


RESEARCH ARTICLE

Open Access



# Detection of long non-coding RNA homology, a comparative study on alignment and alignment-free metrics

Teresa M. R. Noviello<sup>1,2</sup>, Antonella Di Liddo<sup>3</sup>, Giovanna M. Ventola<sup>4</sup>, Antonietta Spagnuolo<sup>5</sup>, Salvatore D'Aniello<sup>5</sup>, Michele Ceccarelli<sup>1,2</sup> and Luigi Cerulo<sup>1,2\*</sup> 

## Abstract

**Background:** Long non-coding RNAs (lncRNAs) represent a novel class of non-coding RNAs having a crucial role in many biological processes. The identification of long non-coding homologs among different species is essential to investigate such roles in model organisms as homologous genes tend to retain similar molecular and biological functions. Alignment-based metrics are able to effectively capture the conservation of transcribed coding sequences and then the homology of protein coding genes. However, unlike protein coding genes the poor sequence conservation of long non-coding genes makes the identification of their homologs a challenging task.

**Results:** In this study we compare alignment-based and alignment-free string similarity metrics and look at promoter regions as a possible source of conserved information. We show that promoter regions encode relevant information for the conservation of long non-coding genes across species and that such information is better captured by alignment-free metrics. We perform a genome wide test of this hypothesis in human, mouse, and zebrafish.

**Conclusions:** The obtained results persuaded us to postulate the new hypothesis that, unlike protein coding genes, long non-coding genes tend to preserve their regulatory machinery rather than their transcribed sequence. All datasets, scripts, and the prediction tools adopted in this study are available at <https://github.com/bioinformatics-sannio/lncrna-homologs>.

**Keywords:** Long ncRNA, Homology, String similarity

## Background

Recent advances in high-throughput sequencing have led to the discovery of a substantial transcriptome portion, across different species, that does not show encoding potential [1]. Long non-coding RNAs (lncRNAs) have emerged as important players in different biological processes, from development and differentiation to multilevel regulation and tumor progression [2]. The rapidly increasing number of evidence relating lncRNAs to important biological roles and diseases [3, 4] increased the interest in developing advanced computational approaches for their

identification and annotation [5–7]. However, despite their abundance and importance, their evolutionary history still remain unclear. As observed in many studies, the sequence conservation of lncRNAs is lower than protein coding genes, especially among distant species, and higher when compared to random or intronic sequences [8–10].

It has also been argued that conservation should be more preserved on RNA secondary structure functional sites than on nucleotide sequences [11]. However, as claimed recently by Rivas et al. [12], in several cases no evidence for selection on preservation of specific secondary structure has been reported till now. Conversely, promoter regions of lncRNAs appear to be generally more conserved than protein-coding genome counterparts, especially in mammalian species [1, 13]. In addition, lncRNA promoters show the presence of common binding

\*Correspondence: [lcerulo@unisannio.it](mailto:lcerulo@unisannio.it)

<sup>1</sup>Dep. of Science and Technology, University of Sannio, via Port'Arsa, 11, 82100 Benevento, Italy

<sup>2</sup>BioGeM, Institute of Genetic Research "Gaetano Salvatore", Camporeale, 83031 Ariano Irpino (AV), Italy

Full list of author information is available at the end of the article



sites for known transcription factors [14, 15], indicating that although the genomic sequences might not be highly conserved, their transcriptional machinery could be. These findings underpin the opportunity to investigate for a sequence similarity measure that is able to capture such kind of conservation, especially in promoter regions, and is computationally efficient for the detection of lncRNA homologs at genomic scale level among different species.

Current homology detection approaches, mainly based on alignment algorithms like Blast, assume the equivalence between homology and nucleotide sequence similarity. Among them, BlastR, a method that uses di-nucleotide conservation in association with BlastP to discover distantly related protein coding homologs [16], has been applied also for lncRNA homology prediction between human and other mammals [17, 18]. Approaches based on Blast-like algorithms are also the basis of lncRNA homology databases pipelines, such as NONCODE<sup>1</sup> and ZFLNC<sup>2</sup>. However, such sets of homologs certainly represent a fraction of the whole set of conserved functions because Blast-like algorithms are designed subsuming the evolution model of proteins that could not work for lncRNAs. So, new algorithms able to capture lncRNA conservation patterns are demanded to solve this gap.

In this study, we investigate whether other kind of sequence similarity metrics, not necessarily based on sequence alignment, can achieve such a task. Our investigation spans from alignment-based metrics, widely used for searching protein coding homologs, to a representative sample of alignment-free metrics, based on information theory, frequency analysis, and data compression. Specifically we consider two alignment-based metrics, Smith–Waterman (SW) and Damerau–Levenshtein (DLevDist) distance (Table 1); and 8 alignment-free metrics (Table 2), including: *n*-gram distance (qgram), Cosine similarity (cosine), Jaccard similarity (jaccard), Base–Base Correlation distance (BBC), Average Common Substring

distance (ACS), Lempel–Ziv complexity distance (LZ), Jensen–Shannon distance (JSD), and Hamming distance (HDist). Alignment-free metrics have been chosen by their popularity in other disciplines and because in our knowledge have never been adopted for homology identification.

We evaluate the metrics in three different species, human (hg38), mouse (mm10), and zebrafish (danRer10), against a manually curated gold-standard, originated from experimentally validated lncRNA homologs collected from the literature with the support of public lncRNA databases, such as lncRNadb [19], LNCipedia [20, 21], and lncRNome [22]. We show that some alignment-free metrics provide a better alternative to pairwise-alignment metrics, such as Smith–Waterman, especially between phylogenetically distant species. Surprisingly, in contrast with protein coding genes, lncRNA homologs exhibit higher alignment-free scores in promoter regions corroborating the hypothesis that lncRNA genes tend to preserve their regulatory machinery rather than their transcribed sequence.

## Results

Given two species  $S_1$  and  $S_2$ , Tables 1 and 2 report the set of metrics, we analyze, to detect whether two genes  $X \in S_1$  and  $Y \in S_2$  are homologs or not. For discussion purposes we consider three main factors that, as expected, could affect homology prediction: i) phylogenetic distance (close or distant), assuming human–mouse as close species, while mouse–zebrafish and human–zebrafish as distant species; ii) kind of transcript (protein coding or long non-coding); and iii) sequence region (promoter or transcript). In the following we report the results obtained with three empirical experiments aimed at evaluating the effectiveness of the proposed metrics: i) evaluation against a manually curated gold-standard originated from experimentally

**Table 1** Definition of the adopted homology metrics (Alignment-based)

Metric	Definition	Description
Smith–Waterman similarity	$SW(X, Y) = \max_{\substack{x \in seq(X) \\ y \in seq(Y)}} \left( \frac{sw(x,y)}{len(x)+len(y)} \right)$	The Smith–Waterman similarity $sw(x,y)$ is given by maximizing a score computed over a number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character [46]. Deletions/insertions (gaps) are penalized with a zero score, matches are rewarded with +5, and substitutions are penalized with -4 (NUC 4.4 substitution matrix). The time complexity is $O(len(x) \cdot len(y))$ .
Damerau–Levenshtein distance	$DLevDist(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} \left( \frac{dl(x,y)}{len(x)+len(y)} \right)$	The Damerau–Levenshtein distance $dl(x,y)$ is given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters [47]. The time complexity is $O(len(x) \cdot len(y))$ .

$X$  and  $Y$  are two candidate long non coding genes,  $seq(X)$  and  $seq(Y)$  are the sets of representative sequences of  $X$  and  $Y$  respectively (promoter or transcript),  $len(x)$  and  $len(y)$  are the lengths of sequences  $x$  and  $y$  respectively. Where applicable a metric is normalized with respect to the sum of sequence length [42] and is minimized (maximized) for distance (similarity) metrics among all couple of transcript sequences  $x \in seq(X), y \in seq(Y)$

**Table 2** Definition of the adopted homology metrics (Alignment-free)

Metric	Definition	Description
n-gram distance	$qgram_n(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} \left( \frac{\sum_i  q_i^x - q_i^y }{len(x) + len(y)} \right)$	A <i>n</i> -gram is a subsequence of <i>n</i> consecutive characters of a string [48]. If $\mathbf{q}^x = (q_1^x, q_2^x, \dots, q_k^x)$ is the <i>n</i> -gram vector of counts of <i>n</i> -gram occurrences in the sequence <i>x</i> the <i>n</i> -gram distance is given by the sum over the absolute differences $ q_i^x - q_i^y $ , where $q_i^x$ and $q_i^y$ are the <i>i</i> -th unique <i>n</i> -grams of <i>x</i> and <i>y</i> respectively obtained by sliding a window of <i>n</i> characters wide over <i>x</i> and <i>y</i> and registering the occurring <i>n</i> -grams. The time complexity is $O(len(x) \cdot len(y))$ .
Cosine similarity	$cosine_n(X, Y) = \max_{\substack{x \in seq(X) \\ y \in seq(Y)}} \frac{\mathbf{q}^x \cdot \mathbf{q}^y}{\ \mathbf{q}^x\  \ \mathbf{q}^y\ }$	The cosine similarity is the cosine of the angle between the two <i>n</i> -gram vectors $\mathbf{q}^x$ and $\mathbf{q}^y$ [40]. The time complexity is $O(len(x) + len(y))$ .
Jaccard similarity	$jaccard_n(X, Y) = \max_{\substack{x \in seq(X) \\ y \in seq(Y)}} \left( \frac{\sum_i (\mathbb{1}_{q_i^x > 0} + \mathbb{1}_{q_i^y > 0})}{\sum_i \mathbb{1}_{q_i^x > 0} \cdot \mathbb{1}_{q_i^y > 0}} - 1 \right)$	The Jaccard coefficient measures the similarity between two finite sets, and is defined as the size of the intersection divided by the size of the union of the sample sets [49]. The size is computed from the set of unique <i>n</i> -grams by means of $\mathbb{1}_{q_i^x > 0}$ , the indicator function having the value 1 if the <i>i</i> -th <i>n</i> -gram is present in <i>x</i> , 0 otherwise. The time complexity is $O(len(x) + len(y))$ .
Base-base correlation distance	$BBC(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} \sqrt{\sum_{i=1}^{16} (V_{x_i} - V_{y_i})^2}$	The Base-base correlation measures the sequence similarity by computing the euclidean distance between two 16-dimensional feature vectors, $V_x$ and $V_y$ , which contain all base pair mutual information [50]. The time complexity is $O(len(x) \cdot len(y))$ .
Average common substring distance	$ACS(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} \frac{1}{2} \left( \sum_{i=1}^{len(x)} \frac{lcs(x(i), y)}{len(x)} + \sum_{i=1}^{len(y)} \frac{lcs(y(i), x)}{len(y)} \right)$	The average common substring is the average lengths of maximum common substrings for constructing phylogenetic trees [51]. Specifically, the $lcs(x(i), y)$ ( $lcs(y(i), x)$ ) is the length of the longest common substring of <i>x</i> ( <i>y</i> ) starting at each position <i>i</i> of <i>x</i> ( <i>y</i> ) and exactly matching some substring in <i>y</i> ( <i>x</i> ). The time complexity is $O(len(x) + len(y))$ .
Lempel-Ziv complexity distance	$LZ(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} \frac{c(xy) - c(x) + c(yx) - c(y)}{\frac{1}{2}[c(xy) + c(yx)]}$	The Lempel-Ziv complexity distance is defined by considering the minimum number of components over all production histories of <i>x</i> and <i>y</i> , $c(x)$ and $c(y)$ and their concatenations, $c(xy)$ and $c(yx)$ [52]. The time complexity is $O(len(x) \cdot len(y))$ .
Jensen-Shannon distance	$JSD(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} \frac{1}{2} KL(V_x, V_M) + \frac{1}{2} KL(V_y, V_M)$	The Jensen-Shannon distance is computed by averaging the Kullback-Leibler Divergence (KL) of $V_x$ with respect to $V_M$ and $V_y$ with respect to $V_M$ , where $V_x$ and $V_y$ are the same 16-dimensional feature vectors defined for BBC, and $V_M = \frac{V_x + V_y}{2}$ [41]. The time complexity is $O(len(x) + len(y))$ .
Hamming distance	$HDist(X, Y) = \min_{\substack{x \in seq(X) \\ y \in seq(Y)}} hd(r(x), r(y))$	The Hamming distance is defined between two strings of the same length as the number of positions in which corresponding values are different. We adopt two bit strings of length <i>n</i> , namely $r(x)$ and $r(y)$ , representing the regulatory transcriptional machinery of <i>x</i> and <i>y</i> respectively, and <i>n</i> is the number of all transcription factors available in JASPAR [24]. Each position <i>i</i> of such bit strings is equal to 1 if the <i>i</i> -th transcription factor binds the promoter while 0 otherwise. The time complexity is $O(n)$ .

*X* and *Y* are two candidate long non coding genes,  $seq(X)$  and  $seq(Y)$  are the sets of representative sequences of *X* and *Y* respectively (promoter or transcript),  $len(x)$  and  $len(y)$  are the lengths of sequences *x* and *y* respectively. Where applicable a metric is normalized with respect to the sum of sequence length [42] and is minimized (maximized) for distance (similarity) metrics among all couple of transcript sequences  $x \in seq(X), y \in seq(Y)$

validated lncRNA homologs (Additional file 4: Table S1), ii) evaluation against NONCODE and ZFIN public annotation databases providing lncRNA homologous associations among different species detected with a Blast like pipeline, and iii) evaluation of functional concordance that looks at protein coding genes localized in the proximity of lncRNAs and measures their Gene Ontology term enrichment.

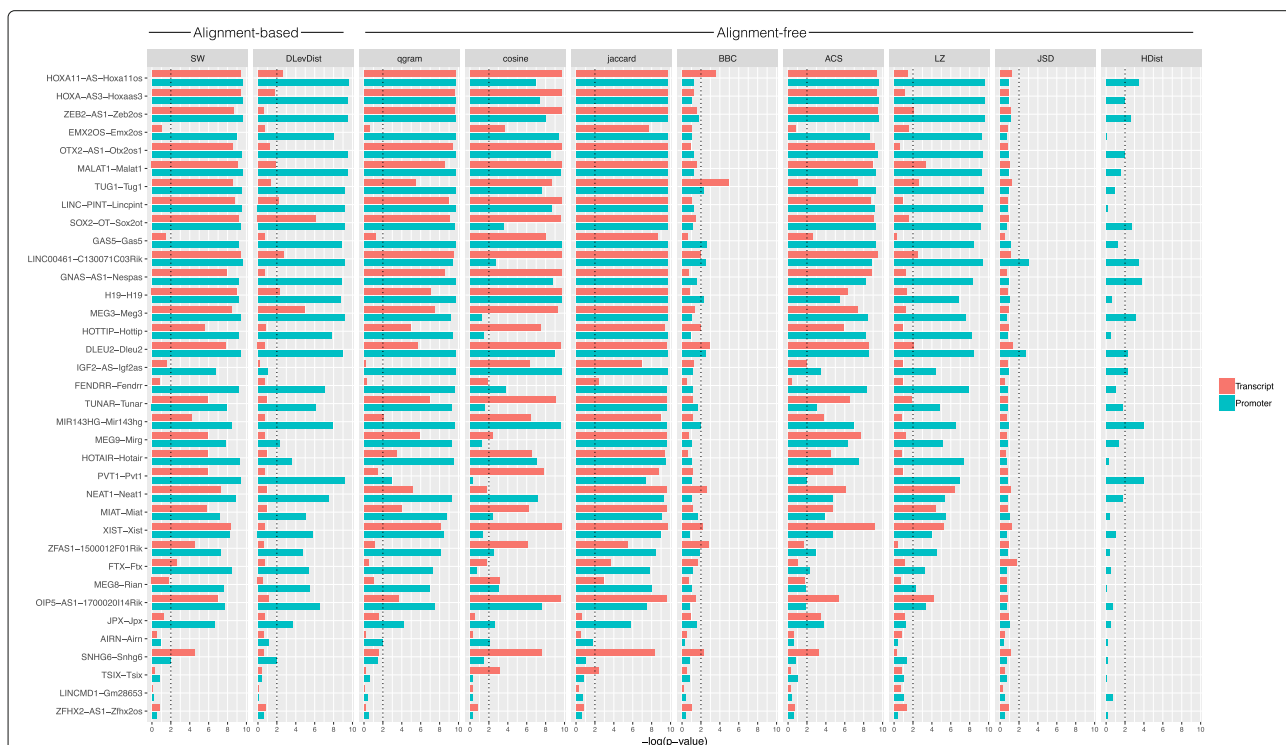
**Metrics evaluation on manually curated gold-standard**

Figures 1, 2 and 3 show, respectively for human-mouse, mouse-zebrafish, and human-zebrafish, the  $-\log(pvalue)$  for each considered metric (Tables 1 and 2) estimated by permutation test over a null distribution of non-homologous pairs randomly selected. The aim is to estimate to which extent a candidate metric is able to separate the true homologous pair from a huge set of random selected non-homologous pairs (permutation test). The set of non-homologous pairs are constructed by fixing a lncRNA candidate in a species and selecting a random set of sequences, approximately of the same length, in the other species known to be not homologous. Metrics depending on parameters were customized accordingly to obtain the best possible results. Specifically, for SW, we estimated the best levels of match

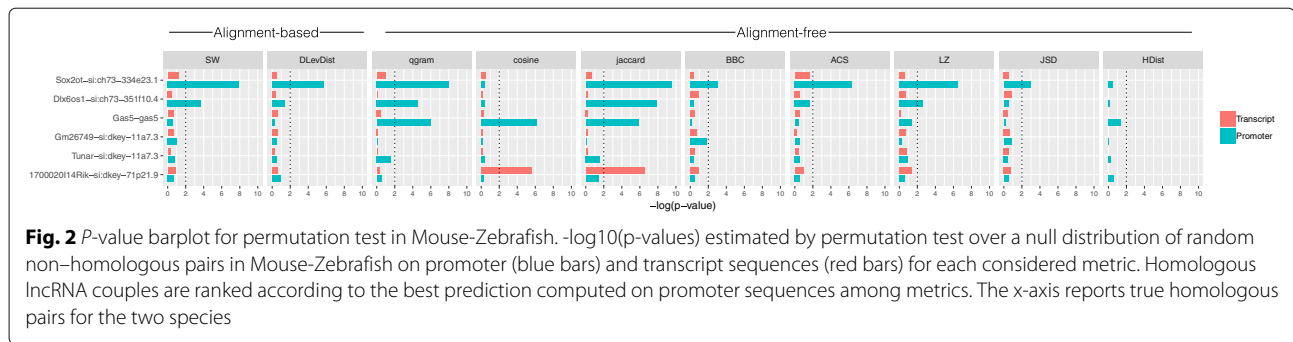
gain and gap/mismatch penalty with a grid searching procedure and for HDist, we adopted the MEME FIMO tool [23] with JASPAR positional frequency matrices (PFMs) [24]. The set of non-homologous pairs is ranked according to the best prediction computed on promoter sequences among metrics.

In closer related species (human-mouse), no distinction can be observed between alignment-based and alignment-free metrics. Figure 1 shows more than 23 out of 36 true homologous pairs with a  $p$ -value  $\leq 0.01$  in both alignment-based and almost all alignment-free metrics. Conversely, alignment-free metrics, especially jaccard and qgram, are more suitable among phylogenetically distant species. Jaccard exhibits a  $p$ -value  $\leq 0.01$  in 3 out of 6 true homologous pairs (Figs. 2 and 3). Instead, some metrics, such as DLevDist, BBC and JSD, are less powerful to detect homologous lncRNAs.

Moreover some couples failed to be detected regardless to the used metrics or sequence region. For example, for ZFHx2-AS1-Zfhx2os (Fig. 1) the literature suggests that a conservation of transcriptional profiles could be observed and that only a small genomic region, which perhaps contains important signals for the antisense transcription, could be considered conserved between human and mouse [25]. Similarly, the conservation of TUNAR



**Fig. 1**  $P$ -value barplot for permutation test in Human-Mouse.  $-\log_{10}(p\text{-values})$  estimated by permutation test over a null distribution of random non-homologous pairs in Human-Mouse on promoter (blue bars) and transcript sequences (red bars) for each considered metric. Homologous lncRNA couples are ranked according to the best prediction computed on promoter sequences among metrics. The x-axis reports true homologous pairs for the two species



involves only a small transcript region (about the 8% of the entire human sequence) that interacts with several RNA-binding proteins (as PTBP1 and hnRNP-K) responsible of functional conservation in all the considered species [26].

The sequence region (transcript vs. promoter) seems to play an important role only in phylogenetically distant species, with the exception of few cases. In Fig. 1 the number of significant true homologous pairs detected by each metric is higher for promoters in 5 cases out of 10 in human-zebrafish (Fig. 2), while such cases are 8 out of 10 in mouse-zebrafish (Fig. 3).

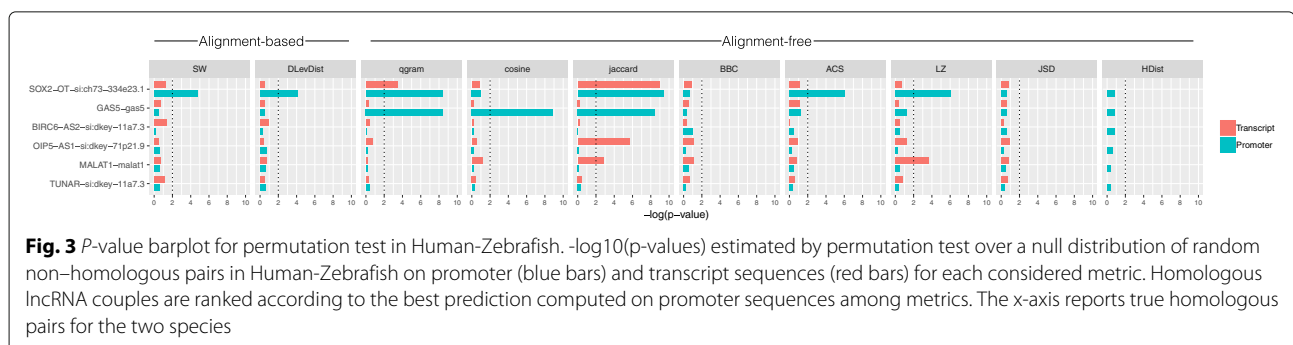
In phylogenetically close species (human-mouse), only few cases are affected by sequence region. For example, promoter sequence seems to be crucial for the functional maintenance of JPX (XIST Activator) in mammal species, differently from TSIX (XIST Antisense RNA), where the transcript provides uniquely the information of conservation. According to the corresponding literature, the promoter of JPX has been shown to interact with the Xist promoter in undifferentiated embryonic stem cells [27], while TSIX seems to be involved in the modulation of chromatin modification status of Xist promoter, suggesting a conserved function in mammals carried by the transcript structure [28].

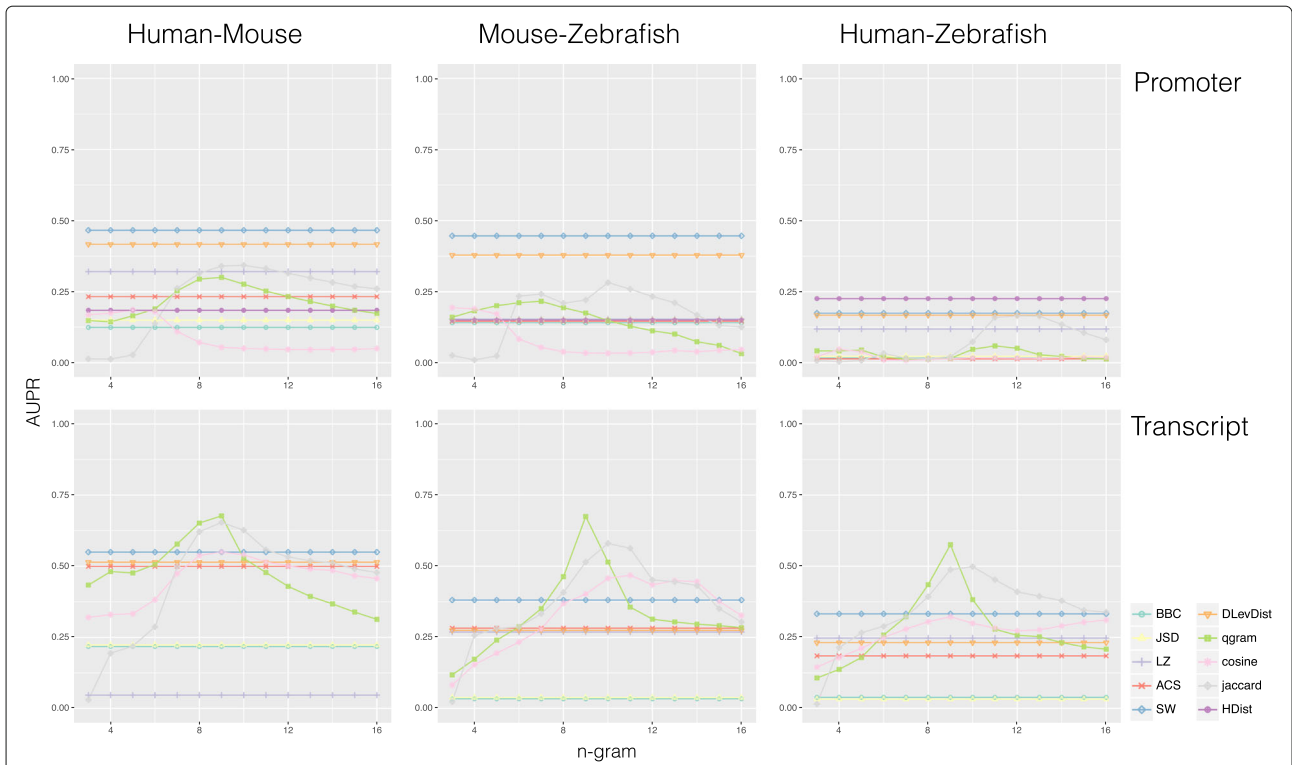
In distant species, alignment-based metrics are able to detect a lower number of homologous lncRNAs. This is probably related to the regulatory machinery that alignment-based metrics are less prone to detect.

### Consensus with NONCODE and ZFLNC pipelines

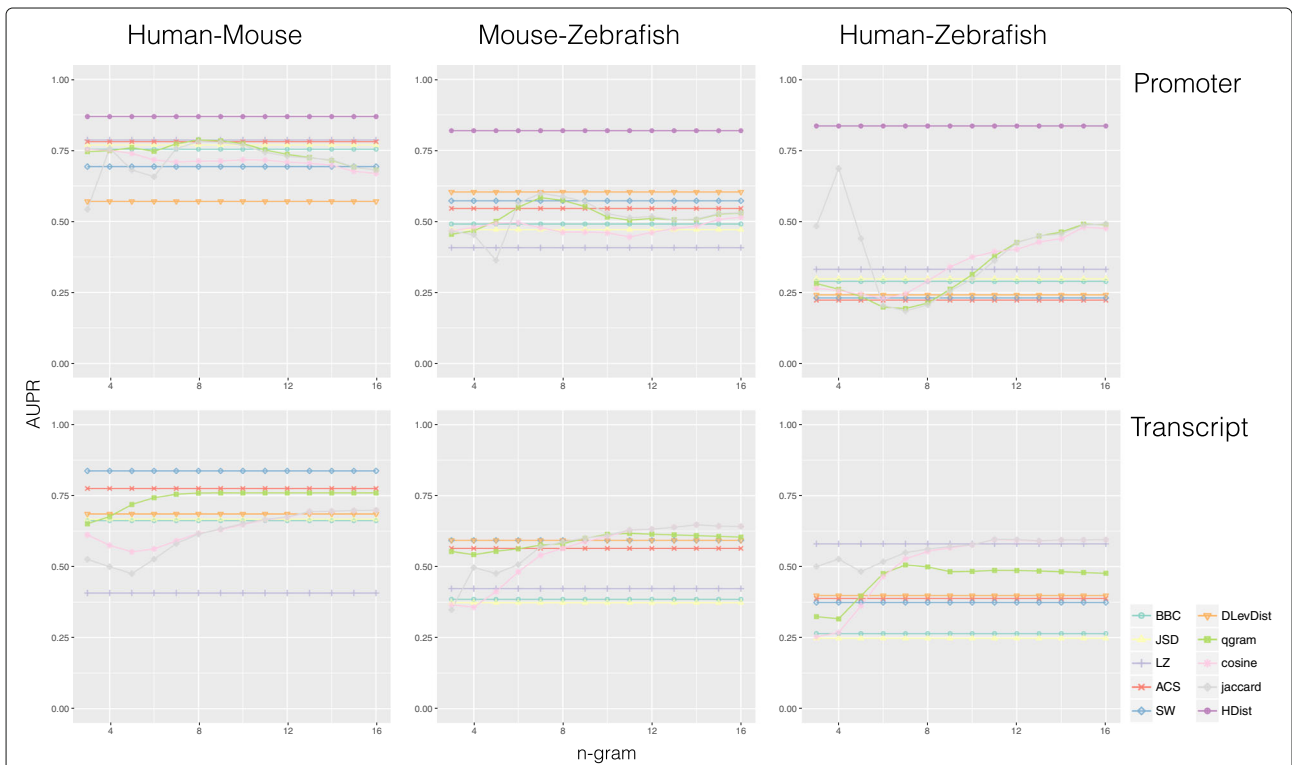
Figures 4 and 5 show the prediction performances, in terms of AUPR (Area under the Precision-Recall curve) plots, obtained by each metric with two database annotations, respectively NONCODE and ZFLNC. The x-axis reports the number *n* of consecutive characters considered for gram-based metrics. This means that remaining metrics are shown as horizontal lines since they do not depend on *n*. As baseline comparison, we computed AUPR also for a random set of protein coding genes (Additional file 1: Figure S1). Additional files 2: Figure S2 and 3: Figure S3 show also the ROC curves obtained respectively in NONCODE and ZFLNC.

SW, jaccard and cosine with *n* greater than 10 perform well when applied to protein coding transcript sequences, confirming that those metrics, in particular SW, are suitable for identifying homologous coding gene in both phylogenetically close and distant species. An opposite behaviour can be observed when comparing promoter sequences. In both phylogenetically close and distant species, the similarity of promoter regions seems to predict better the homology of lncRNAs rather than protein coding genes. In particular, HDist results to be the best predictor in ZFLNC (Fig. 2), reflecting the evidences regarding regulatory programs [29] and conservation status [1, 30] of lncRNAs with respect to protein coding genes. Furthermore, according to the manually curated gold-standard results, some metrics, such as BBC, JSD and LZ, seem to be not suitable for the detection of





**Fig. 4** NONCODE AUPR plots. Metric prediction performance computed on promoter and transcript sequences for NONCODE lncRNA homologs (AUPR on y-axis and  $n$ , the number of consecutive nucleotides in  $n$ -gram metrics, on x-axis)



**Fig. 5** ZFLNC AUPR plots. Metric prediction performance computed on promoter and transcript sequences for ZFLNC lncRNA homologs (AUPR on y-axis and  $n$ , the number of consecutive nucleotides in  $n$ -gram metrics, on x-axis)



homology, both in protein coding genes and in lncRNAs (AUPR less than 0.5 in mouse–zebrafish and less than 0.4 in human–zebrafish).

The conservation degree of lncRNA homologs is mainly affected by evolution distance, reflecting the evidences, shown also in the manually curated gold-standard, that lncRNAs evolve more rapidly. It is possible to observe that AUPR decreases with the increase of species distance for almost all metrics. For example, the AUPR of SW in NONCODE decreases from a 0.55 in human–mouse to 0.45 in mouse–zebrafish and to 0.33 in human–zebrafish (Fig. 1). While, the AUPR of jaccard and cosine in ZFLNC decrease from a 0.78 and 0.77 in human–mouse to 0.64 and 0.61 in mouse–zebrafish and to 0.59 and 0.50 in human–zebrafish, respectively.

Although semi-automatic generated gold-standards present major biases related to underlying automatic pipelines based on BLAST, some of conclusions, drawn with the manually curated gold-standard, are still supported, making the empirical evidence reinforced by a more representative statistical population.

#### Genome functional concordance analysis

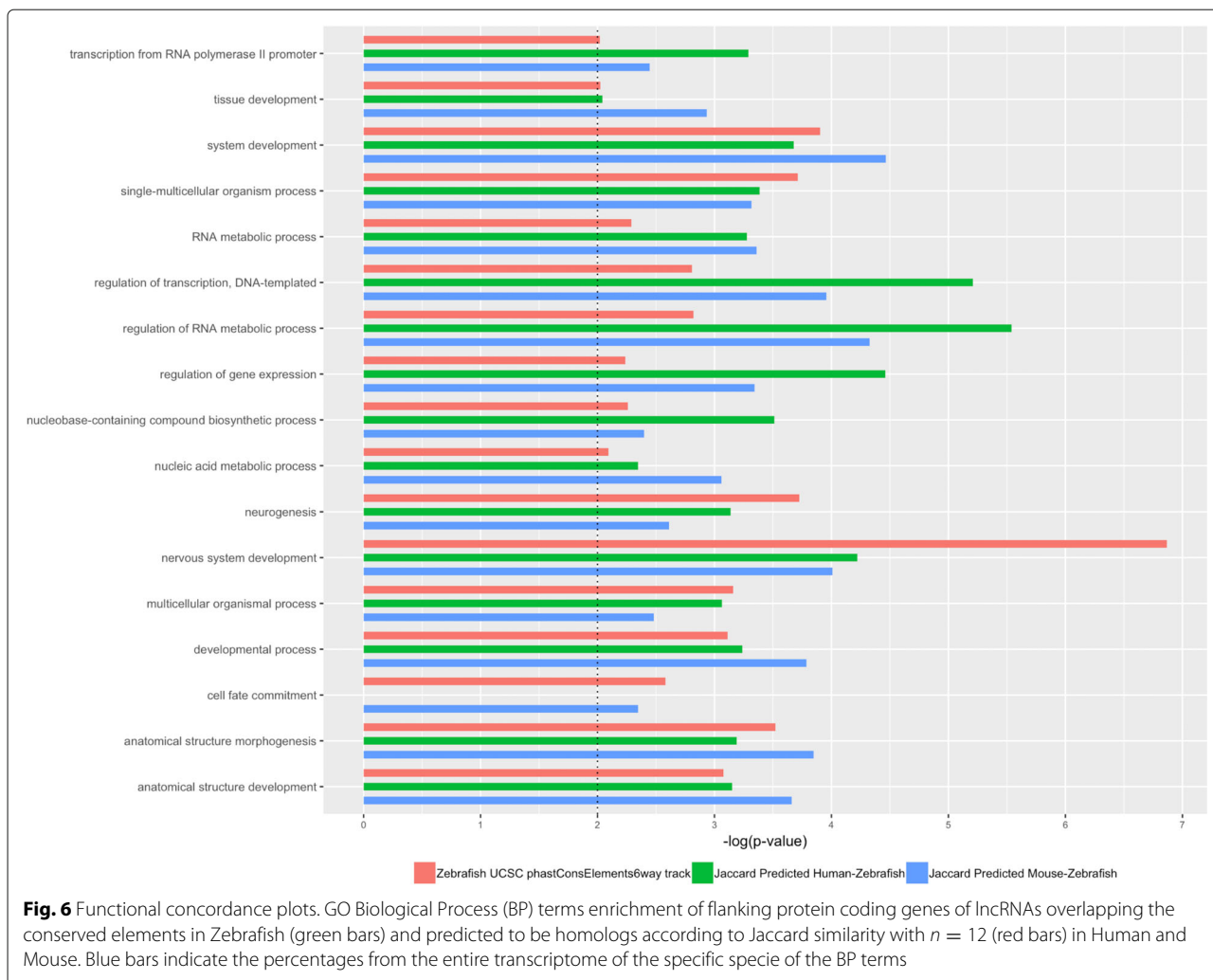
In order to assess the ability of alignment-free metrics to predict conservation of lncRNAs also regarding to their known and preserved biological functionality, we performed a GO enrichment analysis considering the nearest protein coding genes flanking the sets of zebrafish lncRNAs predicted to be orthologs in human and mouse (using jaccard with  $n = 12$ ). We adopted jaccard similarity as predictor since this metric in the previous empirical analyses showed in average a good prediction performance, but similar results can be obtained also with other alignment-free metrics (data not shown). As baseline, we considered the protein coding genes flanking the lncRNAs that overlap the most significantly conserved elements produced by the phastCons program [31] from zebrafish genome. Significantly enriched GO Biological Process (BP) terms ( $p$ -value  $\leq 0.01$ ) were obtained using DAVID functional annotation tool [32] and redundant enriched GO terms were removed using Revigo [33] (Additional file 5: Table S2). For each enriched GO category, the percentages of genes overlapping the most significantly conserved elements are also shown. Figure 6 shows the grouped BP terms that resulted to be enriched in all three considered sets: the jaccard predicted zebrafish lncRNA orthologs in human and mouse, and the phastCons conserved lncRNAs. As expected and in accordance to several studies describing lncRNA functional roles shared by different species [34–37], the enriched categories include development at several stages, regulation of transcription, and metabolic processes. On average, it can be observed an increment in terms of enrichment of the ultra-conserved GO terms considering the

sets of zebrafish lncRNAs predicted to be orthologs in human and mouse. However, it is not surprising that in few cases the GO term enrichment related to the ultra-conserved set is higher than the ones predicted using jaccard similarity. For example, it is known that lncRNAs play critical roles in the development of nervous system (neurogenesis) and that approximately 40% of lncRNAs are expressed in the brain in a tissue specific manner [17]. Moreover, these brain-specific lncRNAs show the highest signals of evolutionary conservation in comparison with those expressed in other tissues [38]. Figure 7 shows the percentages of predicted zebrafish lncRNA orthologs in human and mouse conserved or not with a zebrafish phastCons element and the corresponding percentages of flanking coding genes overlapping or not the same regions of conservation. The observed similarity at functional level in both species given by the GO enrichment analysis is not due to an over-representation of conserved lncRNA orthologs (35% in Human and 36% in Mouse). As expected, the high number of flanking coding genes within the zebrafish phastCons elements reflect the general feature of lncRNAs to be involved in vertebrate shared functional processes through *cis* expression regulation of nearby conserved genes. This result constitutes a further proof that alignment-free metrics, such as Jaccard similarity, work alongside typical approaches based on pure conservation among species, and are able to identify additional orthologs not included in the typical multi-alignment conservation track.

#### Discussion

In this study, we provide a systematic assessment of alignment-based and alignment-free metrics to investigate the conservation of lncRNAs looking at both promoter and transcript sequences in human, mouse and zebrafish. We evaluate the metrics against a manually curated gold-standard of validated lncRNA homologs available in literature. We show how alignment-free metrics could represent a powerful alternative to alignment metrics to detect lncRNA homology, especially in phylogenetically distant species and promoter regions. Despite the under-representation of considered gold-standard, alignment-free metrics, and in particular jaccard, could represent an optimal tradeoff between efficiency and efficacy for large scale genome annotation.

These findings are also supported by an extended empirical evaluation on two semi-automatic generated gold-standard, collected from lncRNA annotation databases as NONCODE and ZFLNC. It is important to specify that, although the necessity of retrieving an increased number of homologous lncRNA couples than that collected in the manually curated gold-standards, the semi-automatic generated gold-standard present several



weaknesses, due to the massive automatic Blast based pipeline biases.

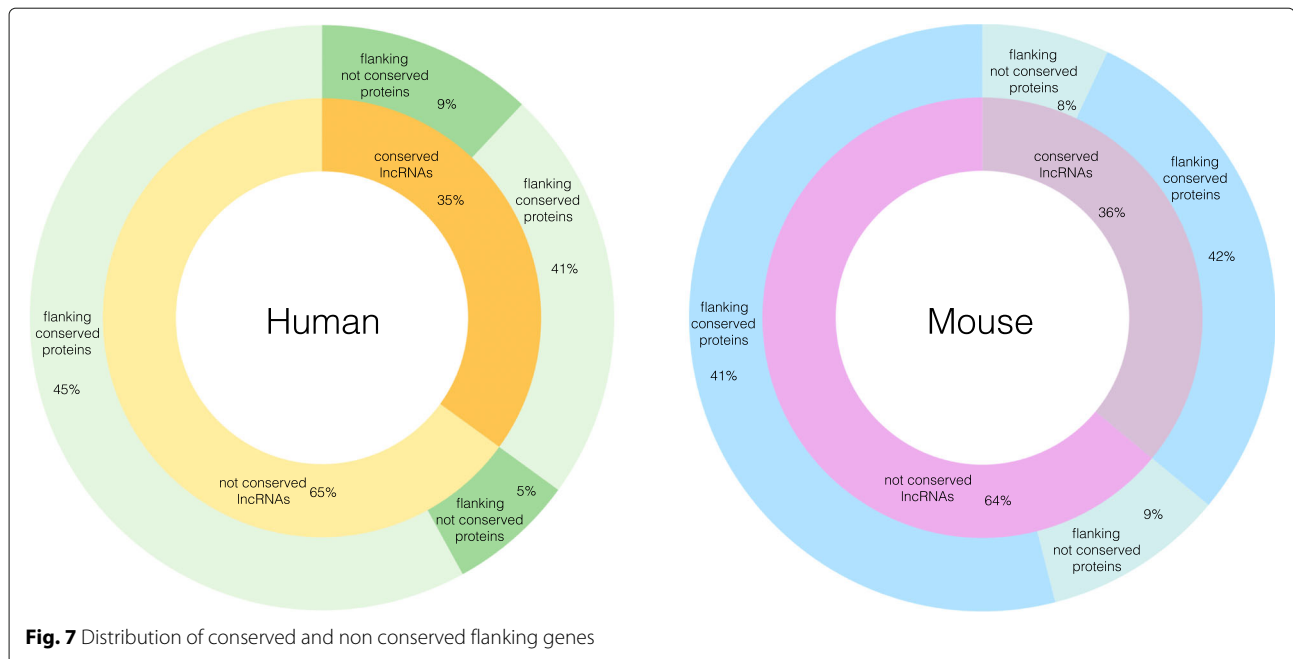
Our results reflect the rapid evolution of lncRNAs, divergent even between closely related species, confirmed by the fact that 81% of lncRNA families are only primate specific [17]. The promoter regions of lncRNA genes are generally more conserved than promoters of protein-coding genes [1] and encode crucial information that is better detected with alignment-free metrics, such as jaccard, suggesting a sustained selective pressure acting on these sequences. The evolution of transcription factor binding sites follow usually patterns marked by relocations and transpositions inside the promoter region. This preserves the regulatory machinery but limit sub-sequence similarity. Alignment-based metrics in preserving the relative order of common sub-sequences are able to detect point mutations, deletion, and insertion of small sequences but are not able to detect re-locations, crossovers, and/or transpositions as alignment-free metrics can do. Genome functional concordance analysis

confirm that conservation captured at promoter level by alignment-free metrics is highly consistent with the preservation of their biological functionality between species carried by coding genomic neighbourhood. This make us to suppose that lncRNA homologs tend to preserve their regulatory relationships more than their transcribed sequence.

### Conclusions

We proposed the use of alignment-free metrics to investigate the mechanism of conservation of long non-coding RNAs in three different species. To some extent, we found that n-gram metrics, when applied to promoter regions, are able to capture lncRNA homology associations between close and distant species. The obtained results persuaded us to formulate a hypothesis of conservation schema that impacts the promoter regions of lncRNAs. This mechanism suggests that lncRNAs tend to preserve the regulatory relationship with transcription factors rather than the information encoded in





their sequence. As our results are limited to the three species, human, mouse, and zebrafish, it is unquestionable that more data on different species and a larger manually curated gold-standard are crucial to generalize the mechanism of conservation governing the evolution of lncRNAs.

## Methods

### Sequence similarity metrics

Given two species  $S_1$  and  $S_2$ , Tables 1 and 2 report the set of metrics, we analyze, to detect whether two genes  $X \in S_1$  and  $Y \in S_2$  are homologs or not. We consider two alignment-based metrics, Smith–Waterman similarity and Damerau–Levenshtein distance (Table 1), widely adopted to detect protein coding homology [39], and several alignment-free metrics (Table 2), including:  $n$ -gram and common substring based distances, adopted in text mining and information retrieval [40]; two factor frequencies distances, Base–base correlation and Jensen–Shannon Divergence test, adopted in genome comparison [41]; Lempel–Ziv complexity distance based on data compression; and Hamming distance adapted to compute the concordance between regulatory transcriptional machinery of promoter sites. To make a measure comparable among sequences with different lengths, where applicable, a metric is normalized with respect to the sum of sequence lengths [42]. A gene  $X$  is modeled as a set of sequences  $seq(X)$  extracted from a genome. In particular, we consider two types of sequence sets: the set of transcribed sequences and the set of promoter regions. A transcribed sequence is constructed by merging all exons

belonging to that transcript, while a promoter region is built by considering the conventionally 2000 bp up and 1000 bp down stream from the transcription starting site. A metric is computed for all possible pairs of sequences belonging to the two sets representing the two candidate genes. Among all measures the minimum is considered if the metric is defined as a distance, instead the maximum if the metric is defined as a similarity.

### Metrics evaluation on manually curated gold-standard

We evaluate the metrics in three different species, human (hg38), mouse (mm10), and zebrafish (danRer10), against a manually curated gold–standard, originated from experimentally validated lncRNA homologs (Additional file 4: Table S1). It has been collected from the literature with the support of: lncRNAdb [19], a database that provides annotations of eukaryotic lncRNAs; LNCipedia [20, 21]; and lncRNome [22], a knowledge-base compendiums of human lncRNAs. Table 3 reports the number of collected lncRNA homologs between human and mouse, mouse and zebrafish, and human and zebrafish.

Due to the limited number of collected homologous pairs, we report to which extend ( $p$ -value) a candidate metric is able to separate the true homologous pair from a huge set of random selected non-homologous pairs (permutation test). The set of non-homologous pairs are constructed by fixing a lncRNA candidate in a species and selecting a random set of sequences, approximately of the same length, in the other species known to be not homologous.

**Table 3** Annotated homologous genes between species in manual curated gold-standard

Gene class Specie1	Gene class Specie2	Human Mouse	Human Zebrafish	Mouse Zebrafish
Antisense	Antisense	12	2	1
Antisense	lincRNA	8	2	0
lincRNA	Antisense	1	1	2
lincRNA	lincRNA	20	2	2
Overlapping	Overlapping	1	1	1
Total lincRNAs		42	8	6
Protein coding	Protein coding	12998	10209	10126

### Consensus with NONCODE and ZFLNC pipelines

NONCODE and ZFLNC are public annotation databases providing lincRNA homologous associations among different species. Such associations are detected by classical sequence homology pipelines based on multi alignment metrics such as those adopted to identify protein coding homologs. Specifically, NONCODE provides conservative and evolutionary status of stored lincRNAs through a genome comparison conservation analysis based on UCSC LiftOver tool; while, ZFLNC provides zebrafish lincRNA functions and homologs identified through a pipeline based on: BLASTn, collinearity with conserved coding gene, and overlap with multi-species ultra-conserved non-coding elements.

Although such databases cannot be adopted as a typical gold-standard because the sample is biased on the similarity metric used in the original discovery pipelines, we still perform an evaluation against database annotations. The aim is to show to which extent alignment-free metrics reproduces the state of art of lincRNA homologs annotated with pipelines based essentially on alignment-based metrics.

From NONCODE we selected 882 human lincRNA sequences having 44 homologous counterparts in zebrafish and 523 in mouse. From ZFLNC we selected 676 zebrafish lincRNA sequences presenting a counterpart both in human and mouse. Prediction accuracy is evaluated with area under the Precision and Recall curve (AUPR), since it gives more information when dealing with highly skewed datasets [43, 44]. Specifically, we provide a normalized version of AUPR that takes into account the unachievable region in PR space, as proposed in Kendrick et al. [44], that allows to compare performances estimated on datasets with different class skews. In additional data we provide also ROC plots.

### Genome functional concordance analysis

It is generally assumed that homologous genes play similar biological roles in different species [45]. Since Gene

Ontology (GO) analysis can be considered as a good in-silico indicator of biological function, we provide an alternative assessment strategy that evaluates the functional concordance of lincRNA homologs candidates. This strategy, adopted similarly in Basu et al. [18], looks at protein coding genes localized in the proximity of lincRNAs (within a window of 1 mb) and measures their GO term enrichment in Biological Processes (BP) with DAVID tool [32].

As case study we evaluate the functional concordance on a set of lincRNA zebrafish homologous candidates predicted from a sample of 1000 random lincRNAs belonging to human and mouse. As baseline, we consider zebrafish lincRNAs belonging to ultra-conserved regions obtained with UCSC phastConsElements6way tracks. This provided us a set of enriched GO terms that can be assumed to be the most conserved biological function among the considered species [34–37]. The idea is to compare the baseline enrichment with the enrichment of predicted lincRNAs flanking protein coding genes. An increment of the latter enrichment means that predicted lincRNAs are able to capture additional flanking proteins not revealed in canonical phastConsElements6way tracks, corroborating the hypothesis that such lincRNAs, in controlling such flanking genes, should contribute to the ultra-conserved biological function.

### Endnotes

<sup>1</sup> <http://www.noncode.org>

<sup>2</sup> <http://www.zflnc.org>

### Additional files

**Additional file 1:** Additional Figure 1. Protein-coding gene AUPR plots. Metric prediction performance computed on promoter and transcript sequences for annotate protein-coding homologs (AUPR on y-axis and  $n$ , the number of consecutive nucleotides in  $n$ -gram metrics, on x-axis). (PDF 158 kb)

**Additional file 2:** Additional Figure 2. NONCODE ROC curves. ROC curves computed on promoter and transcript sequences for NONCODE lincRNA homologs (for  $n$ -gram metrics,  $n = 12$  has been chosen). (PDF 822 kb)

**Additional file 3:** Additional Figure 3. ZFLNC ROC curves. ROC curves computed on promoter and transcript sequences for ZFLNC lincRNA homologs (for  $n$ -gram metrics,  $n = 12$  has been chosen). (PDF 1580 kb)

**Additional file 4:** Additional Table 1. Manually curated gold-standard. Experimentally validated lincRNA homologs for the considered species. (XLSX 13 kb)

**Additional file 5:** Additional Table 2. GO biological process enriched terms. DAVID results for GO enrichment analysis of flanking proteins of Zebrafish lincRNA predicted to be homologous in Human (Sheet 1), Mouse (Sheet 2) and of lincRNA overlapping the conserved elements in Zebrafish (Sheet 3). (XLSX 21 kb)

### Acknowledgements

We would like to thank all reviewers for their valuable suggestions that helped to significantly improve this paper.

### Funding

This work was supported by a research project funded by MiUR (Ministero dell'Università e della Ricerca) under grant FIRB2012-RBFR12QW4I.

### Availability of data and materials

All datasets collected by the authors from public databases (Ensembl, UCSC, NONCODE, ZFLNC, lncRNadb, LNCipedia, and lncRNome), scripts, and the prediction tools adopted in this study are available at <https://github.com/bioinformatics-sannio/lncrna-homologs>.

### Authors' contributions

TMRN conducted the experiments and contributed to conceive the study and design the experiments. ADL contributed to discussions and to construct the gold-standards. GMV contributed to discussions and to construct the manual gold-standard. SDA and AS advised on biological interpretation of results. MC contributed to discussions and coordination of the study. LC conceived the study, designed the experiments, and coordinated the study. All authors accepted the final version of the paper and contributed to writing the manuscript. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Dep. of Science and Technology, University of Sannio, via Port'Arsa, 11, 82100 Benevento, Italy. <sup>2</sup>BioGeM, Institute of Genetic Research "Gaetano Salvatore", Camporeale, 83031 Ariano Irpino (AV), Italy. <sup>3</sup>Buchmann Institute for Molecular Life Sciences, Goethe University, Max-von-Laue-Straße 13, 60438 Frankfurt am Main, Germany. <sup>4</sup>Genomix4Life S.r.l., Via Salvador Allende, 84081 Baronissi (SA), Italy. <sup>5</sup>Dep. of Biology and Evolution of Marine Organisms, Stazione Zoologica "A. Dohrn", Villa Comunale, 80121 Napoli, Italy.

Received: 8 August 2018 Accepted: 19 October 2018

Published online: 06 November 2018

### References

- Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. *Science*. 2005;309(5740):1559–63.
- Mercer TR, Dinger ME, Mattick JS. Long non-coding rnas: insights into functions. *Nat Rev Genet*. 2009;10(3):155–9.
- Wapinski O, Chang HY. Long noncoding rnas and human disease. *Trends Cell Biol*. 2011;21(6):354–61.
- Gong J, Liu W, Zhang J, Miao X, Guo A-Y. lncnasnp: a database of snps in lncrnas and their potential functions in human and mouse. *Nucleic Acids Res*. 2014;43(D1):181–6.
- Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics*. 2013;14(Suppl 2):S7. <https://doi.org/10.1186/1471-2164-14-S2-S7>.
- Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK, DeepInc, a long non-coding rna prediction tool using deep neural network. *Netw Model Anal Health Inform Bioinforma*. 2016;5(1):21.
- Ventola GM, Noviello TM, D'Aniello S, Spagnuolo A, Ceccarelli M, Cerulo L. Identification of long non-coding transcripts with feature selection: a comparative study. *BMC Bioinformatics*. 2017;18(1):187.
- Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? evidence for selection within long noncoding rnas. *Genome Res*. 2007;17(5):556–65.
- Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincrnas in vertebrate embryonic development despite rapid sequence evolution. *Cell*. 2011;147(7):1537–50.
- Ma L, Bajic VB, Zhang Z. On the classification of long non-coding rnas. *RNA Biol*. 2013;10(6):925–34.
- Diederichs S. The four dimensions of noncoding rna conservation. *Trends Genet*. 2014;30(4):121–3.
- Rivas E, Clements J, Eddy SR. Lack of evidence for conserved secondary structure in long noncoding rnas. *Nat Methods*. 2017;14(1):45.
- Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Hanna JH, Regev A, Garber M. Evolutionary analysis across mammals reveals distinct classes of long noncoding rnas. *Genome Biol*. 2016;17(19).
- Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*. 2004;116(4):499–509.
- Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding rnas. *Cell*. 2009;136(4):629–41.
- Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, Beaudouin E, Bucher P, Notredame C. BlaStr—fast and accurate database searches for non-coding rnas. *Nucleic Acids Res*. 2011;39(16):6886–95. <https://doi.org/10.1093/nar/gkr335>.
- Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22(9):1775–89.
- Basu S, Müller F, Sanges R. Examples of sequence conservation analyses capture a subset of mouse long non-coding rnas sharing homology with fish conserved genomic elements. *BMC Bioinformatics*. 2013;14(7):14.
- Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. lncrnadb v2.0: expanding the reference database for functional long noncoding rnas. *Nucleic Acids Res*. 2011;39(Database issue):D146–51. <https://doi.org/10.1093/nar/gkq1138>.
- Volders P-J, Helsens K, Wang X, Menten B, Martens L, Gevaert K, Vandesompele J, Mestdagh P. Lncipedia: a database for annotated human lncrna transcript sequences and structures. *Nucleic Acids Res*. 2013;41(D1):246–51.
- Volders P-J, Verheggen K, Menschaert G, Vandepoel K, Martens L, Vandesompele J, Mestdagh P. An update on lncipedia: a database for annotated human lncrna sequences. *Nucleic Acids Res*. 2015;43(D1):174–80.
- Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B, Jain S, Sati S, Sengupta S, Sachidanandan C, et al. lncname: a comprehensive knowledgebase of human long noncoding rnas. *Database*. 2013;2013:034.
- Grant CE, Bailey TL, Noble WS. Fimo: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–8.
- Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Chèneby J, Kulkarni SR, Tan G, et al. Jasp 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res*. 2017;46(D1):260–6.
- Komine Y, Nakamura K, Katsuki M, Yamamori T. Novel transcription factor zfh-5 is negatively regulated by its own antisense rna in mouse brain. *Mol Cell Neurosci*. 2006;31(2):273–83.
- Lin N, Chang K-Y, Li Z, Gates K, Rana ZA, Dang J, Zhang D, Han T, Yang C-S, Cunningham TJ, et al. An evolutionarily conserved long noncoding rna tuna controls pluripotency and neural lineage commitment. *Mol Cell*. 2014;53(6):1005–19.
- Tsai C-L, Rowntree RK, Cohen DE, Lee JT. Higher order chromatin structure at the x-inactivation center via looping dna. *Dev Biol*. 2008;319(2):416–25.
- Senner CE, Brockdorff N. Xist gene regulation at the onset of x inactivation. *Curr Opin Genet Dev*. 2009;19(2):122–6.
- 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*. 2014;9(10):109443.
- Chiba H, Yamashita R, Kinoshita K, Nakai K. Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs. *BMC Genomics*. 2008;9(1):152.

31. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005;15(8):1034–50.
32. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc.* 2008;4(1):44.
33. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE.* 2011;6(7):21800.
34. Rinn JL, Chang HY. Genome regulation by long noncoding rnas. *Ann Rev Biochem.* 2012;81:145–66.
35. Fatica A, Bozzoni I. Long non-coding rnas: new players in cell differentiation and development. *Nat Rev Genet.* 2014;15(1):7.
36. Kornfeld J-W, Brüning JC. Regulation of metabolism by long, non-coding rnas. *Front Genet.* 2014;5:57.
37. Schmitz SU, Grote P, Herrmann BG. Mechanisms of long noncoding rna function in development and disease. *Cell Mol Life Sci.* 2016;73(13):2491–509.
38. Quan Z, Zheng D, Qing H. Regulatory roles of long non-coding rnas in the central nervous system and associated neurodegenerative diseases. *Front Cell Neurosci.* 2017;11:175.
39. Mount D. *Bioinformatics: Sequence and Genome Analysis*, 2nd. Long Island: Cold Spring Harbor Laboratory Press; 2013.
40. Baeza-Yates RA, Ribeiro-Neto B. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 1999.
41. Lin J. Divergence measures based on the shannon entropy. *IEEE Trans Inf Theory.* 1991;37(1):145–51.
42. Arslan AN, Egecioğlu Ö, Pevzner PA. A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics.* 2001;17(4):327–37.
43. Davis J, Goadrich M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning. ICML '06*. New York, NY, USA: ACM; 2006. p. 233–40.
44. Boyd K, Costa VS, Davis J, Page CD. Unachievable region in precision-recall space and its effect on empirical evaluation. In: *Proceedings of The... International Conference on Machine Learning. International Conference on Machine Learning*, vol. 2012. Edinburgh: NIH Public Access; 2012. p. 349.
45. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997;278(5338):631–7.
46. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981;147(1):195–7.
47. Damerau FJ. A technique for computer detection and correction of spelling errors. *Commun ACM.* 1964;7(3):171–6.
48. Cavnar WB, Trenkle JM, et al. N-gram-based text categorization. *Ann arbor mi.* 1994;48113(2):161–75.
49. Jaccard P. Nouvelles recherches sur la distribution florale. *Bull Soc Vaudense Sci Nat.* 1908;44:223–70.
50. Liu Z, Meng J, Sun X. A novel feature-based method for whole genome phylogenetic analysis without alignment: application to hev genotyping and subtyping. *Biochem Biophys Res Commun.* 2008;368(2):223–30.
51. Ulitsky I, Burstein D, Tuller T, Chor B. The average common substring approach to phylogenomic reconstruction. *J Comput Biol.* 2006;13(2):336–50.
52. Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics.* 2003;19(16):2122–30.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

