Developmental and Comparative Immunology 75 (2017) 38–47

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

Issues with RNA-seq analysis in non-model organisms: A salmonid example

Arvind Sundaram [a], Torstein Tengs [b], Unni Grimholt [b, *]

[a] Department of Medical Genetics, Oslo University Hospital and University of Oslo, 0407 Oslo, Norway
[b] Norwegian Veterinary Institute, Department of Virology, P.O. Box 750 Sentrum, 0106 Oslo, Norway

ARTICLE INFO

Article history:
Received 25 November 2016
Received in revised form
2 February 2017
Accepted 14 February 2017
Available online 20 February 2017

Keywords:
High throughput sequencing
RNA-seq analyses
Comparative immunology
Genomics

ABSTRACT

High throughput sequencing (HTS) is useful for many purposes as exemplified by the other topics included in this special issue. The purpose of this paper is to look into the unique challenges of using this technology in non-model organisms where resources such as genomes, functional genome annotations or genome complexity provide obstacles not met in model organisms. To describe these challenges, we narrow our scope to RNA sequencing used to study differential gene expression in response to pathogen challenge. As a demonstration species we chose Atlantic salmon, which has a sequenced genome with poor annotation and an added complexity due to many duplicated genes. We find that our RNA-seq analysis pipeline deciphers between duplicates despite high sequence identity. However, annotation issues provide problems in linking differentially expressed genes to pathways. Also, comparing results between approaches and species are complicated due to lack of standardized annotation.

1. Introduction to HTS sequencing

Quantifying expression levels of known genes has long been performed using quantitative PCR (qPCR) and microarrays. A recent alternative is RNA sequencing, also called whole transcriptome shotgun sequencing (WTSS) or RNA-seq. This next-generation sequencing of short sequence reads reveal the presence and quantity of RNA in a biological sample at a given moment in time. The key power of RNA-seq lies in the detection and quantification of all transcripts, including previously unknown ones, in a single method. There exists no *one-size-fits-all* approach when it comes to analyzing RNA-seq data and the strategy employed varies based on the organism studied and the research goals.

For higher vertebrates with well-annotated genomes, RNA-seq analyses are fairly straight forward where it is easy to link differentially expressed genes to functional pathways. In species with less complete gene annotation and functional knowledge, this is much more complicated due to lower sequence identity and species-specific evolution of gene families. Advances in sequencing technologies have made it easier to sequence and assemble draft genomes, but annotation is rather tricky. Many lower vertebrate genomes are currently available representing various degrees of finalized annotations.

Mammalian species can rely heavily on functional studies performed in other mammals as the sequence identity and biological build-up is comparable between even quite distant species such as humans and chicken (Hillier L.D.W. et al., 2004). When it comes to teleost species, which diverged from the tetrapod lineage more than 400 million years ago, extrapolating biological function from mammals is more of a challenge. Some proteins are highly conserved in all vertebrates while other molecules such as many immune genes evolve in line with their changing pathogenic environment and thus are quite different even between closely related species.

Many species have experienced gene expansions or genome duplications, adding complexity to a standard diploid genomic content. Polyploidy is particularly common in plants, but can also be found in many other species such as reptiles, amphibians, insects and fish. The evolution of ray-finned fish has been impacted by several whole genome duplication (WGD) events including a teleost-specific WGD event that occurred at the root of the teleost lineage about 350 million years ago (Mya) and more recent WGD events, for instance in salmonids and carp (Macqueen and Johnston, 2014; Taylor et al., 2003; Wang et al., 2012). Salmonids experienced a whole-genome duplication event approximately 94 million years ago (Ss4R) and many of the duplicate regions have

* Corresponding author.
E-mail address: Unni.Grimholt@vetinst.no (U. Grimholt).

http://dx.doi.org/10.1016/j.dci.2017.02.006
0145-305X/

been retained in the genome, provided an added complexity in RNA-seq analyses as sequence identity between duplicates is also high. The duplicate complexity of the Atlantic salmon genome and incomplete functional annotation makes it a good candidate for looking into problematic issues with RNA-seq analysis in a non-model species.

2. Study design and replicates

Study design plays a key role in experiments using data generated from RNA-seq and has been extensively reviewed elsewhere (Conesa et al., 2016). Since the bioinformatic analyses rely on comparing the expression level between samples/conditions, optimum number of biological and technical replicates are essential in obtaining high statistical power. Several studies have tried to address this issue (Busby et al., 2013; Liu et al., 2014), where a recent one by Schurch et al. (Schurch et al., 2016) advocates at least six biological replicates per group. As Illumina sequencing is highly reproducible (Marioni et al., 2008), there is now a consensus on not having any technical replicates mainly due to the high costs involved. Unfortunately, the variance between samples is also influenced by several factors other than the differences in response to treatment between study groups. Environmental factors, handling of samples and library preparation will introduce sampling and technical biases. In fish, one such environmental factor is commonly referred to as a "tank-effect" (Kjoglum S. et al., 2005). Biological variance is also highly problematic when studying non-model lower vertebrates. Producing genetically homogeneous animals eliminating the genetic variation between samples has been undertaken for many species, but clonal lines are currently only available for a few of these species as they are fairly difficult to cultivate (Colleter et al., 2014; Grimholt et al., 2009; Hou et al., 2016; Komen and Thorgaard 2007).

3. Proceeding from biological sample to RNA-seq data

To enable sequencing of RNA molecules, they must first be converted to cDNA, followed by fragmentation and ligation of sequencing adapters to either end of these fragments which is discussed in details elsewhere (Wang et al., 2009). If the intension is to sequence only coding RNAs, fragments containing polyA can be targeted and extracted while rRNA removal kits can be used to target all types of RNA molecules. The former is the most common type of library preparation since rRNA removal kits need to be organism specific and there exits very few kits for non-model teleosts (Abernathy and Overturf, 2016). In order to retain the information about the orientation of transcripts, stranded RNA-seq is preferred over the traditional non-stranded library preparation approach and has become a gold standard in RNA-seq (Zhao et al., 2015). Data generated using this approach vastly reduces the time and computational needs when assembling a de novo transcriptome. Deeper sequencing is required if the study focuses on quantifying the low expressed genes and/or to look at differential isoform expression while one can get away with lower depth/less reads per sample if the study focuses on highly expressed transcripts. Since cost of library preparation and sequencing plays a key role in these studies, one needs to take a balanced decision based on the need and available budget (Busby et al., 2013; Liu et al., 2014).

In this study, we analyzed data from 48 RNA-seq experiments that were part of a bigger study submitted to the NCBI Sequence Read Archive (NCBI Bioproject PRJEB4657). This material was chosen as it represents an optimum number of biological replicates according to Schurch et al. (2016), i.e. has 100 bp paired end sequence data from six animals at two time points for each of two genotypes including a technical replicate for each sample. The material originates from Atlantic salmon fry from two different families consisting of resistant and susceptible genotypes which were challenged with Infectious salmon anemia virus (IPNV) [Appendix A]. Samples were taken at time point zero without challenge and at 24 h post-infection and they were collected in duplicates to reduce technical bias. Unfortunately, we do not have details on how the dataset was generated (despite several attempts to get in contact with the people responsible). Equally, details on the study material and methods employed are not available in GenBank. However, as the dataset just serves as an example on how RNA-seq analyses are handled in a non-model organism and is not a de novo study of genes influencing IPN resistance we make a few assumptions. We assume that the providers have followed standard protocol when genotyping, performing IPN challenge and acquiring samples. We also assumed that the data is unstranded since we do not know if a stranded RNA library kit was used for generating the dataset. Initial testing did not detect any major issues with any particular sample or outliers within this dataset (See 'Differentially expressed transcripts' section for more details), supporting its authenticity.

RNA-seq data for the 48 animals (96 samples in total; 2 technical replicates per animal) were downloaded from SRA using prefetch in the SRA toolkit v2.5.5 (https://www.ncbi.nlm.nih.gov/sra). Data from high-throughput sequencing instruments are generally in fastq format (Cock et al., 2010), but are stored in SRA format at NCBI SRA. These files were converted to fastq format using the fastq-dump tool (option –split_files) from the same toolkit. Raw data has to be checked for quality and undesirable reads have to be either trimmed or removed. After checking the quality of reads using FastQC v0.11.3 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), they were cleaned using Trimmomatic v0.33 (Bolger et al., 2014). Cleaning the raw reads involves three steps: removing reads matching to Illumina adapter sequences; scanning for low quality reads using a 4-base wide sliding window, trimming when the average quality per base drops below 15; retaining only reads that are longer than 36 nucleotides. Illumina sequencing protocol uses PhiX (NCBI accession: NC_001422.1) phage DNA as spike-in and these reads matching the PhiX genome were also removed using bbmap v34.56 (https://sourceforge.net/projects/bbmap/).

4. Reference transcriptome

The cleaned reads must be mapped to a reference transcriptome providing sequence hits constituting the basis for differential expression analysis. This reference transcriptome can either be genome based or a de novo assembly of transcripts. A genome based reference transcriptome results from a genome assembly where available nucleotide transcripts are mapped and annotated in the genome. For a species with an annotated genome, one can use this genome-based reference transcriptome to map and count the short RNA-seq reads.

A de novo assembled reference transcriptome is an assembly of sequence reads produced without a reference genome either originating from the study itself or from other relevant sources. For more details on the various approaches and hurdles in producing such a de novo reference transcriptome see (Martin and Wang, 2011). There are also benefits of using both genome based as well as a de novo approach even if there is a reference genome.

Using a reference genome with reference transcripts enables a direct mapping of the filtered reads to specific transcripts with genomic location and functional annotation. However, this approach also has its hurdles. For most species such as Atlantic salmon, the available genome originates from a completely homozygous or double haploid individual (Sally) and is not

representative for the genetic content and organization of the entire population. New genes, haplotypes with varying number of genes and gene variants not present in the genome sequenced specimen will be identified once more animals are sequenced. The available Atlantic salmon (Sally) genome is also annotated using a reference transcriptome which most likely had flaws. As such, some genes are present in the genome but not annotated as they were missing in the reference transcriptome.

Without a reference genome one can create a reference transcriptome by assembling the reads into transcripts and count the reads mapping to these sequences. Mapping reads and generating data on differential expression will ultimately depend on the quality of the assembled reference transcriptome. And the quality of this reference transcriptome depends on the number of biological samples included, the sequencing depth of each sample, the read length, sequencing platform choice and the overall variance in the material. For a specific review of transcriptome assembly, see (Martin and Wang, 2011), and for comparisons of how choice of assembly software and sequencing depth affect de novo transcriptome assemblies, see (Brautigam et al., 2011; Francis et al., 2013). One advantage with de novo assembly is the identification of genes and gene variants that are not present or annotated in the sequenced genome.

For organisms that have a well-annotated genome, the logical next step in the analysis pipeline is to align the reads to either the genome or the reference transcriptome. In the case of Atlantic salmon, the latest genome ICSASG v2 (NCBI Bioproject: PRJNA287919) is made up of 241,573 scaffolds and is poorly annotated. The next best approach is to use de novo methods to assemble full-length transcripts using the RNA-seq data. There are several tools such as ABySS (Birol et al., 2009; Simpson et al., 2009), Mira (https://sourceforge.net/projects/mira-assembler/), Trinity (Haas et al., 2013), Velvet (Zerbino and Birney, 2008), Oases (Schulz et al., 2012), SOAPdenovo-Trans (Xie et al., 2014), designed to perform de novo transcriptome assemblies using different approaches and algorithms which has been reviewed elsewhere (Robertson et al., 2010). In our study, we used Trinity both to assemble a de novo transcriptome in addition to 'genome-guided' assembly using the latest Atlantic salmon genome. After normalizing the reads to a maximum coverage of 50 (normalization.pl; Trinity v2.0.6), the reads were assembled using the two different approaches. The de novo transcriptome was assembled using data from all 96 samples in Trinity v2.0.6 using default parameters. And the genome guided assembly was performed using the same software, but the input was aligned data against the ICSASG_v2 Salmon genome using tophat2 v2.0.13. The data was provided in BAM format (See for explanation: https://genome.ucsc.edu/goldenpath/help/bam.html).

The de novo assembled transcriptome contained 564,681 transcripts while the genome guided method produced 607,123 transcripts. The trinity pipeline uses a specific nomenclature to name the transcripts and it appends 'TR' to transcripts identified through a de novo approach and 'GG' for the genome-guided approach (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Output-of-Trinity-Assembly). Assemblies produced by the above two approaches were merged along with RNA data from ICSASG_v2 (109,584 sequences) and were refined to decrease redundancy using the tr2aacds pipeline from EvidentialGene package v2013.07.27 (http://arthropods.eugenes.org/EvidentialGene). Evidence Directed Gene Construction for Eukaryotes, abbreviated as EviGene, uses a mixture of well-known tools such as blast, cd-hit and exonerate to compare the probable gene sequences against each other and extracts the best ones that passes several criteria best representing the transcriptome – called 'reference' transcriptome hereon.

Finally, the Core Eukaryotic Genes Mapping Approach (CEGMA) v2.5 (Parra et al., 2007) was used to measure the completeness and contiguity of putative core eukaryotic genes (CEGs). The CEGMA pipeline uses a set of highly conserved protein families that occur in a wide range of eukaryotes and uses profile-hidden Markov models to ensure reliable gene structures such as exon-intron borders in a genome sequence. This could also be used to identify the presence of these core protein coding genes in a transcriptome. Our reference transcriptome is made up of 157,643 transcripts and CEGMA identified 236 (95.16% completeness) of the 248 defined eukaryotic core genes. All but one of these 236 genes were found to be coding for a complete protein.

5. Differentially expressed transcripts

To estimate transcript abundance against a reference transcriptome one may count the number of reads that map uniquely to each transcript. Statistical analysis then identifies transcripts that are differentially expressed between study groups where analysis options include statistical probability testing to eliminate false positives. Reads from our 96 samples were aligned to the reference transcriptome using RSEM v1.2.29 (Li and Dewey, 2011) to count the number of reads aligning to individual transcripts. This count file was used as input to edgeR v3.14.0 (McCarthy et al., 2012) in R v3.3.0 to calculate differential expression which produces a list of significantly up- or down-regulated transcripts. Read counts were normalized within and between the samples using TMM (trimmed mean M-values), the recommended method in edgeR. Normalization is important to make the data from different samples comparable to each other. Multiple dimensional scaling (MDS) plot or principle component analysis along with clustering of samples based on the expression of all genes will identify outliners, which combed with information on how the samples were collected and processed will enable an educated choice on removing or retaining individual samples. Since we knew very little about the experiment, RNA quality and library preparation used, and MDS (Fig. 1a)/clustering (Fig. 1b) did not show any specific outliers, we included all the samples in the subsequent analyses.

Differentially expressed (DE) transcripts were identified within each family and genotype combination. To achieve a picture of how the DE transcripts segregated into genotypes and families, significantly up- and down-regulated transcripts were visualized using Venn diagrams (Fig. 2). This diagram showed that the majority of DE transcripts were unique either to family or to genotype within family. A more restricted number of genes were commonly regulated in both families as well as genotypes, representing general response genes to viral infection. And an even more restricted number of genes were uniquely up- or down-regulated in either the resistant or the susceptible animals. The choice of which genes to focus on then relies on the biological question underlining the study. Narrowing down the list of transcripts to focus on helped us to reduce the computational need to perform annotation – the most time consuming part of RNA-seq analysis.

In our case we chose to focus on the genes DE between the resistant and susceptible animals shared between the two families (Fig. 2). 51 transcripts were up regulated in the resistant genotype while 54 were down-regulated. In the susceptible genotype, 101 transcripts were up-regulated in both families while 78 were down-regulated. In total, we focused on these 284 transcripts for further analyses (Appendix A).

6. Translating DE transcript list to pathways and mechanisms

Of the 284 DE transcripts identified in our example material, many of the genome-based transcripts have a gene identifier while

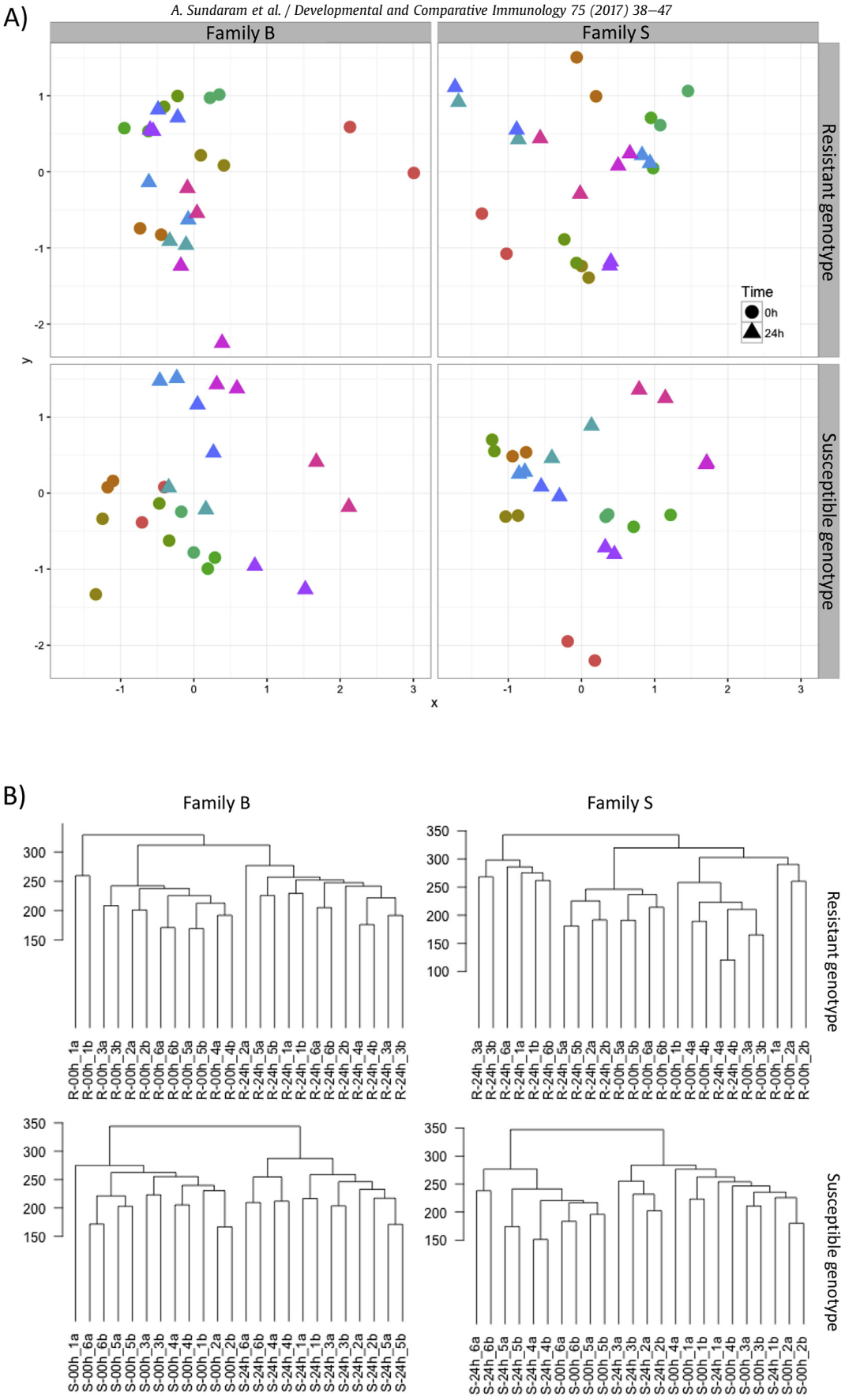

Fig. 1. MDS and clustering of RNA-seq data based on edgeR normalized data. a) Multidimensional scaling plot (MDS) of the samples visualizing variability between the samples. Families and genotypes were analysed separately and the technical replicates are represented by the same colour within each analysis/graph. b) Sample clustering (dendrogram) based on normalized data. An euclidean distance is computed between samples, and the dendrogram is built upon the Ward criterion.

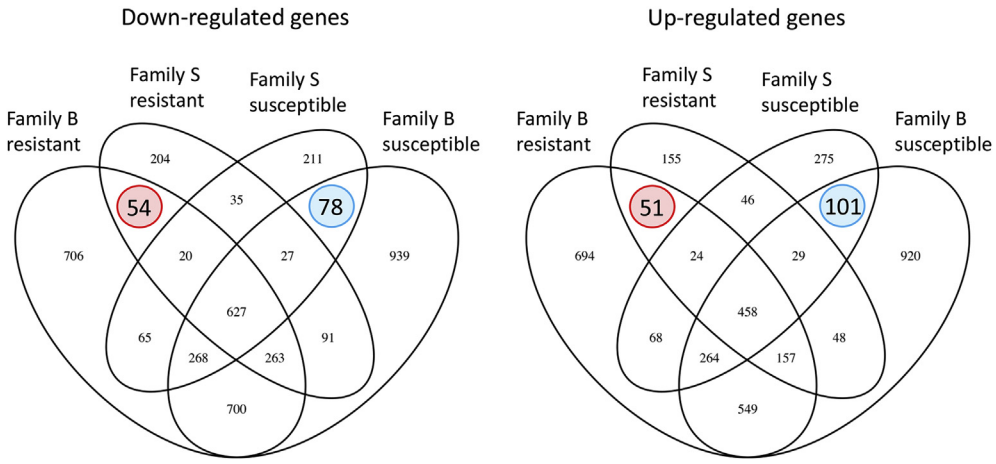

Fig. 2. Venn diagram of DE-regulated transcripts. Venn diagram showing the number of transcripts up-/down-regulated between families and genotypes. Transcript numbers circled in blue constitute 284 transcripts that were further analyzed in this study.

the de novo based transcripts are without a gene annotation. To link these sequences lacking annotation to pathways and/or function they need to be annotated preferentially with a unique gene name. There are many available tools for such analyses e.g. the free online tools DAVID (Huang da et al., 2009), enrich (Kuleshov et al., 2016) or the commercially available Blast2GO (Gotz et al., 2008). We used CateGOrizer (Hu Z.-L. et al., 2008) and REVIGO (Supek et al., 2011) to look for GOs that are enriched in the up- and down-regulated genes (Appendix A).

These 284 transcripts were annotation by BLAST local NCBI-BLAST + v2.2.29 (Camacho et al., 2009) using the NCBI non-redundant (nr) database. The transcripts were also scanned for InterPro domains using Blast2GO v4.0.2 basic free version (Gotz et al., 2008). Both the results were combined and gene ontology (GO) mapping was performed in Blast2GO. 191 transcripts had a unique blast hit against zebrafish cDNAs in Ensembl GRCz10, while 93 transcripts were assigned gene-like names or had insufficient description to warrant an annotation. When comparing the results from this analysis with annotation of the genome-based transcripts there is a high correlation due to the fact that most of the de novo transcripts also had highest match to genome annotated transcripts (data not shown).

When manually analyzing the 93 transcripts that were assigned gene-like names or had insufficient description to warrant an annotation, the majority (56/93) kept the gene_like extension based on blastN and blastP match to annotated teleost transcripts (data not shown). 11 transcripts were identified as repeat sequences using the cGRASP repeat masker (http://lucy.ceh.uvic.ca/repeatmasker/cbr_repeatmasker.py). Only 10 transcripts could be assigned a unique gene identifier while 16 remained unknown, uncharacterized or unresolved due to high sequence identity to more than one gene. In particular the high number of gene_like sequences makes it difficult to link them to functional pathways.

An alternative and currently the most viable approach to biologically categorizing DE transcripts is to use GO ontology terms classified into cellular component, biological process, or molecular function. The results are however, more diffuse than being able to identify specific pathways as exemplified by the figure for the 284 DE transcripts (Appendix B). A third option is using available KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways for other related species such as the cyprinid zebrafish. The KEGG Pathway database is a collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks in various species. Although salmonids and Ostariophysi separated more than 250 million years ago (Near et al., 2012), many gene families are more similar between salmonids and cyprinids than between salmonids and humans. To exemplify, a study of the zebrafish genome revealed that 69% of the zebrafish genes have a human ortholog whereas 31% have no orthologs (Howe et al., 2013). Orthology is here defined using phylogenetic analyses based on sequence identity. Most likely due to the teleost specific genome duplication event, 39% of the orthologous genes had more than one copy in zebrafish as opposed to a single gene in humans. Although gene orthology was not quantified in the Atlantic salmon genome paper (Lien et al., 2016), the orthology between salmon and zebrafish exceeded that between A. salmon and stickleback supporting use of zebrafish as a “model” for A. salmon.

Out of 191 zebrafish transcripts, we were able to identify 110 transcripts having KEGG pathway information using g:Profiler (Reimand et al., 2016). When identifying pathways enriched with the DE transcripts, two pathways stood out i.e the proteasome pathway and the autophagy pathway (Fig. 3). In the IPN resistant animals four proteasome genes (PSMA4, PSMA6, PSMA7 and PSMB6) and one proteasome regulatory particle gene (PSMD7) were down-regulated. In the susceptible animals only one proteasome regulatory particle gene was down-regulated i.e. PSMD11B.

The other pathway identified using the zebrafish KEGG pathway database was regulation of autophagy. In the susceptible animals, both ATG4B and ATG13 were significantly upregulated while in the resistant animals only PIK3R4 (alias VPS15) was significantly upregulated. Both resistant as well as susceptible animals displayed down-regulation of protesome components and up-regulation of autophagy components, although different genes were affected in the two genotypes. Increasing evidence support a close link between these two protein degradation pathways (Lilienbaum, 2013) and the shift between them may be a host response to being infected (Lennemann and Coyne, 2015). As we have no details on how the experiment was performed we refrain from further speculations.

7. What happens to duplicate gene sequences

The proteasome subunits identified as down-regulated in resistant animals when using zebrafish KEGG pathways, i.e. PSMA4, PSMA6, PSMA7 and PSMB6, can act as one example of how duplicate sequences perform in RNA-seq analyses. Only one transcript for each of the four genes was found differentially expressed in our analysis. When we looked for genes in the Atlantic salmon genome

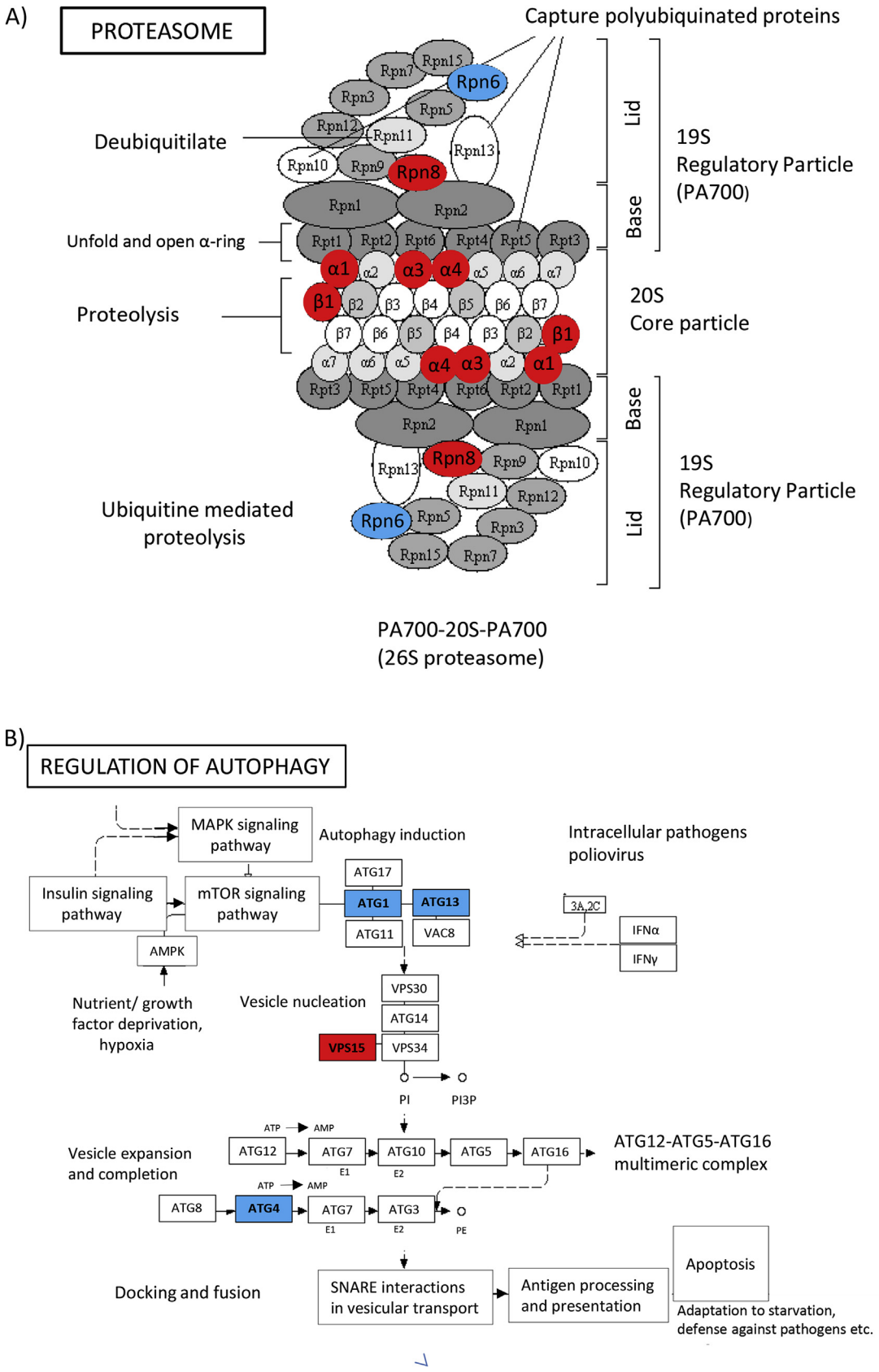

Fig. 3. DE KEGG pathways. a) KEGG pathway "Proteasome". DE transcripts are shown using cyan shading for down-regulated transcripts in susceptible animals and red shading for down-regulated transcripts in resistant animals. Alternative nomenclature for the DE subunits are: Regulatory particle: RPN6 is PSMD7, RPN8 is PSMD11. Core particle: α1 is PSMA6, α3 is PSMA4, α4 is PSMA7/PSMA8 and β1 is PSMB6. b) KEGG pathway "Regulation of autophagy". DE transcripts are shown using cyan shading for up-regulated transcripts in susceptible animals and red shading for up-regulated transcripts in resistant animals. Alternative nomenclature for the DE subunits are: ATG1 is ULK1-ULK3 and VPS15 is PIK3R4.

with high sequence identity to these DE PSMA and PSMB transcripts we found that all four subunits were encoded by more than one gene ranging from two copies of PSMB6 to six copies of PSMA7 (Table 1, Appendix C). As the genes were annotated either gene or gene_like in the genome, thus not discriminating between paralogs, we assigned each gene with our own unique preliminary identifiers. Our suggested nomenclature is in line with previously suggested nomenclature for e.g. MHC and Chemokine receptor nomenclature where duplicates originating from homeolog blocks are assigned the extension a and b while more duplicates receive the extension of 0.1 etc. However, this nomenclature needs to be adapted to what is found in other teleosts. Nucleotide sequence identity in the open reading frame between paralogs ranged from 71% between the PSMA6.3 and PSMA6.1a sequences to 100% between the PSMA4.1a and PSMA4.1b sequences (Appendix C). Obviously, the duplicate PSMA4.1a and .1b sequences will not be separated in this analysis, but the remaining sequences were. Separating so effectively between gene duplicates with high sequence identity may be attributable to paired end data which has an advantage over single end reads since information is obtained from two regions of the transcript. Increased sequence read length also assists in a higher accuracy in matching reads to unique transcripts. Collectively, this helps in deciphering between paralogues, as well as identifying isoforms and splice variants in a much more effective way than previous sequencing approaches (Ozsolak and Milos, 2011).

To assess functional differences between these duplicates in normal Atlantic salmon fry we performed transcriptional profiling of each subunit. All paralogs displayed expression in most animals, and had unique expression profiles (Appendix C). Many also had highest expression levels in reproductive organs with the exception of PSMA6.3, PSMA7a and PSMA7b which displayed low levels in these organs. Such distinct transcription profiles for paralogues have previously been found by us for MHC class I Z and L lineage paralogs (Grimholt et al., 2015b), and chemokine receptor paralogs (Grimholt et al., 2015a). On a larger scale Lien et al. (2016). found that there were far more instances of paralog neo-functionalization than sub-functionalization when they analyzed the fate of Atlantic salmon paralogs measured against orthologous genes from the pre-Ss4R outgroup Northern pike. This is consistent with our findings that only one gene for each of the four proteasome subunits was found differentially expressed in response to IPN infection. The remaining gene copies were not significantly influenced by infection in the DE resistant versus susceptible transcripts analyzed in our study.

8. Comparing results between experiments, approaches and species

A requirement for comparing results between studies or organism is a unified nomenclature. A gene defined as PSMB7 in one species should be an orthologue of a gene defined as PSMB7 in other species. If nomenclature is not standardized between scientific experiments and species, comparing results will create confusion and potentially end in important findings remaining unnoticed.

When we looked closer at teleost PSMA7 and PSMA8 sequences, we found orthologs to the human PSMA8 sequence in all analyzed teleosts while only some species had orthologs to the human PSMA7 sequence (Fig. 4 and Appendix C). Nomenclature for the sequences identified in NCBI databases varied a lot while those in Ensembl were consistent with phylogenetic clustering. The fact that the human PSMA7 and 8 sequences cluster in two clades alongside teleost orthologs suggest that a common ancestor to tetrapods and ray-finned fishes had both these genes. In humans, the α4 subunit of the proteasome is encoded by the two genes, i.e. the PSMA7 and PSMA8 genes, where the PSMA8 subunit has a unique role in spermatoproteasomes while PSMA7 has a housekeeping role (Kniepert and Groettrup, 2014). Based on phylogenetic clustering one would expect a similar functional differentiation for the teleost PSMA7 and PSMA8 sequences. However, it seems odd that the housekeeping subunit PSMA7 is not found in many teleosts. Clear-cut PSMA7 sequences are present in spotted gar, sardines, catfish, Mexican tetra, Northern pike and salmonids, but were not identified in neoteleosts. So potentially the PSMA7 gene has been lost somewhere between salmonids and neoteleosts.

Table 1
Genomic location, accession number and gene names for Atlantic salmon proteasome sequences.

Our name	Name in genome	Accession #	Location in genome
SasaPSMB6a	PSMB6_Like	XP_014051784.1	ssa04; NC_027303.1:35.985.620-35.988.142
SasaPSMB6b	PSMB6	XP_013984457.1	ssa11; NC_027310.1: 77.336.949-77.339.681
SasaPSMA4.1a	PSMA4	NP_001134515.1	ssa26; NC_027325.1:32.939.806-32.943.155
SasaPSMA4.1b	PSMA4_Like	XP_013983107.1	ssa11; NC_027310.1:38.380.642-38.383.977
SasaPSMA4.2	PSMA4	XP_013979513.1	ssa10; NC_027309.1:62.947.128-62.979.393
SasaPSMA6.1a	PSMA6	XP_014060163.1	ssa06 NC_027305.1:48.395.986-48.398.391
SasaPSMA6.1b	PSMA6_Like	XP_013999490.1	ssa15; NC_027314.1:24.839.610-24.842.186
SasaPSMA6.2	PSMA6_Like	XP_014066593.1	ssa09; NC_027308.1:31.342.996-31.358.744
SasaPSMA6.3	PSMA6_Like	XP_014069427.1	ssa09; NC_027308.1:105.413.536-105.420.909
SasaPSMA7a	PSMA7_Like	XP_014000965.1	ssa15; NC_027314.1:60.419.652-60.421.283
SasaPSMA7b	PSMA7	XP_013990556.1	ssa13; NC_027312.1:19.630.566-19.634.058
SasaPSMA8.1a	PSMA7_Like	XP_014024650.1	ssa23; NC_027322.1:14.663.633-14.665.712
SasaPSMA8.1b	PSMA7_Like	XP_013978661.1	ssa10; NC_027309.1:35.937.788-35.939.901
SasaPSMA8.2a	PSMA7_Like	XP_013995708.1	ssa14; NC_027313.1:27.557.984-27.577.773
SasaPSMA8.2b	PSMA7	XP_014046464.1	ssa03; NC_027302.1:27.740.291-27.753.411

Genome nomenclature, our suggested nomenclature, accession numbers and genomic location for Atlantic salmon PSMB6, PSMA4, PSMA6, PSMA7 and PSMA8 genes. Genes residing in homeologous blocks in the Atlantic salmon genome (Lien et al., 2016) are shown using identical shading. DE genes are shaded orange.

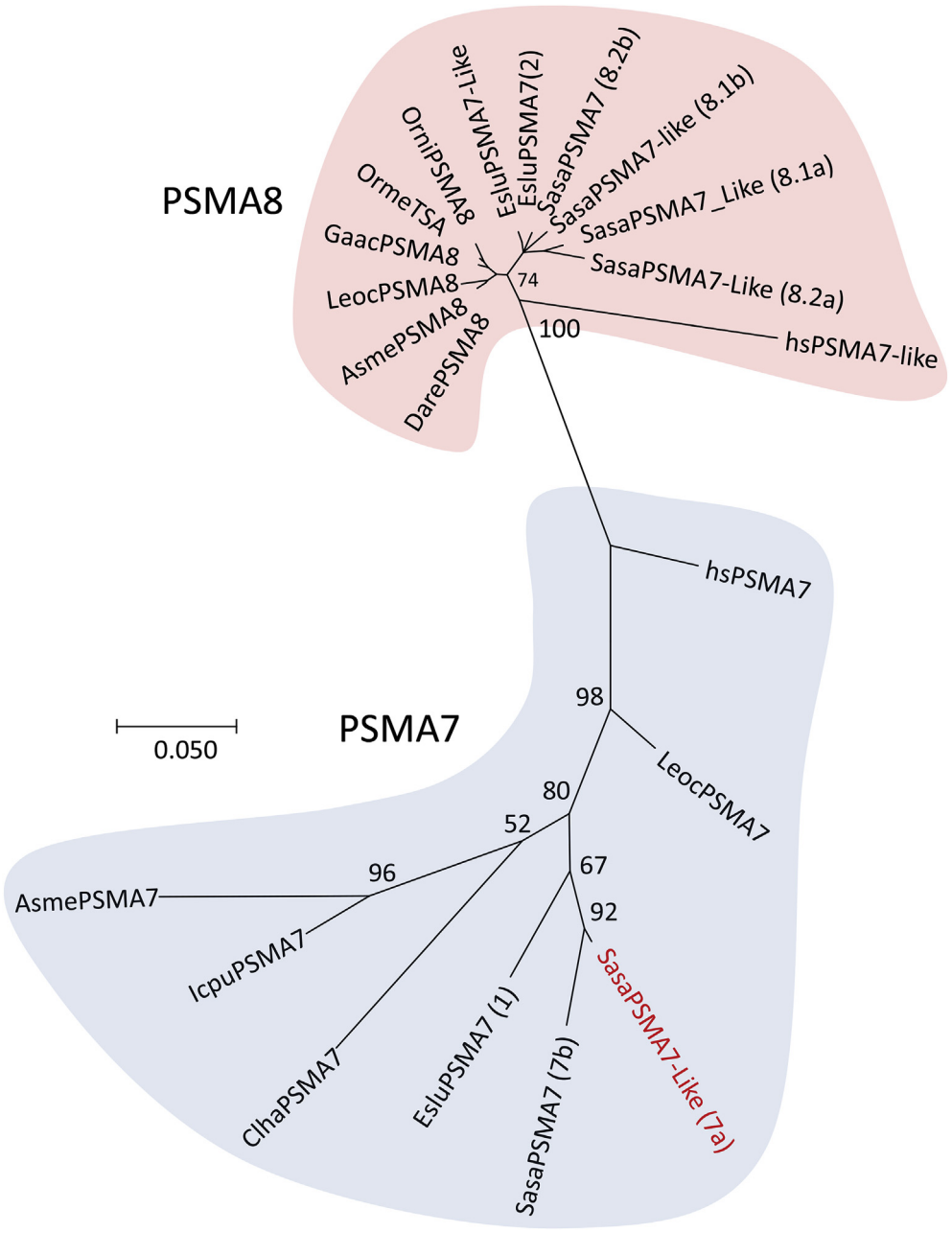

Fig. 4. Phylogeny of selected PSMA7 and PSMA8 amino acid sequences. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). The tree with the highest log likelihood (−1816.5793) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.3346)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 22 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 243 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). Sequence references can be found in Appendix C where hs is human (*Homo sapiens*), Leoc is spotted gar (*Lepisosteus oculeatus*), Eslu is Northern pike (*Esox lucius*), Sasa is Atlantic salmon (*Salmo salar*), Asme is Mexican tetra (*Astyanax mexicanus*), Clha is Atlantic herring (*Clupea harengus*), Orni is Nile tilapia (*Oreochromis niloticus*), Gaac is stickleback (*Gasterosteus oculatus*), Orme is marine medaka (*Oryzias melastigma*) and Dare is zebrafish (*Danio rerio*). The Atlantic salmon DE PSMA7 transcript is shown in red font. Internal bootstrap values for the PSMA8 clade are not shown due to space restrictions. Database sequence names are used where our nomenclature is shown in parenthesis.

Based on the salmon RPKM values (Appendix C) and the lack of PSMA7 sequence variants in neoteleosts one may wonder if teleosts have chosen the PSMA7 variant as the spermatoproteasome subunit while the PSMA8 gene encodes the constitutive subunit. This underlines the need for caution when one assigns function based on phylogenetic clustering.

Reproducibility between experiments will depend on genotype of the experimental animals as well as viral isolate, viral dosage and challenge method. A study comparable to the one analyzed in this manuscript, i.e. immune responses following IPN challenge of Atlantic salmon fry, was recently published where the authors evaluated transcriptional responses 1, 7 and 20 days post-infection using microarrays (Robledo et al., 2016). Microarray is a method where transcripts are hybridized to short oligo-probes representing

selected genes within that species. Robledo et al. (2016) only found a limited number of DE genes at day 1 post-infection and the genes they found to be significantly upregulated in resistant animals i.e. IRF8, IL3R, *M*-CSF, FBXo9 and TCF3 were not identified in our material. Although we cannot guarantee the authenticity of our study material, it is questionable if microarray would separate between the Atlantic salmon PSMA and PSMB duplicate sequences, thus levelling out the differential expression between them.

9. Concluding remarks and future perspectives

RNA-sequencing produces a huge amount of data that are made available in the public domain. However, to be useful for the scientific community, we need these resources to have detailed descriptions of the study material, the study design in addition to a contacting author. NCBI should make this a mandatory requirement for accepting RNA-seq submissions.

Although we make the assumption that biological functions of teleost molecules are identical to their tetrapod orthologs, this is a very simplified view. As we and others have shown that duplicate Atlantic salmon genes may hold different functions, this calls for unique identifiers deciphering between duplicate genes. Atlantic salmon nomenclature must also be conformed to what is found in other teleosts, indicating that international nomenclature panels are needed to agree on common nomenclature. Initially this nomenclature will rely on sequence identity, but this needs to be supported by functional studies. Teleosts may have chosen different functions than their mammalian counterparts, as potentially exemplified by the teleost PSAM7 and PSMA8 genes.

Finally, to exploit the entire potential of RNA-seq data we need pathway databases where nomenclature issues are standardized and unique functional assets of teleost molecules are integrated. Then RNA-seq can reach its full potential in non-model organisms.

Acknowledgements

This study was funded by the Norwegian Research Council program Biotek2021 Project number 244336 (UG), and internal funding from Norwegian Veterinary Institute.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.dci.2017.02.006.

References

Abernathy, J., Overturf, K., 2016. Comparison of ribosomal RNA removal methods for transcriptome sequencing workflows in teleost fish. Anim. Biotechnol. 27, 60–65.
Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A., Jones, S.J., 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25, 2872–2877.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30, 2114–2120.
Brautigam, A., Mullick, T., Schliesky, S., Weber, A.P., 2011. Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C(3) and C(4) species. J. Exp. Bot. 62, 3093–3102.
Busby, M.A., Stewart, C., Miller, C.A., Grzeda, K.R., Marth, G.T., 2013. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics 29, 656–657.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinforma. 10, 421.
Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38, 1767–1771.
Colleter, J., Penman, D.J., Lallement, S., Fauvel, C., Hanebrekke, T., Osvik, R.D., Eilertsen, H.C., D'Cotta, H., Chatain, B., Peruzzi, S., 2014. Genetic inactivation of European sea bass (*Dicentrarchus labrax* L.) eggs using UV-irradiation: observations and perspectives. PLoS One 9, e109572.
Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13.
Francis, W.R., Christianson, L.M., Kiko, R., Powers, M.L., Shaner, N.C., Haddock, S.H., 2013. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics 14, 167.
Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3435.
Grimholt, U., Hauge, H., Hauge, A.G., Leong, J., Koop, B.F., 2015a. Chemokine receptors in Atlantic salmon. Dev. Comp. Immunol. 49, 79–95.
Grimholt, U., Johansen, R., Smith, A.J., 2009. A review of the need and possible uses for genetically standardized Atlantic salmon (*Salmo salar*) in research. Lab. Anim. 43, 121–126.
Grimholt, U., Tsukamoto, K., Azuma, T., Leong, J., Koop, B.F., Dijkstra, J.M., 2015b. A comprehensive analysis of teleost MHC class I sequences. BMC Evol. Biol. 15, 32.
Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512.
Hillier, L.D.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A.M., Delany, M.E., Dodgson, J.B., et al., 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695–716.
Hou, J., Wang, G., Zhang, X., Wang, Y., Sun, Z., Si, F., Jiang, X., Liu, H., 2016. Production and verification of a 2nd generation clonal group of Japanese flounder, *Paralichthys olivaceus*. Sci. Rep. 6, 35776.
Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assuncao, J.A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Elliot, D., Threadgold, G., Harden, G., Ware, D., Begum, S., Mortimore, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Lloyd, C., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Urun, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberlander, M., Rudolph-Geiger, S., Teucke, M., Lanz, C., Raddatz, G., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Schuster, S.C., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M., Enright, A., Geisler, R., Plasterk, R.H., Lee, C., Westerfield, M., de Jong, P.J., Zon, L.I., Postlethwait, J.H., Nusslein-Volhard, C., Hubbard, T.J., Roest Crollius, H., Rogers, J., Stemple, D.L., 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496, 498–503.
Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.
Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.
Kjoglum, S., Grimholt, U., Larsen, S., 2005. Non-MHC genetic and tank effects influence disease challenge tests in Atlantic salmon (*Salmo salar*). Aquaculture 250, 102–109.
Kniepert, A., Groettrup, M., 2014. The unique functions of tissue-specific proteasomes. Trends Biochem. Sci. 39, 17–24.
Komen, H., Thorgaard, G.H., 2007. Androgenesis, gynogenesis and the production of clones in fishes: A review. Aquaculture 269, 150–173.
Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–W97.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.
Lennemann, N.J., Coyne, C.B., 2015. Catch me if you can: the link between autophagy and viruses. PLoS Pathog. 11, e1004685.
Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq

data with or without a reference genome. BMC Bioinforma. 12, 323.
Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Vage, D.I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J., Jonassen, I., Maass, A., Omholt, S.W., Davidson, W.S., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205.
Lilienbaum, A., 2013. Relationship between the proteasomal system and autophagy. Int. J. Biochem. Mol. Biol. 4, 1–26.
Liu, Y., Zhou, J., White, K.P., 2014. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30, 301–304.
Macqueen, D.J., Johnston, I.A., 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc. Biol. Sci. 281, 20132881.
Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509–1517.
Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671–682.
McCarthy, D.J., Chen, Y., Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40, 4288–4297.
Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., Friedman, M., Smith, W.L., 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. U. S. A. 109, 13698–13703.
Ozsolak, F., Milos, P.M., 2011. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87–98.
Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.
Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., Vilo, J., 2016. g: Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 44, W83–W89.
Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., Prabhu, A.L., Tam, A., Zhao, Y., Moore, R.A., Hirst, M., Marra, M.A., Jones, S.J., Hoodless, P.A., Birol, I., 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7, 909–912.
Robledo, D., Taggart, J.B., Ireland, J.H., McAndrew, B.J., Starkey, W.G., Haley, C.S., Hamilton, A., Guy, D.R., Mota-Velasco, J.C., Gheyas, A.A., Tinch, A.E., Verner-Jeffreys, D.W., Paley, R.K., Rimmer, G.S., Tew, I.J., Bishop, S.C., Bron, J.E., Houston, R.D., 2016. Gene expression comparison of resistant and susceptible Atlantic salmon fry challenged with infectious pancreatic necrosis virus reveals a marked contrast in immune response. BMC Genomics 17, 279.
Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092.
Schurch, N.J., Schofield, P., Gierlinski, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., Barton, G.J., 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22, 839–851.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.
Supek, F., Bosnjak, M., Skunca, N., Smuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6, e21800.
Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., Van de Peer, Y., 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 13, 382–390.
Wang, J.T., Li, J.T., Zhang, X.F., Sun, X.W., 2012. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (*Cyprinus carpio*). BMC Genomics 13, 96.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.W., Li, Y., Xu, X., Wong, G.K., Wang, J., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.
Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., Zhang, B., 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics 16, 675.
Hu, Z.-L., Bao, J., Reecy, J.M., 2008. CateGOrizer: a web-based program to batch analyze gene ontology classification categories. Online J. Bioinforma. 9, 108–112.