De novo profiling of long non-coding RNAs involved in MC-LR–induced liver injury in whitefish: discovery and perspectives

Maciej Florczyk (maciej.florczyk@uwm.edu.pl)
Uniwersytet Warminsko-Mazurski https://orcid.org/0000-0003-2277-312X
Paweł Brzuzan
Uniwersytet Warminsko-Mazurski
Maciej Woźny
Uniwersytet Warminsko-Mazurski

Research article

Keywords: lncRNAs, autonomous 3'UTRs, de novo, MALAT1, non-coding RNAs, ceRNAs, drug-induced liver injury, biomarker, liver transcriptome

Posted Date: October 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-62335/v1

Abstract

Background

Microcystin-LR (MC-LR) is a potent hepatotoxin for which a substantial gap in knowledge persists regarding the underlying molecular mechanisms of liver toxicity and injury. Although long non-coding RNAs (lncRNAs) have been extensively studied in model organisms, and their roles have been identified in various cellular processes including participation in regulation of gene expression together with microRNAs, our knowledge concerning the role of lncRNAs in liver injury is limited even in mammals. Given that lncRNAs show low levels of sequence conservation, their role becomes even more unclear in non-model organisms without an annotated genome, like whitefish *(Coregonus lavaretus)*. The objective of this study was to discover and profile aberrantly expressed polyadenylated lncRNAs that are involved in MC-LR–induced liver injury in whitefish.

Results

Using polyA-enriched RNA-Seq data, we *de novo* assembled a high quality whitefish liver transcriptome. This enabled us to find 94 differentially expressed (DE) putative evolutionary-conserved lncRNAs (orthologous to known lncRNAs in other species), such as MALAT1, HOTTIP, HOTAIR or HULC and 4,429 DE putative novel whitefish lncRNAs, which differed from annotated protein-coding transcripts (PCTs) in terms of minimum free energy, GC base-pair content and length. Additionally, we identified DE non-coding transcripts that might be 3’ autonomous untranslated regions of mRNAs (3’UTRs). We found that, in response to MC-LR treatment, these potential 3’UTRs could either be coexpressed with PCTs from the same mRNA, or the 3’UTRs were upregulated while the corresponding PCTs were downregulated, suggesting 3’UTR-dependent gene regulation.

Conclusions

To our knowledge this is the first report on aberrantly expressed lncRNAs in MC-LR–induced liver injury in whitefish. We found both evolutionary conserved lncRNAs as well as novel whitefish lncRNAs that could serve as biomarkers of severe and chronic liver injury. The lncRNA sequence data files and raw sequence files are available in the Dryad Digital Repository and the NCBI Sequence Read Archive, respectively.

Background

A substantial gap in knowledge persists regarding the role of microcystins (MCs) in the underlying molecular mechanisms of organ toxicity and injury. MCs are a group of cyclic heptapeptide hepatotoxins, of which microcystin-LR (MC-LR) is one of the most widely distributed and potent variants. MC-LR is absorbed, transported and accumulated predominantly in liver [1], and it causes drug-induced liver injury (DILI). Studies on the transcriptomic level have revealed various protein coding transcripts (PCTs) involved in the response and progression of MC-LR–induced liver injury in different species [2, 3]. In addition to PCTs, various non-coding RNA transcripts (ncRNAs, NCTs) have been implicated in the

responses to various stressors, including DILI [4]. MC-LR alters the expression levels of small regulatory ncRNAs (shorter than 200 nt) like microRNAs (miRNAs), piwi-associated RNAs (piRNAs) and small interfering RNAs (siRNAs) [5] in various types of tissues and cells [6, 7].

In comparison to knowledge about small regulatory ncRNAs, our understanding of the functions and mechanism of action of long non-coding RNAs (longer than 200 nt; lncRNAs) is still limited. Unlike miRNAs, lncRNAs are poorly conserved among species [8], which hinders research on their function and evolution. However, it has been shown that lncRNAs are involved in a variety of biological processes such as cell proliferation, apoptosis and differentiation [9] by regulating gene expression via a variety of mechanisms including binding (sponging) miRNAs. In sponging, lncRNA competitively binds to miRNA, resulting in changes in the protein level of coding genes at the post-transcriptional level [10]. LncRNAs may function as competing endogenous RNAs (ceRNAs) that share common miRNA-response elements with PCTs [11]. For example, the recently characterized metastasis-associated-in-lung-adenocarcinoma transcript-1 (MALAT1), a well-conserved lncRNA that is implicated in diseases in humans, was shown to bind MiR-34a in melanoma cells thereby lowering MiR-34a levels [12]. However, the role of MALAT1 in DILI has not yet been elucidated, and in general, our knowledge concerning the role of lncRNAs in DILI is still limited even in mammals [13].

Successful applications of RNA-Seq technology for resolving problems pertinent to fish biology and immunology prompted us to use RNA based methods to investigate patterns of MC-LR–induced liver injury in a teleost fish, the whitefish *(Coregonus lavaretus)*. Our previous results showed that repeated exposure of whitefish to MC-LR results in severe liver damage, followed by an unexpected resilience to further exposures to the toxin and regeneration of the damaged liver structure [14]. We showed that in these adaptations, MC-LR regulates several hepatic miRNA signaling pathways and alters the expression profiles of miRNAs over the short-term [15] and long-term [16], suggesting extensive transcriptome rebuilding during these processes. Because lncRNAs have been implicated in species-specific adaptations (e.g. adaptation of zebrafish to cold [17]), a similar adaptation to MC-LR exposure that involved novel lncRNAs may have occurred in the whitefish

Biologically active lncRNAs are present in zebrafish [18] and rainbow trout [19, 20], however, these are species with well-established annotated genomes, and reference genome is not available for the majority of species. In the absence of an annotated genome, separating NCTs from PCTs requires a more challenging bioinformatic approach. As the foundation for our approach to profile lncRNAs in MC-LR–induced liver injury, we *de novo* assembled a whitefish liver transcriptome. Using a step-by-step pipeline designed to filter out redundant contigs, we were able to identify transcripts without coding potential, which were differentially expressed in whitefish liver after exposure to MC-LR. We identified a list of putative lncRNA transcripts, both novel and orthologous to known lncRNAs in other species. In addition, we showed that treatment of fish with MC-LR may affect levels of non-coding transcripts that may be 3’ autonomous untranslated regions of mRNAs. Furthermore, by showing how MC-LR changes expression patterns of putative lncRNAs, including MALAT1, we extended our knowledge regarding the underlying

mechanisms of liver injury in whitefish. Our findings contribute to better understanding of the role of ncRNA in the molecular response to MC-LR–induced liver injury in fish.

Methods

Fish maintenance and exposure

This study is a part of larger project which aimed to examine changes in the liver transcriptome of whitefish exposed to MC-LR and the effects of intervention with the use of microRNA 92b-3p synthetic analogs (mimic and inhibitor) on the transcriptome. Experimental details concerning maintenance and exposure of the individuals used in this project will be described in a separate papers (Brzuzan et al., Woźny et al. in preparation). In the current study, a total number of 52 individuals from the RNA-Seq of the whole project were used in order to assemble the reference transcriptome (Additional File 3). However, only 24 individuals from the project were used to assess MC-LR effects on expression of non-coding transcripts.

Fish maintenance and exposure were conducted at the Department of Salmonid Research in Rutki (Inland Fisheries Institute in Olsztyn, Poland). The department also provided fish for this study. All animal-related procedures were approved by the Local Ethics Committee for Experiments on Animals in Olsztyn, Poland (resolution No. 44/2016 of 30th November 2016). Juvenile whitefish (mean ±standard deviation: 29.9 ±1.6 g, 17.0 ±0.4 cm) were kept in flow-through tanks supplied with well (underground) water. Water temperature in the tanks was 9.3 ±0.2 ºC and oxygen level was 10.3 ±0.6 $mg \cdot L^{-1}$. Throughout the experiment, all the fish were fed with a minimal feeding procedure dependent on the water temperature, caloric content of the feed, and predicted fish mass. However, 1-2 days prior to exposure (intraperitoneal injections) or collection of samples, the fish were deprived of food.

The dose of MC-LR (100 $\mu g \cdot kg^{-1}$ of body mass) and the treatment periods (1, 6, and 9 days) were based on our previous studies on molecular and physiological responses of whitefish to this toxin [14, 16, 21]. MC-LR (purity ≥ 95%; HPLC) was obtained from Enzo Life Sciences (Enzo Biochem, Inc.; USA) and dissolved in phosphate buffer saline (PBS) as a solvent vehicle. Prior to exposure, randomly selected individuals from each group were anesthetized by immersion in MS-222 solution, and then they received an intraperitoneal injection of the MC-LR solution. To maintain continuous exposure, the MC-LR injection (100 $\mu g \cdot kg^{-1}$ of body mass) was repeated after 7 days of the experiment. Fish that received an intraperitoneal injection with pure PBS served as a negative control group. Throughout the exposure period, fish from the different groups were kept in separate tanks. Each experimental group (PBS or MC-LR) in each treatment period (1, 6 and 9 days) consisted of n = 6 individuals, thus in total 24 individuals were used in our MC-LR-treatment study. The number of fish in each experimental group were estimated based on our previous work [16], where we demonstrate that this group size is sufficient to observe a distinct effect of MC-LR on ncRNA expression pattern.

After each exposure period (1, 6, or 9 days), randomly selected individuals from each group were euthanized by the MS-222 anesthetic overdosing (immersion in 300 ppm solution); and fragments of their livers were collected and preserved in RNAlater solution according to the manufacturer's recommendations (Sigma-Aldrich; Germany). All the fish used in this study were euthanised via an overdose of MS-222.

RNA isolation, sequencing and initial *de novo* assembly

Total RNA was extracted from the RNAlater-preserved liver fragments (approximately 20 mg) using a PureLink RNA Mini Kit (Life Technologies) according to the manufacturer's protocol. To remove the genomic DNA residue, the extracted samples were incuabated with TURBO DNAse (Invitrogen, USA) and then purified using the PureLink RNA Mini Kit. RNA integrity was evaluated with an Agilent Bioanalyzer 2100 with an Agilent 6000 Nano Kit and the samples with mean RIN > 8 were taken for library preparation with the Illumina TruSeq Stranded mRNA Library Prep protocol. The libraries were sequenced with an Illumina HiSeq4000 sequencer (250–300 bp insert cDNA size, PE150, 50M reads, 15Gb).

The workflow used to profile changes in expression of putative ncRNAs in MC-LR induced liver injury in whitefish is shown in Figure 1. Quality control of raw sequencing reads was performed with FastQC, version 0.11.8. To remove adapter sequences and low-quality bases, the reads were processed using Trimmomatic, version 0.36 [23]. After quality trimming, every 6th read (starting from the 6th) was selected for downstream analysis. Selected reads were assembled into a reference genome using Trinity, version 2.5.1 with the default parameters [24]. Trimmed reads were mapped back to the reference genome using Bowtie2, version 2.3.5.1 [25].

ncRNA identification pipeline

The following pipeline was based on [26]. First, the Trinity *de novo* assembled genome was filtered for redundant transcripts using the cd-hit-est algorithm of CD-HIT [27] with a sequence identity threshold of 0.9. Filtering by expression was executed with RSEM [28] implemented by the Trinity-provided perl script 'align_and_estimate_abundance'. Transcripts with expression levels below FPKM=1.50 were filtered out from the data set.

To assess the quality of the filtered *de novo* assembled transcriptome, we quantified its completeness by comparing it with a set of highly conserved single-copy orthologs. Using the BUSCO (Benchmarking Universal Single-Copy Orthologs) v2 pipeline [29], we compared our assembly with the predefined set of 3640 Actinopterygian single-copy orthologs from the OrthoDB v10 database [30]. BUSCO analysis calculated the number of orthologs, whose length was within two standard deviations of the mean length of the given BUSCO (complete BUSCOs, C), complete BUSCOs represented by single-copy transcript (single-copy BUSCOs, S), complete BUSCOs evidenced by more than one transcript (duplicated BUSCOs, D), partially recovered BUSCOs (fragmented BUSCOs, F) and not recovered BUSCOs (missing BUSCOs, M). To verify our assembly, we repeated exactly the same procedure with the most complete whitefish whole transcriptome available to date [31].

Next, the transcripts were searched for open reading frames (ORFs) by Transdecoder, version 2.0.1 [32]. To identify protein coding transcripts (PCTs), ORFs and transcripts were searched against the UniProt-Swiss-Prot and Atlantic salmon proteins reference databases (GCF_000233375.1) using blastp and blastx from the BLAST+ suite with a threshold E-value of 1×10^{-3} [33]. Protein family searches were performed with the Pfam 32.0 database [34] using the ORF protein sequences in HMMER, version 3.2.1. Finally, the top BLAST hit based on the bit score, E-value and percent alignment, and all HMMER hits were loaded into Trinotate, version 3.2.1 to generate an annotation report [35]. Based on the report, transcripts that were not PCTs were then filtered against the RFAM database, version 12.0 [36] by the cmscan algorithm implemented by Infernal, version 1.1.3 [37]. Any hit that Infernal considered significant using the default parameters was filtered out (and labeled as a known ncRNA). All remaining putative novel NCTs were further validated by calculating coding potential using CPC [38].

To further verify that the remaining contigs were completely separated from mRNAs, putative novel non-coding transcripts were subjected to a blastn search against the Reference RNA Sequences database (NCBI, refseq_rna). Any transcript that was identified as “mRNA” was set together in a pair with corresponding PCT of the same mRNA. Only transcripts which had a corresponding PCT were subjected to further analysis (putative autonomous 3’UTRs).

At this point all transcripts that were considered to be either known ncRNAs or putative novel ncRNAs, as well as transcripts identified as Atlantic salmon proteins (PCTs) and putative autonomous 3’UTR transcripts, were counted in each sequenced sample using samtools idxstats (Figure 1B) [39].

Free Energy Levels of Non-Coding Transcripts

The minimum free energy of each transcript was calculated using the rnafold algorithm implemented by ViennaRNA, version 2.4.12 [40] using the following options: -p –d2 –noLP. The minimum free energies of the transcripts were then compared to the minimum free energies of a randomly selected set of protein coding transcripts (Fig. 6A).

Functional annotation of putative autonomous 3’UTRs and PCTs of the same mRNA

GO analysis (http://www.geneontology.org) was performed to construct gene annotations. To retrieve GO IDs for particular proteins, we used the Retrieve/ID mapping tool from the UniProt website [41]. WEGO (Web Gene Ontology Annotation Plot) was used to visualize the results [42]. A p-value < 0.05 was considered to indicate a statistically significant difference.

Real-time PCR

To profile known and putative novel lncRNA expression, reverse transcription (RT) was carried out using SuperScript IV Reverse Transcriptase (Thermo Scientific; USA). The reaction contained 1 µg total RNA, 4 µL 5× RT buffer, 1 µL 0.1 M DTT, 1 µL 10 mM dNTP mix, 1 µL Ribonuclease Inhibitor and SuperScript IV RT enzyme, and 1 µL 50 µM Oligo$(dT)_{20}$ primer. The reaction was carried out at 23°C for 10 min, 55°C for

10 min, and 80°C for 10 min. Synthesized cDNA samples were diluted (20×), stored at −80°C, and thawed only once, just before the amplification.

Real-time PCR was used to determine relative expression of known and putative novel lncRNAs in the cDNA samples. Reactions were carried out in final volumes of 20 µL, consisting of 10 µL Power SYBR Green PCR Master Mix (Life Technologies, USA), 0.25 µM of each primer (forward and reverse; Additional File 1), 1 µL cDNA template and 7 µL PCR-grade water. Amplification was performed with a Quant Studio 5 Real-time PCR System (Applied Biosystems; USA) with the following conditions: 95°C for 10 min, then 45 cycles of 95°C for 15 s and 60°C for 1 min. The reaction for each sample was carried out in duplicate. No-template controls (NTCs) were included to test for the possibility of cross-contamination. To check the quality of each PCR product, melting curve analyses were performed after each run. For normalization of data from the treatment (MC-LR) and control (PBS) groups, “uncharacterized transcript” was used as a reference gene. This transcript was selected based on RNA-Seq results. Its stability was confirmed by Real-time PCR (Cq S.D. ±0.84).

Statistical methods

Contigs that were differentially expressed (DE known and novel lncRNAs, DE 3’UTRs, DE PCTs) in the MC-LR-treated and the control (PBS) groups were indicated using the DESeq2 package, version 1.28.0 [43] for R, version 3.6.3 [44]. Adjusted p-values were calculated with Benjamini and Hochberg’s method using the default settings to maintain a nominal false discovery rate of 0.1 [45]. Thresholds of a $\log_2$ fold-change > |2| and an adjusted p-value < 0.001 were used to filter out contigs with the smaller and less statistically significant differences between groups.

Before assessing the differences in minimum free energy and content of GC base pairs, we used histograms and normal Q-Q plots (shown in Additional File 2) to assess the distribution of the data. These methods indicated that, considering the large sample size, any deviations from normality were too small to be important. Thus, we assessed differences between groups using Welch’s t-test, which is robust to violations of the assumption of homoscedasticity. 95% confidence intervals for differences are shown in square brackets in the main text, e.g., [9.5, 11.7]. Statistical calculations were handled in Python’s statistical modules (Sci-Py, StatsModels). Confidence intervals were calculated using the R Tidyverse package.

Statistical calculations for qPCR data were performed using GenEx 7.0. To assess the significance of difference between groups, one-way independent ANOVA was used after log-transforming the data, followed by Dunnett’s post hoc test.

Data availability

The raw data from this study have been submitted to the NCBI SRA database. The accession numbers for data from the individual samples are given in Additional File 3. *De novo* assembled whitefish liver

transcriptome and sequences of transcripts identified in this study have been deposited in The Dryad Digital Repository (https://datadryad.org).

Results

Sequencing results

The number of raw reads in samples ranged from 50797688 to 70885836. The effective rate ranged from 95.88 to 99.25% ([Clean reads/Raw reads]*100%), with a stable base error rate at 0.03% in all samples. Content of GC base pairs ranged from 47.23 to 50.50%. Detailed statistics on the quality of sequencing data for each sample are presented in Additional File 3.

Assessing the quality of the *de novo* assembled liver transcriptome

The number of detected transcripts in our raw *de novo* assembled liver transcriptome was 1,136,890 with an average length of 367 base pairs. After the assembly, we mapped the trimmed reads back to the assembled liver transcriptome. The fraction of aligned reads was between 97 and 98% per sample. The final assembly, which was used in further analyses, was obtained by filtering out too short, redundant and lowly expressed transcripts. The number of transcripts in our final assembly was 420,280 transcripts with an average length of 594 base pairs. The detailed statistics of the final assembly are presented in Additional File 4.

The BUSCO analysis pipeline revealed that of the 3640 Actinopterygian single-copy orthologs searched, our final assembly completely recovered 74.9% and partially recovered 7.5% (Table 1). 17.6% of the single-copy orthologs were reported missing from our liver transcriptome.

De novo transcriptome assembly allowed discovery and classification of non-coding RNAs in whitefish exposed to MC-LR

Using the procedure for identifying ncRNAs in non-model species first described by Harris et al., we managed to first separate PCTs (148,646 contigs) and then discover various long non-coding transcripts in whitefish (209,270) [26]. Subsequent filtering steps allowed us to separate these transcripts into three non-overlapping groups. The first group contained NCTs that showed homology to sequences deposited in the Rfam database (the group of known non-coding transcripts: 20,272 contigs). The second group contained NCTs that had homology with non-coding 3’ untranslated regions of mRNA sequences deposited in the RefSeq database and had associated PCTs (autonomous 3’UTR/PCT group: 104,024 contigs). The third contained NCTs with no homology to any tested database (putative novel long non-coding RNAs: 84,974 contigs). Our filtering process is summarized in the workflow diagram (Figure 1B).

MC-LR exposure altered expression of evolutionary conserved lncRNAs

To obtain the list of evolutionary conserved lncRNAs, transcripts left after removing PCTs were additionally checked for coding potential with the SVM-based Coding Potential Calculator. Non-coding

transcripts were further compared with sequences of transcripts deposited in the Rfam database. This produced a list of 20,272 contigs, which were counted and analyzed for differential expression (Figure 2A-D). Among 4,238 differentially expressed (DE) evolutionary conserved non-coding RNAs, there were 94 known, putatively conserved DE lncRNAs identified, including MALAT1, HOTTIP, HOTAIR and HULC (Figure 2E-H). To show similarities in sequence conservation between whitefish and 17 other species, including human, we aligned the sequence of our putative MALAT1 transcript with seed sequences of MALAT1 deposited in the Rfam database (Additional File 5).

Co-expression of autonomous 3'UTRs and their associated PCTs after MC-LR exposure

To investigate the effects of MC-LR exposure on the expression of non-coding contigs identified as autonomous 3'UTRs, we first set them together with corresponding PCTs of the same mRNA. Then, using the DESeq2 package from Bioconductor, we separately calculated the differential expression of the putative autonomous 3'UTRs and their associated PCTs. Finally, we compared the percentages of the corresponding contigs that were both up- or down-regulated, and those that were regulated in opposing directions (Fig. 3A). We found that, in the majority of cases, if a putative autonomous 3'UTR was DE, the associated PCT was also DE in the same direction. After 1d of the experiment, over 82% of the transcripts pairs were upregulated and only 11% downregulated. In contrast, after 6d and 9d, up- and down-regulated pairs were present about in the same proportions (around 40%). Moreover, we checked whether the pattern of changes in expression of co-expressed PCT/3'UTR pairs was reflected in the pattern of changes in expression of all PCT (paired with 3'UTRs and unpaired). We found that both expression patterns were similar but not identical (Fig. 4).

Gene Ontology analysis indicated similarities in the terms of pairs that were simultaneously upregulated or downregulated after 6 and 9d of MC-LR exposure (Additional File 6). In contrast, upregulated pairs at 1d were enriched in transcription regulator activity and DNA-binding transcription factor activity transcripts when compared with upregulated pairs at 6 and 9d (Fig. 5). Downregulated pairs at 1d were depleted in transcripts involved in enzyme regulator activity processes when compared with downregulated pairs at 6 and 9d of exposure.

MC-LR produced opposing expression profiles of some autonomous 3'UTRs and their associated PCTs

In addition, we found that MC-LR exposure caused contrasting changes in the expression of some 3'UTRs and their associated PCTs (Fig. 3A, green and brown bars). The proportion of these to all DE transcripts was lowest after 1d of MC-LR exposure (7%), and higher at 6 and 9d (19% and 16% accordingly). On the other hand, the proportions of both groups with the opposite expression profiles to each other remained at a similar level throughout all the days of exposure. In the group with 3'UTRs downregulated and PCTs upregulated, gene ontology terms again showed similarities on days 6 and 9, whereas on day 1, the transcripts were comparatively enriched in terms like 'signaling' or 'response to stimulus'. In contrast, in the group with 3'UTRs upregulated and PTCs downregulated, transcripts involved in 'cell', 'cell part', 'membrane' and 'membrane part' were enriched on days 6 and 9.

Putative novel lncRNAs and PCTs differed in terms minimum free energy, GC base-pair content and length

To determine if the novel whitefish lncRNA candidates differed from PCTs in terms of a minimum free-energy (MFE), we used the RNAfold algorithm from the ViennaRNA package. The free energy values of the secondary structures were corrected for the lengths of the sequences (Fig. 6A). The mean length-corrected MFE of the putative lncRNAs was significantly higher than that of the annotated protein coding transcripts (-0.237 kcal/mol/nt ±0.038758 vs. -0.289 kcal/mol/nt ±0.059464 kcal/mol/nt, respectively, $t(3999) = 46.65$, $p < 0.001$, 95 % [0.04963723, 0.05399174]. Moreover the mean content of GC base pairs was significantly higher in PCTs ($t(3999) = 56.16$, $p < 0.001$, [0.06915164, 0.06448719]) (Fig. 6B). Finally, the distribution of transcript lengths differed between the putative lncRNAs and the annotated PCTs, with more PCTs with longer sequence lengths. (Fig. 6C). In summary, structural differences between the putative lncRNAs and the annotated PCTs validated our methodology for discovery of lncRNAs in whitefish.

MC-LR altered the expression profiles of the identified putative novel lncRNAs

Using the DESeq2 package of Bioconductor we investigated whether MC-LR induced changes in the expression profiles of the putative novel lncRNAs discovered in this study. Using an adjusted p-value of 0.001 and a $\log_2$ fold-change of 2 as cutoffs, we identified 1739 and 2689 transcripts that were either up- or down-regulated by MC-LR exposure, respectively. Figure 7 shows volcano plots of up- and downregulated putative novel lncRNAs (Fig. 7A-C) and Venn diagrams with number of transcripts specific to each time period, as well as those which overlap between days of exposure to MC-LR (Fig. 7D,E). In terms of changes in the expression of the identified transcripts, day 6 and day 9 were more similar to each other than to day 1. 31.0% and 34.8% of all up- and downregulated lncRNA transcripts, respectively, were downregulated on both day 6 and day 9, but not on day 1 of the exposure (Fig. 7D, E).

Real-time PCR confirmed aberrant expression of selected transcripts

To validate the RNA-Seq data, we selected three known lncRNAs (Fig. 8) and 10 putative novel lncRNAs (five upregulated and five downregulated, Additional File 7) and designed a qPCR study that re-analyzed their levels. The qPCR data indicated statistically significant changes in the expression of all selected transcripts except MALAT1 transcripts.

Discussion

In this study, using RNA-Seq data, we identified a list of putative lncRNA transcripts involved in MC-LR induced liver injury in whitefish, a non-model species without a reference genome. Further qPCR validation of selected putative lncRNA candidates confirmed the participation of these transcripts in MC-LR induced liver injury in whitefish. We showed that the altered expression profiles of lncRNAs could serve as a potential biomarkers of liver injury in whitefish.

The lack of standardized methodologies for discovery of lncRNAs poses a challenge for the analysis and interpretation of RNA sequencing data. The outcome of pipelines designed to discover lncRNAs in RNA-Seq data strongly depends on factors which precede *in silico* analysis, such as RNA isolation or the method of preparing sequencing libraries [46]. Therefore, methods designed for enriching polyadenylated protein-coding mRNAs may not be optimal for recovering lncRNAs that are present at low levels. On the other hand, designing large-scale RNA-Seq experiment with a large set of samples is demanding and usually some trade-offs must be made [47]. Because the main aim of our RNA-Seq experiment was to analyze the profiles of protein-coding genes, Illumina's TruSeq Stranded mRNA protocol was chosen (Brzuzan et al., in preparation). However, this protocol can also be used for lncRNA discovery [46]. In fact, the majority of biologically functional lncRNAs reported to date are polyadenylated [48], thus approaches based on enriching polyadenylated transcripts to discover functional lncRNAs are common in pipelines based on annotated genomes, as well as those based on *de novo* assembled transcriptomes. For example 54,503 putative lncRNAs were discovered in rainbow trout [19] and 122,969 putative lncRNAs were reported in *Rhinella arenarum* [49]. Our pipeline based on the *de novo* assembly of the liver transcriptome allowed us to obtain 209,270 non-coding transcripts longer than 200 nt of which 84,974 were labeled as putative novel lncRNAs (see further discussion).

Difficulties in comparing results of RNA-Seq *in silico* analysis based on *de novo* assembled transcriptomes may be attributed to multiple factors. For example, poor reproducibility of *de novo* based analyses [50] could come from the source of the reference genome (i.e. selection of tissues), in addition to the methodologies used in preparation of the sequencing libraries. Last but not least, to obtain reliable and comparable results in discovery of lncRNAs, sequencing depth should be considered. It is estimated that, in human samples, >200 million paired-end reads are required to detect the full range of transcripts, including all possible isoforms [51]. However, this number can be much lower for differential expression analyses. For example, if the expectation is that the expression of abundant transcripts changes across conditions, 36 million reads per sample may be sufficient [47]. Because it was expected that (i) MC-LR will drastically change expression profiles of transcripts [16] and (ii) polyadenylated lncRNAs are expressed at higher abundances than non-polyadenylated lncRNAs [52], we sequenced our liver samples with 50 million reads per sample, which was sufficient to identify evolutionary-conserved and novel transcripts.

In organisms without a conclusively proven reference genome, good quality *de novo* assembled transcripts are a prerequisite for obtaining meaningful results in downstream analysis, such as discovery of novel transcripts [53]. We assembled whitefish liver transcriptome from 52 liver samples that originated from different experimental groups, including some that were part of another study, which extended the scope of available transcripts used for the assembly (Brzuzan et al., in preparation, Additional File 3). BUSCO analysis, which estimates assembly quality based on evolutionary-informed expectations of gene content from orthologues selected from OrthoDB, showed 74.9% completeness of Actinopterygian core genes (OrthoDB v10). For comparison, the current best assembly of a whitefish full-length transcriptome based on a whole fish homogenate showed only slightly higher completeness (76.6%, OrthoDB v10, Table 1) [31]. Importantly, BUSCO recovery estimates tend to be higher in full organism assemblies than in those assembled from a set of separate tissues. For example, a whitefish

tissue-based full-length transcriptome deposited in the PhyloFish database showed only 26% completeness [31]. Because our assembly of a whitefish liver transcriptome showed completeness similar to the current best whitefish whole transcriptome assemblies, we believe that it not only provided a solid foundation for our analysis, but it could also extend the completeness of current and future assemblies of whitefish transcriptomes.

Designing an accurate step-by-step pipeline for discovery of lncRNAs is an essential step for producing high-quality results. This is even more important for non-model species without a reference genome, as redundancy tends to be higher in those analyses. As a consensus state-of-the-art integrated pipeline does not yet exist, we decided to base our pipeline on a previously described pipeline for identification of lncRNAs in non-model species [26]. The core aim of this pipeline was to remove known PCTs and ncRNAs in a sequence of filtration steps. As a result, the pipeline predicts novel lncRNAs, then uses software that employs Support Vector Machine in an attempt to validate the novel transcripts by assesing protein-coding potential. Unfortunately, as both the pipeline and the validation process use BLAST results with varying levels of confidence, this validation is in fact only pseudo-independent, and thus the presented transcripts are predictions at best, and only experimental evidence can validate their true function [26]. However, a lack of an assessment of the sensitivity or the specificity of a pipeline is common among current studies aiming to classify novel lncRNA transcripts, particularly in non-model species without well-annotated genomes. Here, we show that the putative lncRNAs differ substantially form the PCTs in terms of minimum length-corrected free energy, GC content and distribution of transcripts lengths (Fig. 6). These factors are considered crucial for RNA secondary-structure stability, and our results are in line with previous reports from studies of fish, in which lncRNAs differed from PCTs in the same manner [19]. Additionally, to validate the results of our pipeline, we performed a qPCR validation of selected putative lncRNA transcripts (Additional File 7).

Based on the similarity of our transcripts to the non-coding sequences deposited in the Rfam database, we identified putative whitefish liver lncRNAs whose expression was changed after MC-LR administration (known differentially-expressed lncRNAs). We found that MC-LR altered expression of potent regulators of genes, such as HOTAIR, HOTTIP, HULC or MALAT1 (Fig. 2E-H). Generally, lncRNAs are poorly conserved among species [8], but some of them, for example MALAT1, are conserved in mammals [54] as well as in other vertebrates, such as zebrafish [55], suggesting that they have important conserved biological roles. Moreover, MALAT1 is one of the most abundantly expressed lncRNAs in normal tissues [56, 57], even similar in this regard to many protein-coding genes, such as GAPDH [58]. MALAT1 expression has been shown to be either upregulated (lung cancer, hepatocellular carcinoma) or downregulated (colorectal cancer, breast cancer) [57], indicating its role is either cancer-promoting or tumor-suppressing. At the functional level, MALAT1 has been shown to bind various miRNAs, which promoted [59] or decreased cancer progression [60]. For example, knocking down MALAT1 in melanoma cells significantly upregulated the expression of tumor suppressing MiR34a [61]. Interestingly, our previous results demonstrated that MiR34a was upregulated in whitefish liver after MC-LR exposure [62].

Because our RNA-Seq and qPCR results showed that MALAT1 was present at a high level in normal whitefish liver, and most likely its expression was downregulated at 6 and 9d after MC-LR exposure, this could indicate that decreased abundance of MALAT1 may also be linked with upregulated levels of MiR34a in whitefish liver after MC-LR exposure. Although the qPCR data on MALAT1 was not statistically conclusive, it was consistent with the RNA-Seq data, in that both techniques suggested or indicated, respectively, downregulation of this transcript 6 and 9d after MC-LR exposure. Although it is too soon to speculate on whether and how this effect could potentially underlie the molecular mechanisms of MC-LR–induced liver injury, this finding adds to the little that is known about possible lncRNA and miRNA interactions in liver cells.

Previous research on MALAT1 variability has demonstrated that the majority of its transcripts are not polyadenylated [63]. However, it has been revealed that, in mammals, MALAT1 not only has a highly abundant short isoform that lacks a poly(A) tail, but also has a long isoform that is present at a much lower level and has a genomically encoded poly(A) tract. The longer isoform is further processed by RNAse P and RNAse Z to the shorter isoform. In this study, using a library preparation method that enriched only polyadenylated transcripts, we might have detected mainly polyadenylated MALAT1 transcripts, thus confirming the presence of that isoform in fish. To better study the biological function of these lncRNA, further research is required to investigate both the structural diversity and gene regulation of MALAT1.

After removal of PCTs and known NCTs, a large group of the remaining NCTs still mapped to the mRNAs deposited in the Reference Sequence database (RefSeq). However, after closer examination of particular BLAST hits, we noticed that the vast majority of our remaining NCTs mapped not to the coding parts of matched mRNA sequences, but to their non-coding 3’UTR regions. The presence of autonomous 3'UTR transcripts separated from their associated mRNAs has been documented in studies on mouse and human cells [64–66]. Because 3’UTR regions that were considered to be a part of the canonical transcripts are in fact biologically significant autonomous units participating in post-transcriptional regulation [66], we analyzed how expression of our putative autonomous 3’UTR transcripts corresponded to that of the PCT from the same mRNA. We found that, in the normal (unchallenged) condition, almost the same number of mRNAs had a significantly higher number of 3’UTR transcripts (48% of differentially expressed (DE) mRNAs) as those which had a higher number of PCTs (52% of DE mRNAs). This may indicate that, in normal whitefish liver, expression of those transcripts remains in a stable relationship. Moreover, we showed that exposure to MC-LR increased the abundance of PCTs while decreasing that of 3’UTR transcripts form the same mRNA (Fig. 3B). In pairs in which expression was changed after MC-LR exposure, 60% of the pairs had more PCTs than 3’UTR transcripts. In contrast, the same DE pairs showed the opposite pattern of expression in the control samples (i.e. 60% of the same pairs had more 3’UTR transcripts), indicating that MC-LR changed PCT/3’UTR ratio in about 20% of DE mRNAs (depending on the duration of exposure). This also indicated that, after MC-LR exposure, expression of our putative 3’UTR transcripts changed independently of PCTs expression.

In the majority of cases, if a putative autonomous 3'UTR was DE after MC-LR exposure, the associated PCT was also DE in the same direction. Our results show that, out of all DE PCT/3'UTR pairs after 1d of MC-LR exposure, over 82% of the paired transcripts were upregulated (Fig. 3A). In contrast, at after 6 and 9d of exposure, there was no longer a majority of PCT/3'UTR pairs that were DE in the same direction. Moreover, gene ontology terms analysis showed that, after 1d of exposure, the co-upregulated pairs were enriched in transcription regulators, suggesting that these transcripts play roles in regulating the response to severe liver damage (Fig. 5). Additionally, our results show that DE PCT/3'UTR pairs at 6 and 9d of the exposure are similar in terms of function and direction of expression changes (Additional File 6). The observed changes in the liver transcriptome between the 1st and the 6th or 9th day of exposure may support our previous finding that challenging whitefish with MC-LR results in severe liver damage, which is followed by resilience to further exposures to the toxin, allowing for regeneration of the damaged organ [14]. It should be investigated whether this apparent shift in the transcriptome profile reflects remodeling of liver cell processes for repair of the tissue.

Moreover, we also found that, in response to MC-LR exposure, some putative autonomous 3'UTRs and PCTs from the same mRNA were differently expressed in opposite directions, i.e. one was upregulated while the other was downregulated (Fig. 3A green and brown bars). We were particularly interested in pairs where the 3'UTR transcripts were upregulated and the PCTs were downregulated, as this could be attributed to a recently discovered mechanism in which the 3'UTR is cleaved and shortened [66]. The discoverers of that mechanism hypothesized that it serves as a global regulatory tool that works by increasing the effectiveness of miRNA binding sites upstream of the cleavage site. This would cause levels of 3'UTRs to rise after cleavage, while levels of the corresponding PCTs would drop as a result of more effective miRNA binding. Our previous study showed that miRNAs that play roles in transcription regulation are also aberrantly expressed after exposure to MC-LR [16]. Because there is a possible crosstalk between various types of NCTs participating in post-transcriptional regulation of PCTs in MC-LR–induced liver injury, our results suggest the necessity of adopting a wider perspective when investigating the effects of MC-LR–induced liver damage. Researchers looking to discover all aspects of transcriptome regulation by ncRNAs in MC-LR toxicity should focus on decoding the crosstalk between NCTs and PCTs, i.e. competition between endogenous RNAs [67]. It has been shown recently that this strategy can be applied to successfully investigate mechanisms of MC-LR toxicity in other tissues. For example, Meng et. al showed a transcriptomic regulatory network of miRNAs, piRNAs, circular RNAs, lncRNAs and mRNAs that were simultaneously involved in the cytotoxicity of MC-LR in testicular tissues in mice [5].

Although we showed that the putative 3'UTR contigs were expressed independently of PCTs, some of the autonomous 3'UTRs paired with PCTs from the same mRNA could in fact be fragmented or incomplete mRNAs which were not assembled properly. However, even if a *de novo* transcriptome assembly approach for detecting various types of non-coding RNAs is not without flaws, it can be used as an extension to analyses based on annotated genomes. Because the detection of autonomous 3'UTR transcripts using an annotated genome needs additional enrichment of 3'UTR transcripts in the RNA isolation or library preparation steps, it requires additional sequencing runs. Alternatively, pipelines

designed to analyze autonomous 3'UTRs based on the *de novo* assembled transcriptome may be used in a preliminary research, eventually leading to more focused sequencing runs. Furthermore, the huge collection of deposited transcriptomic data may be reused for additional *de novo* analysis by teams seeking direction for their studies on non-coding RNAs.

The aforementioned group of 84,974 putative novel lncRNAs contained transcripts longer than 200 nucleotides with no homology to any tested database. We showed that MC-LR upregulated expression levels of 1739 putative novel lncRNAs and downregulated 2690 (Fig. 7). We found that MC-LR–induced change in expression of putative novel lncRNAs was also reflected in change in expression of PCTs (Fig. 4). Because, as for now, co-expression of lncRNAs with PCTs is the most common approach for identifying potential target genes of lncRNAs [12], this will allow to identify potential target genes of lncRNAs, and to investigate their biological function in MC-LR–induced liver injury in whitefish. Moreover, putative novel lncRNAs might also serve as potential biomarkers for early detection of severe liver injury in whitefish (1d), as well as the fish's recovery form exposure to the toxin, including regeneration of liver tissues (6d, 9d). As demonstrated in certain types of cancers [68], lncRNAs have highly specific expression patterns and relatively stable secondary structures, and are efficiently detected in blood, plasma and urine. With these properties they have the potential to serve as novel noninvasive biomarkers for drug-induced liver injury. For example, our previous results suggests that the non-coding MiRNA-122 can be a non-invasive biomarker for detecting liver damage in fish, and promising alternative to current gold-standard hepatotoxicity markers [69].

The adverse effects of MC-LR in whitefish are not limited only to the liver. For example, we previously showed that MC-LR caused brain injury in whitefish [22]. Although plasma levels of brain-specific MiR124-3p were not altered, it is possible that some brain-specific lncRNAs could serve as biomarkers of brain injury. Interestingly, MALAT1 shows relatively high expression in the brain, where it is involved in regulating synaptogenesis [70]. MALAT1 was downregulated in glioma [71] and is able to regulate levels of MiR124 in various diseases, including Parkinson's disease [72]. Future studies will advance understanding of the roles of lncRNAs in MC-LR toxicity and will likely reveal novel biomarkers and targets for treatment. It will also be important to determine whether the aberrantly expressed lncRNAs detected in this study can be detected in noninvasively collected samples.

Conclusions

In this study, we detected differentially expressed polyadenylated lncRNAs in whitefish exposed to MC-LR. To achieve this, we constructed an extensive liver transcriptome that can be used to complete the current whitefish genome assemblies or to curate ones that are developed in the future. We obtained a dataset that provides a starting point for future studies on the role of lncRNAs in MC-LR–induced liver injury and subsequent liver regeneration. Among the detected DE transcripts, we identified novel, uncharacterized contigs that could potentially be used as non-invasive biomarkers of MC-LR–induced liver injury in whitefish. In addition, we detected transcripts with homology to lncRNAs previously described in other species. Current understanding of lncRNAs is limited [73], and we believe that our work is a step toward

better understanding lncRNA expression and functions in the context of MC-LR toxicity, and the mechanisms of MC-LR toxicity in general. Lastly, we believe that to fully understand the molecular functions of lncRNAs, studies should adopt a broader perspective, including simultaneous analysis of all aspects of the network of competing endogenous RNAs. For that purpose, a combination of different methods of library preparation for RNA sequencing should be considered, which also may allow new types of RNA to be uncovered.

Abbreviations

BLAST: Basic local alignment search tool

BUSCO: Benchmarking set of Universal Single-Copy Orthologs

bp: base pair

DE: Differentially expressed

DILI: Drug-induced liver injury

GO: Gene ontology

lncRNA: long non-coding RNA

MC: microcystin

MC-LR: microcystin-LR

NCBI: National Centre for Biotechnology Information

ncRNAs: non-coding RNAs

NCTs: non-coding transcripts

miRNA: microRNA

ORF: Open Reading frame

PCTs: protein-coding transcripts

piRNA: piwi-associated RNA

RIN: RNA integrity number

RT: Reverse transcription

siRNA: small interfering RNA

SVM: Support Vector Machine

Declarations

Ethics approval and consent to participate

All animal-related procedures were approved by the Local Ethics Committee for Experiments on Animals in Olsztyn (National Ethics Committee for Experiments on Animals, Poland; resolution No. 44/2016 of 30th November 2016). The authors obtained written consent to use the fish in this study from the Department of Salmonid Research in Rutki (Inland Fisheries Institute in Olsztyn, Poland).

Consent for publication

Not applicable.

Availability of data and materials

The raw data from this study have been submitted to the NCBI SRA database. The accession numbers for data from the individual samples are given in Additional File 3. *De novo* assembled whitefish liver transcriptome and sequences of transcripts identified in this study have been deposited in The Dryad Digital Repository (https://datadryad.org).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Wrote the manuscript: MF; designed study: PB, MW; edited the manuscript: PB, MW; carried out experiments: PB, MW, MF; data analysis: MF. All authors read and approved the final manuscript.

Funding

The project was funded by the National Science Centre of Poland (NCN; decision number: 2016/21/B/NZ9/03566). Paweł Brzuzan was the main recipient (Principal Investigator) of the NCN OPUS 11 research grant. The funder (NCN) was not involved in study design, collection and analysis of data, preparation of the manuscript, or decision to publish.

Acknowledgements

We thank Stefan Dobosz, Janusz Krom, Rafał Różyński, and Tomasz Zalewski for their excellent assistance in hatchery operations.

References

1. Fischer WJ, Dietrich DR. Pathological and Biochemical Characterization of Microcystin-Induced Hepatopancreas and Kidney Damage in Carp (Cyprinus carpio). Toxicology and Applied Pharmacology. 2000;164:73–81.
2. Weng D, Lu Y, Wei Y, Liu Y, Shen P. The role of ROS in microcystin-LR-induced hepatocyte apoptosis and liver injury in mice. Toxicology. 2007;232:15–23.
3. Wei L, Sun B, Song L, Nie P. Gene expression profiles in liver of zebrafish treated with microcystin-LR. Environmental Toxicology and Pharmacology. 2008;26:6–12.
4. Shah N, Nelson JE, Kowdley KV. MicroRNAs in Liver Disease: Bench to Bedside. Journal of Clinical and Experimental Hepatology. 2013;3:231–42.
5. Meng X, Peng H, Ding Y, Zhang L, Yang J, Han X. A transcriptomic regulatory network among miRNAs, piRNAs, circRNAs, lncRNAs and mRNAs regulates microcystin-leucine arginine (MC-LR)-induced male reproductive toxicity. Science of The Total Environment. 2019;667:563–77.
6. Feng Y, Ma J, Xiang R, Li X. Alterations in microRNA expression in the tissues of silver carp (Hypophthalmichthys molitrix) following microcystin-LR exposure. Toxicon: Official Journal of the International Society on Toxinology. 2017;128:15–22.
7. Xu L, Qin W, Zhang H, Wang Y, Dou H, Yu D, et al. Alterations in microRNA expression linked to microcystin-LR-induced tumorigenicity in human WRL-68 Cells. Mutation Research. 2012;743:75–82.
8. Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, et al. Neutral evolution of “non-coding” complementary DNAs. Nature. 2004;431:1–2.
9. Zhao L, Wang W. miR-125b suppresses the proliferation of hepatocellular carcinoma cells by targeting Sirtuin7. International Journal of Clinical and Experimental Medicine. 2015;8:18469–75.
10. Paraskevopoulou MD, Hatzigeorgiou AG. Analyzing MiRNA-LncRNA Interactions. Methods in Molecular Biology (Clifton, NJ). 2016;1402:271–86.
11. Beermann J, Piccoli M-T, Viereck J, Thum T. Non-coding RNAs in Development and Disease: Background, Mechanisms, and Therapeutic Approaches. Physiological Reviews. 2016;96:1297–325.
12. Li G, Shi H, Wang X, Wang B, Qu Q, Geng H, et al. Identification of diagnostic long non-coding RNA biomarkers in patients with hepatocellular carcinoma. Molecular Medicine Reports. 2019;20:1121–30.
13. Wen C, Yang S, Zheng S, Feng X, Chen J, Yang F. Analysis of long non-coding RNA profiled following MC-LR-induced hepatotoxicity using high-throughput sequencing. Journal of Toxicology and Environmental Health, Part A. 2018;81:1165–72.
14. Woźny M, Lewczuk B, Ziółkowska N, Gomułka P, Dobosz S, Łakomiak A, et al. Intraperitoneal exposure of whitefish to microcystin-LR induces rapid liver injury followed by regeneration and resilience to subsequent exposures. Toxicology and Applied Pharmacology. 2016;313:68–87.
15. Brzuzan P, Woźny M, Wolińska L, Piasecka A. Expression profiling in vivo demonstrates rapid changes in liver microRNA levels of whitefish (Coregonus lavaretus) following microcystin-LR exposure. Aquatic Toxicology (Amsterdam, Netherlands). 2012;122-123:188–96.

16. Brzuzan P, Florczyk M, Łakomiak A, Woźny M. Illumina Sequencing Reveals Aberrant Expression of MicroRNAs and Their Variants in Whitefish (Coregonus lavaretus) Liver after Exposure to Microcystin-LR. PloS One. 2016;11:e0158899.
17. Jiang N, Meng X, Mi H, Chi Y, Li S, Jin Z, et al. Circulating lncRNA XLOC_009167 serves as a diagnostic biomarker to predict lung cancer. Clinica Chimica Acta. 2018;486:26–33.
18. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Research. 2012;22:577–91.
19. Al-Tobasei R, Paneru B, Salem M. Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. PLoS ONE. 2016;11.
20. Paneru B, Al-Tobasei R, Palti Y, Wiens GD, Salem M. Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in response to infection with Flavobacterium psychrophilum. Scientific Reports. 2016;6.
21. Brzuzan P, Woźny M, Ciesielski S, Łuczyński MK, Góra M, Kuźmiński H, et al. Microcystin-LR induced apoptosis and mRNA expression of p53 and cdkn1a in liver of whitefish (Coregonus lavaretus L.). Toxicon: Official Journal of the International Society on Toxinology. 2009;54:170–83.
22. Florczyk M, Brzuzan P, Łakomiak A, Jakimiuk E, Woźny M. Microcystin-LR-Triggered Neuronal Toxicity in Whitefish Does Not Involve MiR124-3p. Neurotoxicity Research. 2019;35:29–40.
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). 2014;30:2114–20.
24. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology. 2011;29:644–52.
25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357–9.
26. Harris ZN, Kovacs LG, Londo JP. RNA-seq-based genome annotation and identification of long-noncoding RNAs in the grapevine cultivar “Riesling”. BMC Genomics. 2017;18.
27. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
28. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
29. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
30. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research. 2019;47:D807–11.

31. Carruthers M, Yurchenko AA, Augley JJ, Adams CE, Herzyk P, Elmer KR. De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. BMC Genomics. 2018;19:32.
32. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494–512.
33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC bioinformatics. 2009;10:421.
34. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Research. 2019;47:D427–32.
35. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Reports. 2017;18:762–76.
36. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research. 2018;46:D335–42.
37. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
38. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, et al. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research. 2007;35 Web Server issue:W345–349.
39. Li H, Xie P, Li G, Hao L, Xiong Q. In vivo study on the effects of microcystin extracts on the expression profiles of proto-oncogenes (c-fos, c-jun and c-myc) in liver, kidney and testis of male Wistar rats injected i.v. With toxins. Toxicon: Official Journal of the International Society on Toxinology. 2009;53:169–75.
40. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology. 2011;6:26.
41. Consortium TU. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Research. 2019;47:D506–15.
42. Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, et al. WEGO: A web tool for plotting GO annotations. Nucleic Acids Research. 2006;34 Web Server issue:W293–297.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:1–21.
44. R Core Team. R: A Language and Environment for Statistical Computing. 2020.
45. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57:289–300.

46. Chao H-P, Chen Y, Takata Y, Tomida MW, Lin K, Kirk JS, et al. Systematic evaluation of RNA-Seq preparation protocol performance. BMC Genomics. 2019;20.
47. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics. 2014;15:121–32.
48. Zhang Y, Yang L, Chen L-L. Life without A tail: New formats of long noncoding RNAs. The International Journal of Biochemistry & Cell Biology. 2014;54:338–49.
49. Ceschin DG, Pires NS, Mardirosian MN, Lascano CI, Venturino A. The Rhinella arenarum transcriptome: De novo assembly, annotation and gene prediction. Scientific Reports. 2020;10:1–8.
50. Wolfien M, Rimmbach C, Schmitz U, Jung JJ, Krebs S, Steinhoff G, et al. TRAPLINE: A standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinformatics. 2016;17.
51. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Research. 2011;21:2213–23.
52. Kashi K, Henderson L, Bonetti A, Carninci P. Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2016;1859:3–15.
53. Eldem V, Zararsiz G, Taşçi T, Duru IP, Bakir Y, Erkan M. Transcriptome Analysis for Non-Model Organism: Current Status and Best-Practices. Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health. 2017.
54. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC genomics. 2007;8:39.
55. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011;147:1537–50.
56. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC genomics. 2007;8:39.
57. Sun Y, Ma L. New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis. Cancers. 2019;11.
58. Zhang B, Arun G, Mao YS, Lazar Z, Hung G, Bhattacharjee G, et al. The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell Reports. 2012;2:111–23.
59. Chen L, Yao H, Wang K, Liu X. Long Non-Coding RNA MALAT1 Regulates ZEB1 Expression by Sponging miR-143-3p and Promotes Hepatocellular Carcinoma Progression. Journal of Cellular Biochemistry. 2017;118:4836–43.
60. Xia C, Liang S, He Z, Zhu X, Chen R, Chen J. Metformin, a first-line drug for type 2 diabetes mellitus, disrupts the MALAT1/miR-142-3p sponge to decrease invasion and migration in cervical cancer cells. European Journal of Pharmacology. 2018;830:59–67.

61. Li F, Li X, Qiao L, Liu W, Xu C, Wang X. MALAT1 regulates miR-34a expression in melanoma cells. Cell Death & Disease. 2019;10:1–11.
62. Łakomiak A, Brzuzan P, Jakimiuk E, Florczyk M, Woźny M. Molecular characterization of the cyclin-dependent protein kinase 6 in whitefish (Coregonus lavaretus) and its potential interplay with miR-34a. Gene. 2019;699:115–24.
63. Wilusz JE, Freier SM, Spector DL. 3’ end processing of a long nuclear-retained non-coding RNA yields a tRNA-like cytoplasmic RNA. Cell. 2008;135:919–32.
64. Mercer TR, Wilhelm D, Dinger ME, Soldà G, Korbie DJ, Glazov EA, et al. Expression of distinct RNAs from 3' untranslated regions. Nucleic Acids Research. 2011;39:2393–403.
65. Kocabas A, Duarte T, Kumar S, Hynes MA. Widespread Differential Expression of Coding Region and 3’ UTR Sequences in Neurons and Other Tissues. Neuron. 2015;88:1149–56.
66. Malka Y, Steiman-Shimony A, Rosenthal E, Argaman L, Cohen-Daniel L, Arbib E, et al. Post-transcriptional 3-UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments. Nature Communications. 2017;8.
67. Kartha RV, Subramanian S. Competing endogenous RNAs (ceRNAs): New entrants to the intricacies of gene regulation. Frontiers in Genetics. 2014;5.
68. Yarmishyn AA, Kurochkin IV. Long noncoding RNAs: A potential novel class of cancer biomarkers. Frontiers in Genetics. 2015;6.
69. Florczyk M, Brzuzan P, Krom J, Woźny M, Łakomiak A. miR-122-5p as a plasma biomarker of liver injury in fish exposed to microcystin-LR. Journal of Fish Diseases. 2016;39:741–51.
70. Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. The EMBO Journal. 2010;29:3082–93.
71. Han Y, Wu Z, Wu T, Huang Y, Cheng Z, Li X, et al. Tumor-suppressive function of long noncoding RNA MALAT1 in glioma cells by downregulation of MMP2 and inactivation of ERK/MAPK signaling. Cell Death & Disease. 2016;7:e2123.
72. Liu W, Zhang Q, Zhang J, Pan W, Zhao J, Xu Y. Long non-coding RNA MALAT1 contributes to cell apoptosis by sponging miR-124 in Parkinson disease. Cell & Bioscience. 2017;7.
73. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Reports. 2015;11:1110–22.

Tables

Table 1 Summary of the complete, duplicated, fragmented and missing orthologs inferred from Benchmarking Universal Single-Copy Orthologs (BUSCO) search against the orthologs for *Actinopterygii*

BUSCO Statistic	Whitefish liver OrthoDBv10 (this study)	European whitefish whole transcriptome OrthoDBv10 (Carruthers et al. 2018)
Complete BUSCOs (C)	2725 (74.9%)	2786 (76.6%)
Complete and single-copy BUSCOs (S)	1780 (48.9%)	1713 (47.1%)
Complete and duplicated BUSCOs (D)	945 (26.0%)	1073 (29.5%)
Fragmented BUSCOs (F)	272 (7.5%)	320 (8.8%)
Missing BUSCOs (M)	643 (17.6%)	534 (14.6%)
Total BUSCO groups searched	3640	3640

Figures

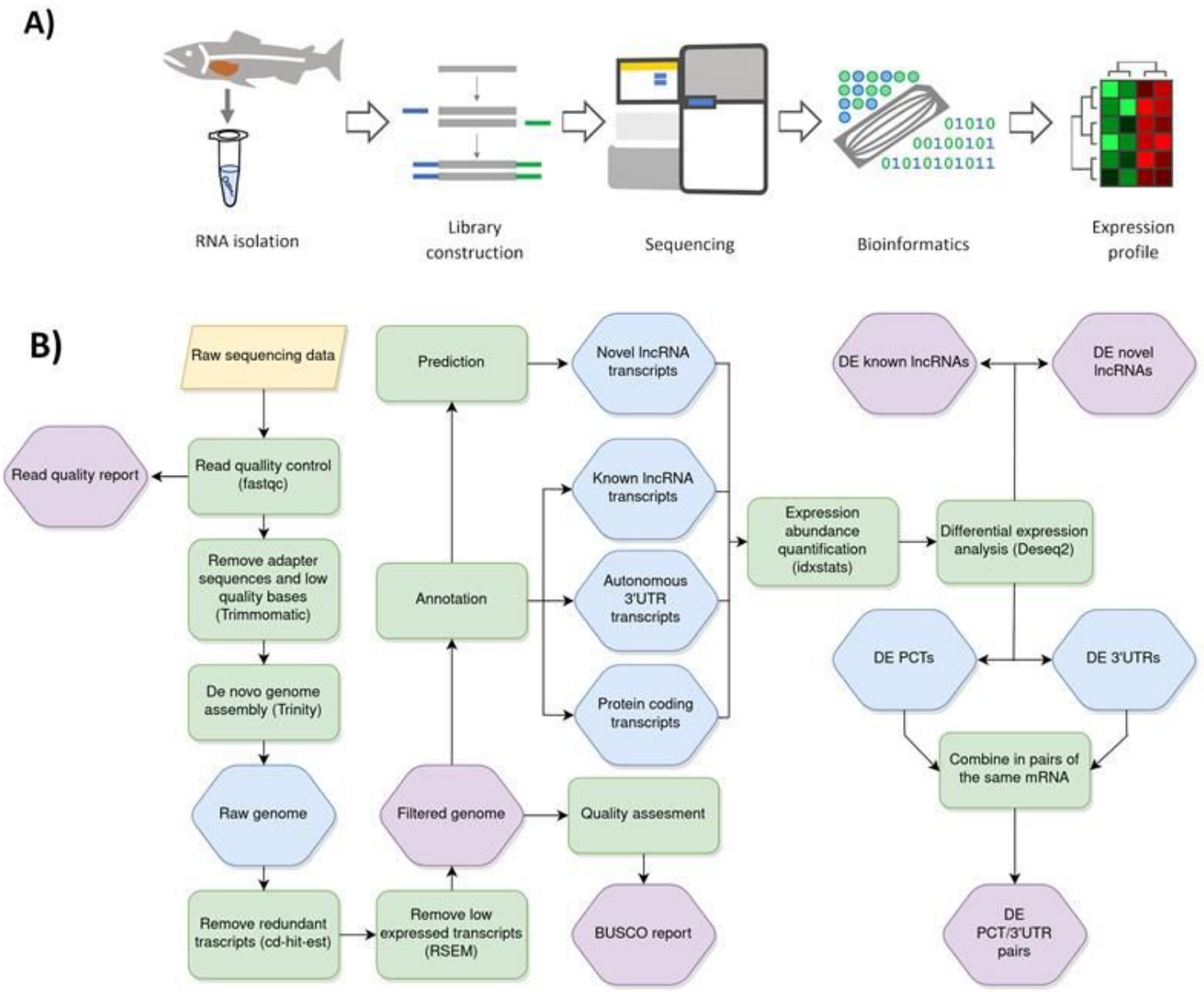

Figure 1

Schematic representation of pipeline used to profile changes in expression of putative long non-coding RNAs (lncRNAs) in microcystin-LR (MC-LR) induced liver damage in whitefish. (A) Overview of the experimental procedure. (B) Bioinformatic analysis workflow. Yellow shapes indicate pipeline input; green shapes indicate action step taken in analysis; blue shapes indicate output of an action; purple shapes indicate final output. DE, differentially expressed; PCTs, protein coding transcripts;

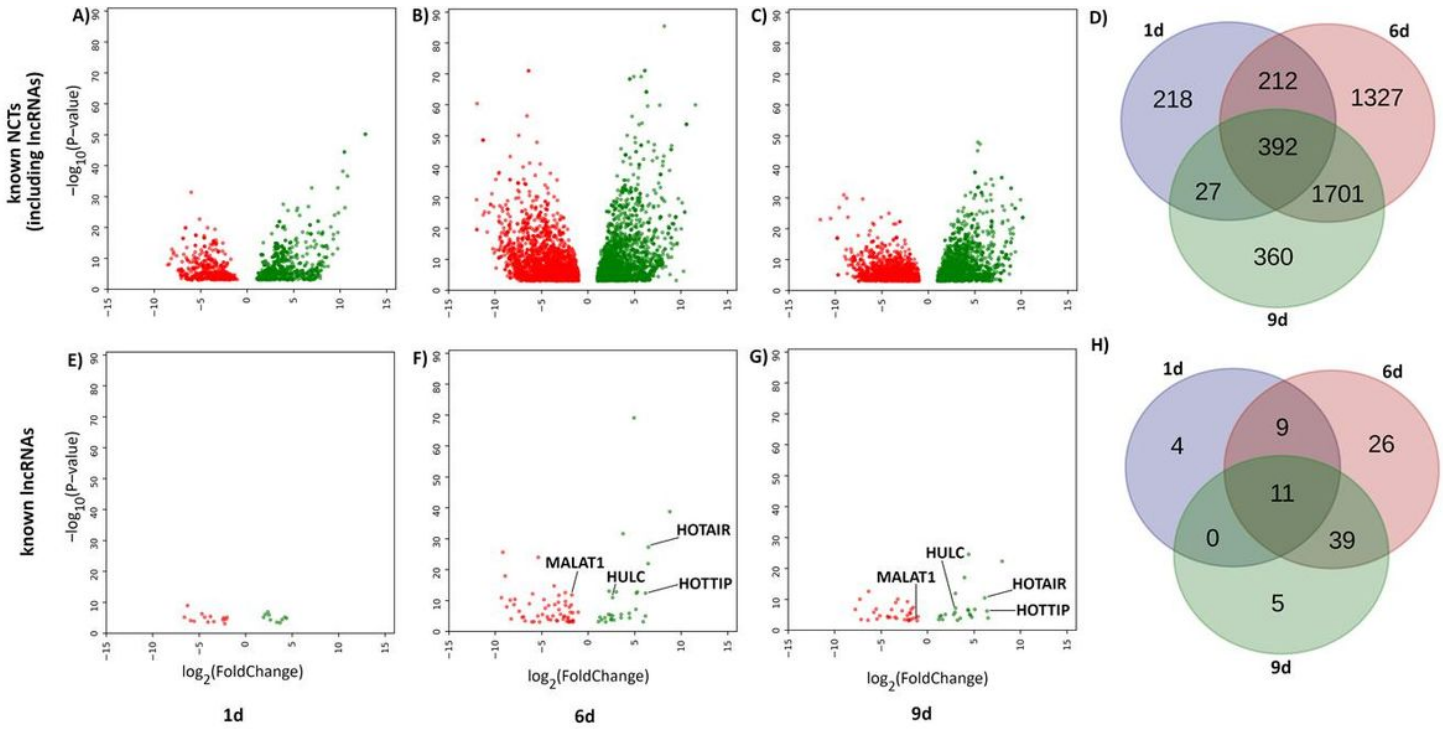

Figure 2

Differentially expressed putative known non-coding RNAs after MC-LR exposure. (A-D) Volcano plots and Venn diagram of differentially expressed (DE) transcripts with homology to any transcript deposited in Rfam database (including lncRNAs). (E-H) Vol-cano plots and Venn diagram of DE transcripts with homology to transcripts labeled as lncRNAs in Rfam database. Expression of putative MALAT1 transcript was downregu-lated at 6d and 9d of microcystin-LR (MC-LR) exposure.

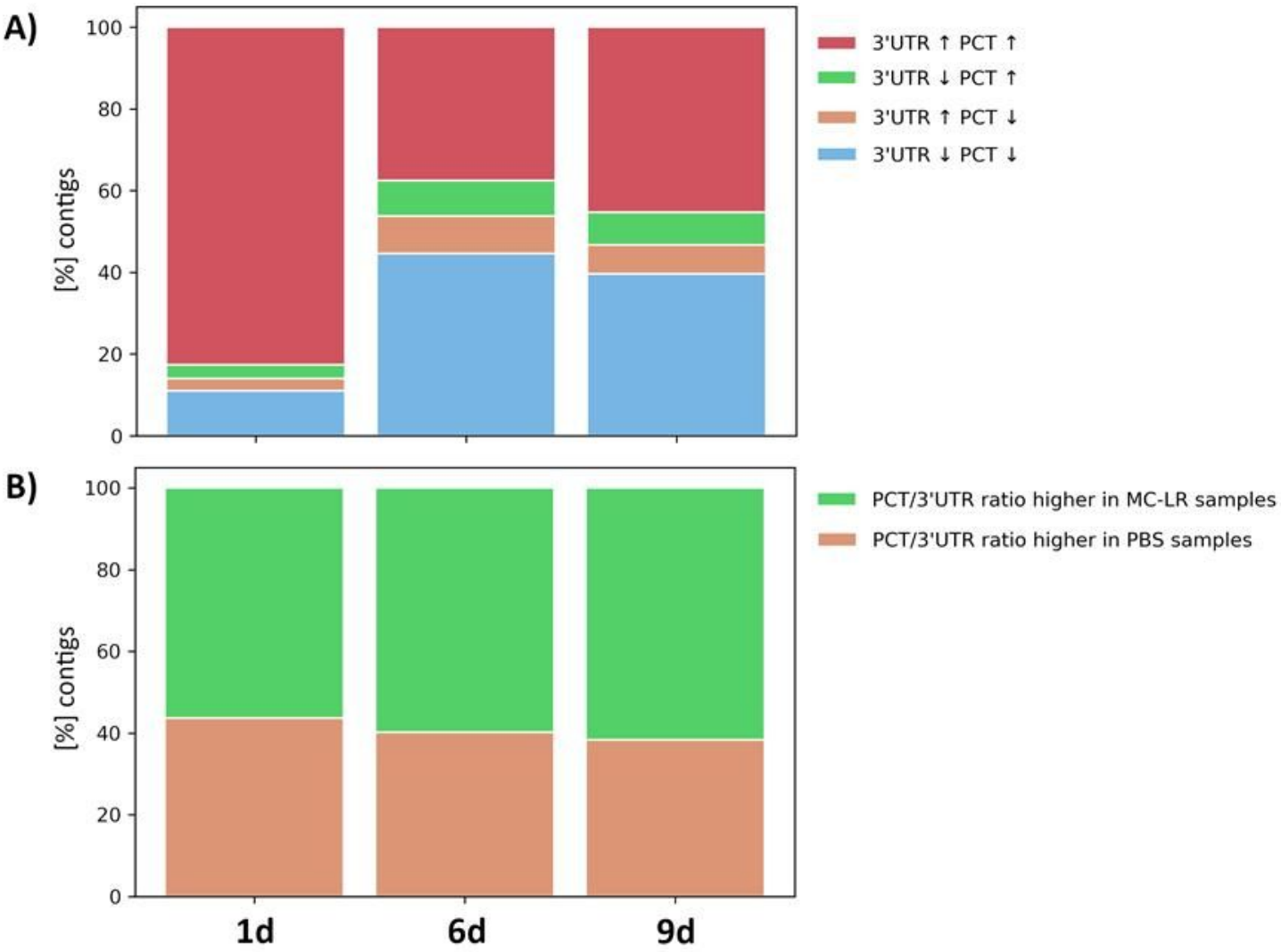

Figure 3

Pairs of differentially expressed (DE) putative autonomous 3’UTRs and protein coding transcripts (PCTs) from the same mRNA in whitefish liver after microcysitin-LR (MC-LR) exposure. (A) Percentages of DE contigs from the same mRNA that both were upregulated (red) or down-regulated (blue), as well as those with opposing expression profiles. (B) Percentages of contigs from the same mRNA for which the ratio of PCT to putative autonomous 3’UTR transcripts was higher in the control (PBS) than in the ex-posed (MC-LR) groups (brown), as well as those for which the ratio was higher in the exposed than in the control groups (green).

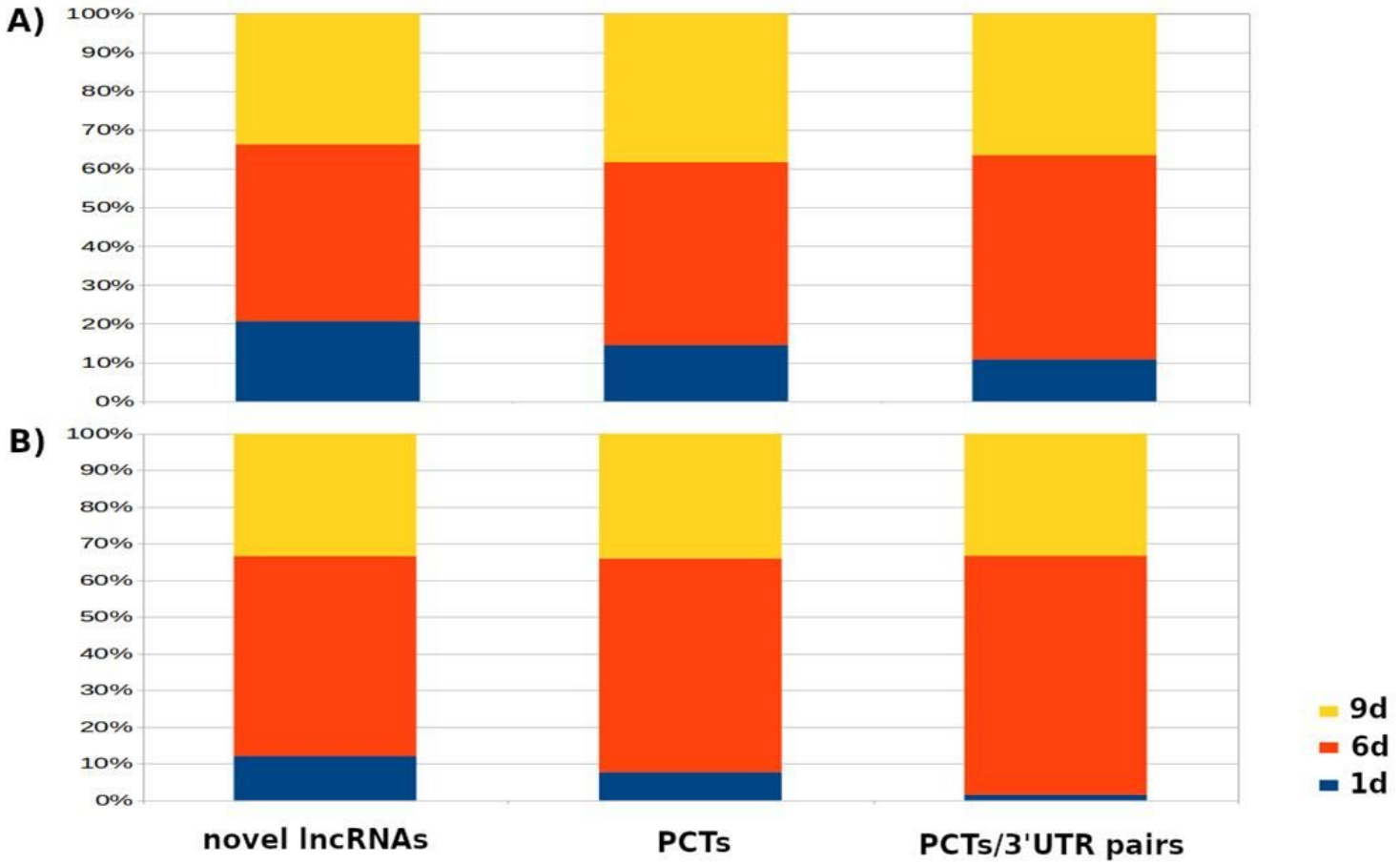

Figure 4

Percentages of differentially expressed (DE) putative novel lncRNAs, protein-coding transcripts (PCTs) and putative PCT/3'UTR pairs, that were either upregulated (A) or downregulated (B) after microcystin-LR exposure. Similarities between all three groups indicate co-expression in response to MC-LR–induced liver injury.

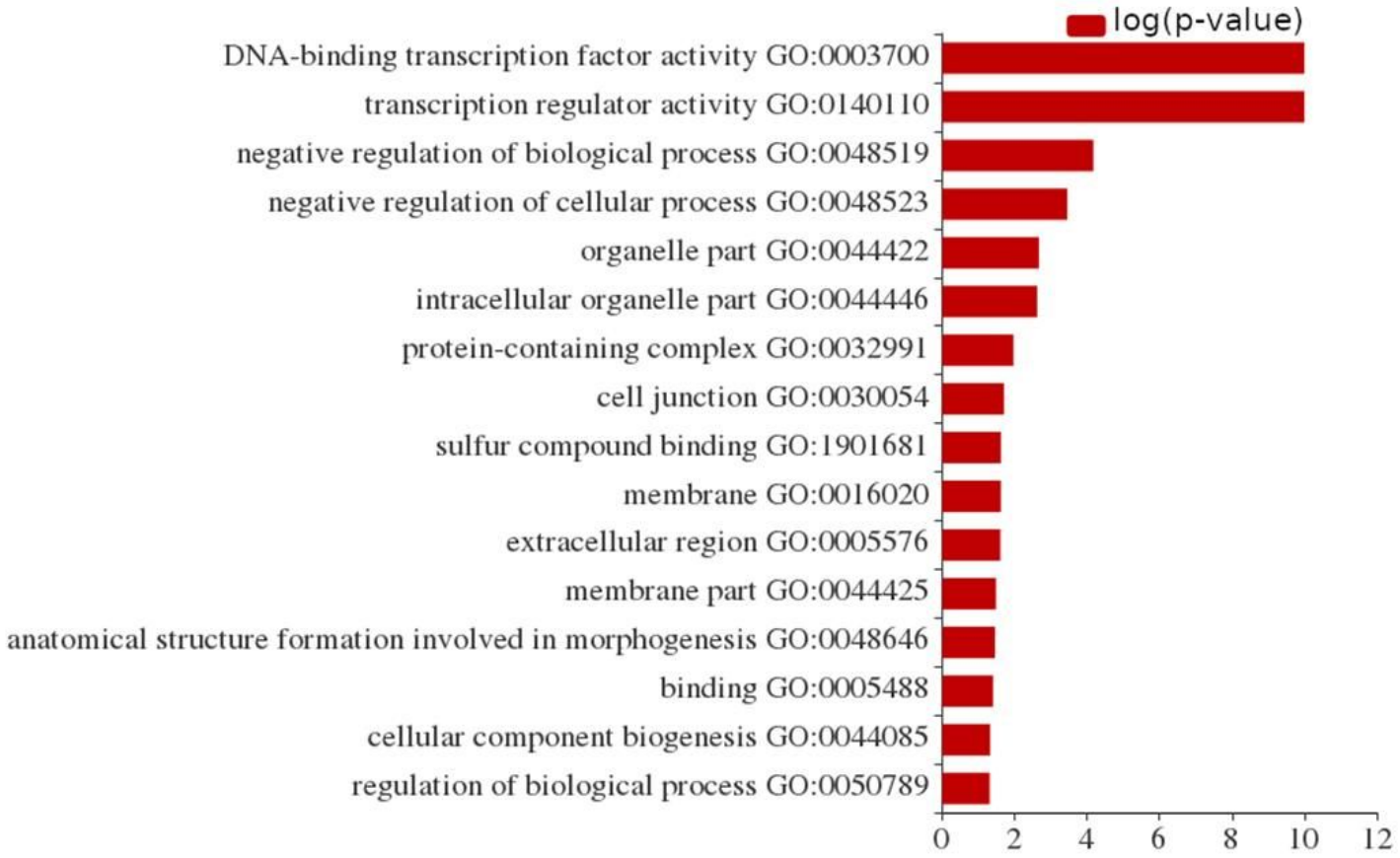

Figure 5

Gene ontology of co-expressed pairs of putative autonomous 3'UTRs and protein coding transcripts (PCTs) from the same mRNA that were upregulated after microcystin-LR (MC-LR) exposure. Figure shows enriched gene ontology terms for pairs that were upregulated on d1 compared to d6 and d9, $p < 0.05$.

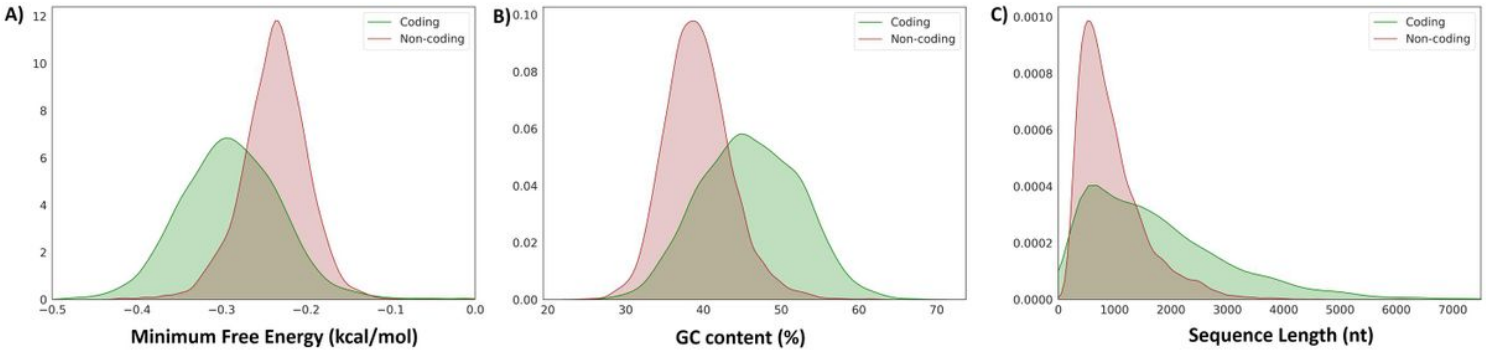

Figure 6

Distributions of length-corrected minimum free energy, content of GC base pairs and transcript lengths differ between protein coding transcripts (PCTs) and putative novel lncRNAs. (A) lncRNA transcripts have a higher mean length-corrected minimum free energy (−0.237 kcal/ mol/nt) than PCTs (−0.289 kcal/ mol/nt); (B) lncRNA transcripts have a lower mean GC base pair content (0.393) than PCTs (0.460); (C)

Distribution of sequence lengths differs between lncRNAs and PCTs. Note that the distribution of PCT sequence lengths includes many longer sequence lengths.

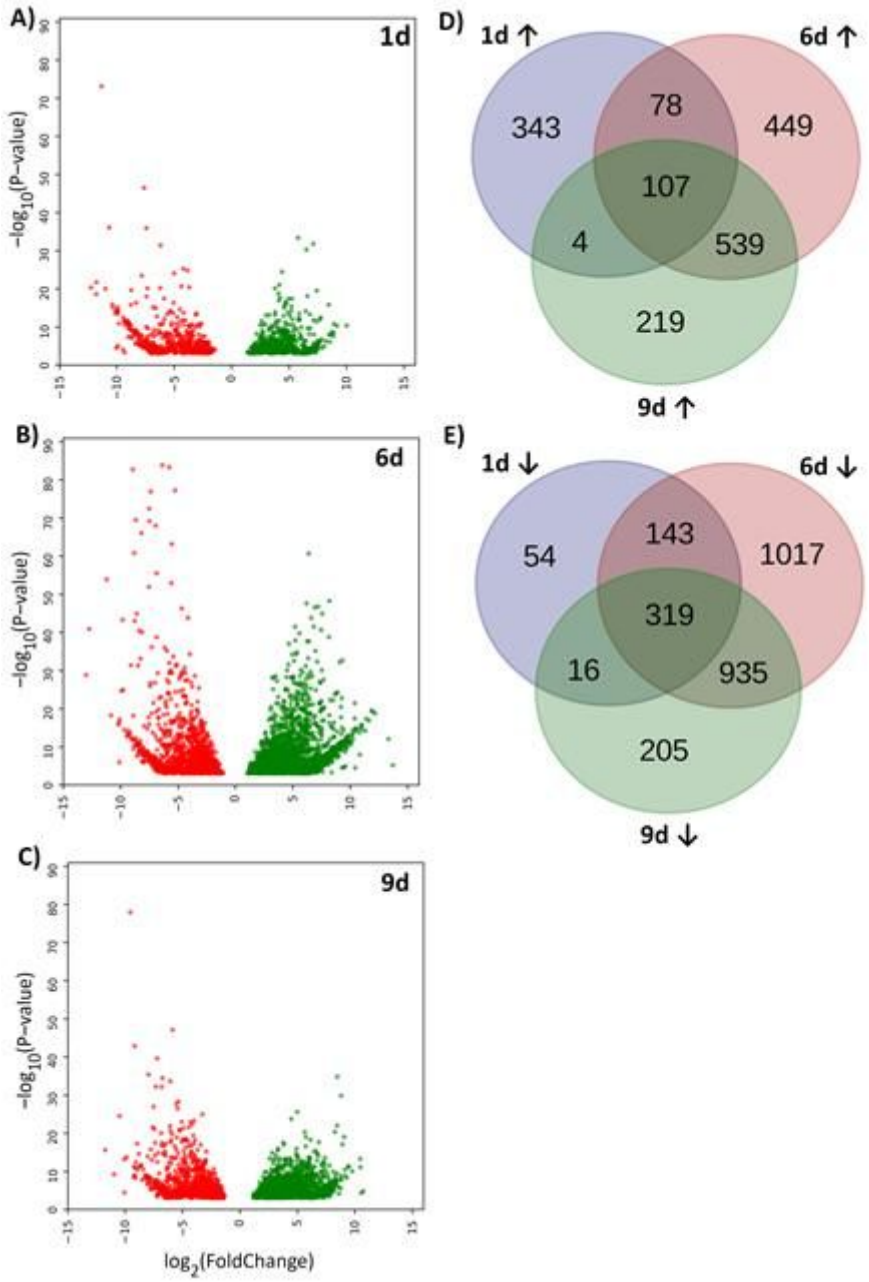

Figure 7

Microcystin-LR (MC-LR) induced differentially expressed (DE) putative novel lncRNAs in whitefish liver. Volcano plots after 1d (A), 6d (B) and 9d (C) of exposure. Venn diagrams of upregulated (D) and downregulated (E) putative novel lncRNAs after MC-LR exposure. Note that days 6 and 9 are more similar to each other than to day 1 in terms of which lncRNAs were DE on those days.

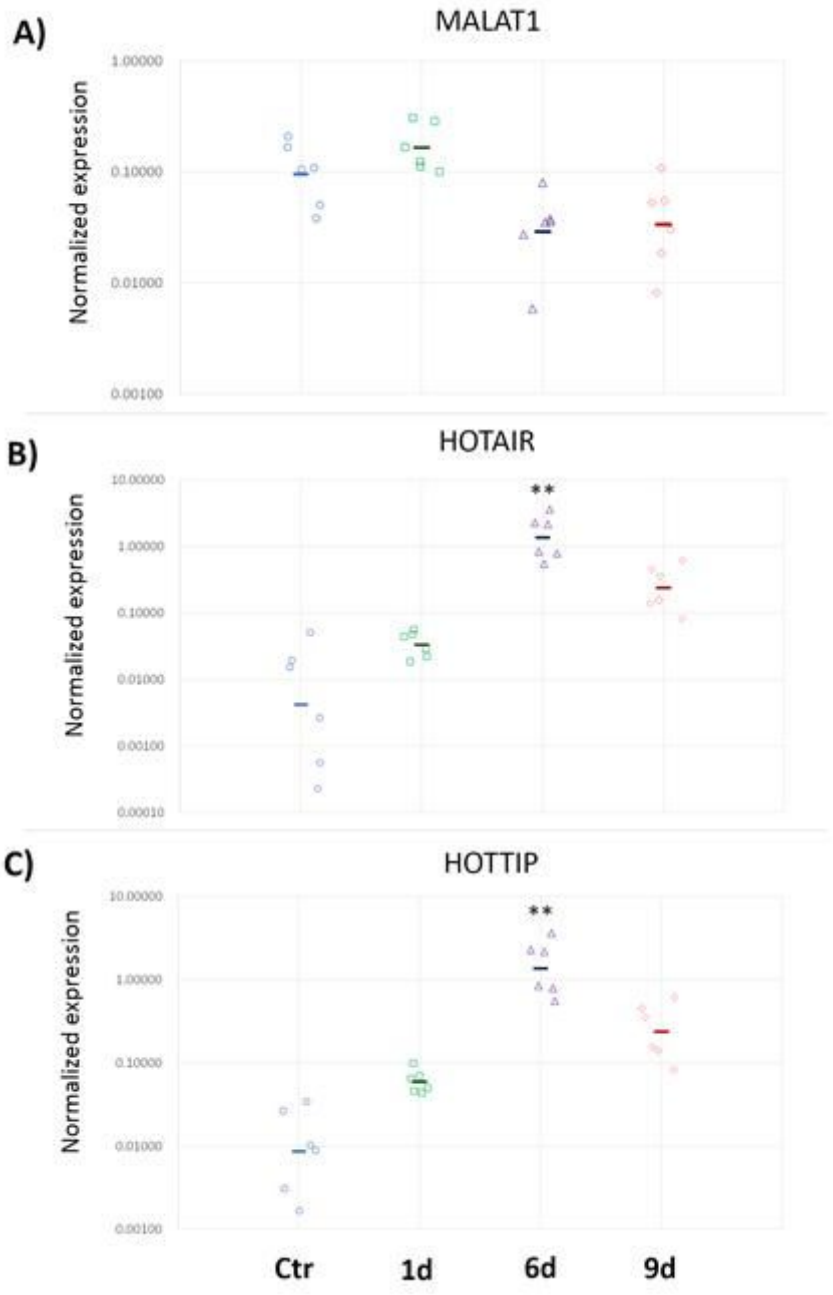

Figure 8

Expression of putative MALAT1 (A), HOTAIR (B) and HOTTIP (C) in whitefish liver after 1d, 6d and 9d of MC-LR exposure quantified using RT-qPCR. ** $p < 0.01$. Ctr - control, unchallenged group. Points represent individual fish in respective experimental group.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AuthorChecklistFullMF.pdf

- AdditionalFile7.tif
- AdditionalFile6.tif
- AdditionalFile5.doc
- AdditionalFile4.doc
- AdditionalFile3.xls
- AdditionalFile2.tif
- AdditionalFile1.doc