirus HKU and Simian-Human immunodeficiency SIV have the most significant matches with HIV (structural distance 0.078; p-value < $10^{-6}$; for SIV; structural distance 0.093 p-value < $10^{-4}$ for HKU). This finding is particularly

relevant since SIV and HIV share many similarities in terms of pathogenicity and evolution [20]. Indeed, previous studies already reported a similarity in terms of secondary structures between HIV and SIV that is not explained by sequence similarity [21].

In addition, we found that HIV 5' region is structurally similar to a strain of Ebola virus (Tai Forest; **Supplementary Table 2**). In agreement with this observation, previous studies indicate that HIV and Ebola have the same mechanisms of egress, taking contact with the cellular protein Tsg101 [22]. Moreover, HIV 5' is the most conserved region in all ssRNA viruses (**Figure 6A)**. This result indicates that the secondary structure of this region is not only necessary for HIV encapsidation [23], but is also essential for the activity of other viruses.

We also compared structural distances and sequence similarities of all HIV strains (4804; see **Material and Methods**). We found two clusters (brown and red; **Figure 6B**) that are similar in terms of structure (~0.06 structural distance; p-value < $10^{-6}$) and sequence (95%-80% sequence similarity). Other clusters (red and green; **Figure 6B**) showed significant distance in structure (from 0.06 to 0.09 of structural distance; p-value < $10^{-6}$) that is not identifiable by sequence similarity (~85% sequence similarity). This result suggests that HIV could have evolved maintaining a similar sequence but different structures, as previously reported in literature [24].

## Discussion

We developed the *CROSSalign* method based on the combination of the CROSS algorithm to predict the RNA secondary structure at single-nucleotide resolution [10] and the Dynamic Time Warping (DTW) algorithm to align profiles of different lengths [11]. DTW has been previously applied in different fields, especially pattern recognition and data mining [25,26], but has never been used to investigate structural alignments. Since CROSS has no sequence length restrictions and shows strong performances on both coding and non-coding RNAs [10] the combination with DTW allows very accurate comparisons of structural profiles. Other thermodynamic approaches, such as RNAstructure [14] or RNAfold [12] cannot be directly used for such task since they are restricted on the sequence length (typically <1000 nt) [27].

We applied *CROSSalign* to investigate the structural conservation of lncRNAs in different species and the complete genomes of ssRNA viruses. We found that the algorithm is able to find structural homologues between thousands of possible matches correctly identifying the regions of similarity between profiles of different length. The results of our analysis reveal a structural conservation between know lncRNA domains including *XIST* RepA (1/8176 best hit; 95% overlap with the query region) and *HOTAIR* D2 (1/8176 best hit; 78% overlap with the query region), but also identify structural similarities between regulatory regions of HIV and other ssRNA viruses, opening new questions regarding similar mechanisms mediated by the secondary structure.

Our webserver is available at http://service.tartaglialab.com//new_submission/crossalign (documentation and tutorials are at the webpages http://service.tartaglialab.com/static_files/algorithms/crossalign/documentation.html and http://service.tartaglialab.com/static_files/algorithms/crossalign/tutorial.html) and allows to predict structural similarities between two or more RNA molecules. *CROSSalign* can be interrogated to search structural similarity between thousands of lncRNA molecules and identifies regions using a specific DTW algorithm (open begins and ends *OBE*).

As shown in the examples presented, *CROSSalign* is a very versatile algorithm able to simplify the complex search for structural similarity among RNA molecules and shows great potential for the study of lncRNAs.

# Materials and Methods

### *Prediction of the RNA secondary structure: CROSS*

Secondary structure profiles were generated using CROSS [10]. We trained CROSS on data from high-throughput experiments (PARS: yeast and human transcriptomes [3,28] and icSHAPE: mouse transcriptome [29]) as well as on low-throughput SHAPE [19] and high-quality NMR/X-ray data [30]. Since each approach has practical limitations and a different range of applicability, we combined the five models into a single algorithm, *Global Score,* which provides a *consensus* prediction.

The consensus model *Global Score* was trained and tested on independent sets of NMR/X-ray structures (11'670 training fragments, 5'475 testing fragments [12,31]). In the testing phase, single and double-stranded nucleotides were recognized with an AUC of 0.72 and a PPV of 0.74. Comparing the structures with experimental SHAPE data, we observed similar performances (AUC of 0.76 and PPV of 0.76; see http://service.tartaglialab.com/static_files/algorithms/cross/documentation.html#5]

Comparisons between CROSS and other algorithms have been reported in our previous publication [10]. In addition, as done with experimental SHAPE data, *Global Score* can be used as a constraint in *RNAstructure* [13,14]. On our test set [12], *Global Score* increases the PPV of *RNAstructure* from 0.68 to 0.72, with remarkable improvements in 13 cases (from 0.44 to 0.72) and decreases the PPV in only three cases for which real SHAPE data does not improve performances. Moreover, using the partition

function computed with *RNAstructure*, we calculated the AUC for each structure with and without CROSS constraints and observed an improvement from 0.81 to 0.86 when CROSS is integrated in the algorithm. On the test set [12], we found a similar trend using *RNAfold* [12] (the PPV increases from 0.67 to 0.70 using *Global Score* and the AUC remains at 0.85).

In this study, all the profiles were computed using the *Global Score* module without smoothing: nucleotides with a score higher then 0 are predicted to be double-stranded and structured, while nucleotides with a score lower than 0 are single-stranded. Since the algorithm has no sequence length restriction and shows strong performances on both coding and non-coding RNAs [10] it was combined with DTW for pairwise comparison of structural profiles. Thermodynamic approaches, such as RNAstructure [14] or RNAfold [12], could not be directly used for such task since they are restricted on the sequence length (maximum ~1000nt).

***Comparison of structural profiles: DTW***

To compare two CROSS profiles, we used the Dynamic Time Warping (*DTW*) algorithm available in the R package *dtw* [11]. The open begin and end (*OBE-DTW*) algorithm was employed to compare profiles of different lengths. Indeed, the *DTW* method imposes the same begins and ends to the two profiles that are compared, while *OBE-DTW* searches the profile of shorter length within the other one.

I*DTW* is used to compare profile of similar length (i.e., one sequence is less than 3 times longer than the other), while *OBE-DTW* is preferred to search modules inside

bigger profiles (e.g., RepA inside the complete *XIST* sequence; ~45 times bigger). The structural distance is computed with an asymmetric pattern and using the Manhattan distance, which is optimal for comparing profiles of different length. To avoid biases regarding the length of the profiles, the final structural distance is normalized for the length of both profiles using the internal function *normalizedDistance*. We also tested different normalizations to DTW outputs (including length of the shorter or longer profile) and we found that the normalization based on the length of both profiles is optimal. The function *index* was used to visualize the optimal path and to extract the matching coordinates between the two profiles.

***Statistical analysis***

To compute the significance of a specific DTW score, we analyzed the statistical distributions generated using human lncRNAs of different lengths (200, 500, 1000, 5000 nucleotides). 100 molecules for each class were employed to compute the structural distance between the classes. The distributions are set as a reference to compute the p-values in new analyses (**Supplementary Table 3**)

***Datasets***

- lncRNA sequences were downloaded from ENSEBLE 82 using Biomart and specifying lincRNAs, for a total of 4427 sequences for mouse and 8176 for human.
- The complete viral genomes were downloaded from NCBI selecting ssRNA

viruses having as host human or primates (for SIV), for a total of 292 complete genomes.

- The complete rRNA sequences were downloaded from NCBI.
- RepA, D2 and D4 were selected from the data publicly available from the work of Rivas et al., 2016 [15]. To keep consistency between the results we tried to select the same species between the two sets of multialignments. When this was not possible we selected similar species (rat and mouse, orangutan and chimpanzee, lemur and sloth).
- The HIV strains were downloaded from HIV databases (https://www.hiv.lanl.gov/), selecting only complete genomes for a total of 4804 sequences processable by CROSS.

***Sequence alignment***

To compute the sequence alignments we used the browser version of EMBOSS-needleall, public available at http://www.bioinformatics.nl/cgi-bin/emboss/needleall.

The tool was used with standard settings to speed-up the calculation. The sequence identity was retrieved from the corresponding field from EMBOSS multiple output.

## *Algorithm Description*

***Availability of Data***

The code is publicly available under an open source license compliant with Open

Source Initiative at *https://github.com/armaos/algorithm-crossalign*. The source code is deposited in a DOI-assigning repository *https://doi.org/10.5281/zenodo.1168294.*

***Input***

The user should paste one or two RNA sequences in FASTA format into the dedicated form, providing an email address (optional) to receive a notification when the job is completed. The algorithm can be launched in 4 different modes, each of them being a specific variation of the DTW algorithm (**Supplementary Figure 1**).

- The *standard-DTW* is recommended to compare profile of similar length RNAs (i.e., one sequence is less than 2 times longer than the other).
- *OBE-DTW* (open begins and ends) is a specific mode to search a smaller profile inside a bigger one. This is the recommended modality when comparing profile of very different sizes (i.e., one sequence is more than 5 times longer than the other). Please keep in mind that the sequence in the form of RNA input 1 will be searched in RNA input 2, so the sequence in RNA input 1 should be smaller than the other.
- The *fragmented OBE-DTW* is a specific modality to search unknown secondary structure domains of one profile inside the other. The secondary structure of RNA input 1 will be fragmented with a non-overlapping window of 200 nucleotides [optimal size to search secondary structure domains in large RNAs [32,33]] Each fragment of one sequence be then searched against the other sequence. This approach is the recommended mode when the user is not interested in the global similarity between two secondary structure profiles,

but wants to search an unknown domain conserved in both sequences. A minimum length of 600 nucleotides is recommended for fragmentation.

- The *dataset* mode allows the user to search a single sequence inside all the lincRNAs of a specific organism. The shorter profile of each couple will be searched in the larger one following the *OBE-DTW* procedure. The organisms available are Human, Mouse, Rat, Macaque and Zebrafish. The lincRNAs where downloaded using Biomart (Ensemble 82). New organisms and updated versions will be regularly added.

***Output***

We report the *CROSSalign* score that measures the structural distance between two structures. The closer the score is to 0, the higher the similarity in terms of secondary structure. According to our statistical analysis, RNA molecules with a structural distance of 0.10 or higher are to be considered different in term of secondary structure (see ***Documentation***).

The main image shows the overall structural similarity of the two profiles employed to calculate *CROSSalign* score (**Supplementary Figure 2A**). On the two axes the user will see the structural profiles obtained with CROSS for the two RNA sequences in input (score >0 means a double-stranded nucleotide; <0 single-stranded;). For a better visualization, the profiles are smoothed using a function previously defined [10]

The similarity is represented by the red path in the figure, obtained with the index function of the *dtw* package. The closer the path is to the diagonal, the more similar

are the profiles. Vertical or horizontal paths are to be considered gaps, while diagonal paths highlight similar regions of the two profiles.

Since *OBE-DTW* allows the identification of the optima starting/ending points of a match, the optimal match is reported in term of coordinates relative to the larger profile (RNA input 2). The main plot shows the CROSS profiles of the optimal matching region selected by the *OBE-DTW* algorithm (**Supplementary Figure 2B**). In order to keep the gaps introduced by the *OBE-DTW* algorithm, the two profiles are not smoothed.

The *fragmented OBE-DTW* is a particular form of *OBE-DTW* optimized to search all the possible structural domains of a particular sequence in another one. The main output is a scrolling table reporting the structural score, the beginning of the match, the end of the match and the p-value (**Supplementary Figures 2C**). All the values are computed with the same procedures used for *OBE-DTW*. The table can also be downloaded as a .txt file. The same output is used for the *dataset* mode, but in this case the table can only be downloaded.

## ACKNOWLEDGMENTS

We thank Philipp Germann, Fernando Cid and the other members of our group for useful comments.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), through the European Research Council, under grant agreement RIBOMYLOME_309545 (Gian Gaetano Tartaglia), and from the Spanish Ministry of Economy and Competitiveness (BFU2014-55054-P and BFU2017-86970-P). We also acknowledge support from AGAUR (2014 SGR 00685), the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017' (SEV-2012-0208). We also thank the CRG fellowship to SM.

## COMPETING INTERESTS

The authors declare no competing financial or non-financial interests

## AUTHORS CONTRIBUTIONS

GGT designed the work, RDP implemented the approach, AA and RDP developed the server. GGT, RDP and SM wrote the paper. All authors reviewed the manuscript.

## SUPPLEMENTARY MATERIAL

See accompanying document.

# Figures

**Figure 1.** Validation of *CROSSalign* method. (**A**) Structural distances correlation between CROSS and crystallographic profiles computed on 22 structures (*standard-DTW*). (**B**) Correlation between structural distances computed with SHAPE and crystallographic data on the same data (*standard-DTW*). (**C**) Correlation between structural distances (crystallographic profiles) and sequence similarity. Clusters of similar structures and different sequences (sequence similarity < 40%) are highlight in brown (structural score < 0.2), orange (structural score < 0.1) and (red structural score < 0.05). (**D**) Correlation between structural distances (CROSS) and sequence similarity. The clusters previously identified for crystallographic data are shown in the plot. (**E**) Correlation between structural distances (SHAPE) and sequence similarity. In this case the cluster previously identified are disrupted.

**Figure 2.** Structural conservation of *XIST* RepA within primates. (**A**) Structural distances of Orangutan RepA are computed against all human lincRNAs. The structural distance is calculated using *OBE-DTW* and plotted against sequence similarity. Orangutan RepA is identified as the best match (colored in red). (**B**) Structural distances of baboon RepA against all the lincRNAs of human. Baboon RepA is identified as the best match (colored in red).

**Figure 3.** Structural similarities between human and mouse *XIST* RepA (**A**) Structural similarities of human RepA against all mouse lincRNAs. The structural distance was calculated using CROSS (*OBE-DTW*) and plotted against sequence similarity. Human RepA is not the best match ($5^{th}$ best hit; colored in red). (**B**) Structural similarities of

human RepA against all mouse lincRNAs using double-stranded nucleotides (nucleotide with CROSS score < 0 are set to 0). Human RepA is identified as the best match (colored in red), which highlights the importance of the structural content for the regulatory domains of the lncRNAs. (**C**) Secondary structure profile of human RepA, obtained as optimal path with *OBE-DTW*, compared with the best match in mouse lincRNAs (*Mirg*; ENSMUSG00000097391). The two secondary structure profiles show a strong a correlation (0.92).

**Figure 4.** Structural conservation D2 *HOTAIR* in different species. (**A**) Structural similarities of orangutan D2 against all human lincRNAs. The structural distance was obtained using *OBE-DTW* and plotted against with the sequence similarity. The D2 of orangutan is identified as the best match (colored in red). (**B**) Structural similarities of human D2 against all mouse lincRNAs. Human D2 is identified as the best match (colored in red).

**Figure 5.** Structural conservation of D4 *HOTAIR* in different species. (**A**) Structural similarities of Orangutan D4 against all human lincRNAs. The structural distance was obtained using *OBE-DTW* and plotted against sequence similarity. Orangutan D4 is identified as the best match (colored in red). (**B**) Structural similarities of human against all mouse lincRNAs. Human D4 is not identified as the best match ($1849^{th}$ best hit; colored in red;).

**Figure 6.** Structural analysis of HIV transcriptome. (**A**) Structural conservation of HIV genome (divided in 10 not overlapping regions) compared with the complete genome of 292 ssRNA viruses. The region spanning the first 1000nt (including 5' UTR) is the most

conserved among all the viruses. (**B**) Structural distances of complete HIV genome against the complete genomes of 4884 HIV strains. Using analysis of primary and secondary structures, we identified four main clusters (red, green, brown and yellow). Red and green boxes indicate strains whose structural difference cannot be identified through sequence analysis, while brown and red boxes as well as green and yellow boxes identify strains with similar structures and different sequences.

# Supplementary Figures and Tables

**Supplementary Figure 1.** Main page of *CROSSalign* webserver. The user can upload fasta sequences and select a DTW mode for the task. The algorithm measures structural distances between two or more RNA sequences. Specifically, it 1) compares profiles of similar length (*standard-DTW*), 2) searches domains of a short profile in a large one (*OBE-DTW*), 3) fragments a long sequence (*fragmented OBE-DTW*) or 4) searches a profile in all the lncRNAs of a specific organism (*dataset*). See the **Tutorial** for more details.

**Supplementary Figure 2.** Outputs of *CROSSalign* webserver. (**A**) *Standard-DTW*. On the two axes there are the secondary structure profiles obtained with CROSS, while in the main plot it is reported the optimal path (highlight the regions of similarity. (**B**) *OBE-DTW* overlap of the best matching region between the two secondary structure profiles (obtained with CROSS). (**C**) *Fragmented OBE-DTW* output table showing structural distance, starting and ending positions of the match and the p-values for all the fragments of the input profile. See the **Tutorial** for more details.

**Supplementary Figure 3.** (**A**) Structural distances correlation between *RNAstructure* and crystallographic profiles (22x22 structures). (**B**) Structural distances correlation between *RNAfold* (Vienna suite) and crystallographic profiles (22x22 structures).

**Supplementary Figure 4.** (**A**) Correlation between structural distances (*RNAstructure*) and sequence similarity. In this case the cluster previously identified are disrupted. (**B**) Correlation between structural distances (*RNAfold*) and sequence similarity. In this case the cluster previously identified are disrupted.

**Supplementary Figure 5.** (**A**) Structural difference with respect to human (structural distance *100) for *XIST* RepA in 10 different species. The primates cluster together. (**B**) Sequence distances from human calculated as (100-sequence similarity)% for 10 different species. The primates cluster cannot be identified by primary sequence.

**Supplementary Figure 6.** Dendogram showing the species tree for *XIST* RepA obtained using structural distance.

**Supplementary Figure 7.** Cumulative distribution function (CDF) of the structural content of all the human lincRNAs predicted by CROSS. The structural contents of the D domains of *HOTAIR* are reported on the curve.

**Supplementary Figure 8.** (**A**) Structural difference with respect to human (structural distance *100) for *HOTAIR* D2 of 10 different species. The primates tend to cluster together. (**B**) Sequence difference with respect to human [(100-sequence similarity)%] for 10 different species. The primates cluster can also be identified by primary sequence.

**Supplementary Figure 9.** (**A**) Differences in structure from human (structural distance *100) for the D4 of 10 different species. The primates tend to cluster together. (**B**) Sequence distances from human calculated as (100-sequence similarity)% for 10 different species. The primates cluster can be identified by primary sequence.

**Supplementary Table 1.** Table summarizing the results for the lncRNAs reported in our work.

**Supplementary Table 2.** Table reporting the 3 best candidates for each HIV domain. The results discussed in the main text are reported in bold.

**Supplementary Table 3.** (**A**) Means and standard deviations for structural distances of reference distributions. The *standard-DTW* was used to compare profiles of similar length, while *OBE-DTW* for the other cases. (**B**) P-values at 1% of the structural distances. The *standard-DTW* was used for profiles of similar length, while *OBE-DTW* for the all others.

1. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
2. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8,** 444–445 (2011).
3. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (2014).
4. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15,** 509–524 (2014).
5. Bokov, K. & Steinberg, S. V. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* **457,** 977–980 (2009).
6. Petrov, A. S. *et al.* History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15396–15401 (2015).
7. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **17,** 601–614 (2016).
8. Diederichs, S. The four dimensions of noncoding RNA conservation. *Trends Genet.* **30,** 121–123 (2014).
9. Chursov, A., Frishman, D. & Shneider, A. Conservation of mRNA secondary structures may filter out mutations in Escherichia coli evolution. *Nucleic Acids Res.* **41,** 7854–7860 (2013).
10. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile RNA structure. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1094
11. Giorgino, T. Computing and Visualizing Dynamic Time Warping Alignments in *R*□: The **dtw** Package. *J. Stat. Softw.* **31,** (2009).
12. Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F. & Wolfinger, M. T. SHAPE

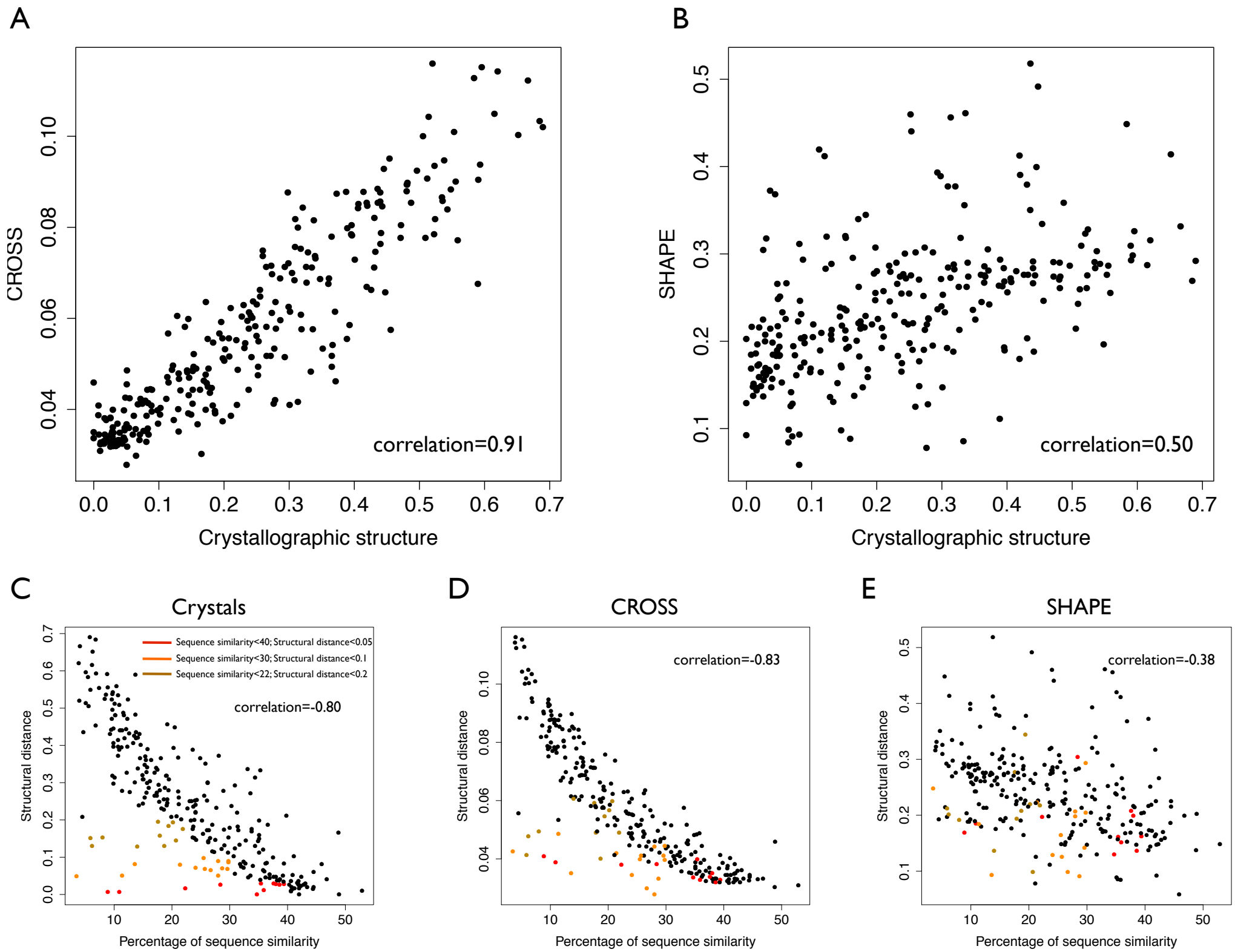
directed RNA folding. *Bioinforma. Oxf. Engl.* **32,** 145–147 (2016).

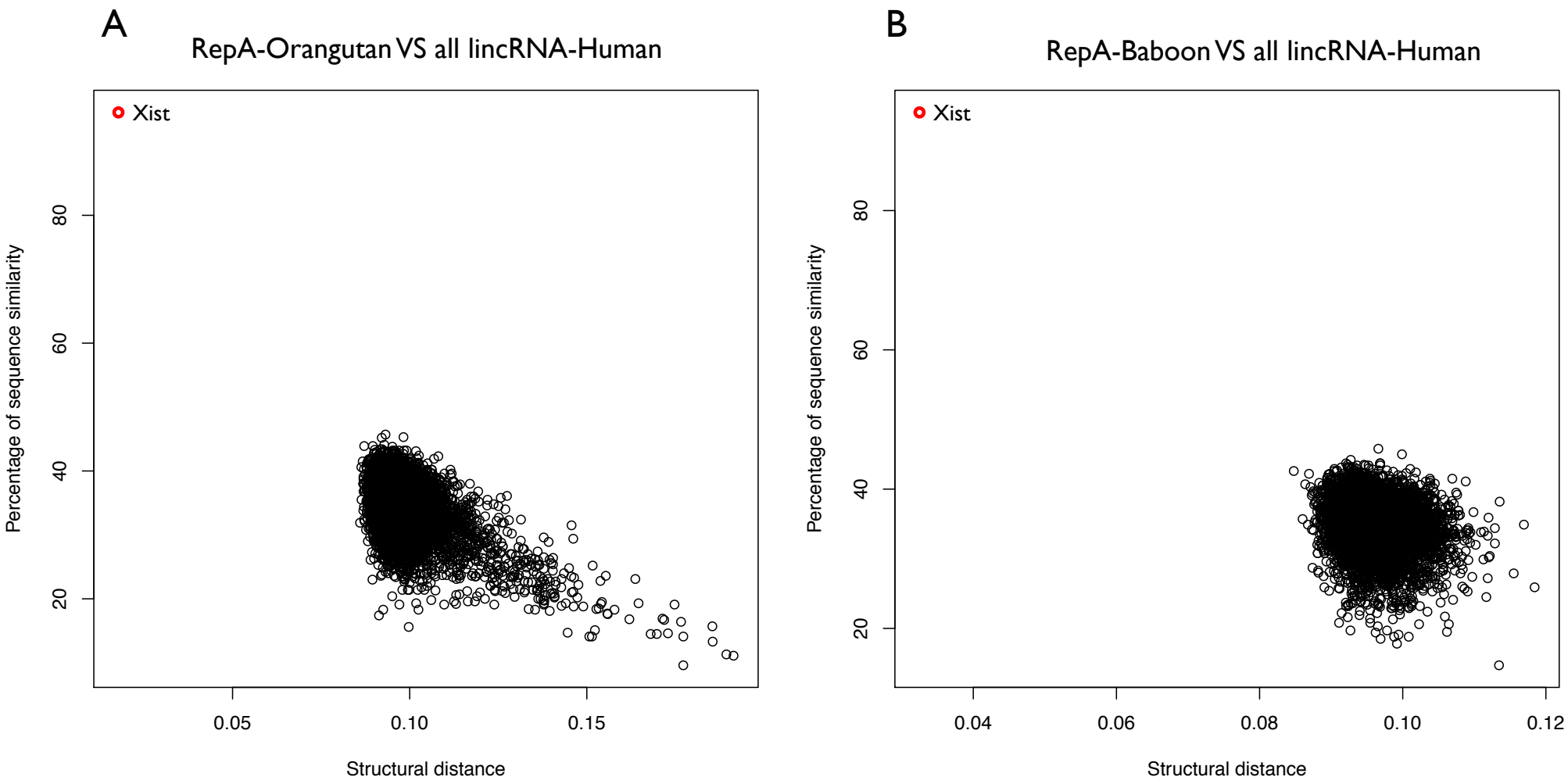
13. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288,** 911–940 (1999).
14. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11,** 129 (2010).
15. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14,** 45–48 (2017).
16. Breschi, A., Gingeras, T. R. & Guigó, R. Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* **18,** 425–440 (2017).
17. Yamada, N. *et al.* Xist Exon 7 Contributes to the Stable Localization of Xist RNA on the Inactive X-Chromosome. *PLOS Genet.* **11,** e1005430 (2015).
18. Schmitt, B. M. *et al.* High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Res.* **24,** 1797–1807 (2014).
19. Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460,** 711–716 (2009).
20. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS Pandemic. *Cold Spring Harb. Perspect. Med.* **1,** a006841–a006841 (2011).
21. Rizvi, T. A. & Panganiban, A. T. Simian immunodeficiency virus RNA is efficiently encapsidated by human immunodeficiency virus type 1 particles. *J. Virol.* **67,** 2681–2688 (1993).
22. Martin-Serrano, J., Zang, T. & Bieniasz, P. D. HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate

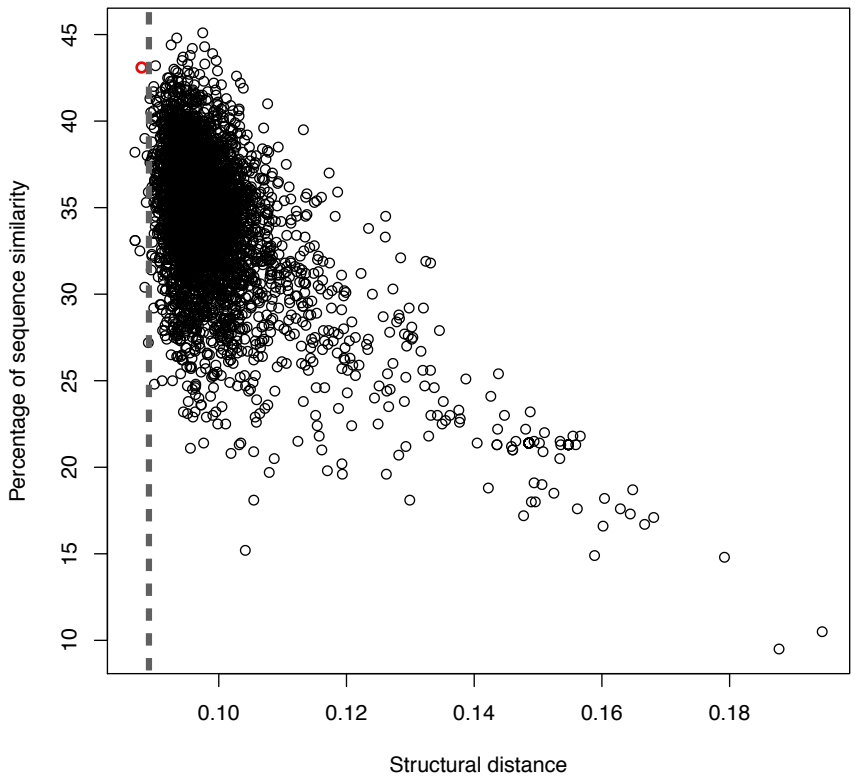
egress. *Nat. Med.* **7,** 1313–1319 (2001).

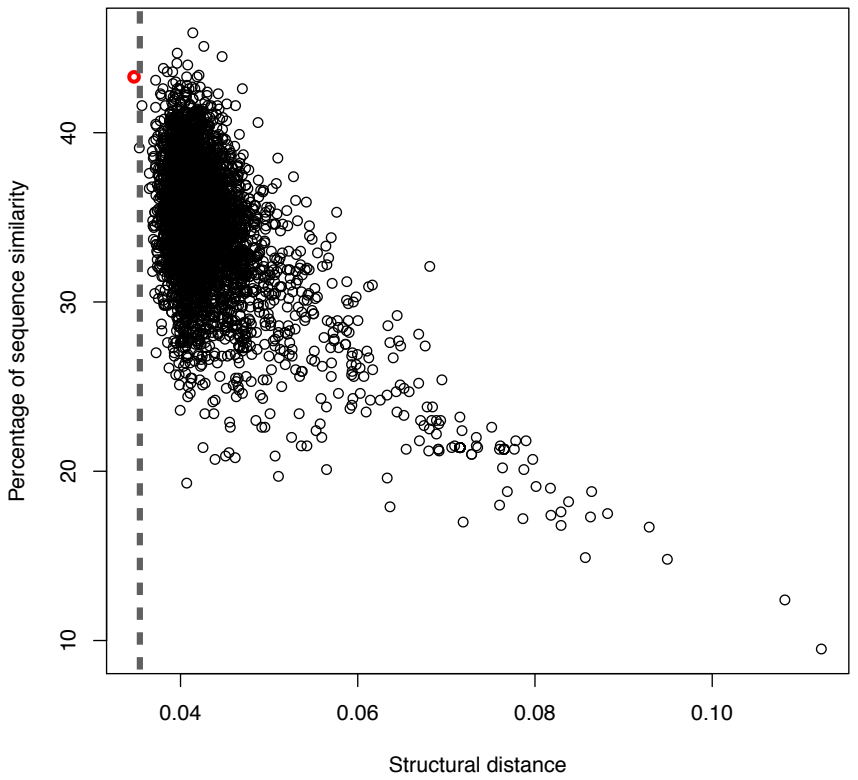
23. Lu, K., Heng, X. & Summers, M. F. Structural Determinants and Mechanism of HIV-1 Genome Packaging. *J. Mol. Biol.* **410,** 609–633 (2011).
24. Pollom, E. *et al.* Comparison of SIV and HIV-1 Genomic RNA Structures Reveals Impact of Sequence Evolution on Conserved and Non-Conserved Structural Motifs. *PLoS Pathog.* **9,** e1003294 (2013).
25. Rath, T. M. & Manmatha, R. Word image matching using dynamic time warping. in **2,** II-521-II-527 (IEEE Comput. Soc, 2003).
26. Keogh, E. J. & Pazzani, M. J. Scaling up dynamic time warping for datamining applications. in 285–289 (ACM Press, 2000). doi:10.1145/347090.347153
27. Cirillo, D. *et al.* Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. Methods* **14,** 5–6 (2017).
28. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
29. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519,** 486–490 (2015).
30. Andronescu, M., Bereg, V., Hoos, H. H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* **9,** 340 (2008).
31. Wu, Y. *et al.* Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res.* **43,** 7247–7259 (2015).
32. Lange, S. J. *et al.* Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.* **40,** 5215–5226 (2012).
33. Agostini, F., Cirillo, D., Bolognesi, B. & Tartaglia, G. G. X-inactivation:

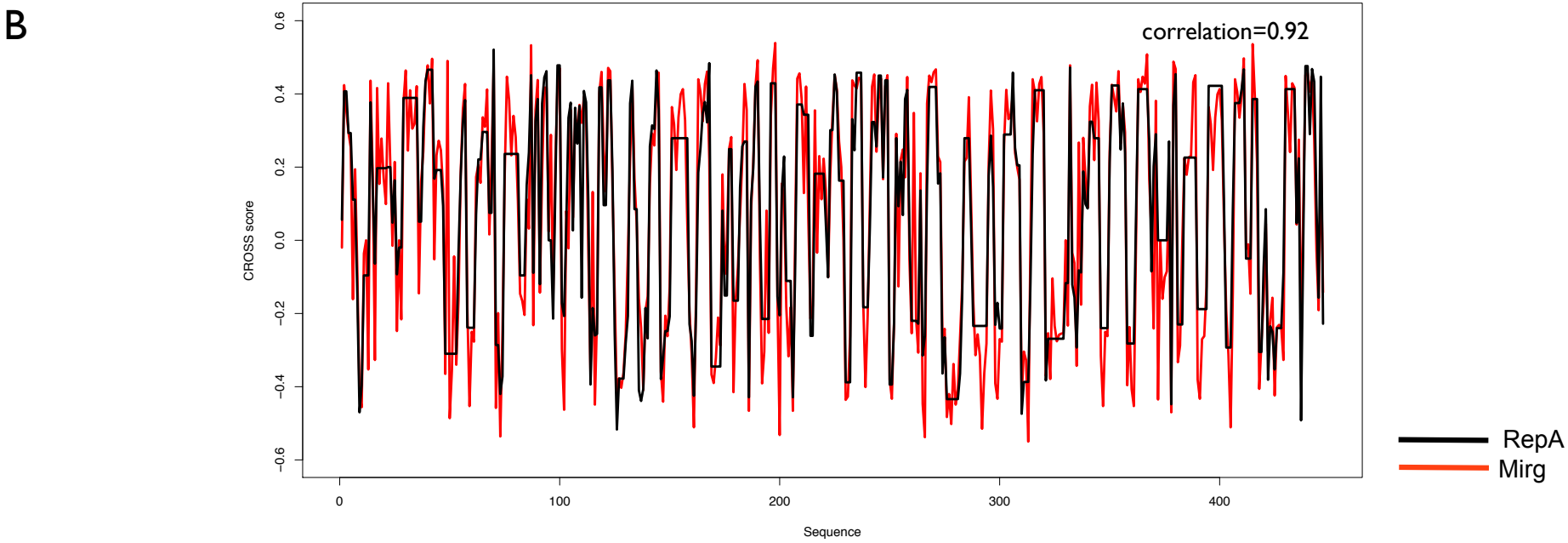
quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res*. **41,** e31 (2013).

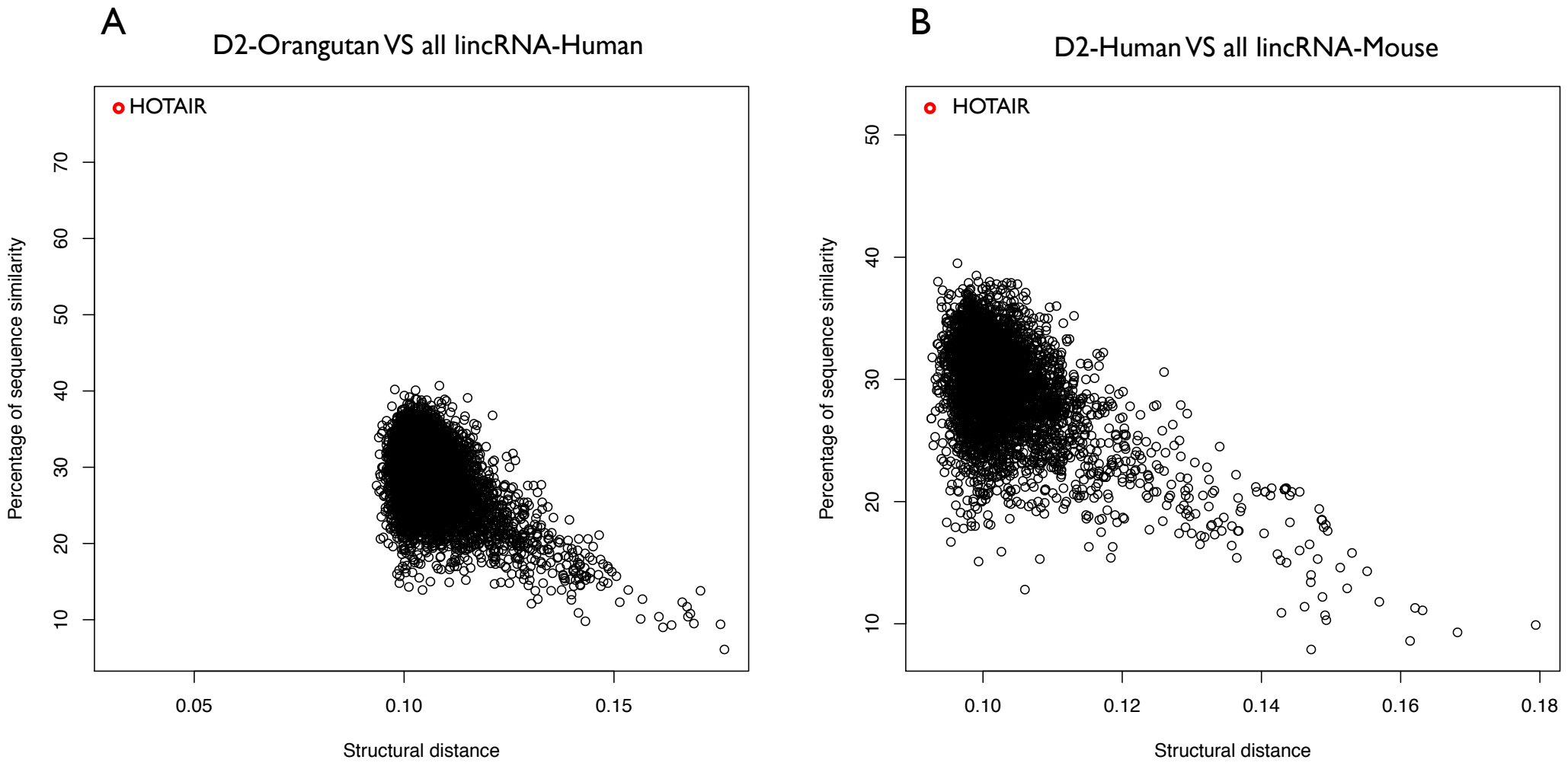

A
CROSS
0.04
0.06
0.08
0.10
correlation=0.91
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Crystallographic structure
B
SHAPE
0.1
0.2
0.3
0.4
0.5
correlation=0.50
Crystallographic structure
C
Crystals
Sequence similarity<40; Structural distance<0.05
Sequence similarity<30; Structural distance<0.1
Sequence similarity<22; Structural distance<0.2
correlation=-0.80
Structural distance
Percentage of sequence similarity
D
CROSS
correlation=-0.83
Structural distance
Percentage of sequence similarity
E
SHAPE
correlation=-0.38
Structural distance
Percentage of sequence similarity

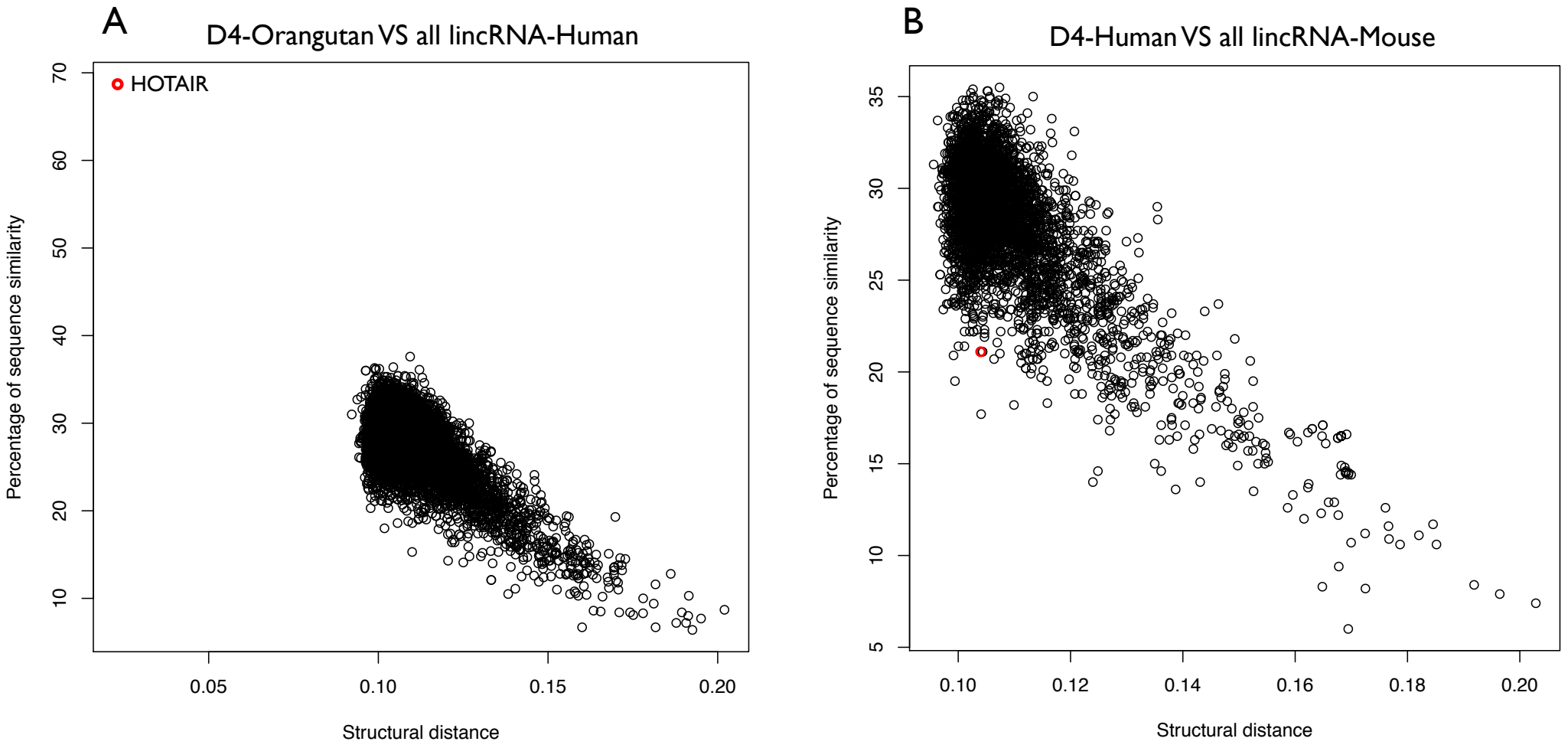

A
RepA-Orangutan VS all lincRNA-Human
Xist
Percentage of sequence similarity
20
40
60
80
0.05
0.10
0.15
Structural distance
B
RepA-Baboon VS all lincRNA-Human
Xist
Percentage of sequence similarity
20
40
60
80
0.04
0.06
0.08
0.10
0.12
Structural distance

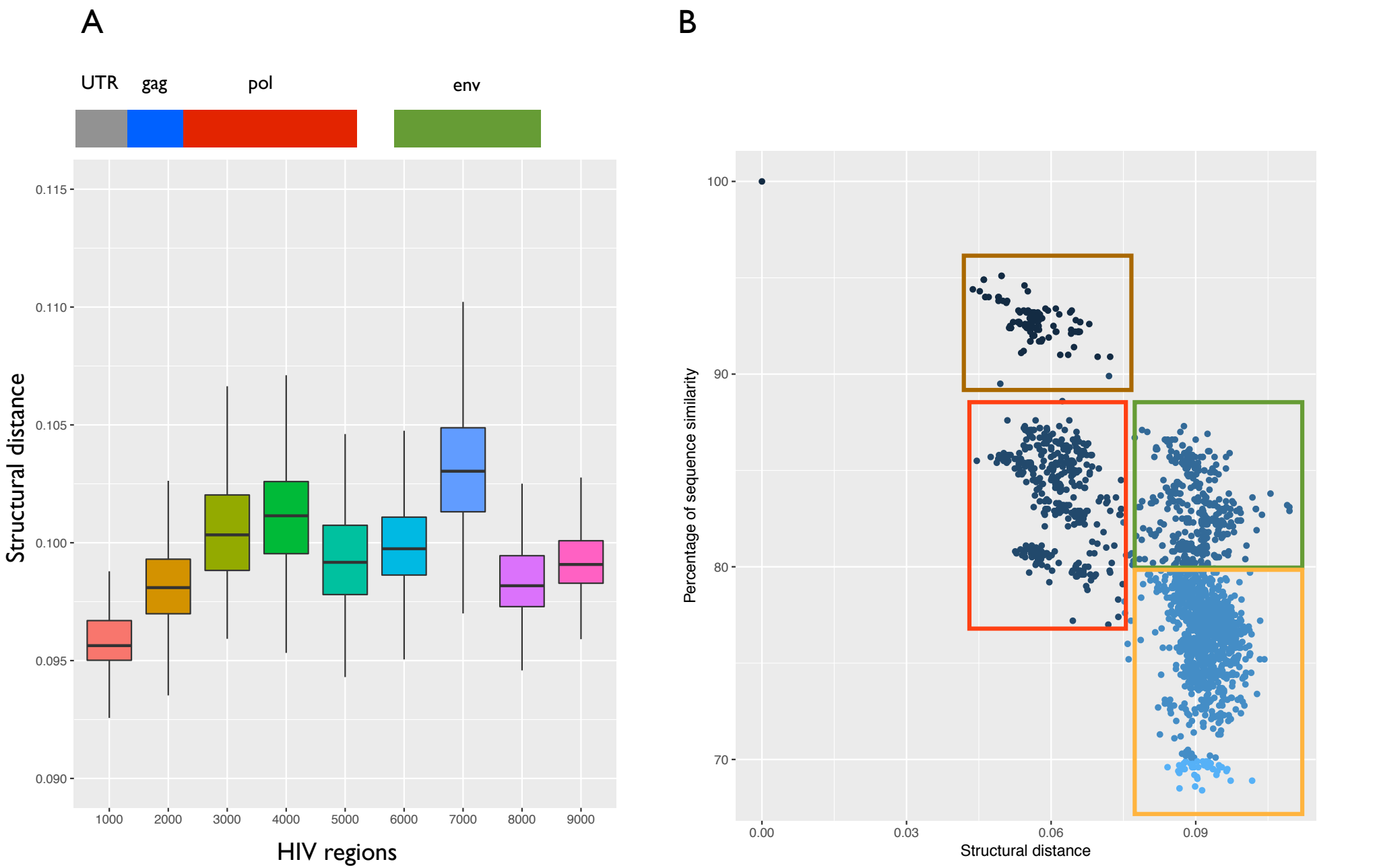

A
RepA-Human VS all lincRNA-Mouse
Percentage of sequence similarity
Structural distance
10
15
20
25
30
35
40
45
0.10
0.12
0.14
0.16
0.18


C
RepA-Human VS all lincRNA-Mouse structure only
Percentage of sequence similarity
Structural distance
10
20
30
40
0.04
0.06
0.08
0.10


B
correlation=0.92
CROSS score
Sequence
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0
100
200
300
400
RepA
Mirg


A
D2-Orangutan VS all lincRNA-Human
HOTAIR
Percentage of sequence similarity
10
20
30
40
50
60
70
0.05
0.10
0.15
Structural distance
B
D2-Human VS all lincRNA-Mouse
HOTAIR
Percentage of sequence similarity
10
20
30
40
50
0.10
0.12
0.14
0.16
0.18
Structural distance


A
D4-Orangutan VS all lincRNA-Human
HOTAIR
Percentage of sequence similarity
Structural distance
10
20
30
40
50
60
70
0.05
0.10
0.15
0.20
B
D4-Human VS all lincRNA-Mouse
Percentage of sequence similarity
Structural distance
5
10
15
20
25
30
35
0.10
0.12
0.14
0.16
0.18
0.20


A
UTR
gag
pol
env
Structural distance
0.115
0.110
0.105
0.100
0.095
0.090
1000
2000
3000
4000
5000
6000
7000
8000
9000
HIV regions
B
Percentage of sequence similarity
100
90
80
70
0.00
0.03
0.06
0.09
Structural distance