RNA sequencing analysis

Processing of next-generation sequencing followed an updated pipeline for gene-level analysis (Anders et al. 2013; Garcia et al. 2018). Briefly, reads were evaluated by FastQC v0.11.3 (Andrews 2015) to detect major sequencing problems and then trimmed for quality control with Skewer v0.2.2 (Jiang et al. 2014) to remove adapters and ends of reads with low mean Phred quality score (Skewer options: `-x FRC_adapters.fa --mode any --end-quality 30 --mean-quality 30 --min 30 --format auto --compress --threads 9`, FRC_adapters.fa). RNA-seq alignment and quantification proceeded with Bowtie2 v2.2.3 (Langmead et al. 2009; Langmead and Salzberg 2012) being used to build HISAT2 (Kim et al. 2015; Kim et al. 2019) genome index files from the Genome Reference Consortium Zebrafish Build 10 (GRCz10) genome downloaded from Ensembl (release 91, ftp://ftp.ensembl.org/pub/release-91/fasta/danio_rerio/dna/Danio_rerio.GRCz10.dna.toplevel.fa.gz). To enhance the genome index files, HISAT2 scripts, hisat2_extract_splice_sites.py and hisat2_extract_exons.py, were used to extract splice site and exon information, respectively, from the Ensembl 91 gene transfer format (GTF) file (ftp://ftp.ensembl.org/pub/release-91/gtf/danio_rerio/Danio_rerio.GRCz10.91.gtf.gz), and a modified version of the script, hisat2_extract_snps_haplotypes_VCF.py, was created to allow for extraction of SNPs and haplotypes from the Ensembl 91 variant call format (VCF) file (ftp://ftp.ensembl.org/pub/release-91/variation/vcf/danio_rerio/danio_rerio.vcf.gz, hisat2_extract_snps_haplotypes_VCF_CD.py). The extracted genomic information was incorporated into HISAT2 index files, using the `hisat2-build` command. After building enhanced genome index files, trimmed reads were aligned with HISAT2 v2.1.0 using the following options: `-q --phred33 --new-summary --downstream-transcriptome-assembly --threads 9`. Samtools v1.4 (Li et al. 2009) was used to convert aligned reads to binary alignment style (BAM) and then to sort BAM reads by position using the "`view -u`" and "`sort -@9`" commands, respectively. Gene counts were estimated using the `htseq-count` command from HTSeq v0.9.1 (Anders et al. 2015) with the GRCz10 Ensembl 91 GTF annotation (options: `--format=bam --stranded=yes --type=exon --idattr=gene_id --additional-attr=gene_name --mode=intersection-nonempty`).

Analysis of gene counts was conducted using R v3.6.0 (R Core Team 2019) and Bioconductor v3.9 (Gentleman et al. 2004; Huber et al. 2015) packages in the RStudio v1.2.1335 (RStudio Team 2018) integrated development environment with a customized script (Differential_gene_analysis_custom_script.R) based on a maintained Bioconductor workflow package from the Gordon Smyth lab, https://www.bioconductor.org/packages/release/workflows/vignettes/RnaSeqGeneEdgeRQL/inst/doc/edgeRQL.html (Chen et al. 2016). The Bioconductor package, edgeR v3.26.0, was used to normalize gene counts and determine differential expression (Lun et al. 2016; McCarthy et al. 2012; Robinson and Smyth

2007, 2008; Robinson et al. 2010; Robinson and Oshlack 2010). Briefly, genes were filtered to exclude those with low counts across libraries, only keeping genes expressed in a minimum of four samples with average counts per million reads per sample above 0.97, which corresponds to a minimum read count of 10-20 (Chen et al. 2016; Lun et al. 2016). Filtered genes were then normalized across samples using the trimmed mean of M values (TMM) method to minimize composition bias between libraries (Robinson and Oshlack 2010). Differential expression above a logarithmic fold change threshold of $\log_2(1.5)$ between experimental and control samples was determined with functions from edgeR, which uses the negative binominal generalized linear model extended by quasi-likelihood methods to fit the count data, the Cox-Reid profile-adjusted likelihood method to calculate dispersions, empirical Bayes quasi-likelihood F-tests to calculate differential expression, and a version of the TREAT method to determine significance above a threshold (Chen et al. 2014; Lun et al. 2016; McCarthy and Smyth 2009; McCarthy et al. 2012). The '`robust=TRUE`' option was used to protect the empirical Bayes estimates against the possibility of outlier genes with wide-ranging individual dispersions. Genes with a Benjamini-Hochberg (BH) adjusted $p \leq 0.05$ were considered significantly differentially expressed. The biomaRt v2.40.0 package was used to connect Ensembl gene identifier information to Ensembl BioMart annotation information (e.g. gene symbols, biotypes, human orthologs, etc.) To explore the data, interactive multidimensional scaling (MDS), volcano, and mean-difference (MD) plots were created using the Glimma v1.12.0 package (Su et al. 2017). To understand the functional consequences of chemical exposure, we performed biological process network enrichment analysis and Gene Ontology (GO) term enrichment on the significant differentially expressed human orthologs including protein complexes using GeneGo MetaCore version-19.3 build-69800 from Clarivate Analytics, as described in Garcia *et al*. (2018; The Gene Ontology Consortium 2019). Only enriched biological process networks and GO terms with a false discovery rate (FDR) adjusted $p \leq 0.05$ were considered significant.

miRNA sequencing analysis

Small RNA sequencing reads underwent two initial rounds of quality control and adapter trimming with Skewer v0.2.2 (Jiang et al. 2014) to remove excess adapters and low quality bases (Skewer 1 and 2 options: `-x FRC_miRNA_adapters.fa --mode any --end-quality 30 --mean-quality 30 --format auto --threads 9`), and was followed by a third round that also excluded reads above 30 bases (Skewer 3 options: `-x FRC_miRNA_adapters.fa --mode any --end-quality 30 --mean-quality 30 --max 30 --format auto --threads 9`; FRC_miRNA_adapters.fa). To prepare reads for miRDeep2 (Friedlander et al. 2012) analysis, FastX-Toolkit v0.0.13 (Gordon and Hannon 2010) was used to convert fastq files to fasta format (options: `-Q33 -n`) and whitespace was removed from fasta read files using the miRDeep2 v2.0.0.8 script, remove_white_space_in_id.pl. To prepare genome index files for read mapping, whitespace was removed from the GRCz10 genome using the remove_white_space_in_id.pl script provided with miRDeep2 v2.0.0.8, and then Bowtie v1.2.1.1

(Langmead et al. 2009) was used to build Bowtie index files from the whitespace-free GRCz10 genome with default bowtie-build settings. The whitespace-free fasta read files were then mapped to the genome using a modified miRDeep2 v2.0.0.8 mapper.pl script, which is basically a wrapper that maps reads using custom Bowtie options and then outputs them into desired formats for use with miRDeep2 (options: `FRCs_miRNA_config_file.txt -d -c -m -v -p <GRCz10_no_whitespace> -q -o 30`; mapper_CLD_Q33.pl, FRCs_miRNA_config_file.txt).

To aid in miRNA identification and quantification, mature and hairpin miRNA reference files (ftp://mirbase.org/pub/mirbase/22/mature.fa.gz, ftp://mirbase.org/pub/mirbase/22/hairpin.fa.gz) were downloaded from miRBase release 22 (Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008; Kozomara and Griffiths-Jones 2011, 2014; Kozomara et al. 2019). The miRDeep2 v2.0.0.8 script, extract_miRNAs.pl, was used to select zebrafish mature and hairpin miRNAs, as well as to select other mature miRNAs from *Tetraodon nigroviridis*, *Fugu rubripes*, *Xenopus tropicalis*, *Mus musculus*, and *Homo sapiens* to help the novel prediction algorithm. The base miRDeep2 v2.0.1.2 script, miRDeep2.pl, was modified to replace the internal make_html.pl part of the script with make_html_2012_CLD.pl, which allows Rfam Bowtie index files to be created in the working directory instead of the installation directory (make_html_2012_CLD.pl, miRDeep2_2012_CLD.pl), and this modified script was used to quantify and predict new miRNAs (options: `<READS_FILE.fa> <GENOME_FILE.fa> <READS_FILE.arf> <EXTRACTED_ZEBRAFISH_MATURE_miRNAS.fa> <EXTRACTED_OTHER_MATURE_miRNAs.fa> <EXTRACTED_ZEBRAFISH_HAIRPIN_miRNAS.fa> -a 10 -t Zebrafish -P`).

Analysis of miRNA counts was conducted using R v3.6.0 (R Core Team 2019) and Bioconductor v3.9 (Gentleman et al. 2004; Huber et al. 2015) packages in the RStudio v1.2.1335 (RStudio Team 2018) integrated development environment with a customized script (Differential_miRNA_analysis_custom_script.R). The Bioconductor package, edgeR v3.26.0, was used to normalize gene counts and determine differential expression (Chen et al. 2016; Lun et al. 2016; McCarthy et al. 2012; Robinson and Smyth 2007, 2008; Robinson et al. 2010; Robinson and Oshlack 2010). After averaging the count values of miRNAs derived from multiple precursors, miRNAs were filtered to exclude those with low counts across libraries, only keeping miRNAs expressed in a minimum of four samples with average counts per million reads per sample above 1.51, which corresponds to a minimum read count of 10-20 (Chen et al. 2016; Lun et al. 2016). Filtered genes were then normalized across samples using the TMM method to minimize composition bias between libraries (Robinson and Oshlack 2010). Differential expression between experimental and control samples was determined with functions from edgeR, which uses the negative binominal generalized linear model extended by quasi-likelihood methods to fit the count data, the Cox-Reid profile-adjusted likelihood method to calculate dispersions, and empirical Bayes quasi-likelihood F-tests to calculate differential expression (Chen et al. 2014; Lun et al. 2016; McCarthy et al. 2012). The '`robust=TRUE`' option was used to protect the

empirical Bayes estimates against the possibility of outlier genes with wide ranging individual dispersions. Genes with a BH-adjusted $p \leq 0.05$ were considered significantly differentially expressed. To explore the data, interactive MDS, volcano, and MD plots were created using the Glimma v1.12.0 package (Su et al. 2017).

miRNA Target Identification

The Bioinformatics Resource Manager (BRM) was used to determine if significantly differentially expressed genes identified in the RNA-seq data overlapped with the potential targets of significantly differentially expressed miRNAs (Brown et al. 2019; Tilton et al. 2012). The BRM, an online tool that facilitates genomic identifier retrieval and integration, includes a workflow for connecting miRNAs with target genes. It leverages computationally predicted miRNA target data from MicroCosm (Griffiths-Jones et al. 2006), TargetScan (Agarwal et al. 2015), and microRNA.org (Betel et al. 2008), and experimentally validated miRNA target data from miRTarBase (Chou et al. 2018). Repressive interactions were identified by uploading significantly differentiated upregulated miRNAs to the miRNA Targets workflow, selecting all target databases with required hits from at least one of the four databases, and then merging the results of the miRNA targets database search with the list of significantly differentiated downregulated genes identified from the RNA-seq data. Similarly, derepressive interactions were identified by uploading significantly differentiated downregulated miRNAs to the miRNA Targets workflow, selecting all target databases with required hits from at least one of the four databases, and then merging the results of the miRNA targets database search with the list of significantly differentiated upregulated genes identified from the RNA-seq data. Of the four selected databases, only MicroCosm, TargetScan, and miRTarBase include zebrafish data; therefore, high confidence interactions were selected when met by one of the following criteria: 1) inclusion in miRTarBase, 2) prediction by both MicroCosm and TargetScan, 3) prediction by MicroCosm with $p \leq 0.01$, or 4) prediction by TargetScan with a weighted context++ score ≥ 50.

References

Agarwal V, Bell GW, Nam JW, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 4: e05005. https://doi.org/10.7554/eLife.05005.

Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature Protocols. 8(9): 1765-1786. https://doi.org/10.1038/nprot.2013.099.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a python framework to work with high-throughput sequencing data. Bioinformatics. 31(2): 166-169. https://doi.org/10.1093/bioinformatics/btu638.

Andrews S. 2015. FastQC: A quality control tool for high throughput sequence data. 0.11.9. Cambridge, UK: Babraham Bioinformatics. Online: 10 June 2020. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Betel D, Wilson M, Gabow A, Marks DS, Sander C. 2008. The microRNA.Org resource: Targets and expression. Nucleic Acids Research. 36(Database issue): D149-153. https://doi.org/10.1093/nar/gkm995.

Brown J, Phillips AR, Lewis DA, Mans MA, Chang Y, Tanguay RL, et al. 2019. Bioinformatics Resource Manager: A systems biology web tool for microRNA and omics data integration. BMC Bioinformatics. 20(1): 255. https://doi.org/10.1186/s12859-019-2805-6.

Chen Y, Lun ATL, Smyth GK. 2014. Differential expression analysis of complex RNA-seq experiments using edgeR. In: Statistical analysis of next generation sequencing data, (Datta S, Nettleton D, eds). Cham, Switzerland:Springer International Publishing, 51-74.

Chen Y, Lun AT, Smyth GK. 2016. From reads to genes to pathways: Differential expression analysis of RNA-seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 5: 1438. https://doi.org/10.12688/f1000research.8987.2.

Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, et al. 2018. miRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. Nucleic Acids Research. 46(D1): D296-D302. https://doi.org/10.1093/nar/gkx1067.

Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. Mirdeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research. 40(1): 37-52. https://doi.org/10.1093/nar/gkr688.

Garcia GR, Shankar P, Dunham CL, Garcia A, La Du JK, Truong L, et al. 2018. Signaling events downstream of AHR activation that contribute to toxic responses: The functional role of an AHR-dependent long noncoding RNA (slincR) using the zebrafish model. Environ Health Persp. 126(11): 117002. https://doi.org/10.1289/EHP3281.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology. 5(10): R80. https://doi.org/10.1186/gb-2004-5-10-r80.

Gordon A, Hannon G. 2010. FASTX-toolkit. Version 0.0.13Cold Spring Harbor Laboratory. http://hannonlab.cshl.edu/fastx_toolkit/index.html.
Griffiths-Jones S. 2004. The microRNA registry. Nucleic Acids Research. 32(Database issue): D109-111. https://doi.org/10.1093/nar/gkh023.
Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research. 34(Database issue): D140-144. https://doi.org/10.1093/nar/gkj112.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: Tools for microRNA genomics. Nucleic Acids Research. 36(Database issue): D154-158. https://doi.org/10.1093/nar/gkm952.
Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods. 12(2): 115-121. https://doi.org/10.1038/nmeth.3252.
Jiang H, Lei R, Ding SW, Zhu S. 2014. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 15: 182. https://doi.org/10.1186/1471-2105-15-182.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 12(4): 357-360. https://doi.org/10.1038/nmeth.3317.
Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37(8): 907-915. https://doi.org/10.1038/s41587-019-0201-4.
Kozomara A, Griffiths-Jones S. 2011. miRBase: Integrating microRNA annotation and deep-sequencing data. Nucleic Acids Research. 39(Database issue): D152-157. https://doi.org/10.1093/nar/gkq1027.
Kozomara A, Griffiths-Jones S. 2014. miRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research. 42(Database issue): D68-73. https://doi.org/10.1093/nar/gkt1181.
Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: From microRNA sequences to function. Nucleic Acids Research. 47(D1): D155-D162. https://doi.org/10.1093/nar/gky1141.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 10(3): R25. https://doi.org/10.1186/gb-2009-10-3-r25.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods. 9(4): 357-359. https://doi.org/10.1038/nmeth.1923.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25(16): 2078-2079. https://doi.org/10.1093/bioinformatics/btp352.

Lun AT, Chen Y, Smyth GK. 2016. It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. Methods in Molecular Biology. 1418: 391-416. https://doi.org/10.1007/978-1-4939-3578-9_19.

McCarthy DJ, Smyth GK. 2009. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 25(6): 765-771. https://doi.org/10.1093/bioinformatics/btp053.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Research. 40(10): 4288-4297. https://doi.org/10.1093/nar/gks042.

R Core Team. 2019. R: A language and environment for statistical computing. Version 3.6.0. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson MD, Smyth GK. 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 23(21): 2881-2887. https://doi.org/10.1093/bioinformatics/btm453.

Robinson MD, Smyth GK. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 9(2): 321-332. https://doi.org/10.1093/biostatistics/kxm030.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1): 139-140. https://doi.org/10.1093/bioinformatics/btp616.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 11(3): R25. https://doi.org/10.1186/gb-2010-11-3-r25.

RStudio Team. 2018. RStudio: Integrated development for R. Version 1.2.1335. Boston, MA: RStudio, Inc., PBC. http://www.rstudio.com/.

Su S, Law CW, Ah-Cann C, Asselin-Labat ML, Blewitt ME, Ritchie ME. 2017. Glimma: Interactive graphics for gene expression analysis. Bioinformatics. 33(13): 2050-2052. https://doi.org/10.1093/bioinformatics/btx094.

The Gene Ontology Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research. 47(D1): D330-D338. https://doi.org/10.1093/nar/gky1055.

Tilton SC, Tal TL, Scroggins SM, Franzosa JA, Peterson ES, Tanguay RL, et al. 2012. Bioinformatics Resource Manager v2.3: An integrated software environment for systems biology with microRNA and cross-species analysis tools. BMC Bioinformatics. 13(1): 311. https://doi.org/10.1186/1471-2105-13-311.