Environmental regulation of gene expression mediated by Long non-coding RNAs

Jingliang Kang[1], Arthur Chung[1], Sneha Suresh[1], Lucrezia L. Bonzi[1], Jade M. Sourisse[1], Sandra Ramirez[1], Daniele Romeo[1], Natalia Petit-Marty[1], Cinta Pegueroles[2], Celia Schunter[1*]

[1]Swire Institute of Marine Science, School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong SAR

[2]Department of Genetics, Microbiology and Statistics, Institute for Research on Biodiversity (IRBio), University of Barcelona, Barcelona, Spain

*Correspondence: celiaschunter@gmail.com

ABSTRACT

The majority of the transcribed genome does not have coding potential but is composed of non-coding transcripts that are involved in transcriptional and post-transcriptional regulation of protein-coding genes. Regulation of gene expression is important in determining the response of organisms to changes in the environment, and therefore their persistence as population or species under global change. However, long non-coding RNAs (lncRNAs) are scarcely studied especially in non-model organisms due to the lack of a reliable pipeline for their accurate identification and annotation. Here, we present a pipeline which uses a combination of alignment-dependent and independent methods for the identification of conserved and species-specific lncRNAs from RNA-Seq data. Validation of this pipeline was performed using existing RNA-Seq data from *Acanthochromis polyacanthus* brain tissue, identifying a total of 4,728 lncRNAs across the genome, the majority of which (3,272) are intergenic. To investigate the possible implications of these intergenic lncRNAs (lincRNAs), we estimated the expression changes of lincRNAs and coding genes in response to ocean acidification. We found lincRNAs which neighbour or possibly trans-regulate differentially expressed coding genes related to pH regulation, neural signal transduction and ion transport, which are known to be important in the response to ocean acidification in fish. Overall, this pipeline enables the use of existing RNA sequencing data to reveal additional underlying molecular mechanisms involved in the response to environmental changes by integrating the study of lncRNAs with gene expression.

Keywords: epigenetic regulation, annotation, RNA sequencing, lncRNAs, environmental change, ocean acidification

INTRODUCTION

Among pervasive genomic regions that can be transcribed, some encode long non-coding RNAs (lncRNAs), defined as RNAs longer than 200 nucleotides that are not translated into functional proteins (Statello et al., 2021). LncRNAs represent a highly heterogeneous class of transcripts mainly transcribed by RNA polymerase II and, usually, inefficiently spliced. While some lncRNAs remain in the nucleus, others are polyadenylated and exported to the cytoplasm (Statello et al., 2021). Although initially considered to be products of transcriptional noise or spurious transcription (Struhl, 2007), lncRNAs are now known to play critical roles in diverse biological processes, including DNA repair, proliferation, and embryonic development (Fernandes et al., 2019; Li et al., 2019; Vance & Ponting, 2014). LncRNAs are involved at many levels, including chromatin modifications, pre-transcription, transcription, and post-transcription through regulation of the associated gene expression (Gardini & Shiekhattar, 2015; Kornfeld & Brüning, 2014; Necsulea et al., 2014; Ulitsky, 2016; H. Xu et al., 2019; Zhu et al., 2013). In addition, lncRNAs are involved in specific physiological processes and gene dysfunction and can be used as biomarkers of certain diseases (Beck et al., 2018; Fernandes et al., 2019; Jiang et al., 2016).

LncRNAs are not as conserved as protein-coding sequences due to their fast turn-over (Lopez-Ezquerra et al., 2017; Pegueroles et al., 2019), and they tend to be cell and tissue specific (Cabili et al., 2015; Derrien et al., 2012). In addition, many lncRNAs are lowly expressed (Guo et al., 2020; Hezroni et al., 2015; Necsulea et al., 2014; Quinn et al., 2016). As such, despite the growing interest in lncRNAs and the availability of RNA sequencing data in many species, lncRNAs remain unexplored in most species, particularly in non-model species. To date, several bioinformatic tools have been developed to detect lncRNAs by computing coding potential score (CPS) of transcripts, either by using sequence alignments (alignment-dependent) or detecting intrinsic features of the input RNA sequences (alignment-free). In the alignment-dependent methods, such as PhyloCSF (Lin et al., 2011) and CPC (Kong et al., 2007), sequences are either aligned between species or to protein databases, which may be biased toward misclassifying species-specific or lowly conserved coding and non-coding transcripts. In contrast, although the alignment-free methods like CPAT (Wang et al., 2013) and FEELnc (Wucher et al., 2017) are useful to discriminate species-specific lncRNA, these tools use different intrinsic features (such as the length and integrity of the longest open reading frame) of the input RNA sequences, which can result in differences in detecting lncRNAs.

Nevertheless, most studies related to lncRNA identification used only one method (Boltaña et al., 2016; Dettleff et al., 2020; Mu et al., 2016; Paneru et al., 2018; D. Quan et al., 2020; Ren et al., 2020; J. Xu et al., 2019). Hence, a pipeline combining both alignment-dependent and alignment-free strategies is essential to detect conserved and species-specific lncRNAs.

Environmental perturbations produce molecular responses in organisms needed to maintain cellular homeostasis and function, and lncRNAs may play a role in some of these responses by regulating gene expression (Paneru et al., 2018; Sarangdhar et al., 2018) through cis- and trans-regulation (Gil & Ulitsky, 2020; Luo et al., 2019; Nadal-Ribelles et al., 2014; D. Quan et al., 2020). A large proportion of lncRNAs with functional implications are in fact cis-regulatory and hence affect the regulation of the neighbouring protein-coding genes allowing for plasticity in gene networks across time (Engreitz et al., 2016; Gil & Ulitsky, 2020). In the model-plant genus *Arabidopsis*, for instance, a lncRNA tightly controls the expression of highly conserved transcription factors that promote cold tolerance in many plant species (Kindgren et al., 2018). In vertebrates, such as fish, lncRNAs identified in rainbow trout (*Oncorhynchus mykiss*) can potentially mediate regulation of the heat stress response (J. Quan et al., 2020) and one lncRNA affects the mucosal immunity of turbot (*Scophthalmus maximus*) in response to bacterial infection (N. Yang et al., 2020). The involvement of lncRNAs in the regulation of the response to environmental change reveals its importance as a potential regulator of (phenotypic) plasticity and in turn adaptive potential in rapidly changing environments. Hence, there is a need for more studies to clarify the role of lncRNAs in the molecular response of species to changing environment.

To this end, we propose a comprehensive roadmap and guideline for the *de novo* identification of lncRNAs and use a case study to evaluate the involvement of lncRNAs and related coding genes in the response to an environmental change. For this, we use the spiny chromis *Acanthochromis polyacanthus*, a common coral reef fish in the Western Pacific, extensively studied for fish acclimation processes to environmental changes in the last decade. Generous RNA sequencing has been performed on this species (Bernal et al., 2022; Kang et al., 2022; Schunter et al., 2021, 2016) to interpret its transcriptional responses to ocean environmental changes. Despite the abundance of transcriptional information nothing is known about the potential of lncRNAs acting as modulators of gene expression in response to environmental stress. This extensive knowledge on the transcriptional processes to ocean acidification in *A.*

polyacanthus allows for a broad evaluation of the involvement of lncRNAs and provides a guideline for future studies also in other systems. As lncRNAs can tune neighbouring coding gene expression by directly affecting nuclear architecture or indirectly affecting their transcription or translation (Ransohoff et al., 2018) we aim to combine the study of lncRNAs with coding gene expression to evaluate the transcriptional and epigenetic factors that define the biological response of organisms to environmental changes, providing information on the adaptive potential of the species.

METHODS

Discovery and Annotation of lncRNAs

To identify lncRNAs in the *Acanthochromis polyacanthus* genome, 226 RNA-seq samples were compiled from published studies investigating the effect of elevated CO_2 on the brain transcriptome of *A. polyacanthus* (Kang et al., 2022; Schunter et al., 2018, 2016; PRJNA658203). This allows for a comprehensive annotation across the genome of lncRNAs associated with the exposure to elevated CO_2 in laboratory and natural settings. We processed all samples to identify a reliable set of lncRNAs across the genome (Figure 1 & S1). The RNAseq reads are paired-end and non-strand specific, as most RNA sequencing datasets in molecular ecological studies. In our pipeline, the raw reads quality was first assessed with FASTQC v0.11.9 (Andrews, 2010), and then adapter sequences and low-quality sections of reads were removed by Trimmomatic v0.39 (Bolger et al., 2014) using the following parameters: "ILLUMINACLIP:2:30:10 LEADING:4 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:40". The resulting clean reads were aligned to the reference genome of *A. polyacanthus* (NCBI database, accession number: ASM210954v1) with HISAT2 v2.1.0 (Kim et al., 2019) using default parameters and with "*--known-splicesite-infile*" to provide known splice sites.

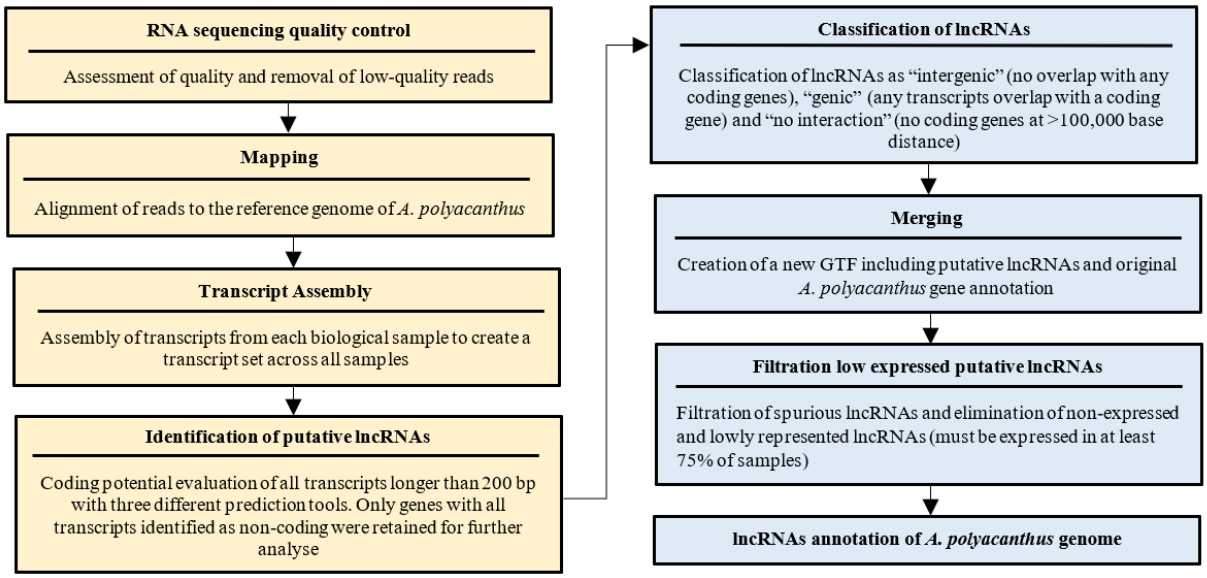

Figure 1: Flowchart illustrating the process for long non-coding RNA (lncRNA) identification in Acanthochromis polyacanthus *genome. In yellow are the steps for the initial identification of putative lncRNAs, while in blue are the post-processing analyses for the creation of the final high confidence lncRNA set.*

StringTie v2.1.5 (M. Pertea et al., 2015) was then applied to assemble transcripts. All resulting transcripts among the 226 fish samples were merged for a unified set of non-redundant transcripts (M. Pertea et al., 2016).

To estimate the coding potential of all transcripts, we extracted the transcript sequences from the *A. polyacanthus* genome sequence using GffRead v0.12.7 (G. Pertea & Pertea, 2020) with the parameters "*-w, -g*". Three tools, including one alignment-dependent (CPC v0.9-r2, Kong et al., 2007) and two alignment-free methods (CPAT v1.2.4, Wang et al., 2013; FEELnc v0.2, Wucher et al., 2017), were used to calculate the coding potential of all transcripts. CPC v0.9-r2 (Kong et al., 2007) was used with default parameters to estimate coding potential based on sequence similarity between transcripts and Uniref90 protein database, and the transcripts with a coding potential score < 0 were considered as putative lncRNAs. For the two alignment-free tools, CPAT v1.2.4 (Wang et al., 2013) was run with default parameters using the coding mRNA sequences of zebrafish as training model, and transcripts with a coding probability < 0.38 were considered as putative lncRNAs, which is the cut-off value suggested by CPAT (Wang et al., 2013). The second alignment-free tool we used was FEELnc v0.2 (Wucher et al., 2017), which annotates lncRNAs based on a Random Forest model trained with general features such as multi k-mer frequencies and relaxed open reading frames. We ran the program with FEELnc_filter.pl to filter out spurious coding transcripts using "-b transcript_biotype=protein_coding, --monoex=-1", and then FEELnc_codpot.pl with default parameters to compute the coding potential score for each transcript by setting the mode of the lncRNA simulation to intergenic. The putative lncRNAs were then identified based on the best coding potential cutoff of receiver operating characteristic curve plot (Sing et al., 2005). For all three methods, only genes with all transcripts identified as non-coding were retained for further analyses.

For a more conservative approach in avoiding spurious lncRNA annotations we only retained the transcripts that were identified as lncRNAs by at least two tools. Using a window size 100 kilobase (kb) (Wucher et al., 2017), these putative lncRNAs were categorized by computing interactions with their proximal coding transcripts using default parameters of FEELnc_classifier.pl (Wucher et al., 2017). A lncRNA is classified as "intergenic" (i.e., lincRNA) if all transcripts of this lncRNA have no location overlap with any neighbouring. Due to the fact that intergenic lncRNA (lincRNA) gene expression patterns, sequence

conservation and perturbation outcomes are easier to interpret than those of transcripts from other lncRNAs, for instance overlapping with coding genes ("genic" lncRNAs; Ulitsky & Bartel, 2013), only lncRNA genes classified as intergenic were used in further analyses to investigate their potential involvement in the response to ocean acidification in fish.

The expression levels of the putative lincRNAs and mRNAs were quantified using FeatureCounts v2.0.0 (Liao et al., 2014) allowing for multi-mapped reads to be counted fractionally. Only transcripts with reads number > 0 in at least 75% of the samples were retained for further analysis. The read numbers of remaining transcripts were normalized using DESeq2 v1.32.0 (Love et al., 2014) and lincRNAs with a normalised expression < 1 in 10% samples were removed to obtain the final list of candidate lincRNA genes. This provided us with a final set of high confidence lincRNAs in both location and expression. LincRNAs with more than 500 normalized reads were considered as highly expressed lincRNAs.

Case study: Expression patterns of lincRNAs in fish living in CO_2 seeps

Using RNA sequencing data from wild *A. polyacanthus* samples collected in Papua New Guinea (Kang et al., 2022; NCBI Bioproject PRJNA691990), we performed a case study to evaluate the expression changes of lincRNAs in different environmental conditions. Seven brain samples were collected from fish from a coral reef situated in a naturally bubbling CO_2 seep with elevated CO_2 levels close to the predicted levels for the end of this century due to ocean acidification (pH= 7.77, pCO_2= 843 µatm; IPCC, 2022). Further eleven fish were sampled from an adjacent control reef approximately 500 m away from the CO_2 seep (pH= 8.01, pCO_2= 443 µatm; Kang et al., 2022). There was no significant difference in temperature and salinity between the CO_2 seep and the control site (Fabricius et al., 2011) allowing for the evaluation of effects of long-term ocean acidification conditions on fish.

To evaluate differential gene expression of lincRNAs and coding genes between fish living at different CO_2 levels, we performed differential gene expression analysis using DESeq2 v1.34 (Love et al., 2014) in R v.3.5.1. Between samples from control and CO_2 seep, lincRNAs and coding genes were considered as differentially expressed (DE) with an FDR adjusted p-value $\leq$ 0.05, and the average of the normalized count values (basemean) $\geq$10 as well as Log2FoldChange $\geq$ 0.3. A principal component analysis (PCA) was performed using the log 2-fold normalized expression of the samples from the two different CO_2 level sites.

To identify lincRNAs and coding gene co-expression modules with significant correlation with CO_2 levels, a weighted gene co-expression network analysis (WGCNA; Langfelder & Horvath, 2008) was also applied on our fish samples. The co-expression similarity of gene modules was obtained by selecting a *signed* network adjacency type with a soft thresholding power of 10 calculated according to the scale-free topology criteria. We established the expression similarity between nodes of genes that are co-expressed by calculating the Topological Overlap Measure (TOM). These co-expression networks were grouped into different coloured modules based on their eigengenes values. Pearson correlation was then calculated to evaluate the correlation between gene modules and samples from control and CO_2 seep. Genes of significantly correlated gene module ($p < 0.01$) were used in the following analysis.

Prediction of potential cis- and trans-acting lincRNAs

To investigate the potential role of cis-acting lincRNAs in the response to elevated CO_2 we focused on lncRNAs which are neighbouring to coding genes with a maximum of 100 kilobase (kb) distance (Wucher et al., 2017). We performed several analyses to investigate a) neighbouring coding genes of highly expressed lincRNAs, b) neighbouring differentially expressed (DE) coding genes of DE lincRNAs, and c) neighbouring coding genes of lincRNAs found co-expressed within the same WGCNA module.

For potential trans-acting lincRNAs, the reads number of all 226 individuals was normalized by calculating transcripts per kilobase million (TPM), and the genes with TPM ≥ 1 among more than 90% individuals were kept. Based on the TPM values, except for the lincRNAs that have neighbouring coding genes, we estimated the association relationship between each lincRNA-coding gene pair for d) DE lincRNAs and DE coding genes, and e) lincRNAs and coding genes within the same WGCNA module using Spearman's correlation test (Tsai et al., 2021) in R version 3.6.3. The lincRNAs that showed Spearman's correlation coefficient $|rho| \geq 0.9$ and p value ≤ 0.01 with a coding gene, were considered as potential trans-acting lincRNAs. Such correlated coding genes were used for functional enrichment analysis.

For all subsets, functional enrichment analyses were performed in OmicsBox v2.0.36 (BioBam Bioinformatics, 2019) with all the annotated genes in *A. polyacanthus* genome as reference.

RESULTS

LncRNA annotation pipeline

Brain *Acanthochromis polyacanthus* RNA sequencing samples had on average 32,7 million high-quality paired-end reads (Table S1) and on average, 87.5% of these reads mapped to the reference genome (Table S2). The assembly steps resulted in a total of 116,237 transcripts across the genome. We used different algorithms to predict the coding potential score (CPS) of these transcripts. This resulted in 29,127 non-coding transcripts with CPAT, 28,865 with CPC and 11,897 with FEELnc, belonging to 17,434, 7,970 and 3,633 putative long non-coding RNA genes (lncRNAs) respectively. Of these, 9,209 putative lncRNAs were identified by at least two programs (Figure 2A). After filtering out lowly expressed putative lncRNAs, we obtained a final set of 4,728 lncRNAs, consisting of 1,313 genic, 3,272 intergenic and 143 non-neighbouring lncRNAs.

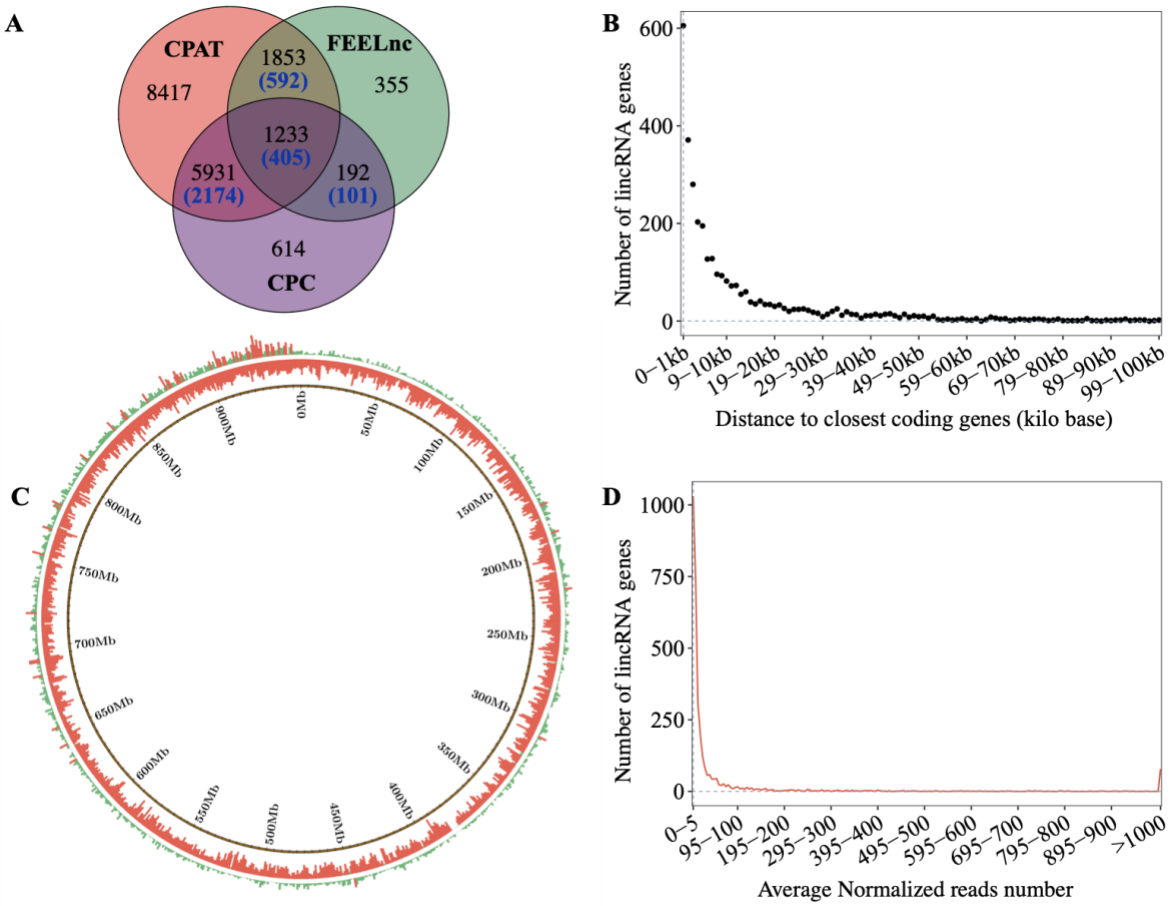

Figure 2: Intergenic lncRNA genes (lincRNAs) along the Acanthochromis polyacanthus genome. ***A****. Venn diagram of putative lncRNAs identified by CPAT, CPC and FEELnc*

programs. Blue numbers in brackets indicate overlapping lincRNAs among the total lncRNAs. ***B****. Distance ranges between the identified 3,272 lincRNAs and their neighbouring coding genes.* ***C****. Distribution of lincRNAs (first outer circle) and coding genes (second outer circle) along the whole genome (concatenating all scaffolds together from longest to shortest). Genome sliding windows are indicated in the inner black circle. Bar height and color are indicative of the number of genes found in each 1Mb sliding window, with red and green bars indicating more and less than 10 lincRNAs or coding genes per window, respectively.* ***D.*** *LincRNAs levels of expression across different expression level ranges.*

The intergenic lncRNA genes (lincRNAs; Table S3) are significantly shorter in length than coding genes (two-sample wilcoxon rank sum test, $p < 2.2e-16$) with 1,386 lincRNAs with a length less than 400 nucleotides (Figure S2A). Of all intergenic lncRNA genes (Table S3), 2,143 (65.5%) lincRNAs are monoexonic, while 1,129 (34.5%) lincRNAs have at least two exons (Figure S2B). In comparison, 30,282 coding genes (88.6%) have at least two exons, which are significantly more than lincRNAs (two-sample wilcoxon rank sum test, $p < 2.2e-16$). The GC content (43.1%) of lincRNAs is significantly lower than coding genes (50.0%, two-sample wilcoxon rank sum test, $p < 2.2e-16$, Figure S2C). LincRNAs (Table S3) are mostly located within a short distance to their neighbouring coding genes (Figure 2B) with 1,654 (50.6%) of lincRNAs located at a distance smaller than 5kb from their neighbour coding genes. Of these lincRNA, 354 are antisense, 243 sense lincRNAs and 1,057 strand-unknown lincRNAs (Table S3). Among the 354 antisense lincRNAs, four lincRNAs (MSTRG.12220, MSTRG.40264, MSTRG.36665, MSTRG.10872) contain both convergent and divergent transcripts, while 146 and 204 only contained divergent and convergent transcripts, respectively (Figure S3). LincRNAs were dispersed throughout the whole genome (Figure 2C), however, the overall lincRNAs density was significantly lower than in protein coding genes (two-sample wilcoxon rank sum test, $p < 2.2e-16$). Among 955 genomic sliding windows of 1Mb in the *A. polyacanthus* genome, 829 regions included more than one lincRNA and 53 regions included more than ten lincRNAs. Protein coding genes, by contrast, displayed a higher density (934 regions with more than 10 coding genes).

Most lincRNAs exhibited expression with less than 500 normalized reads, and only 125 lincRNAs had elevated expression levels (average normalized reads > 500; Figure 2D; Table S4). The neighbouring coding genes of the 125 most expressed lincRNAs were involved in signal transduction, GABA-A receptor, circadian rhythm, and ion transport (Figure 3A; Table

S4). The highest expressed lincRNA (MSTRG.9528; expression > 190,000 normalized reads) neighbours the coding gene peripherin-2 (PRPH2).

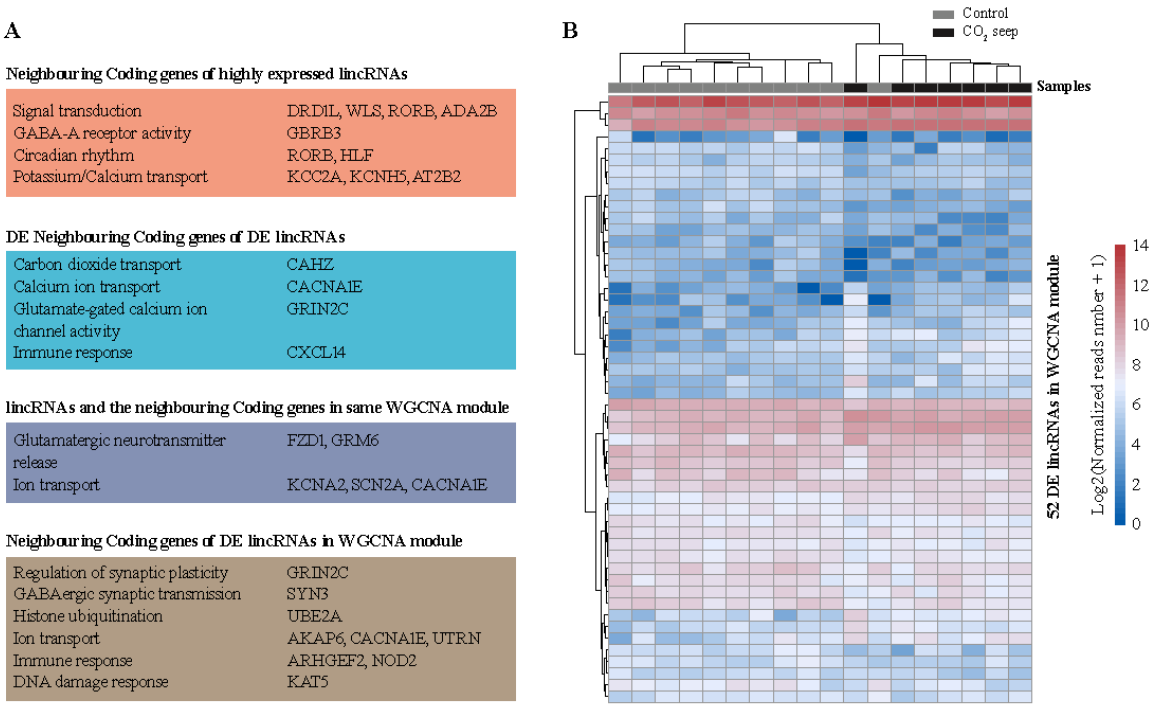

Figure 3: Functions of neighbouring coding genes of lincRNAs (A) and gene expression pattern of 52 differential expressed lincRNAs which were also related to pH level in module turquoise by WGCNA analysis (B).

<u>Case study: response to ocean acidification in fish</u>

By applying our pipeline to already existing transcriptomic data from wild *A. polyacanthus* collected from a CO_2 seep and a control reef site in Papua New Guinea, we were able to evaluate the involvement of genomic regulations by lincRNAs in fish brain in response to an environmental factor. Fish from the CO_2 seep displayed a distinct brain expression pattern, both when the analysis was run using all genes (coding genes and lincRNAs; Figure 4A) as well as only lincRNAs (Figure 4B). Between samples from CO_2 seep and control site, 3,431 coding genes (Figure S4A) and 97 lincRNAs (Table 1; Table S5) were significantly differentially expressed by DESeq2 v1.34 (Love et al., 2014).

WGCNA results showed that three modules were significantly correlated with environmental CO_2 levels ("turquoise", p = 0.001, Table S7; "sky-blue", p= 0.009, Table S8; "dark-green", p= 0.006, Table S9). Turquoise and sky-blue modules were negatively correlated with CO_2 levels and included 329 lincRNAs and 6,259 coding genes (Table 1; Table S7), and 26 lincRNAs and 136 coding genes (Table 1; Table S8), respectively. Dark green module was conversely positively correlated with CO_2 levels and consisted of 6 lincRNAs and 231 coding genes (Table 1; Table S9).

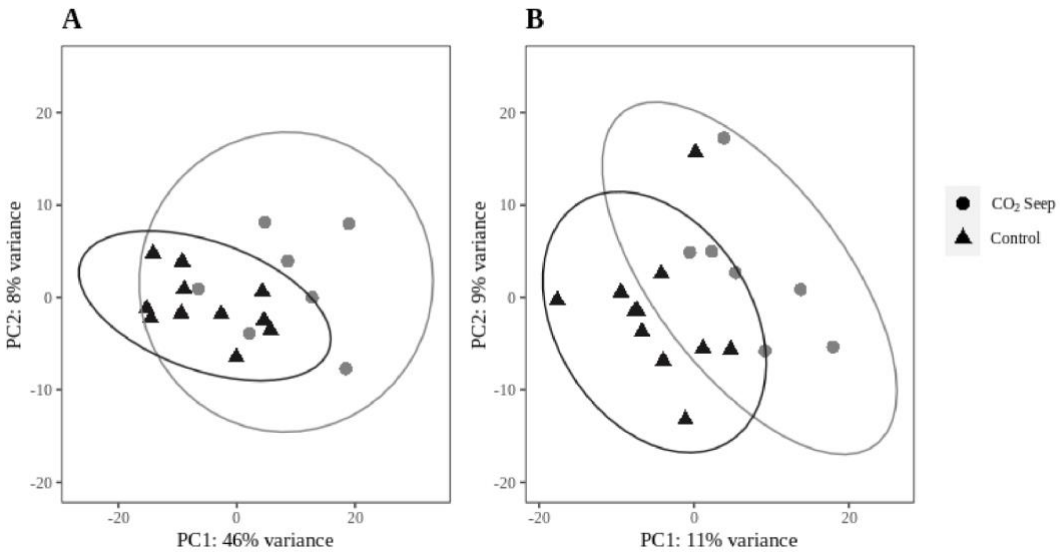

Figure 4: Principal component analyses (PCAs) of log2-fold normalized expression values of (A) lincRNA and coding genes or (B) lincRNAs only in the brain of wild collected Acanthochromis polyacanthus from a CO_2 seep (circles) and an adjacent control reef (triangles). Ellipse areas represent a 95% confidence level.

Table 1. Differentially expressed lincRNAs and coding genes identified through differential expression analysis (DESeq2) and weighted gene co-expression network analysis (WGCNA). In the "lincRNAs with neighbouring coding gene" row are the numbers of lincRNAs for which the neighbouring coding genes were also differentially expressed or found in the same co-expression module.

Gene type	**Differential Expression**	**Weighted co-expression network**		
		turquoise	sky-blue	dark-green
lincRNAs	97	329	26	6
coding genes	3431	6,259	136	231
lincRNAs with neighbouring coding gene	13	65	0	0

Cis-regulatory roles of lincRNAs in response to elevated CO_2

Based on the significantly differentially (DE) genes detected by DESeq2, DE coding genes with neighbouring lincRNAs had significantly higher expression changes (Pearson's Chi-squared test, p = 3.425e-05). 3.5% (7/198) of the DE coding genes with neighbouring lincRNA exhibit high expression difference (|log2FoldChange| > 2), whereas only 0.7% (23/3233) of the DE coding genes without neighbouring lincRNA display elevated expression (Figure S4B). Among the neighbouring coding genes of the 97 DE lincRNAs, 13 coding genes also displayed significant differential expression (Table S6). These DE neighbouring coding genes are related to carbon dioxide transport and hypotonic salinity response (CAHZ: carbonic anhydrase), calcium ion transport (CACNA1E: calcium voltage-gated channel subunit alpha1 E), glutamate-gated calcium ion channel activity (GRIN2C: glutamate ionotropic receptor NMDA type subunit 2C) or immune system response (CXCL14: C-X-C motif chemokine 14; Figure 3A).

Among the significantly correlated co-expression networks, only the turquoise module was found with 68 lincRNAs whose neighbouring coding genes were also included in the same co-expression module (Tables S7 & S10). The functional implications of these neighbouring coding genes show involvement in, for instance, glutamatergic neurotransmitter release (FZD1: frizzled class receptor 1; GRM6: glutamate metabotropic receptor 6) and ion transport

(KCNA2: potassium voltage-gated channel subfamily A member 2; SCN2A: sodium voltage-gated channel alpha subunit 2; CACNA1E: calcium voltage-gated channel subunit alpha1 E; Figure 3A).

Among the lincRNAs in WGCNA modules, turquoise module showed 52 lincRNAs which were also significantly differentially expressed by DESeq2 (Figure 3; Tables S7 & S11). The neighbouring coding genes of these lincRNAs were related to the regulation of synaptic plasticity (GRIN2C: glutamate receptor ionotropic NMDA 2C), modulation on GABAergic synaptic transmission (SYN3: synapsin-3), histone ubiquitination (UBE2A: ubiquitin-conjugating enzyme E2 A), regulation of ion transport (AKAP6: A-kinase anchor protein 6; CACNA1E: calcium voltage-gated channel subunit alpha1 E; UTRN: utrophin), innate immune response (ARHGEF2: rho guanine nucleotide exchange factor 2; NOD2: nucleotide-binding oligomerization domain-containing protein 2) and DNA damage response (KAT5: histone acetyltransferase KAT5; Figure 3A). Of the other two significantly correlated modules, sky-blue and dark green, only two and one lincRNAs, respectively, were also found to be differentially expressed (Table S8 & S9).

Regulatory roles of potential trans-acting lincRNAs in response to elevated CO_2

Between DE lincRNAs and DE coding genes, 148 lincRNA-coding gene pairs exhibited high and significant correlation with $|rho| \geq 0.9$ and p value ≤ 0.01, which include 14 lincRNAs and 129 coding genes (Table 2). With the co-expression networks, only the turquoise module revealed 219 lincRNA-coding gene pairs with high and significant correlation made of 14 lincRNAs and 129 coding genes (Table 2). In summary, 21 lincRNAs may trans-regulate 186 coding genes (Table S12), including genes involved in ion transport (KCMA1: Calcium-activated potassium channel subunit alpha-1; SCN8A: Sodium channel type 8 subunit alpha; CACNA1A: Voltage-dependent P Q-type calcium channel subunit alpha-1A; RYR3: Ryanodine receptor 3; NALCN: Sodium leak channel non-selective), glutamate receptor activity (GRM4: Glutamate receptor 4; NMDE1: Glutamate receptor NMDA 2A), and immune response (BCL11A: B-cell lymphoma leukemia 11A; BCL11B: B-cell lymphoma leukemia 11B; BCL9: B-cell CLL lymphoma 9).

Table 2. Significantly correlated pairs of lincRNA and coding genes, which were differentially expressed (DE) lincRNAs and DE coding genes from DESeq2, and lincRNAs and coding genes in the same modules from the weighted correlated co-expression network analysis.

	LincRNA-coding gene pairs	LincRNAs	Coding genes
DESeq2	148	14	129
WGCNA	219	14	129

DISCUSSION

De novo *lncRNAs discovery pipeline*

The boost in RNA sequencing studies including a large variety of non-model organisms provides an excellent resource that could be (re-)used to obtain a more complete picture of lncRNAs regulations on coding genes. Here we present a solid pipeline that can make use of already existing data to detect conserved as well as species-specific lncRNAs. This pipeline can be easily applied to any species with an available reference genome. The number of published genomes has been increasing drastically the last few years even in non-model organisms owing to several international initiatives such as the Earth BioGenome project (EBP), Darwin Tree of Life Project, The Vertebrate Genomes Project, 1000 Fungal Genomes Project and ERGA (European Reference Genome Atlas), among others. However, to answer ecological and evolutionary relevant questions we need not only the complete sequence of genomes but also accurate annotations including regulatory genes such as lncRNAs.

Most studies related to lncRNA identification rely on only one method of coding potential detection (Boltaña et al., 2016; Dettleff et al., 2020; Mu et al., 2016; Paneru et al., 2018; J. Quan et al., 2020; Ren et al., 2020; H. Xu et al., 2019). In contrast, our results based on an alignment-dependent (CPC) and two alignment-free methods (CPAT, FEELnc), indicate that 92.0%, 51.7% and 90.2% lncRNAs detected by CPC, CPAT and FEELnc respectively were called as lncRNAs by at least two tools. Different tools revealed to have varying performances (Schneider et al., 2017; Wucher et al., 2017) and the three tools used in our study perform well in predicting lncRNAs also in other species (Duan et al., 2021). Thus, lncRNAs predicted using our pipeline could be reliable resources for future studies to investigate lncRNAs functions. Our pipeline was integrated as an automatic tool, which can assemble transcripts based on the sequence alignment data, detect the lncRNAs from the reference genome through CPC, CPAT and FEELnc, and classify the lincRNAs detected by at least two methods. Hence, our pipeline is easily applied to other studies involving RNA sequencing datasets to investigate into lncRNAs detection and analysis.

We identified a total of 9,209 putative lncRNAs, of which 49% were expressed in at least 75% of our *A. polyacanthus* brain samples. While the numbers found in mammal species are higher (6,010 - 96,411 from the rhesus macaque to human; NONCODE v5.0, Fang et al., 2018) it has

to be noted that in our case we only investigate one tissue, and the inclusion of more tissues is likely to increase the number of lncRNAs. Nonetheless, teleost model species show similar numbers of expressed lncRNAs. Zebrafish lncRNAs identified so far are 3,503 (Zhao et al., 2021) and between three- and six thousand lncRNAs have been found to be expressed across different tissues also in rainbow trout (Al-Tobasei et al., 2016), coho (Leiva et al., 2020) and Atlantic salmon (Boltaña et al., 2016). Other studies found higher numbers of lncRNAs in some teleost species, such as 31,984 in the brain of *Larimichthys crocea* (Liu et al., 2018) using CPC or 14,614 across several tissues of *Genypterus chilensis* (Dettleff et al., 2020) using CPAT. However, these studies only used one method to evaluate coding potential and did not ensure that all transcripts of a lncRNA were non-coding, which are less stringent criteria compared to what was applied here.

Overall, the majority of lncRNAs we identified throughout the genome of *A. polyacanthus* had relatively low expression. This observation is in accordance with many previous studies that report a typical low expression of lncRNAs across species and tissues (Guo et al., 2020; Hezroni et al., 2015; Necsulea et al., 2014; Quinn et al., 2016), as lncRNAs are constantly submitted to various regulation mechanisms and they are typically short-lived (Necsulea et al., 2014; Wu et al., 2014). Despite low expression levels, these lncRNAs can be functional since some regulatory mechanisms do not require high concentration of effector molecules (Aprea & Calegari, 2015). However, some lncRNAs can also be expressed at elevated levels and highly expressed intergenic lncRNAs (lincRNAs), for instance, are associated with diseases (Wan et al., 2013; F. Yang et al., 2011). In the brains of our coral reef fish, we also found some highly expressed lincRNAs with neighbouring coding genes involved in fundamental functions such as a $GABA_A$ receptor, circadian rhythm genes and genes related to signal transduction (such as DRD1L and ADA2B), important for normal brain function and neuronal activity (Bhat et al., 2010; Logan & McClung, 2019). In humans, rodents, and zebrafish, a lncRNA promotes the expression of homeobox transcription factors required for the development of GABAergic neurons Y-Aminobutyric acid (or GABA) which is the main inhibitory neurotransmitter in the vertebrate brain (Feng et al., 2006). Furthermore, lncRNAs have been shown to regulate core circadian rhythm genes in mammals (Mosig & Kojima, 2021). With a substantial fraction of lncRNAs affecting the gene expression of their neighbouring coding genes (Engreitz et al., 2016), it is no surprise to see elevated expression in lncRNAs neighbouring $GABA_A$ receptor and core circadian rhythm genes as these are known to play important roles in the

transcriptional response to ocean acidification (Kang et al., 2022; Schunter et al., 2021, 2016). As our RNAseq data is from ocean acidification experiments, this may suggest a regulatory involvement of lncRNAs in these key functions. Hence, our pipeline of lncRNA annotation and expression analysis allows for the discovery of new lncRNAs that may be involved in the response to an environmental change in a wild coral reef fish, but it also recovers conserved lncRNAs found to possibly be involved in essential functions also in other species.

Functional responses of lncRNAs to environmental change

Our pipeline facilitates the investigation of lncRNAs as a potential regulatory mechanism in response to environmental change and opens doors to fields like environmental science and ecology. In our case study, we made use of a published RNA sequencing set of *Acanthochromis polyacanthus* brains from a volcanic CO_2 seep, representing future ocean acidification conditions, and a control site. We detected differentially expressed lincRNAs and mRNAs in the brain of *A. polyacanthus* as a response to natural environmental differences (CO_2 levels). Interestingly, for the coding genes that have a neighbouring lincRNA in close proximity, we found larger expression differences between fish from CO_2 seeps and control sites suggesting a potential regulation of gene expression by neighbouring lncRNAs with the exposure to elevated CO_2. For the differentially expressed lincRNAs, the set of neighbouring protein coding genes that were also differentially expressed are involved in functions related to pH regulation in fish. When fish respond to ocean acidification, carbonic anhydrase (CAHZ) is often upregulated, as in our data here, to catalyze the hydration of CO_2 and play an essential role in acid-base and ion regulatory functions (Heuer & Grosell, 2014). A further reduction of intracellular pH in fish is necessary to prevent acidosis when faced with elevated CO_2 levels in fish (Schmidt, 2019), and this process performed by coding gene GRIN2C may potentially be regulated by differentially expressed lincRNAs.

Further biological processes are known to be involved in the response to ocean acidification in fish brains. One of them is synaptic transmission and the downregulation of a lincRNA and the neighbouring coding gene CACNA1E (voltage-gated calcium channel complex) can contribute to synaptic transmission (Berecki et al., 2014), which was also found to behave similarly to a previous RNAseq *A. polyacanthus* data (Schunter et al., 2018). Some lincRNAs may trans-regulate the expression of ion transporters, such as calcium-activated potassium channel subunit alpha-1 (KCMA1), sodium leak channel non-selective (NALCN), and voltage-

dependent P Q-type calcium channel subunit alpha-1A (CACNA1A), which play critical roles in the neural signal transduction (Gadsby, 2009) and were also reported in our previous study (Kang et al., 2022). In addition, immune responses are commonly associated with ocean acidification in fish including our study species (De Souza et al., 2014; Kang et al., 2022; Machado et al., 2020), and here we find differential expression of lincRNAs neighbouring and possibly trans-regulating a variety of immune response genes that show co-expression in a gene network significantly correlated with CO_2 levels. This suggests that lincRNAs annotated in our coral reef fish show differential expression in response to an environmental factor, CO_2 levels, and that some of these differentially expressed lincRNAs are in close proximity to coding genes that are known to have functional implications in the response to ocean acidification in fishes.

We identified candidate lincRNAs to be involved in acclimation to elevated CO_2 levels and future genetic manipulation experiments would allow us to verify the direct effect of a lincRNA on the nearby coding gene (Akay et al., 2019; Rodriguez-Lopez et al., 2022). CRISPR/Cas9 based genome editing is a promising technique to experimentally evaluate the role of lincRNAs, although it is not readily available for most non-model organisms. Our study is a start to understand regulatory elements potentially involved in the regulation of the gene expression patterns observed with an environmental change. This case study presents a pipeline that can be applied by using RNA sequencing reads, now frequently incorporated into experimental studies, and adds an epigenetic mechanistic aspect with no additional sequencing effort needed. We demonstrate that lncRNAs can be easily annotated de novo in non-model organisms and encourage more molecular ecologists to discover lncRNAs in their study species and investigate the mechanistic role of lncRNA in response to environmental change.

ACKNOWLEDGMENTS

We would like to thank Timothy Ravasi and Philip L. Munday for for helping with the collection and sequencing of the large RNA sequencing datasets produced in previous studies. We are grateful to Helen Leung, Ho Wu Cheuck and Kam Yan Chit for their support with this project.

DATA AVAILABILITY

The RNA sequencing raw data used in this study is found in the Bioprojects: PRJNA691990; PRJNA311159; PRJNA658203 (Reviewer link as the last is not published yet: https://dataview.ncbi.nlm.nih.gov/object/PRJNA658203?reviewer=47r5c4kubkjn124c770t5vvt60)

The lncRNA annotation of *Acanthochromis polyacanthus* is available here: 10.6084/m9.figshare.20045780. Reviewer link: https://figshare.com/s/33802daaa28dc7ccd876

CODE AVALABILITY

The scripts have been deposited in Github (https://github.com/jinglkang/lncRNAs_detect).

SUPPLEMENTARY FIGURES

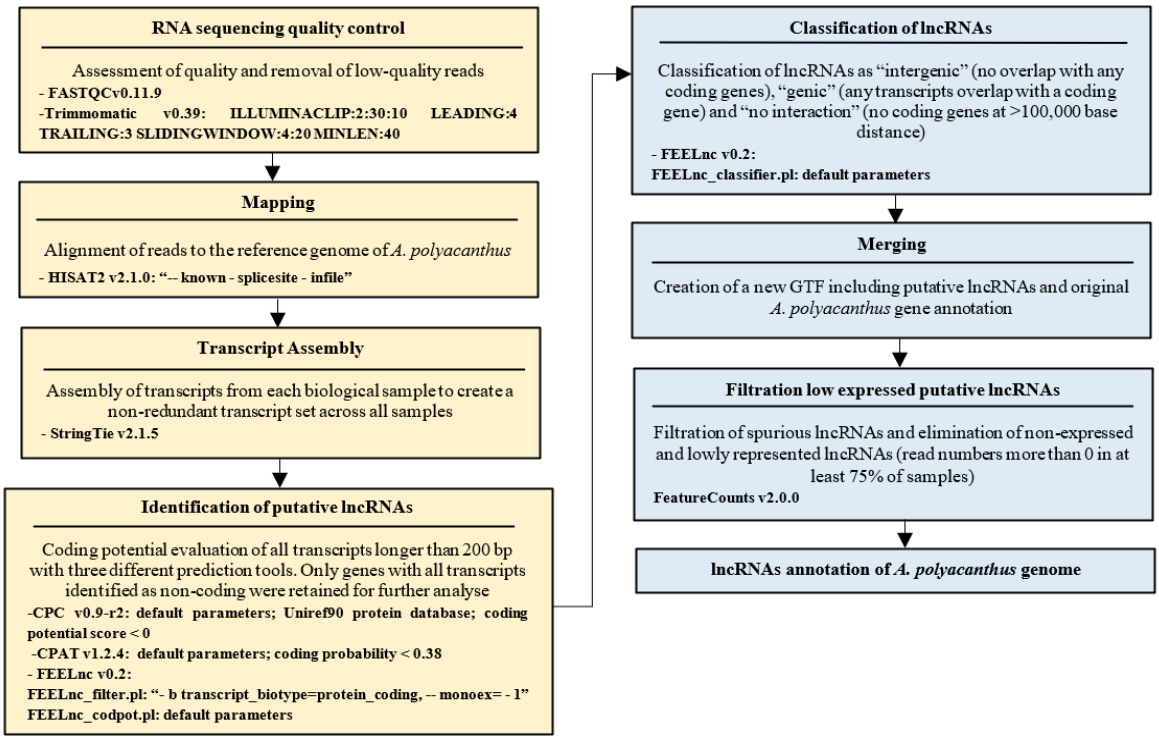

Figure S1: Detailed flowchart illustrating the process for long non-coding RNA (lncRNA) identification including the programs and parameters used

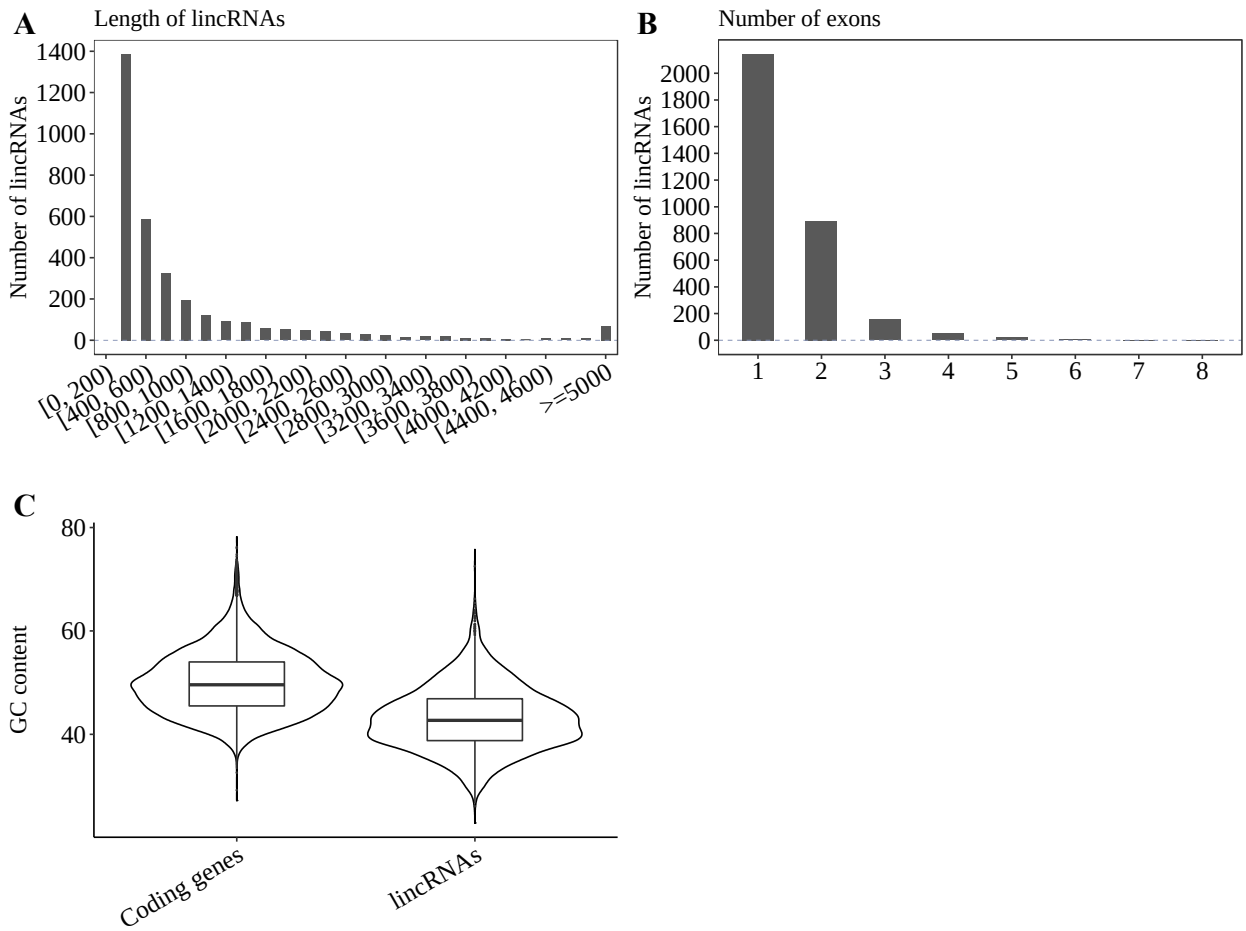

Figure S2: Sequence features of intergenic long non-coding RNAs (lincRNAs) in *Acanthochromis polyacanthus*. A. Length distributions of lincRNAs. B. Distributions of the number of exons in lincRNAs. C. The GC content in coding genes and lincRNAs. The GC content of lincRNAs were significantly lower (two-sample wilcoxon rank sum test, $p < 2.2e-16$) than coding genes.

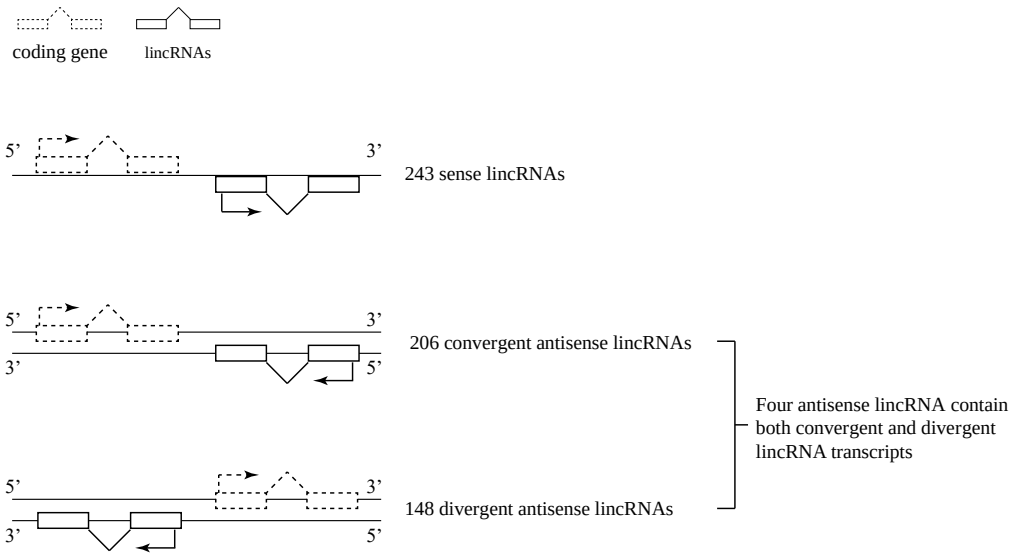

Figure S3: The orientation of intergenic long non-coding RNAs (lincRNAs) in *Acanthochromis polyacanthus*.

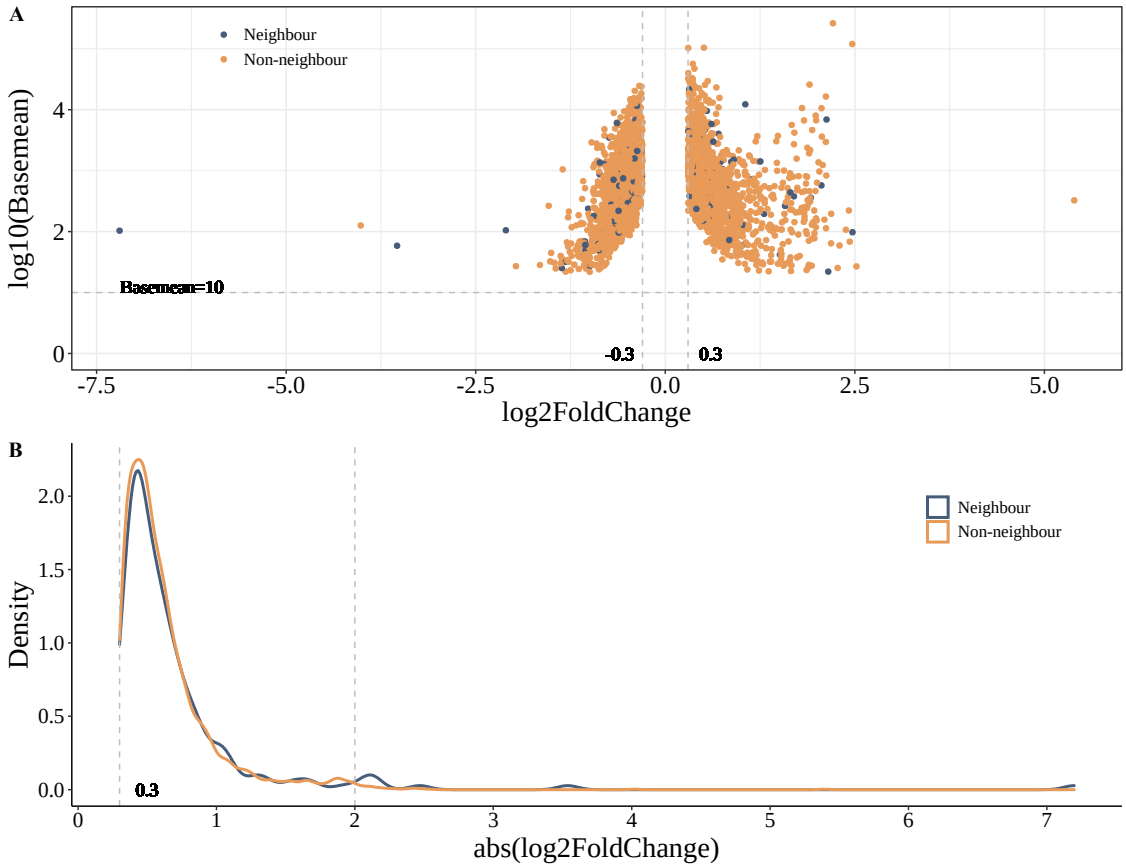

Figure S4: Expression pattern comparison between coding genes with (blue) and without (orange) neighbouring lincRNAs. A. Volcano plot of all expressed coding genes. The gene expression log2 fold change and log10(basemean) between samples from CO_2 seep and control

site are reported on the X and the Y axes, respectively. B. Density of log2FoldChange absolute value of gene expression between samples from CO_2 seep and control site.

REFERENCES

Akay, A., Jordan, D., Navarro, I. C., Wrzesinski, T., Ponting, C. P., Miska, E. A., & Haerty, W. (2019). Identification of functional long non-coding RNAs in C. elegans. *BMC Biology*, *17*(1), 1–14.

Al-Tobasei, R., Paneru, B., & Salem, M. (2016). Genome-wide discovery of long non-coding RNAs in rainbow trout. *PLoS ONE*, *11*(2).

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved May 24, 2022, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aprea, J., & Calegari, F. (2015). Long non-coding RNA s in corticogenesis: deciphering the non-coding code of the brain. *The EMBO Journal*, *34*(23), 2865–2884.

Beck, D., Thoms, J. A. I., Palu, C., Herold, T., Shah, A., Olivier, J., … Pimanda, J. E. (2018). A four-gene LincRNA expression signature predicts risk in multiple cohorts of acute myeloid leukemia patients. *Leukemia*, *32*(2), 263–272.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Berecki, G., McArthur, J. R., Cuny, H., Clark, R. J., & Adams, D. J. (2014). Differential Cav2.1 and Cav2.3 channel inhibition by baclofen and α-conotoxin Vc1.1 via $GABA_B$ receptor activation. *Journal of General Physiology*, *143*(4), 465–479.

Bernal, M. A., Ravasi, T., Rodgers, G. G., Munday, P. L., & Donelson, J. M. (2022). Plasticity to ocean warming is influenced by transgenerational, reproductive, and developmental exposure in a coral reef fish. *Evolutionary Applications*, *15*(2), 249–261.

Bhat, R., Axtell, R., Mitra, A., Miranda, M., Lock, C., Tsien, R. W., & Steinman, L. (2010). Inhibitory role for GABA in autoimmune inflammation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2580–2585.

BioBam Bioinformatics. (2019). OmicsBox – Bioinformatics made easy. *BioBam Bioinformatics*. Retrieved from https://www.biobam.com/omicsbox/

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Boltaña, S., Valenzuela-Miranda, D., Aguilar, A., Mackenzie, S., & Gallardo-Escárate, C. (2016). Long noncoding RNAs (lncRNAs) dynamics evidence immunomodulation during ISAV-Infected Atlantic salmon (*Salmo salar*). *Scientific Reports*, *6*(1), 1–13.

Cabili, M. N., Dunagin, M. C., McClanahan, P. D., Biaesch, A., Padovan-Merhar, O., Regev, A., … Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biology*, *16*(1).

De Souza, K. B., Jutfelt, F., Kling, P., Förlin, L., & Sturve, J. (2014). Effects of increased CO_2 on fish gill and plasma proteome. *PLoS ONE*, *9*(7), e102901.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789.

Dettleff, P., Hormazabal, E., Aedo, J., Fuentes, M., Meneses, C., Molina, A., & Valdes, J. A. (2020). Identification and Evaluation of Long Noncoding RNAs in Response to Handling Stress in Red Cusk-Eel (*Genypterus chilensis*) via RNA-seq. *Marine Biotechnology*, *22*(1), 94–108.

Duan, Y., Zhang, W., Cheng, Y., Shi, M., & Xia, X. Q. (2021). A systematic evaluation of bioinformatics tools for identification of long noncoding RNAs. *RNA*, *27*(1), 80–98.

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., … Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and

splicing. *Nature*, *539*(7629), 452–455.
Fabricius, K. E., Langdon, C., Uthicke, S., Humphrey, C., Noonan, S., De'ath, G., … Lough, J. M. (2011). Losers and winners in coral reefs acclimatized to elevated carbon dioxide concentrations. *Nature Climate Change*, *1*(3), 165–169.
Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., … Zhao, Y. (2018). NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research*, *46*(D1), D308–D314.
Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P., & Kohtz, J. D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes and Development*, *20*(11), 1470–1484.
Fernandes, J. C. R., Acuña, S. M., Aoki, J. I., Floeter-Winter, L. M., & Muxel, S. M. (2019). Long non-coding RNAs in the regulation of gene expression: Physiology and disease. *Non-Coding RNA*, *5*(1), 17.
Gadsby, D. C. (2009). Ion channels versus ion pumps: The principal difference, in principle. *Nature Reviews Molecular Cell Biology*, *10*(5), 344–352.
Gardini, A., & Shiekhattar, R. (2015). The many faces of long noncoding RNAs. *FEBS Journal*, *282*(9), 1647–1657.
Gil, N., & Ulitsky, I. (2020). Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews Genetics*, *21*(2), 102–117.
Guo, C. J., Ma, X. K., Xing, Y. H., Zheng, C. C., Xu, Y. F., Shan, L., … Chen, L. L. (2020). Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell*, *181*(3), 621-636.e22.
Heuer, R. M., & Grosell, M. (2014). Physiological impacts of elevated carbon dioxide and ocean acidification on fish. *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, *307*(9), R1061–R1084.
Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports*, *11*(7), 1110–1122.
IPCC. (2022). Summary for Policymakers. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Retrieved from www.ipcc.ch
Jiang, X., Lei, R., & Ning, Q. (2016). Circulating long noncoding RNAs as novel biomarkers of human diseases. *Biomarkers in Medicine*, *10*(7), 757–769.
Kang, J., Nagelkerken, I., Rummer, J. L., Rodolfo-Metalpa, R., Munday, P. L., Ravasi, T., & Schunter, C. (2022). Rapid evolution fuels transcriptional plasticity to ocean acidification. *Global Change Biology*, *00*, 1–16.
Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915.
Kindgren, P., Ard, R., Ivanov, M., & Marquardt, S. (2018). Transcriptional read-through of the long non-coding RNA SVALKA governs plant cold acclimation. *Nature Communications*, *9*(1).
Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., & Gao, G. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(SUPPL.2), 345–349.
Kornfeld, J. W., & Brüning, J. C. (2014). Regulation of metabolism by long, non-coding RNAs. *Frontiers in Genetics*, *5*(MAR), 57.
Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 1–13.
Leiva, F., Rojas-Herrera, M., Reyes, D., Bravo, S., Garcia, K. K., Moya, J., & Vidal, R. (2020).

Identification and characterization of miRNAs and lncRNAs of coho salmon (Oncorhynchus kisutch) in normal immune organs. *Genomics*, *112*(1), 45–54.

Li, Z., Zhao, W., Wang, M., & Zhou, X. (2019). The Role of Long Noncoding RNAs in Gene Expression Regulation. In *Gene Expression Profiling in Cancer* (pp. 1–17).

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930.

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, *27*(13), 275–282.

Liu, W., Liu, X., Wu, C., & Jiang, L. (2018). Transcriptome analysis demonstrates that long noncoding RNA is involved in the hypoxic response in *Larimichthys crocea*. *Fish Physiology and Biochemistry*, *44*(5), 1333–1347.

Logan, R. W., & McClung, C. A. (2019). Rhythms of life: circadian disruption and brain disorders across the lifespan. *Nature Reviews Neuroscience*, *20*(1), 49–65.

Lopez-Ezquerra, A., Harrison, M. C., & Bornberg-Bauer, E. (2017). Comparative analysis of lincRNA in insect species. *BMC Evolutionary Biology*, *17*(1), 1–11.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12).

Luo, Y., Chen, J. J., Lv, Q., Qin, J., Huang, Y. Z., Yu, M. H., & Zhong, M. (2019). Long non-coding RNA NEAT1 promotes colorectal cancer progression by competitively binding miR-34a with SIRT1 and enhancing the Wnt/β-catenin signaling pathway. *Cancer Letters*, *440*, 11–22.

Machado, M., Arenas, F., Svendsen, J. C., Azeredo, R., Pfeifer, L. J., Wilson, J. M., & Costas, B. (2020). Effects of Water Acidification on Senegalese Sole Solea senegalensis Health Status and Metabolic Rate: Implications for Immune Responses and Energy Use. *Frontiers in Physiology*, *11*, 26.

Mosig, R. A., & Kojima, S. (2021). Timing without coding: How do long non-coding RNAs regulate circadian rhythms? *Seminars in Cell and Developmental Biology*.

Mu, C., Wang, R., Li, T., Li, Y., Tian, M., Jiao, W., … Bao, Z. (2016). Long Non-Coding RNAs (lncRNAs) of Sea Cucumber: Large-Scale Prediction, Expression Profiling, Non-Coding Network Construction, and lncRNA-microRNA-Gene Interaction Analysis of lncRNAs in *Apostichopus japonicus* and *Holothuria glaberrima* During LPS Challenge. *Marine Biotechnology*, *18*(4), 485–499.

Nadal-Ribelles, M., Solé, C., Xu, Z., Steinmetz, L. M., DeNadal, E., & Posas, F. (2014). Control of Cdc28 CDK1 by a Stress-Induced lncRNA. *Molecular Cell*, *53*(4), 549–561.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., … Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, *505*(7485), 635–640.

Paneru, B., Ali, A., Al-Tobasei, R., Kenney, B., & Salem, M. (2018). Crosstalk among lncRNAs, microRNAs and mRNAs in the muscle “degradome” of rainbow trout. *Scientific Reports*, *8*(1), 8416.

Pegueroles, C., Iraola-Guzmán, S., Chorostecki, U., Ksiezopolska, E., Saus, E., & Gabaldón, T. (2019). Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus *Caenorhabditis*. *RNA Biology*, *16*(3), 320–329.

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Quan, D., Chen, K., Zhang, J., Guan, Y., Yang, D., Wu, H., … Lv, L. (2020). Identification of lncRNA NEAT1/miR-21/RRM2 axis as a novel biomarker in breast cancer. *Journal of Cellular Physiology*, *235*(4), 3372–3381.

Quan, J., Kang, Y., Luo, Z., Zhao, G., Ma, F., Li, L., & Liu, Z. (2020). Identification and characterization of long noncoding RNAs provide insight into the regulation of gene expression in response to heat stress in rainbow trout (*Oncorhynchus mykiss*). *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, *36*, 100707.

Quinn, J. J., Zhang, Q. C., Georgiev, P., Ilik, I. A., Akhtar, A., & Chang, H. Y. (2016). Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. *Genes and Development*, *30*(2), 191–207.

Ransohoff, J. D., Wei, Y., & Khavari, P. A. (2018, November 15). The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology*, *19*(3), 143–157.

Ren, X., Chen, C., Luo, Y., Liu, M., Li, Y., Zheng, S., … Chen, R. (2020). LncRNA-PLACT1 sustains activation of NF-κB pathway through a positive feedback loop with IκBα/E2F1 axis in pancreatic cancer. *Molecular Cancer*, *19*(1), 1–19.

Rodriguez-Lopez, M., Anver, S., Cotobal, C., Kamrad, S., Malecki, M., Correia-Melo, C., … Bähler, J. (2022). Functional profiling of long intergenic non-coding RNAs in fission yeast. *ELife*, *11*.

Sarangdhar, M. A., Chaubey, D., Srikakulam, N., & Pillai, B. (2018). Parentally inherited long non-coding RNA Cyrano is involved in zebrafish neurodevelopment. *Nucleic Acids Research*, *46*(18), 9726–9735.

Schmidt, M. (2019). Behavioural disturbances and underlying neurophysiological mechanisms during ocean acidification and warming in *Gadus morhua* and *Boreogadus saida*, (Doctoral dissertation, Universität Bremen).

Schneider, H. W., Raiol, T., Brigido, M. M., Walter, M. E. M. T., & Stadler, P. F. (2017). A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts. *BMC Genomics*, *18*(1), 804.

Schunter, C., Jarrold, M. D., Munday, P. L., & Ravasi, T. (2021). Diel pCO_2 fluctuations alter the molecular response of coral reef fishes to ocean acidification conditions. *Molecular Ecology*, *30*(20), 5105–5118.

Schunter, C., Welch, M. J., Nilsson, G. E., Rummer, J. L., Munday, P. L., & Ravasi, T. (2018). An interplay between plasticity and parental phenotype determines impacts of ocean acidification on a reef fish. *Nature Ecology and Evolution*, *2*(2), 334–342.

Schunter, C., Welch, M. J., Ryu, T., Zhang, H., Berumen, M. L., Nilsson, G. E., … Ravasi, T. (2016). Molecular signatures of transgenerational response to ocean acidification in a species of reef fish. *Nature Climate Change*, *6*(11), 1014–1018.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941.

Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology*, *22*(2), 96–118.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural and Molecular Biology 14*(2), 103–105.

Tsai, C. H., Lin, T. C., Chang, Y. H., Tsai, H. K., & Huang, J. H. (2021). Identification and comparative analysis of long non-coding RNAs in the brain of fire ant queens in two different reproductive states. *BMC Genomics*, *22*(5), 1–14.

Ulitsky, I. (2016). Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics*, *17*(10), 601–614.

Ulitsky, I., & Bartel, D. P. (2013, July 3). LincRNAs: Genomics, evolution, and mechanisms. *Cell*, *154*(1), 26–46.

Vance, K. W., & Ponting, C. P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in Genetics*, *30*(8), 348–355.

Wan, G., Hu, X., Liu, Y., Han, C., Sood, A. K., Calin, G. A., … Lu, X. (2013). A novel non-coding RNA lncRNA-JADE connects DNA damage signalling to histone H4 acetylation. *EMBO Journal*, *32*(21), 2833–2847.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., & Li, W. (2013). CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*(6), e74–e74.

Wu, Z., Liu, X., Liu, L., Deng, H., Zhang, J., Xu, Q., … Ji, A. (2014). Regulation of lncRNA expression. *Cellular and Molecular Biology Letters*, *19*(4), 561–575.

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., … Derrien, T. (2017). FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, *45*(8), e57–e57.

Xu, H., Cao, L., Sun, B., Wei, Y., & Liang, M. (2019). Transcriptomic analysis of potential "lncRNA-mRNA" interactions in liver of the marine teleost *Cynoglossus semilaevis* fed diets with different DHA/EPA ratios. *Frontiers in Physiology*, *10*(APR), 331.

Xu, J., Meng, Q., Li, X., Yang, H., Xu, J., Gao, N., … Chen, R. (2019). Long noncoding RNA MIR17HG promotes colorectal cancer progression via miR-17-5p. *Cancer Research*, *79*(19), 4882–4895.

Yang, F., Zhang, L., Huo, X. S., Yuan, J. H., Xu, D., Yuan, S. X., … Sun, S. H. (2011). Long noncoding RNA high expression in hepatocellular carcinoma facilitates tumor growth through enhancer of zeste homolog 2 in humans. *Hepatology*, *54*(5), 1679–1689.

Yang, N., Wang, B., Yu, Z., Liu, X., Fu, Q., Cao, M., … Li, C. (2020). Characterization of a novel lncRNA (SETD3-OT) in turbot (*Scophthalmus maximus L.*). *Fish and Shellfish Immunology*, *102*, 145–151.

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., … He, S. (2021). NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Research*, *49*(D1), D165–D171.

Zhu, J. J., Fu, H. J., Wu, Y. G., & Zheng, X. F. (2013). Function of lncRNAs and approaches to lncRNA-protein interactions. *Science China Life Sciences*, *56*(10), 876-885.