Transcriptome-wide comparisons and virulence gene polymorphisms of host-associated genotypes of the cnidarian parasite Ceratonova shasta in salmonids

Gema Alama-Bermejo*,1,3,5, Eli Meyer2, Stephen D. Atkinson1, Astrid S. Holzer3, Monika M. Wiśniewska3, Martin Kolisko3-4, Jerri L. Bartholomew1

1 Department of Microbiology, Oregon State University, Nash Hall 226, 97331 Corvallis, OR, USA

2 Department of Integrative Biology, Oregon State University, Cordley Hall 3106, 97331 Corvallis, OR, USA

3 Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 1160/31, 37005 České Budějovice, Czech Republic

4 Department of Molecular Biology and Genetics, Faculty of Science, University of South Bohemia, Branišovská 1760, 37005, České Budějovice, Czech Republic

5 Centro de Investigación Aplicada y Transferencia Tecnológica en Recursos Marinos Almirante Storni (CIMAS), CCT CONICET – CENPAT, Güemes 1030, 8520 San Antonio Oeste, Argentina

*Corresponding author: Gema Alama-Bermejo gema.alama@gmail.com

Eli Meyer eli.meyer@oregonstate.edu

Stephen Douglas Atkinson atkinsos@oregonstate.edu

Astrid Sibylle Holzer astrid.holzer@paru.cas.cz

Monika Magdalena Wiśniewska wisniem95@gmail.com

Martin Kolisko kolisko@paru.cas.cz

Jerri Lee Bartholomew bartholj@science.oregonstate.edu

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Abstract

*Ceratonova shasta* is an important myxozoan pathogen affecting the health of salmonid fishes in the Pacific Northwest of North America. *C. shasta* exists as a complex of host-specific genotypes, some with low to moderate virulence, and one that causes a profound, lethal infection in susceptible hosts. High throughput sequencing methods are powerful tools for discovering the genetic basis of these host/virulence differences, but deep sequencing of myxozoans has been challenging due to extremely fast molecular evolution of this group, yielding strongly divergent sequences that are difficult to identify, and unavoidable host contamination. We designed and optimized different bioinformatic pipelines to address these challenges. We obtained a unique set of comprehensive, host-free myxozoan RNA-seq data from *C. shasta* genotypes of varying virulence from different salmonid hosts. Analyses of transcriptome-wide genetic distances and maximum likelihood multigene phylogenies elucidated the evolutionary relationship between lineages and demonstrated the limited resolution of the established Internal Transcribed Spacer marker for *C. shasta* genotype identification, as this marker fails to differentiate between biologically distinct genotype II lineages from coho salmon and rainbow trout. We further analyzed the datasets based on polymorphisms in two gene groups related to virulence: cell migration and proteolytic enzymes including their inhibitors. The developed SNP-calling pipeline identified polymorphisms between genotypes and demonstrated that variations in both motility and protease genes were associated with different levels of virulence of *C. shasta* in its salmonid hosts. The prospective use of proteolytic enzymes as promising candidates for targeted interventions against myxozoans in aquaculture is discussed. We developed host-free transcriptomes of a myxozoan model organism from strains that exhibited different degrees of virulence, as a unique source of data that will foster functional gene analyses.
and serve as a base for the development of potential therapeutics for efficient control of these parasites.

**Keywords:** Myxozoa, aquaculture, cell migration/motility, proteases, SNPs

**Introduction**

Myxozoans are a group of cnidarians that emerged as parasites of aquatic invertebrate hosts (annelids and byrozoans) ~650 million years ago (Holzer et al., 2018). Following the emergence of vertebrates, myxozoans acquired these as secondary hosts, predominantly fish, which fostered massive host-associated diversification and led to the distinct success of the Myxozoa. Their presently known diversity accounts for approximately a fifth of all cnidarian species (Atkinson et al. 2018a) but may be greatly underestimated (Hartikainen et al., 2016). Myxozoans are extremely reduced in size and body plan, produce spores as transmission stages, but retain the nematocysts (polar capsules) present in all Cnidaria, both free-living and parasitic (Holland et al., 2011, Shpirer et al., 2018). Myxozoans are especially known for the diseases they can cause in wild and cultured fish.

Salmonid fishes are of significant economic and cultural value in the Pacific Northwest of North America. In many rivers these fish are exposed to the myxozoan parasite, *Ceratonia shasta* (Noble, 1950) (syn. *Ceratomyxa shasta*), which affects wild and artificially reared salmon and trout. This microscopic cnidarian has a life cycle that alternates between actinospores, which develop in the annelid host, and myxospores, which develop in the vertebrate host - various species of Pacific salmon and trout (Bartholomew et al., 1997). The parasite penetrates through the gills of the fish, enters the bloodstream, then invades all layers of the gut, where it proliferates, and can cause
severe enteronecrosis with gross lesions of swollen, necrotic and hemorrhagic intestine (Bjork and Bartholomew, 2009a, 2010). *Ceratonova shasta* is one of the most virulent myxozoans known, with mortalities approaching 100% in susceptible species/stocks (Hallett and Bartholomew, 2012), with an infectious dose of a single parasite actinospore capable of causing a lethal infection (Bjork and Bartholomew, 2009a).

Several abiotic and biotic factors affect the severity of disease caused by *C. shasta*: host stock origin (sympatric or allopatric with the parasite), water temperature, water flow, actinospore density (Bjork and Bartholomew, 2009a, b; Hallett et al., 2012; Ray et al., 2012), and parasite genotype (Hurst and Bartholomew, 2012). The parasite occurs across the Pacific Northwest in at least three host-specific genotypes (0, I & II; Atkinson et al., 2018b; Stinson et al., 2018), which have different levels of virulence (*i.e.* mortality, proliferation and pathogenicity) in different salmonid species and strains. These genotypes are presently identified based on single nucleotide polymorphisms (SNPs) in the Internal Transcribed Spacer region 1 (ITS-1) of the ribosomal DNA array (rDNA) (Atkinson and Bartholomew, 2010a, b). Although dependent on parasite dose and fish strain, the most virulent genotypes are type I (specific for Chinook salmon, *Oncorhynchus tschawytscha*) and type II, a generalist that can infect up to six different species but is dominant in coho salmon (*Oncorhynchus kisutch*) (Stinson et al., 2018).

Two biotypes of genotype II, IIC and IIR, are differentiated by host specificity and virulence: both cause disease in susceptible allopatric rainbow trout (*Oncorhynchus mykiss*), but IIC causes a dose-dependent mortality in sympatric coho salmon, while IIR can only infect coho salmon to a limited extent (Hurst and Bartholomew, 2012); there are no ITS-1 molecular differences to distinguish these biotypes. The most genetically distinct strain, genotype 0, is less virulent and produces chronic infections in *O. mykiss* stocks, both sympatric and allopatric with the parasite, with low proliferation rates and
few clinical signs (Atkinson and Bartholomew, 2010b; Stinson et al., 2018). The mechanisms accounting for these virulence differences in *C. shasta* genotypes are unknown.

Virulence factors are molecules responsible for the pathogenicity of an organism – its ability to cause disease and mortality in the host (*e.g.* Casadevall and Pirofski, 1999). These factors are important in different aspects of the life history of a parasite *e.g.* adhesion, invasion, migration or host immune evasion (*e.g.* McKe row et al., 2006; Bouzid et al., 2013). Motility genes and proteases are generally considered candidate virulence factors (*e.g.* Barragan and Sibley, 2002, McKe row et al., 2006; Bouzid et al., 2013). Virulence factors vary both among and within species of parasites: different factors may be expressed by different genotypes or strains, with parasite populations often able to be ranked from highly virulent to avirulent (*e.g.* Ali et al., 2007; Dardé, 2008). Transcriptomic and genomic technologies are enabling novel approaches to compare parasite strains and identify potential virulence factors (*e.g.* Eichenberger et al., 2017), which can then become targets for drug and chemotherapeutic strategies for disease control (Seib et al., 2009, Mennerat et al., 2010).

Only recently have genomic and transcriptomic data become available for myxozoans (*e.g.* Yang et al., 2014; Chang et al., 2015; Foox et al., 2015). Myxozoans are unculturable, obligate endoparasites, thus unambiguous sequencing and data analysis have proven extremely difficult, given the presence of contaminating host tissue. We solved several fundamental challenges of host contamination in our ‘omics studies of different genotypes of *C. shasta* by taking advantage of a novel aspect of its biology: that less-contaminated, metabolically active parasite material is present in ascitic fluid, produced in systemic infections of type IIR in susceptible rainbow trout. Herein, we report a novel method for removal of host contamination in a two-step bioinformatic
process, at both read and transcript levels. We then used this workflow to characterize genetic variation among RNAseq data from genotypes I, IIC and IIR, with the aim to identify: 1. if there are genetic differences that support the separation of biotype II into IIR and IIC, and 2. which category of transcript (proteases or motility factors) is most variable between genotypes, and thereby identify which category of genes is under highest selection pressure. For broader context, we reconstructed the evolutionary history of *C. shasta* genotypes based on phylogenomic analyses.

**Material and methods**

**Sampling and ITS-1 rDNA genotyping**

We collected different *C. shasta* genotypes from infected fish: allopatric rainbow trout *O. mykiss* (Roaring River Hatchery stock, ascitic fluid, genotype II, biotype IIR); sympatric Chinook salmon *O. tshawytscha* (Iron Gate Hatchery stock, intestine, genotype I); sympatric coho salmon, *O. kisutch* (Iron Gate Hatchery stock, intestine, genotype II, biotype IIC) (Table 1). All fish were euthanized with an overdose of buffered MS-222. Ascitic fluid was collected by syringe from the abdominal cavity (Figure 1) and intestinal tissue was dissected out. Wet mounts of tissues were examined by microscope to confirm presence of parasite developmental stages, *i.e.* pre-sporegonic, sporogenic and/or mature spores of *C. shasta*. Ascitic fluid was centrifuged at 8000 rpm for 2 minutes. The supernatant was removed, and the pellet was suspended in RNA later (Qiagen, Valencia, California) and stored at -80°C until RNA extraction. Similarly, ~10 mm sections of intestines were stored in RNA later. Replicates of all samples were taken for DNA analyses, and stored at -20°C until extraction, using the DNeasy Blood and Tissue kit (Qiagen). The parasite ITS-1 genotype composition in each sample was confirmed by PCR amplification and Sanger sequencing (GenBank accession numbers: MN173024 Cs-genotype-I, MN173025 Cs-genotype II; Atkinson
and Bartholomew, 2010a). This study was carried out in accordance to the recommendations of Oregon State University Institutional Animal Care and Use Committee (IACUC) under approval ACUP #4666.

**Transcriptome library prep and next generation sequencing (NGS)**

RNA from each of the five fish samples was extracted using a High Pure RNA Tissue Kit (Roche, Switzerland) and treated on-column with DNAsen I to remove genomic DNA. RNA was quantified using a spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and 1 µg of total RNA was used downstream. Samples (Table 1) were analyzed in two sets: a deeply sequenced *C. shasta* genotype IIR reference transcriptome from one rainbow trout (RBT6; Willamette River) and low coverage *C. shasta* transcriptomes from genotypes I (Chinook salmon, Lower Klamath River), IIC (Coho salmon, Lower Klamath River), 2x IIR (RBTC16 – Upper Klamath River; RBTJ7 – Lower Klamath River, both from rainbow trout).

**Reference transcriptome of *C. shasta* genotype IIR**

RNA (RBT6) from rainbow trout was submitted to Oregon State University’s Center for Genome Research and Biocomputing (OSU CGRB) for directional library preparation and sequencing. A cDNA library was made using the PrepXTM mRNA Strand Specific Library Prep Kit with poly-A selection in an Apollo 324 NGS Library Prep System (Wafergen, Fremont, California, USA). The library was quantified and size-checked using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and Bioanalyzer 2100 (Agilent Technologies, Santa Clara, California, USA). The reference library was pair-end sequenced on 1 lane of HiSeq 3000 (Illumina, San Diego, California, USA).

**Lower-coverage transcriptomes of genotypes I, IIC and 2 x IIR**

http://mc.manuscriptcentral.com/gbe
Four normalized custom libraries were prepared following Meyer et al., 2009. In summary, first strand cDNA was synthesized using SuperScript II Reverse Transcriptase (Life Technologies, Carlsbad, California, USA), then normalized using DSN (double-strand-specific-nuclease, Evrogen, Moscow, Russia) to decrease the prevalence of the most abundant, repeated transcripts and increase discovery of rare transcripts; we considered that this was important due to contamination by host RNA. Following normalization, the cDNA was amplified and fragmented using NEBNext dsDNA Fragmentase (New England Biolabs, Ipswich, Massachusetts, USA). Oligonucleotide adaptors were ligated to the fragmented cDNA. Constructs were amplified and barcoded with HiSeq oligos and selected by size (350 - 550 bp) on a 2% TAE agarose gel. The four libraries were pair-end sequenced by Illumina HiSeq 2000 at CGRB, using 1/6th of a lane for each.

Custom scripts used in this study were written by E.M. (available at https://github.com/Eli-Meyer or in Supplementary Material 1). Analyses were performed on the OSU CGRB computing cluster. We used bit score thresholds instead of e-values for our BLAST searches to avoid potential problems with different database sizes (Pearson, 2013). Our bioinformatic pipeline is summarized in Figure 2.

**Removal of host contamination at read level - filtering and mapping of reads**

Raw Illumina pair-end reads from the five libraries were filtered for adaptor sequences (score 15 - 20), low quality base-calls (> 20 - 60 bases with quality score below 20) and homopolymer poly-A (> 50 homopolymer repeats), using custom Perl scripts. We then filtered out host reads prior to assembly or SNPs calling, by mapping against two references: 1) combined *Oncorhynchus mykiss* genome (Berthelot et al., 2014; http://www.genoscope.cns.fr/trout/data/; BioProject Acc. Number PRJEB4421 version GCA_900005704.1) and the mitochondrial genome (Zardoya et al., 1995; BioProject...
Acc. Number PRJNA11824 version NC_001717.1); 2) our reference *C. shasta* genome, which was obtained from purified myxospores and pre-filtered using BLASTn (score = 150) against the same host reference (Version Velvet2015-93, 14,586 sequences, 185-452,519 bp length, N50 =36,283, total size 69.8 Mb; Atkinson et al., in preparation).

We mapped the high-quality, filtered transcriptome reads against the references using gmapper (Shrimp v 2.0; David et al., 2011), reporting one hit per read and using the local alignment option. We then filtered the resulting .sam file to exclude ambiguous, short, and weak matches: minimum match length 45 bp in any 50 bp alignment. We compared the filtered .sam files to determine if reads best matched host, parasite, or neither host nor parasite (NHP), and binned reads into respective .fastq files. *C. shasta* and NHP reads from each of the five datasets were used for subsequent analyses; host reads were excluded.

**Reference transcriptome de novo assembly and contaminant removal at the contig level**

We generated two reference assemblies of the deeply sequenced IIR dataset (RBT6): a more conservative assembly of only reads that best matched the *C. shasta* genome assembly, and a less conservative version with reads that matched the *C. shasta* genome, plus the NHP reads (Figure 2). The less conservative assembly was made to recover parasite transcripts missed due to lack of completeness/coverage of the *C. shasta* genome. Reads were *de novo* assembled into contigs using the Trinity v2.2.0 (2016) pipeline (Haas et al., 2013).

We used a second level of contaminant filtering, by BLASTn comparison of our assembled transcriptomes against the same reference genome databases used for read filtering (rainbow trout and *C. shasta*). We used a custom script that identified the most likely contig origin (bit score >100). The script assigned contigs to one of three bins.
based on the best match: 1) target (= *C. shasta*); 2) contaminant (= rainbow trout); 3) neither host nor parasite (NHP; = bit score below threshold for both databases).

Additionally, we used taxon ID annotations to remove contigs of prokaryotes, viruses, fungi, algae and protists. For this purpose, we downloaded the nr database (03.03.2014) and removed all taxa belonging to the before-mentioned groups by filtering on phylum level (level 4 in NCBI), using a >50 bit score.

FPKM (Fragments per kilobase of transcript per million mapped reads) and TPM (Transcripts per million) were calculated using an alignment-based quantification method, RSEM (Li and Dewey, 2011).

**Reference transcriptome annotation and completeness**

We used three methods to annotate contigs in the two assemblies: gene names (putative functional annotations), Gene Ontology terms, and taxon identification. Gene names were assigned by comparison with the UniProt database (Release 2015), using BLASTx (NCBI; cut-off e-value $10^{-6}$). We used UniProt over larger databases such as nr because it has more accurate and fewer redundant gene function annotations. We assigned gene names to contigs based on the best BLAST match after excluding database hits whose annotations included uninformative terms *e.g.* 'unknown', 'uncharacterized', ‘predicted protein’ (ambiguous terms filtering step; full list specified in Supplementary Material 2). We assigned Gene Ontology terms (GO; [http://geneontology.org/](http://geneontology.org/)) using the same best BLAST hits from UniProt. UniProt is equipped with GO terms and annotations, hence it was efficient to use one database for both gene and GO annotations. We annotated contigs with available taxon identification to phylum level (level 4 in NCBI) using the best hits from a BLASTx search of the NCBI nr (non-redundant) database (bit score >50).
We ran the CEGMA analysis (Parra et al., 2007) using the 248 core eukaryotic proteins to test for completeness of the reference transcriptome, with default settings, except for specifying max_intron size of 2630 to account for known intron size information of the myxozoan *Myxobolus cerebralis* (see Chang et al., 2015).

**Phylogenomic analyses**

The raw reads were checked for quality using FastQC (Andrews, 2010) and adapters and low-quality sites were removed from the reads using Trimmomatic v0.38. The transcriptomic datasets for *Kudoa iwatai* (SRR1300899), *Thelohanellus kitauei* (SRR1103279) and *Polypodium hydriforme* (SRR1336770) were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra). All the reads were assembled using rnaSPAdes v3.13.1 using default settings. Protein coding sequences were predicted using default settings in program “LongestOrfs” implemented in TransDecoder v5.5.0.

The phylogenomic tree was built using a dataset of 78 ribosomal protein coding genes comprising diverse eukaryote taxa (Huchon et al., 2020, https://datadryad.org/stash/dataset/doi:10.5061/dryad.v15dv41sm). 51 genes of these 78 candidate sequences from our transcriptomes were identified by a tBLASTn search, e-value threshold 1e-10. These selected nucleotide sequences were then added to the respective gene alignments and aligned individually using the translation align function in Geneious Prime 2019.2.1 using MAFFT algorithms and the translation frame 1 (Katoh and Standley, 2013). Poorly aligned regions were then removed using trimAl v1.2 (with 0.01 gap rate cut-off) and preliminary trees were constructed using RAxML v8.2.12 with the GTRGAMMA model and 100 rapid bootstrap replicates. Single gene trees were manually inspected to select orthologous sequences, and paralogous sequences (see Supplementary Material 3 for more detail) and host contamination were removed from...
the datasets. The remaining orthologs were realigned following the previously described method and concatenated into a supermatrix using catfasta2phyml (Nylander, 2016). The resulting alignment was subjected to maximum likelihood phylogenetic analysis in RAxML v8.2.12 using the GTRGAMMA model and 100 rapid bootstrap replicates.

**Candidate virulence gene searches**

We chose cell migration genes and proteases/inhibitors as candidate virulence factors, due to their relevance in host-pathogen interactions (e.g. Barragan and Sibley, 2002; McKerrow et al., 2006; Bouzid et al., 2013). Given the low amount of functional (GO) annotation in our largest assembly, *C. shasta* + NHP (see Results), we used BLASTx to search for homologues of genes of interest in two custom concatenated databases that comprised the Cell Migration Knowledge Database (CMKB) which includes proteins, families and complexes involved in cell migration (http://www.cellmigration.org/index.shtml; ~7,600 protein sequences), and the MEROPS database, which consists of a non-redundant library of full length sequences of peptidases and peptidase inhibitors (http://merops.sanger.ac.uk/; ~450,000 sequences; Rawlings et al., 2016). We searched using the longest representatives for each gene in the *C. shasta* + NHP assembly (23,418 contigs; IIR-RBT6) then parsed matches (bitscore > 100), and posteriorly classified homologues to proteases or motility genes matching the specific databases. We then selected only genes with the same annotation in UniProt (determined using the same standards: no ambiguous terms filtering, bitscore > 100) to confirm gene identity. Due to disagreements between annotations (CMKB vs UniProt, and MEROPS vs UniProt), we curated gene lists manually, removing genes that matched one or more of the following criteria: 1) no genetic distances available (only available for reference); 2) disagreements between annotations from the different databases, after checking for synonyms or function similarity; 3) annotations that
contained terms from UniProt “Uncharacterized protein” and “Predicted protein” (ambiguous terms), and whose identity could not be confirmed; 4) annotations that contained non-specific terms, such as “heat shock protein” or “ribosomal protein”. For genetic distance analyses and inference of SNPs-based phylogenetic trees we created lists of homologs that met different sets of the above criteria (strict: 1-4 criteria; permissive-filtered: 1, 2 & 4; permissive: 1 & 4 criteria).

**Analysis of genetic variation between *C. shasta* genotypes using SNPs from RNAseq data**

The pipeline for finding (“calling”) SNPs from RNAseq data employs samtools v1.9 (Li et al., 2009) and bcftools v1.9 (Li, 2011) and is described in EM’s wiki page (http://sites.science.oregonstate.edu/~meyere/wiki/index.php/RNASeqSNPs). To quantify genetic variation between different genotypes of *C. shasta*, parasite and NHP reads of the different transcriptomes (genotype I, IIC and the three IIR) were mapped against the non-redundant assembly (23,418 sequences/genes) of the reference IIR-RBT6 transcriptome (*C. shasta* + NHP, filtered for host and other contaminants). We mapped reads using gmapper as detailed before, but reported only the 3 top hits per read, followed by more stringent filtering of the .sam file with a minimum 99 bp matches over a 101 bp length sequence alignments (allowing for only 2 mismatches), to recover only true SNPs. Highly expressed genes were resampled to avoid genotyping errors, using a maximum coverage of 100×. We used samtools and bcftools (see above) to convert .sam files to .bam files, then sorted, indexed, and called SNPs from the alignments generating .vcf (variant call format) files. Genotypes were called from nucleotide frequencies with a minimum coverage of 5× and heterozygososity threshold of 0.25. We combined results into tables, which consisted of a list of loci (labelled homozygous or heterozygous according to the parameters selected) for each *C. shasta*
transcriptome. These tables were combined as ten pairwise comparisons. Overall
genetic transcriptome-wide distances were calculated as 1 – (proportion of shared
alleles) (Bowcock et al. 1994). We estimated these distances for all genes (22,755), for
genes available across the different transcriptomes (593) and for cell migration (141)
and proteases and inhibitors (41) strictly curated datasets. Genetic distances of selected
genes from the cell migration and proteases and inhibitors datasets were parsed and
hierarchical cluster analyses were performed using hclust function in Rstudio (v3.6.0,
Inc. Boston, MA, USA) (Supl. MM for the R code). Intra-transcriptomic variation was
calculated by calling genotypes from the same alignments with at least 20 reads, in
order to confirm genotype results and to resolve any genotypic variation in the C. shasta
population infecting the sampled fish. The heterozygosity threshold was set at 0.1 (at
least 2 reads out of 20 to call a locus heterozygous). A ratio was calculated for each
transcriptome: heterozygous loci divided by total sites (homozygous +
heterozygous). Due to the limited number of replicates for each genotype, no statistical
comparison was performed between transcriptomes.

Using the strictly curated lists of homologues (of motility and proteases genes), we
parsed the genetic distances, then calculated the relative frequency (%) of genes for two
categories: perfectly conserved genes and genes with genetic divergence. We calculated
which subset of genes of interest was less conserved by subtracting the overall relative
frequency (%) of perfectly conserved genes for each transcriptome’s pairwise
comparison, from the relative frequency (%) of perfectly conserved cell migration or
proteases/inhibitors genes.

SNPs-based phylogenetic analyses
Using the genetic variation data, we explored whether SNPs-based phylogenetic analyses would resolve phylogenetic relationships between *C. shasta* genotypes. SNP genotypes matrices of all transcriptomes were converted to a FASTA-format alignment using the “gt2fas.pl” (https://github.com/em-bellis/2brad_utilities). The mask alignment option in Geneious Prime was used to remove identical bases from previously obtained alignments and a maximum likelihood trees were constructed using RAxML v8.2.12 (ASC_GTRGAMA model) with 1000 bootstrap replicates. Up to 8 different trees were generated from the datasets with different levels of completeness or curation mentioned above.

**Results**

**Sequencing and reads composition - removal of host reads**

Read data was archived in the NCBI Sequence Read Archive (Table 1; accession number PRJNA241036). Reference transcriptomes were assembled from a single deeply-sequenced sample (RBT6): comprising 788 million raw, paired 101 bp reads, with only 3.8% of reads filtered out as low quality, homopolymer repeats or adaptor sequences. The remaining 759 million high quality reads were mapped separately against both host and parasite to assess contamination: 59.6% (452 million) mapped best to *C. shasta*, 24.1% (182 million) matched neither of the two databases, 0.8% (5 million) were ties and 15.6% (118 million) had host origin and were discarded (Table 1, Figure 3).

The other four isolates (Table 1) were sequenced at lower coverage and yielded on average 29.6 (15.6 - 44.3) million paired 101 bp reads each. Only 19.3 – 40.0% of reads from these libraries were kept after quality and adaptor filtering steps. The two ascitic fluid-derived datasets (IIR) had the highest proportions of parasites reads (32.6 -
60.4%), whereas intestine-derived transcriptomes had the lowest: I (4.5 %), IIC (12.8%). The proportion of reads in the NHP bins were similar across transcriptomes (24.9 - 35.1%). Host contamination was 9.5 - 70.2%, with genotypes I and IIC having the highest amount (56.8 - 70.2%) (Figure 3).

Reference transcriptome assembly and functional annotations

We produced two reference transcriptomes from IIR-RBT6, using either C. shasta reads, or C. shasta reads + NHP reads (Figure 2). The 452 million reads that mapped best to the C. shasta genome were assembled into 44,986 transcripts (“C. shasta only” assembly), and the 635 million reads that mapped best to both the C. shasta genome and NHP were assembled into 75,087 transcripts (“C. shasta + NHP” assembly) (Table 2).

BLAST homology searches against UniProt annotated 18.8 / 21.2 % of transcripts (C. shasta only / C. shasta + NHP; Table 2; Supplementary File 1) with annotations for 7.6 / 13.1 % of transcripts < 400 bp, 20.9 / 21.1% of transcripts 400 - 1000bp, and 59.8 / 61.2 % of transcripts > 1000 bp. We assigned GO terms to 61.8 / 70.5% of transcripts that had matches with UniProt, which represented only 11.6 / 14.9% of the total assembly (Table 2; Supplementary File 1). Taxonomic origin obtained by BLAST search to NCBI nr showed that approximately one-third (35.2 / 39.4%; 15,831 / 29,560) of transcripts matched taxa in the nr database: 9.7 % (1,530) were annotated as Cnidaria in the C. shasta only assembly, whereas 7.2% (2,126 sequences) were annotated as Cnidaria in the C. shasta + NHP assembly (Table 2; Supplementary File 1).

Host and other contamination removal at contig level

Using Taxon ID annotation, we removed non-fish contamination (prokaryotes, viruses, fungi, protists, algae) from the final datasets: in C. shasta only we removed 12.1% (5,465 transcripts); in C. shasta + NHP we removed 8.9% (6,669 transcripts).
Secondary host filtering at the transcript level, by comparing transcripts from both reference assemblies against *C. shasta* and *O. mykiss* genomes, showed that the *C. shasta* only assembly had 96.4% (43,395) of transcripts matching the parasite genome (Figure 4A), with only 0.3% (115) transcripts matching fish; 3.3% (1,476) of transcripts had no match. *C. shasta* transcripts were the longest (av. length 600 bp, range 201 - 10,626 bp) while the fish transcripts still present after assembly were short (av. length 234 bp, range 201 - 481bp), and no match transcripts were also short (av. length 278 bp, 201 - 1,136 bp).

In the *C. shasta* + NHP assembly, 65% (48,824) of transcripts matched the parasite genome and 15.5% (11,628) matched rainbow trout (Figure 4B). No match transcripts represented 19.5% (14,635). Size distribution showed that both *C. shasta* and no match transcripts had similar length distributions: av. length 609 bp (range 201 - 18,696 bp), and 582 bp (201 - 12,037 bp) respectively, whereas fish transcripts were shorter: 262 bp (201 - 2,278 bp).

After these filtering steps, the final reference transcriptomes had the following compositions: *C. shasta* only: 39,407 transcripts (av. length 569 bp, 201 - 10,626 bp, N50 825); *C. shasta* + NHP: 56,876 transcripts (av. length 579 bp, 201 - 18,696 bp, N50 870). UniProt and GO terms annotations rates remained similar to non-fish + other contaminants filtered assemblies. By size, long transcripts were more highly represented in the annotated genes proportion (UniProt), due to the removal of short *O. mykiss* transcripts (6.5 / 7.4% of transcripts < 400 bp in length, 19.6 / 20.1% of transcripts 400 / 1000 bp in length and 65.5 / 66.4% of transcripts > 1000 bp) (Figure 4). Taxon ID annotations rate was lower (> 10%), due to removal of non-fish contamination (microorganisms).
CEGMA analysis of the *C. shasta* only assembly identified 46% complete core eukaryotic proteins present (50% if partial matches were included). For the cleaned *C. shasta* + NHP assembly, CEGMA identified 71% complete core eukaryotic proteins (76% if partial matches were included).

**Phylogenomic analyses of *Ceratonova shasta* genotypes**

ML tree reconstruction based on 51 genes from 8 taxa and 29,730 sites (Figure 5A) showed that *C. shasta* genotypes cluster in two well supported clades (96-100 bootstrap support), independent from geographic origin: one containing genotypes I and IIC from coho and Chinook salmon and a monophyletic group accommodating all transcriptomes of genotype IIR from rainbow trout. Within the IIR clade, IIR genotypes from the Klamath river (LKR and UKR) appear to be more closely related to each other than to IIR genotype from Willamette river. SNP-based ML analyses showed similar clustering when using the full SNP dataset (22,755 genes, 918 SNPs; Figure 5B) or a subset of genes present in all 10 pairwise comparisons (593 genes, 235 SNPs; Figure 5C).

**Overall transcriptome-wide genetic distances between genotypes**

The number of genes used to calculate genetic distances between transcriptomes varied depending on library completeness (657 – 3,857; Supplementary Table 1; Supplementary File 2). We found the highest genetic transcriptome-wide distances between genotypes I and II (2.1 x 10^{-3} - 2.4 x 10^{-3}). IIC was closest to IIR (1.7 x 10^{-3} - 1.8 x 10^{-3}), but the distances were the same magnitude between IIC and I (2.1 x 10^{-3}). IIR intra-genotype genetic distances were an order of magnitude lower (1.1 x 10^{-4} to 1.5 x 10^{-4}) compared to the distances observed between genotypes I and II (22,755 genes, see Supplementary Table 2 for genetic distances of the other genes datasets).

**Within transcriptome/genotype diversity**
Heterozygosity levels varied 2 - 10 fold between transcriptomes. The transcriptome with the highest variability was genotype I with 1 variant in every 208 loci (63 heterozygous loci/13,116 total sites) followed by the IIR (RBT6) with 1/432 variant loci (3,348 heterozygous loci/1,445,297 total sites). IIC showed 1/614 variant loci (113 heterozygous loci/69,226 total sites). The least variable libraries were the two IIR transcriptomes from the Klamath River, IIR (RBTC16) with 1/1,869 variant loci (48 heterozygous loci/9,670 total sites) and IIR (RBTJ7), with the lowest heterozygous loci ratio, with 1/2,159 variant loci (25 heterozygous loci/53,948 total sites).

**Genetic distances and SNPs-based phylogenetic analyses of putative virulence genes in different *C. shasta* genotypes**

From the longest representatives for each gene in the *C. shasta* + NHP assembly (23,418 contigs), 431 matched cell migration genes in CMKB and 507 had homologues to proteases and inhibitors in the MEROPS database. We identified many discrepancies between annotations from MEROPS and UniProt (without ambiguous terms filtering). Manual curation produced a subset of 41 homologous genes to MEROPS and 14 SNPs (strictly curated) for proteases and inhibitors, 56 genes, 15 SNPs (permissive_filtered dataset), and 110 genes, 26 SNPs (permissive dataset), listed in Supplementary File 3 & 4. Fewer discrepancies were observed between CMKB and UniProt annotations, but we removed genes with annotations seemingly unrelated to cell migration (*e.g.* heat shock proteins, ribosomal proteins). After curation, 164 gene homologues to CMKB, 40 SNPs remained in the strictly annotated dataset, 271 genes, 71 SNPs in the permissive_filtered one and 325 genes, 83 SNPs in the permissive one; (Supplementary File 3 & 4). We compared subsets of these among genotypes, depending on the completeness of the transcriptomes (Supplementary Table 1).
Relative frequencies of perfectly conserved transcripts (i.e. with no SNPs) are given in Supplementary Table 3. Relative frequency values of conserved genes of interest (cell migration or proteases/inhibitors) against overall conserved genes in each transcriptome, are shown in Figure 6A, Supplementary Table 4. Cell migration genes were less conserved between genotypes I and II (a negative relative value in I/IIC and I/IIR comparisons), but more conserved in IIC/IIR comparisons, and in intra-genotype IIR comparisons (Figure 6A; positive relative values). Conversely, proteases and inhibitors were more conserved for I vs. IIC, but less conserved for I vs. IIR (Figure 6A), but these results varied depending on which IIR dataset was used (Supplementary Figure 1). I/IIR-RBT6 and I/IIR-RBTC16 showed negative relative values, while I/IIR-RBTJ7 showed slightly positive value. Protease and inhibitor genes were in general less conserved for IIC/IIR (negative value) except for IIC/IIR-RBTC16, which showed a slight positive relative frequency. We observed lower relative frequencies between IIR transcriptomes, except for IIR-RBTJ7/IIR-RBTC16 (Supplementary Figure 1).

Cell migration SNP-based phylogenetic analyses (strictly curated to permissive datasets, Figure 6B and Suppl. Figure 2) showed the same tree topology as the larger transcriptomic SNP dataset of all genes (22,755 & 593 genes, respectively, Figure 5B & 5C). Topology had high nodal support, and indicated a close relationship of genotypes I and IIC, with all IIR transcriptomes forming a monophyletic sister clade to these two genotypes. The subset of proteases and inhibitors genes revealed an unstable tree topologies with both the strictly curated (Figure 6C), and the permissive_filtered (Suppl. Figure 2) datasets, having low nodal support likely due to the low number of genes and fewer informative sites within them. ML analyses of the permissive cell proteases/inhibitors dataset resulted in the same clustering as in phylogenomic analyses and motility-based SNP analyses, though with low nodal support.
SNP analyses highlight specific virulence gene candidates

When comparing cell migration subsets in I vs II, we observed non-zero genetic distances (based on SNPs) in actin cytoskeleton proteins (ARP2/3, several actin isoforms, F-actin capping protein), tubulin alpha chain and microtubule force-producing proteins (dynamin), actin binding (coronin), actin and tubulin folding molecules (T-complex protein 1) and Rho family (small GTPases involved in migration regulation) genes (Supplementary File 3, Supplementary Figure 3).

For the protease subset, up to 10 α and β subunits of the proteasome complex had genetic distances between all genotypes except for comparisons between the different IIR isolates. A cathepsin Z-like cysteine protease, homologous to cathepsin X of *M. cerebralis*, showed genetic differences between all I and II comparisons (except for any comparison to transcriptome IIR-RBTJ7), with the highest genetic distance for I/IIR-RBT6. A methionyl aminopeptidase (homologue to methionyl aminopeptidase 2 of *Hydra vulgaris syn. Hydra magnipapillata*) had genetic differences between I and two of the IIR datasets (IIR-RBTJ7 and IIR-RBT6). Other methionyl aminopeptidase showed differences between IIC/IIR-RBTC16 and IIC/IIR-RBTJ7. Genetic differences were observed for a serpin or serine protease inhibitor between I and II (except IIR-RBTJ7), IIR/IIC and IIR-RBTJ7/IIR-RBT6 (Supplementary File 3, Supplementary Figure 4).

Discussion

The challenge of myxozoan ‘omics: parasite identification and host filtration

Myxozoans are unculturable, and throughout most of their life cycle are composed of only a few cells. Only some species have spore-forming stages with macroscopic plasmodia, which represent clonal parasite material that can be physically separated
from host tissue; many other species are more broadly distributed within hosts, and
intimately associated with host cells. These properties make myxozoan ‘omics research
fundamentally challenging. Thus most genome and transcriptome data from myxozoans
are from spores or spore-forming stages (Jiménez-Guri et al., 2007; Holland et al., 2011;
Nesnidal et al., 2013; Yang et al., 2014; Foon et al., 2015; Chang et al., 2015). In a
parallel study we have sequenced the Ceratonova shasta genome from purified
myxospores, which had been separated physically from contaminating host cells, to
obtain a high proportion of parasite DNA (Atkinson et al., in preparation). In the
present study, we investigated transcriptome-wide genetic differences among closely
related Ceratonova shasta genotypes from different host fish. Given that C. shasta is an
obligate parasite that does not form cysts, its metabolically active developmental stages
typically occur in intimate contact with host intestinal tissues, which presents an
inherent challenge to obtaining samples with high proportions of parasite transcripts.
We succeeded in obtained high quality RNA from C. shasta stages, utilizing a unique
feature of the parasite’s biology: virulent genotype IIR can cause systemic infection in
allopatric stocks of rainbow trout and induce the production of ascitic fluid in the body
cavity (Figure 1). This ascitic fluid is rich in manifold C. shasta developmental stages
(visibly motile pre-sporogonic and sporogonic stages, and mature myxospores) and thus
provided excellent material for transcriptomics. The data that we obtained from a single
ascitic fluid sample (RBT6) had sufficient coverage and sequence depth to build a
reference transcriptome, against which we then compared the lower coverage
transcriptome data from the other genotypes, sampled from intestinal tissues, and which
presented the expected challenges of host contamination and low read depth.
Given the mixed host-parasite read data, we developed bioinformatic pipelines to
separate transcripts by species origin. Host genomes and transcriptomes, especially for
fish species important to aquaculture, are available (e.g. Berthelot et al., 2014; Tine et al., 2014; Xu et al., 2014; Lien et al., 2016). Two myxozoan studies have included host filtering steps in assembly pipelines (Chang et al., 2015; Foorx et al., 2015). Foorx et al. (2015) filtered the transcriptome of Myxobolus pendula using an in silico hybridization pipeline of iterative BLAST searches against a filtered myxozoan/cnidarian query set and a close relative fish host. Chang et al. (2015) filtered genomic and transcriptomic data using BLAST searches of the specific fish host genomes. We attempted post-assembly filtration using predicted proteins of these available myxozoans/cnidarians vs fish hosts and found this gave only a poor recovery rate of C. shasta sequences (data not shown). However, as we had genomes of both host (O. mykiss; public data) and parasite (C. shasta; our draft data, unpublished), we used these in a novel approach of combined pre-assembly read filtering, and post-assembly contig filtering. Removal of host reads prior to assembly greatly reduces assembly of host contigs, and minimizes creation of chimeric transcripts (Daly et al., 2015). Myxozoans have some of the most derived genomes of all metazoans (Chang et al., 2013; Holzer et al., 2018), and host-parasite sequence chimeras can be a frequent and insidious component of contaminated assemblies, hence a host removal step is essential for comparative approaches and may be further improved once long-read sequences become available. The requirement for host and parasite reference genomes may be considered a limitation in our pipeline but we found that this approach gave improved yields of transcriptome libraries with maximum expressed gene recovery and minor host contamination. During post-assembly filtering of contigs using BLAST, we created distinct bins of host- or parasite-transcripts, and a third bin of neither-host-nor-parasite (NHP), which allowed us to recover additional contigs that could not be identified by mapping against the references (i.e. either true parasite genes missing from the draft C. shasta genome, or non-parasite,
non-host contaminants). Interestingly, the similar size distribution of transcripts in both 
*C. shasta* and NHP bins suggested parasite origin for the majority of NHP transcripts.  
CEGMA analyses showed that the combined *C. shasta*+NHP assembly had more matches (71%) than *C. shasta* alone (46%), further supporting the validity of combining both contigs that matched to the parasite genome and non-fish matches. The difference in completeness of the *C. shasta* only and *C. shasta* + NHP assemblies revealed a limitation of our pipeline, and demonstrated the need to include NHP reads in our analyses to avoid losing parasite genes, which could not be matched to our presumably incomplete reference genome. The CEGMA completeness of *C. shasta* + NHP (71-76%) of identifiable core eukaryotic genes was similar to results from other cnidarian transcriptomes: 77-84% for *Kudoa iwatai*, 76-90% for five corallimorpharians, 87% for *H. vulgaris* syn. *H. magnipapillata*, 91% for *Calliactis polypus*, and 90-92% for three scleractinian corals (Chang et al., 2015; Rodrigues et al., 2016; Kenkel and Bay, 2017; Lin et al., 2017; Stewart et al., 2017).

We then used taxon ID assignments from NCBI database searches to further remove contaminating contigs attributable to microorganisms. This is the first time that multiple bioinformatics filters have been used to remove significant, unavoidable contaminants from myxozoan high throughput sequence datasets. Overall, we consider that our bioinformatic pipeline is optimal for generating transcriptomes of non-model organisms within host tissues, as it prevented both the assembly of chimeric sequences and the loss of unknown and highly derived parasite genes.

**Functional myxozoan annotations**

Myxozoan genomes are highly divergent from both their free-living cnidarian relatives and all other metazoans, even considering conserved genes (Hartigan et al., 2016). Thus myxozoan ‘omics datasets can be functionally annotated only poorly at present: *e.g.*
5.6% (Foxx et al., 2015), 19% (Chang, 2013), 21% (Yang et al., 2014; specified as 3,500/16,638 proteins), 11-13% (this study). Even annotation of genes by comparison with other myxozoan species is difficult, given that they are as strongly divergent from each other as from other cnidarians, and other metazoans. Development of a comprehensively annotated myxozoan reference will still require considerable manual effort to identify more unambiguous myxozoan genes and provide meaningful functional annotations.

The lack of functional annotations for C. shasta made enrichment analysis or parsing of the data using GO terms unfeasible, as several functional categories of interest did not have any hits in the transcriptome. We then switched to publicly available databases to parse genes of interest, however we observed discrepancies between UniProt and MEROPS annotations for our protease and inhibitor gene subsets of interest. This was in concordance with previous reports of miss-annotations from public databases, e.g. Schnoes et al. (2009) who show that >80% of contigs in 10 out of 37 enzyme families were mis-annotated. The main sources of these errors have been linked to over-prediction and error propagation (Jones et al., 2007, Schnoes et al., 2009), particularly in non-model species (Clark and Greenwood, 2016). As C. shasta, and myxozoans in general, are only distantly related to any model organism (even the free-living cnidarian Nematostella vectensis), we were not surprised by both the low rate and error-prone nature of the annotation using general organismal databases. For the cell migration database, CMBK, we obtained annotations that agreed substantially with those from UniProt (which was expected given CMBK is a subset of the NCBI database). However, we found that CMBK contained genes with a questionable link to cell motility (e.g. ribosomal genes, heat shock proteins), which suggested that present curation of this database is too permissive. Thus we regarded manually curated subsets
of genes as the best option for exploring our data. Future work should concentrate on better functional annotations of myxozoan genes of interest (virulence, structure) and cnidarian genes in general.

**C. shasta genotypes: transcriptomic data elucidate inter-genomic relationships and pave the way to new markers**

*C. shasta* represents an excellent myxozoan model parasite for several reasons: it affects both wild and cultured salmonid fishes of economic and cultural value; its life cycle is known and can be maintained in the laboratory (Bartholomew et al., 1997); and it is the only myxozoan species for which host-specific genotypes have been characterized and associated with different virulence patterns (Atkinson and Bartholomew, 2010a, b; Hurst and Bartholomew, 2012; Stinson et al., 2018). While *C. shasta* genotypes are characterized by ITS sequence differences, this marker has several drawbacks: it does not resolve observed virulence differences between coho salmon and rainbow trout (“biotypes” IIC and IIR; Hurst and Bartholomew, 2012); it can have intraorganism variation (Atkinson et al., 2018b); and it cannot differentiate between geographic isolates (Stinson et al., 2018). Additional markers for resolution of different *C. shasta* strains are needed to better characterize spatial and temporal variation of the parasite in environmental samples, particularly when assessing risk for endangered coho salmon populations in the Pacific Northwest (Williams et al., 2016).

Our phylogenomic analyses of *C. shasta* provided unexpected results regarding the relationships between the different genotypes. We provide, for the first time, significant support for the separation of IIC and IIR into two independent genotypes, which are not resolved by ITS data alone. Phylogenomics showed that genotypes I and IIC share a closer relationship with each other than with IIR, which correlates with the evolutionary distance between their fish hosts, with coho (IIC) and Chinook (I) salmon being more
closely related to each other than to rainbow trout (IIR; Domanico et al., 1997; Crête-Lafrenière et al., 2012). We hypothesize that the high virulence of IIR in a naïve host (allopatric rainbow trout) is a consequence of a relatively recent host switch from a common ancestor of genotypes I and IIC.

Though based on a limited number of isolates, we also observed some phylogeographic signal, in addition to host association, in clustering of the genotypes. The two sympatric IIR isolates from the Klamath River clustered together, while the IIR isolate from the Willamette River (some 500 km distant) was distinct (Figure 5). This signal of sympatry is consistent with data from studies showing that salmonids show strong site fidelity (e.g. Minakawa and Kraft, 2005; Dittman et al., 2010), which can lead to establishment of local pathogen strains, for example in IHN virus (Kurath et al., 2003) and in another myxozoan, Parvicapsula minibicornis (Atkinson et al., 2011).

Characterizing virulence in C. shasta genotypes

Heterozygosity: We examined patterns of C. shasta genotypic diversity and virulence by comparing heterozygosity levels between isolates from different geographic localities and hosts. We found no obvious relationship between parasite heterozygosity and a particular geographic site, though our analyses were limited by geographic sampling from only three sites (Lower and Upper Klamath River, Willamette River). We observed isolates of low (IIR-RBTJ7), medium (IIC) and high (I) heterozygosity within the Klamath River. High virulence of pathogen strains is related to higher levels of heterozygosity (Cogliati et al., 2012). This may be the case for C. shasta genotype I, which is highly diverse and relatively pathogenic in a native host (Chinook salmon), however, genotype IIR, which is highly virulent in allopatric rainbow trout stocks, had low (IIR-RBTJ7 and IIR-RBTC16) or moderately high diversity (IIR-RBT6). Hence we
found no evidence to suggest that greater heterozygosity correlates with the higher virulence of *C. shasta* genotype IIR.

**Cell migration**: Cell-motility in parasites is important for host invasion, adhesion to host cells and migration through tissues (Barragan and Sibley, 2002; Lentini et al., 2015) and hence can affect virulence. For example, the capacity of *Toxoplasma gondii* for migration, both across tissues and over long distances, characterizes its virulence (Barragan and Sibley, 2002). Developmental stages of *C. shasta* are motile and capable of producing different cell protrusions (filopodia, lamellipodia and blebs), with specific functions such as anchoring/adhesion, crawling and blebbing (Alama-Bermejo et al., 2019). For *C. shasta*, we hypothesized that differences in parasite migration and proliferation strategies underpin differences in virulence (proliferation rate, spore production, cumulative mortality rate), as observed previously between genotypes I/IIC and IIR (Hallett et al., 2012; Stinson and Bartholomew, 2012). *In vivo* observations reveal that IIR infection is characterized by rapid proliferation, fast amoeboid bleb-based motility and high adhesion, with significant differential expression (by qPCR) of key motility and adhesive factors between type 0 (low virulent) and IIR (highly virulent) genotypes in rainbow trout (Alama-Bermejo et al., 2019). In the current study, we curated a “motility” dataset based on genes with functions related to migration, proliferation, cytoskeleton and cell division. While our genetic distance analyses showed there was a higher frequency of perfectly conserved cell migration genes between IIC and IIR (Figure 6A), our SNPs-based phylogenetic analyses (Figure 6B) indicated a closer evolutionary relationship between I and IIC than either of them with IIR. This was in concordance with the observed general phylogenomic clustering of the genotypes, and suggested that these genes are likely under positive selection pressure and linked to virulence. The genetic variation and phylogenetic relationships of SNPs in
cell migration genes that we observed provides a valuable genetic framework to understand the observed differences in *C. shasta* virulence. Future work could explore silencing of selected motility factors in *C. shasta* to determine their importance and roles in parasite virulence.

**Proteolytic enzymes and inhibitors:** Proteases and their inhibitors comprise 1-5% of infectious organism genomes (Tyndall et al. 2005), and this appears similar in myxozoans: ~2.5% (422 proteins) in the *T. kitauei* proteome (Yang et al., 2014) and 2.6% of expressed genes in *S. molnari* (Hartigan et al. under review). Our relative frequency analyses showed that these genes are generally less conserved/more variable in *C. shasta* than cell migration genes. We observed that they had more variation between genotypes I and IIR, and IIC and IIR, but were more conserved between I and IIC, again showing that IIR was distinct. We consider this evidence that proteolytic enzymes are important contributors of virulence in *C. shasta*. This pattern was not well supported by the SNP data as they did not produce a well-supported SNP tree, but this is likely an artifact of the limited genes and positions available for analyses in the curated and permissive_filtered dataset (only 41 and 56 genes respectively), while the largest dataset (permissive, 110 genes) supported the same phylogenetic clustering of genotypes as in the phylogenomic study and the cell migration dataset (≥164 genes). Whether the observed SNP-based distances are relevant at the functional level is unknown and requires functional analyses of specific genes. Our study is the first to examine genetic distances between biologically different genotypes of a myxozoan and, given associations between relative SNP frequencies and virulence, it is likely that at least some of these genetic differences are translated into differences in protein regulation and activities that affect virulence in different fish hosts.
Biological relevance of candidate virulence genes in *C. shasta* – pinpointing molecules for future interference studies

Pathogens evolve with their hosts in a stepwise “arms race” of virulence and resistance, mediated by genetic changes (Coletta-Filho et al., 2015). We identified molecular differences between genes related to parasite motility and proteolysis in *C. shasta* host-associated genotypes. Specifically, we observed that genes encoding several molecules important to the formation and functioning of the cytoskeleton (ARP2/3, F-actin capping protein) and dynamin, had smaller genetic distances between them in the two II genotypes than between either genotype II and I. These proteins are involved in growth, morphogenesis, migration, cell-to-cell spread and virulence of bacteria (Choe and Welch, 2016), fungal plant pathogens (González-Rodríguez et al., 2016), and parasites (Bookwalter et al., 2017). Only a few motility genes in our *C. shasta* transcriptomes showed smaller genetic distances between genotypes I and IIC, than to IIR, for example coronin, which is an actin filament-binding protein. This protein is essential for the transmission of *Plasmodium*, as individuals without coronin are unable to invade the salivary glands of mosquitos (Bane et al., 2016). During the invasion of tissues in the systemic phase of *C. shasta* infections, small changes in the coronin structure could be responsible for a more efficient/faster colonization of tissues by genotype IIR, making it more virulent than IIC and I.

We identified several proteases and protease-inhibitors with smaller genetic distances between I and IIR, than between IIC and IIR, particularly proteasome subunits. The proteasome is a protease complex that plays a central role in regulating cellular processes, including cell cycle, apoptosis and differentiation, modulation of immune response and regulation of gene expression in eukaryotic cells (Konstantinova et al., 2008; Tanaka, 2009). In trypanosomes, protein degradation during parasite cell
differentiation is primarily proteasome-dependent, and proteasome inhibition impedes
differentiation and transformation from non-infectious epimastigotes to infective
trypomastigotes (Cardoso et al., 2011; Gupta et al., 2018). Proteasome inhibitors are
used routinely in cancer therapy (e.g. multiple myeloma) (Manasanch and Orlowski,
2017) and are considered both disease markers and therapeutic targets (Muñoz et al.,
2015; Morais et al., 2017; Varga et al., 2017). We observed marked sequence variation
in *C. shasta* proteasome subunit mRNA sequences between different genotypes, which
suggested that they are important to *C. shasta* pathogenesis and they should similarly be
considered as both potential molecular markers for *C. shasta* genotypes, and drug
targets for myxozoans. Other proteolytic enzymes and inhibitors with potentially
important differences between genotypes were: 1. cathepsin Z, whose homolog in *M.
cerebralis* is thought to be involved in tissue invasion and lysis or the initiation of
sporogenesis (Kelley et al., 2003); 2. methionyl aminopeptidases, which are enzymes
whose inhibition results in antiparasitic activity (Chen et al., 2006, Zheng et al., 2015),
e.g. by growth inhibition; 3. protease inhibitors (serpins) which are common in blood-
feeding parasites such as ticks, for example, the cattle tick *Rhipicephalus microplus*
encodes at least 24 serpins that inhibit pro-inflammatory and pro-coagulatory proteases
of the host (Tirloni et al., 2014), processes that are likely to be important in blood and
tissue dwelling organisms such as myxozoans. Future molecular and functional studies
should test the suitability of these genes as virulence markers and determine their
specific role in parasite migration and proteolysis.

**Conclusions**

The advent of genomics has facilitated the *in silico* identification of parasite virulence
factors, a key objective for therapeutant design. Yet for obligate parasites, creation of
high quality reference genomic and transcriptomic assemblies are made challenging by
host contamination, and uncertainties of parasite gene identification and annotation. For myxozoan ‘omics studies, the production of well-curated genomic and transcriptomic databases has been a major hurdle to progress in targeted vaccine approaches for this parasite group. Hence, we designed an optimized bioinformatics pipeline for processing high throughput sequencing data of myxozoans, which recovers a maximum number of putative parasite genes while filtering out sequences of the host and other contaminant organisms. We recommend this repeat-filtration pipeline as a method for cleaning up sequence data of derived non-model organisms such as myxozoans. This workflow allowed us to produce transcriptomic datasets of host-associated genotypes of *C. shasta* having different virulence in their respective hosts. We analyzed genetic distances and SNPs between the transcriptomes, with a focus on candidate virulence factors, motility genes and proteolytic enzymes including their inhibitors. Phylogenomic analyses supported the observed host-associated clustering of genotypes, and suggested an evolutionary history based on host switches, rather than geographic variations. Altogether these results support characterization of genotype II into sub-types IIC (coho salmon) and the derived type IIR (rainbow trout). In the Klamath River, we speculate that IIR evolved as a result of host switching to introduced rainbow trout when its natural coho salmon host was excluded with the construction of barrier dams (Hurst and Bartholomew, 2012; Hurst et al., 2012). In this naïve host, genotype IIR is highly virulent and we hypothesize that this indicates a relatively recent host-switch event with insufficient time elapsed for purifying selection among virulence factors, *i.e.* selective removal of gene variants that are deleterious in this genotype, which would occur over time as a result of mutual host-parasite adaptation. We identified variation in genes that are essential for the biology of the parasite and play a role in shaping *C. shasta* virulence, for example the observed SNPs in motility and protease genes clearly

http://mc.manuscriptcentral.com/gbe
correlated with the virulent genotype. These findings represent a crucial step towards characterizing the connections between genotype and pathology. By cataloguing the pan-genomic SNP diversity of *C. shasta*, we have created a valuable resource for the development of diagnostic tools and as future targets for therapeutic intervention and vaccine design in salmonid enteronecrosis for this genetically complex parasite.

**Acknowledgments**

We are grateful to Ruth Milston-Clements, Ryan Craig and Richard Holt for fish husbandry at the John L. Fryer Aquatic Animal Health Laboratory; to Charlene Hurst (OSU) for her help with RNA extractions; and to Mark Dasenko at OSU CGRB for library preparation and next generation sequencing. We would like to thank to Tamar Lotan (University of Haifa) for critical feedback on this manuscript. We acknowledge the following funding agencies: Czech Science Foundation (G.A.B. Postdoctoral project 14–28784P and ECIP 505/12/G112; A.S.H. EXPRO grant 19-28399X), Consellería de Educación, Investigación, Cultura y Deporte, Valencia, Spain (G.A.B. #APOSTD/2013/087), Czech Academy of Sciences -Fellowship Purkyne (M.K.) and by European Regional Development Fund (M.K. CZ.02.1.01/0.0/0.0/16_019/0000759). Funds for this project were provided in part by the Bureau of Reclamation, US Department of Interior through Inter-agency Agreement #R15PG00065, as part of its mission to manage, develop, and protect water and related resources in an environmentally and economically sound manner in the interest of the American public. The views in this report are the authors’ and do not necessarily represent the views of Bureau of Reclamation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Mention of trade names does not imply U.S. Government endorsement.

**Data availability statement**

http://mc.manuscriptcentral.com/gbe
The datasets analysed during the current study are available in the SRA-NCBI repository under SRA accession number SRP040506, BioProject PRJNA241036. Biosamples represent reads from 3 genotypes with single (genotype I and IIC) or triplicate (genotype IIR) isolates: SAMN12275264 (IIR_RBT6; reference transcriptome), SAMN12275265 (IIR_RBTC16), SAMN12275266 (IIR_RBTJ7), SAMN12275267 (IIC), SAMN12275268 (I).

The following intermediate and final files of all analyses performed in this study are deposited in DRYAD (https://doi.org/10.5061/dryad.tx95x69tt): 1) Host filtered parasite and neither reads lists of I, IIC, IIR_RBT7, IIR_RBTC16 and IIR_RBT6 (20 .list files); 2) Genotype IIR_RBT6 reference assembled transcriptomes (non-filtered) (2 .fasta files); 3) Host and other contaminants (other microorganisms) filtered IIR_RBT6 assemblies (2 .fasta files); 4) Longest representatives reference IIR_RBT6 cs+neither assembly used for SNPs analyses (1 .fasta file); 5) SNPs tables (Genotypes called from nucleotide frequencies with a) minimum coverage of 5 and 0.25 heterozygosity threshold (5 .tab files) and b) minimum coverage of 20 reads and 0.1 heterozygosity threshold (5 .tab files)); 6) Phylogenomic and SNPs-based phylogenetic alignments (9 .nex files and 51 individual genes alignments .fasta used for phylogenomics).

**Competing interests**

The authors declare that they have no competing interests.

**Authors contributions statement**

GAB, SA, AH & JB conceived and designed the experiments; GAB & SA collected and processed the samples; EM wrote the scripts and designed filtering pipeline; GAB, EM & SA analyzed the data; MW & MK performed phylogenetic analyses; GAB & AH drafted the manuscript. All authors edited and approved the final manuscript.
References

Alama-Bermejo, G., Holzer, A.S., Bartholomew, J.L., 2019. Myxozoan Adhesion and Virulence: *Ceratonova shasta* on the Move. Microorganisms 7, 397. https://doi.org/10.3390/microorganisms7100397.

Ali, I.K.M., Mondal, U., Roy, S., Haque, R., Petri, W.A., Clark, C.G., 2007. Evidence for a link between parasite genotype and outcome of infection with *Entamoeba histolytica*. J. Clin. Microbiol. 45, 285-289.

Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data. Available: http://www.bioinformatics.babraham.ac.uk/?/projects/fastqc/. Accessed 2019 August 14.

Atkinson, S.D., Bartholomew, J.L., 2010a. Disparate infection patterns of *Ceratomyxa shasta* (Myxozoa) in rainbow trout *Oncorhynchus mykiss* and Chinook salmon *Oncorhynchus tshawytscha* correlate with internal transcribed spacer-1 sequence variation in the parasite. Int. J. Parasitol. 40, 599-604. https://doi.org/10.1016/j.ijpara.2009.10.010.

Atkinson, S.D., Bartholomew, J.L., 2010b. Spatial, temporal and host factors structure the *Ceratomyxa shasta* (Myxozoa) population in the Klamath River basin. Infect. Genet. Evol. 10, 1019-1026. https://doi.org/10.1016/j.meegid.2010.06.013.

Atkinson, S.D., Bartholomew, J.L., Lotan, T., 2018a. Myxozoans: Ancient metazoan parasites find a home in phylum Cnidaria. Zoology 129, 66-68. https://doi.org/10.1016/j.zool.2018.06.005.

Atkinson, S.D., Hallett, S.L., Bartholomew, J.L., 2018b. Genotyping of individual *Ceratonova shasta* (Cnidaria: Myxosporea) myxospores reveals intra-spore ITS-1
variation and invalidates the distinction of genotypes II and III. Parasitology. 27, 1-6. https://doi.org/10.1017/S0031182018000422.

Atkinson, S.D., Jones, S.R., Adlard, R.D., Bartholomew, J.L., 2011. Geographical and host distribution patterns of Parvicapsula minibicornis (Myxozoa) small subunit ribosomal RNA genetic types. Parasitology 138, 969-977. https://doi.org/10.1017/S0031182011000734.

Bane, K.S., Lepper, S., Kehrer, J., Sattler, J.M., Singer, M., Reinig, M., et al., 2016. The Actin Filament-Binding Protein Coronin Regulates Motility in Plasmodium Sporozoites. PLoS Pathog. 12:e1005710. https://doi.org/10.1371/journal.ppat.1005710.

Barragan, A., Sibley, L.D., 2002. Transepithelial migration of Toxoplasma gondii is linked to parasite motility and virulence. J. Exp. Med. 195, 1625-1633.

Bartholomew, J.L., Whipple, M.J., Stevens, D.G., Fryer, J.L., 1997. The life cycle of Ceratomyxa shasta, a myxosporean parasite of salmonids, requires a freshwater polychaete as an alternate host. J. Parasitol. 83, 859-868.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nature Commun. 5:3657. https://doi.org/10.1038/ncomms4657.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L., 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457. https://doi.org/10.1038/368455a0
Bjork, S.J., Bartholomew, J.L., 2009a. Effects of Ceratomyxa shasta dose on a susceptible strain of rainbow trout and comparatively resistant Chinook and coho salmon. Dis. Aquat. Org. 86, 29-37.

Bjork, S.J., Bartholomew, J.L., 2009b. The effects of water velocity on the Ceratomyxa shasta infectious cycle. J. Fish Dis. 32, 131-142. http://doi.org/10.1111/j.1365-2761.2008.00964.x.

Bjork, S.J., Bartholomew, J.L., 2010. Invasion of Ceratomyxa shasta (Myxozoa) and comparison of migration to the intestine between susceptible and resistant fish hosts. Int. J. Parasitol. 40, 1087-1095. doi: 10.1016/j.ijpara.2010.03.005.

Bookwalter, C.S., Tay C.L., McCorrie, R., Previs, M.J., Lu, H., Krementsova, E.B., et al., 2017. Reconstitution of the core of the malaria parasite glideosome with recombinant Plasmodium class XIV myosin A and Plasmodium actin. J. Biol. Chem. 292, 19290-19303. doi: 10.1074/jbc.M117.813972.

Bouzid, M., Hunter, P.R., Chalmers, R.M., Tyler, K.M., 2013. Cryptosporidium pathogenicity and virulence. Clin. Microbiol. Rev. 26, 115-134.

Cardoso, J., Lima, C.D.P., Leal, T., Gradia, D.F., Fragoso, S.P., Goldenberg, S., et al., 2011. Analysis of Proteasomal Proteolysis during the In Vitro Metacyclogenesis of Trypanosoma cruzi. PLoS ONE 6:e21027. https://doi.org/10.1371/journal.pone.0021027.

Casadevall, A., Pirofski, L., 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. Infect. Immun. 67, 3703-3713.
Chang, E., 2013. Transcriptomic evidence that enigmatic parasites *Polypodium hydriforme* and Myxozoa are cnidarians. [dissertation/PhD’s thesis]. [Kansas (KS)]: University of Kansas.

Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., Cartwright, P., 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. P. Natl. Acad. Sci. USA. 112, 14912-14917.
https://doi.org/10.1073/pnas.1511468112.

Chen, X., Chong, C.R., Shi, L., Yoshimoto, T., Sullivan, D.J., Liu, J.O., 2006. Inhibitors of *Plasmodium falciparum* methionine aminopeptidase 1b possess antimalarial activity. P. Natl. Acad. Sci. USA 103, 14548-14553. https://doi.org/10.1073/pnas.0604101103.

Choe, J.E, Welch, M.D., 2016. Actin-based motility of bacterial pathogens: mechanistic diversity and its impact on virulence. Pathogens and Disease, 74, ftw099.
https://doi.org/10.1093/femspd/ftw099.

Clark, K.F., Greenwood, S.J., 2016. Next-Generation Sequencing and the Crustacean Immune System: The Need for Alternatives in Immune Gene Annotation, Integr. Comp. Biol. 56, 1113–1130. https://doi.org/10.1093/icb/icw023.

Cogliati, M., Barchiesi, F., Spreghini, E., Tortorano, A.M., 2012. Heterozygosis and Pathogenicity of *Cryptococcus neoformans* AD-Hybrid Isolates. Mycopathologia 173: 347. https://doi.org/10.1007/s11046-011-9467-x.

Coletta-Filho, H.D., Bittleston, L.S., Lopes, J.R.S., Daugherty, M.P., Almeida, R.P.P., 2015. Genetic distance may underlie virulence differences among isolates of a bacterial plant pathogen. J. Plant Pathol. 97, 465-469. http://dx.doi.org/10.4454/JPP.V97I3.022
Crête-Lafrenière, A., Weir, L.K., Bernatchez, L., 2012. Framing the salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. PLoS ONE 7:e46662. https://doi.org/10.1371/journal.pone.0046662.

Daly, G.M., Leggett, R.M., Rowe, W., Stubbs, S., Wilkinson, M., Ramirez-Gonzalez, R.H., et al., 2015. Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data. PLoS ONE 10:e0129059. https://doi.org/10.1371/journal.pone.0129059.

Dardé, M.L., 2008. Toxoplasma gondii, "new" genotypes and virulence. Parasite 15, 366-371.

David, M., Dzamba, M., Lister, D., Ilie, L., Brudno, M. 2011. SHRiMP2: Sensitive yet practical short read mapping. Bioinformatics 27, 1011-1012. https://doi.org/10.1093/bioinformatics/btr046.

Dittman, A.H., May, D., Larsen, D.A., Moser, M.L., Johnston, M., Fast, D., 2010. Homing and Spawning Site Selection by Supplemented Hatchery- and Natural-Origin Yakima River Spring Chinook Salmon. T. Am. Fish. Soc. 139, 1014-1028. https://doi.org/10.1577/T09-159.1.

Domanico, M.J., Phillips, R.B., Oakley, T.H., 1997. Phylogenetic analysis of Pacific salmon (genus Oncorhynchus) using nuclear and mitochondrial DNA sequences. Can. J. Fish. Aquat. Sci. 54, 1865-1872. https://doi.org/10.1006/mpev.1995.1034

Eichenberger, R.M., Ramakrishnan, C., Russo, G., Deplazes, P., Hehl, A.B., 2017. Genome-wide analysis of gene expression and protein secretion of Babesia canis during virulent infection identifies potential pathogenicity factors. Sci. Rep. 7:3357. https://doi.org/10.1038/s41598-017-03445-x.

http://mc.manuscriptcentral.com/gbe
Foolx, J., Ringuette, M., Desser, S.S., Siddall, M.E., 2015. *In silico* hybridization enables transcriptomic illumination of the nature and evolution of Myxozoa. BMC Genomics. 16:840. https://doi.org/10.1186/s12864-015-2039-6.

González-Rodríguez, V.E., Garrido, C., Cantoral, J.M., Schumacher, J., 2016. The F-actin capping protein is required for hyphal growth and full virulence but is dispensable for septum formation in *Botrytis cinerea*. Fungal Biol.-UK 120, 1225-1235. https://doi.org/10.1016/j.funbio.2016.07.007.

Gupta, I., Aggarwal, S., Singh, K., Yadav, A., Khan, S., 2018. Ubiquitin Proteasome pathway proteins as potential drug targets in parasite *Trypanosoma cruzi*. Sci. Rep. 8:8399. https://doi.org/10.1038/s41598-018-26532-z.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., et al., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494-1512.

Hallett, S.L., Ray, R.A., Hurst, C.N., Holt, R.A., Buckles, G.R., Atkinson, S.D., Bartholomew, J.L., 2012. Density of the Waterborne Parasite *Ceratomyxa shasta* and Its Biological Effects on Salmon. Appl. Environ. Microb. 78, 3724-3731. http://doi.org/10.1128/AEM.07801-11.

Hallett, S.L., Bartholomew, J.L., 2012. “*Myxobolus cerebralis* and *Ceratomyxa shasta.*” in Fish parasites: pathobiology and protection, ed. P. T. K. Woo, K. Buchmann (CABI), 141-172. https://doi.org/10.1111/faf.12000

Hartigan, A., Estensoro, I., Vancová, M., Bílý, T., Patra, S., Eszterbauer, E., Holzer, A. S., 2016. New cell motility model observed in parasitic cnidarian *Sphaerospora molnari* (Myxozoa:Myxosporea) blood stages in fish. Sci. Rep. 6:39093. https://doi.org/10.1038/srep39093.

http://mc.manuscriptcentral.com/gbe
Hartikainen, H., Bass, D., Briscoe, A. G., Knipe, H., Green, A. J., Okamura, B., 2016. Assessing myxozoan presence and diversity using environmental DNA. Int. J. Parasitol. 46, 781-792. https://doi.org/10.1016/j.ijpara.2016.07.006.

Holland, J.W., Okamura, B., Hartikainen, H., Secombes, C.J., 2011. A novel minicollagen gene links cnidarians and myxozoans. P. Roy. Soc. B-Biol. Sci. 278, 546-553. https://doi.org/10.1098/rspb.2010.1301

Holzer, A.S., Bartošová-Sojková, P., Born-Torrijos, A., Lővy, A., Hartigan, A., Fiala, I., 2018. The joint evolution of the Myxozoa and their alternate hosts: A cnidarian recipe for success and vast biodiversity. Mol. Ecol. 27, 1651-1666. https://doi.org/10.1111/mec.14558.

Huchon, Dorothée et al. (2020), An animal without a mitochondrial genome: the salmon parasite Henneguya salminicola (Cnidaria: Myxozoa), Dryad, Dataset, https://doi.org/10.5061/dryad.v15dv41sm

Hurst, C.N., Bartholomew, J.L., 2012. Ceratomyxa shasta genotypes cause differential mortality in their salmonid hosts. J. Fish Dis. 35, 725-732. https://doi.org/10.1111/j.1365-2761.2012.01407.x.

Hurst, C.N., Holt, R.A., Bartholomew, J.L., 2012. Dam removal and implications for fish health: Ceratomyxa shasta in the Williamson River, Oregon, USA. N. Am. J. Fish Manage. 32, 14-23. https://doi.org/10.1080/02755947.2012.655843.

Jiménez-Guri, E., Philippe, H., Okamura, B., Holland, P.W.H., 2007. Buddenbrockia is a cnidarian worm. Science 317, 116-118. https://doi.org/10.1126/science.1142024
Jones, C.E., Brown, A.L., Baumann, U., 2007. Estimating the annotation error rate of curated go database sequence annotations. BMC Bioinform. 8:170. https://doi.org/10.1186/1471-2105-8-170.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780. doi:10.1093/molbev/mst010

Kelley, G.O., Adkison, M.A., Leutenegger, C.M., Hedrick, R.P., 2003. *Myxobolus cerebralis*: identification of a cathepsin Z-like protease gene (MyxCP-1) expressed during parasite development in rainbow trout, *Oncorhynchus mykiss*. Exp. Parasitol. 105, 201-210. https://doi.org/10.1016/j.exppara.2003.12.004.

Kenkel, C.D., Bay, L.K., 2017. Novel transcriptome resources for three scleractinian coral species from the Indo-Pacific. GigaScience 6:1-4. https://doi.org/10.1093/gigascience/gix074.

Konstantinova, I.M., Tsimokha, A.S., Mittenberg, A.G., 2008. Role of proteasomes in cellular regulation. Int. Rev. Cell Mol. Biol. 267,59-124. https://doi.org/10.1016/S1937-6448(08)00602-3.

Kurath, G., Garver, K.A., Troyer, R.M., Emmenegger, E J., Einer-Jensen, K., Anderson, E.D., 2003. Phylogeography of infectious haematopoietic necrosis virus in North America. J. Gen. Virol. 84,803-814. doi: 10.1099/vir.0.18771-0.

Lentini, G., Kong-Hap, M., El Hajj, H., Francia, M., Claudet, C., Striepen, B., et al., 2015. Identification and characterization of *Toxoplasma* SIP, a conserved apicomplexan cytoskeleton protein involved in maintaining the shape, motility and virulence of the parasite. Cell. Microbiol. 17, 62-78. http://doi.org/10.1111/cmi.12337.
Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. 12, 323. https://doi.org/10.1186/1471-2105-12-323.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27, 2987-2993. https://doi.org/10.1093/bioinformatics/btr509.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al., 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., et al., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200-205. https://doi.org/10.1038/nature17164.

Lin, M.F., Moya, A., Ying, H., Chen, C.A., Cooke, I., Ball, E.E., et al., 2017. Analyses of corallimorpharian transcriptomes provide new perspectives on the evolution of calcification in the Scleractinia (Corals). Genome Biol. Evol. 9:150-160. https://doi.org/10.1093/gbe/evw297.

Manasanch, E.E., Orlowski, R.Z., 2017. Proteasome inhibitors in cancer therapy. Nat. Rev. Clin. Oncol. 14:417-433. https://doi.org/10.1038/nrclinonc.2016.206.

McKerrow, J., Caffrey, C., Kelly, B., Loke, P., Sajid, M., 2006. Proteases in parasitic diseases. Annu. Rev. Pathol. 1, 497-536. https://doi.org/10.1146/annurev.pathol.1.110304.100151

Mennerat, A., Nilsen, F., Ebert, D., Skorping, A., 2010. Intensive farming: evolutionary implications for parasites and pathogens. Evol. Biol. 37, 59–67. https://doi.org/10.1007/s11692-010-9089-0
Meyer, E., Aglyamova, G.V., Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J.K., et al., 2009. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFLx. BMC Genomics 10:219. https://doi.org/10.1186/1471-2164-10-219.

Minakawa, N., Kraft, G., 2005. Homing behaviour of juvenile coho salmon (Oncorhynchus kisutch) within an off-channel habitat. Ecol. Freshw. Fish 14:197-201. https://doi.org/10.1111/j.1600-0633.2005.00093.x.

Morais, E.R., Oliveira, K.C., Paula, R.Gd., Ornelas, A.M.M., Moreira, É.B.C., Badoco, F.R., et al., 2017. Effects of proteasome inhibitor MG-132 on the parasite Schistosoma mansoni. PLoS ONE 12:e0184192. https://doi.org/10.1371/journal.pone.0184192.

Muñoz, C., Francisco, J.S., Gutiérrez, B., González, J., 2015. Role of the Ubiquitin-Proteasome Systems in the Biology and Virulence of Protozoan Parasites. BioMed Res. Int. 2015, 141526. http://dx.doi.org/10.1155/2015/141526.

Nesnidal, M.P., Helmkampf, M., Bruchhaus, I., El-Matbouli, M., Hausdorf, B., 2013. Agent of whirling disease meets orphan worm: phylogenomic analyses firmly place Myxozoa in Cnidaria. PLoS ONE 8,e54576. https://doi.org/10.1371/journal.pone.0054576.

Nylander, J., 2016. catfasta2phyml. Available at: https://github.com/nylander/catfasta2phyml/. Accessed 2019 September 12.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061-1067. https://doi.org/10.1093/bioinformatics/btm071.
Pearson, W.R., 2013. An introduction to sequence similarity (‘homology’) searching. Curr. Protoc. Bioinformatics. Chapter 3: Unit 3.1. https://doi.org/10.1002/0471250953.bi0301s42.

Rawlings, N.D., Barrett, A.J., Finn, R.D., 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 44, D343-350. https://doi.org/10.1093/nar/gkv1118.

Ray, R.A., Holt, R.A., Bartholomew, J.L., 2012. Relationship between temperature and Ceratomyxa shasta -induced mortality in Klamath River salmonids. J. Parasitol. 98, 520-526. https://doi.org/10.1645/JP-GE-2737.1.

Rodrigues, M., Ostermann, T., Kremeser, L., Lindner, H., Beisel, C., Berezikov, E., et al., 2016. Profiling of adhesive-related genes in the freshwater cnidarian Hydra magnipapillata by transcriptomics and proteomics. Biofouling 32,1115-1129. https://doi.org/10.1080/08927014.2016.1233325.

Schnoes, A.M., Brown, S.D., Dodevski, I., Babbitt, P.C., 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. 5:e1000605. https://doi.org/10.1371/journal.pcbi.1000605.

Seib, K.L., Dougan, G., Rappuoli, R., 2009. The key role of genomics in modern vaccine and drug design for emerging infectious diseases. PLoS Genet. 5:e1000612. https://doi.org/10.1371/journal.pgen.1000612.

Shpirer, E., Diamant, A., Cartwright, P., Huchon, D. 2018. A genome wide survey reveals multiple nematocyst-specific genes in Myxozoa. BMC Evol. Biol. 18, 138. https://doi.org/10.1186/s12862-018-1253-7.
Stewart, Z.K., Pavalovic, A., Hock, D.H., Prentis, P.J., 2017. Transcriptomic investigation of wound healing and regeneration in the cnidarian Calliactis polypus. Sci. Rep. 7:41458. https://doi.org/10.1038/srep41458.

Stinson, M.E.T., Atkinson, S.D., Bartholomew, J.L., 2018. Widespread distribution of Ceratonova shasta (Cnidaria:Myxosporea) genotypes indicates both evolutionary adaptation to its salmonid fish hosts. J. Parasitol. 104, 645-650. doi:10.1645/18-79.

Stinson, M.E., Bartholomew, J.L., 2012. Predicted redistribution of Ceratomyxa shasta genotypes with salmonid passage in the Deschutes River, Oregon. J. Aquat. Anim. Health. 24, 274-280. https://doi.org/10.1080/08997659.2012.716012.

Tanaka, K., 2009. The proteasome: Overview of structure and functions. P. Jpn. Acad. B-Phys. 85,12-36. https://doi.org/10.2183/pjab.85.12.

Tine, M., Kuhl, H., Gagnaire, P.A., Louro, B., Desmarais, E., Martins, R.S., et al., 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat. Commun. 5:5770. https://doi.org/10.1038/ncomms6770.

Tirloni, L., Reck, J., Terra, R.M.S., Martins, J.R., Mulenga, A., Sherman, N.E., et al., 2014. Proteomic Analysis of Cattle Tick Rhipicephalus (Boophilus) microplus Saliva: A Comparison between Partially and Fully Engorged Females. PLOS ONE 9:e94831. https://doi.org/10.1371/journal.pone.0094831.

Tyndall, J.D.A., Nall, T., Fairlie, D.P., 2005. Proteases Universally Recognize Beta Strands In Their Active Sites. Chem. Rev. 3: 973-1000. https://doi.org/10.1021/cr040669e
Varga, G., Mikala, G., Kiss, K.P., Kosóczki, É., Szabó, E., Meggyesi, N., et al., 2017. Proteasome Subunit Beta Type 1 P11A Polymorphism Is a New Prognostic Marker in Multiple Myeloma. Clin. Lymphoma Myeloma Leuk. 17, 734-742. https://doi.org/10.1016/j.clml.2017.06.034.

Williams, T.H., Spence, B.C., Boughton, D.A., Johnson, R.C., Crozier, L.G., Mantua, N.J., et al., 2016. Viability assessment for Pacific salmon and steelhead listed under the Endangered Species Act: Southwest. U.S. Department of Commerce, NOAA Technical Memorandum. NMFS-SWFSC-564. doi:10.7289/V5/TM-SWFSC-56.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., et al., 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nat. Genet. 46:1212-1219. https://doi.org/10.1038/ng.3098.

Yang, Y., Xiong, J., Zhou, Z., Huo, F., Miao, W., Ran, C., et al., 2014. The genome of the myxosporean *Thelohanellus kitauei* shows adaptations to nutrient acquisition within its fish host. Genome Biol. Evol. 6, 3182-3198. https://doi.org/10.1093/gbe/evu247.

Zardoya, R., Garrido-Pertierra, A., Bautista, J.M., 1995. The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss*. J. Mol. Evol. 41, 942-951.

Zheng, J., Jia, H., Zheng, Y., 2015. Knockout of leucine aminopeptidase in *Toxoplasma gondii* using CRISPR/Cas9. Int. J. Parasitol. 45, 141-148. https://doi.org/10.1016/j.ijpara.2014.09.003.

**Figure Legends**

http://mc.manuscriptcentral.com/gbe
**Fig. 1.** (A) Rainbow trout infected with *Ceratonova shasta* genotype IIR, showing swollen abdomen due to accumulation of ascitic fluid in the visceral cavity. (B) Ascitic fluid rich in *C. shasta* pre-sporogonic, sporogonic and spore stages.

**Fig. 2.** The workflow developed during this study for host contaminant filtration, assembly and annotation. We filtered out host contamination at both the read level before assembly, using read mappings to reference genomes of the parasite *Ceratonova shasta* and host *O. mykiss*, and at the contig level after assembly, using BLASTn against the same reference genomes. We produced two versions of the reference transcriptome: 1) *C. shasta* only, which is the more conservative assembly from only those reads that mapped to the parasite genome; and 2) *C. shasta* + NHP, which was assembled from both the reads that mapped to the *C. shasta* genome and those that mapped to neither host nor parasite (NHP).

**Fig. 3.** Results of first stage filtering: the percentages of reads that best matched our draft *Ceratonova shasta* genome, rainbow trout genome or NHP (neither host nor parasite) for each of the five *C. shasta* genotype transcriptome libraries.

**Fig. 4.** Results of second stage filtering: the percentages and size distributions of assembled contigs that best matched to the *Ceratonova shasta* genome, rainbow trout genome or NHP (neither host nor parasite) for the two versions of the reference transcriptome: (A) *C. shasta* only and (B) *C. shasta* + NHP.

**Fig. 5.** Maximum likelihood phylogenetic trees. (A) Phylogenomic tree showing *Ceratonova shasta* genotypes relationships. Alignment based on 51/79 genes from Chang et al. (2015). *Polypodium hydriforme* was set as the outgroup. (B-C) SNPs-based trees of transcriptomes of *C. shasta* genotypes using different subset of genes: (B) all
genes \((N=22,755)\), \(C\) only genes present across all ten transcriptomes pairwise comparisons \((N=593)\). Values at nodes indicate bootstrap support.

**Fig. 6.**-Relative conservation of motility and protease/inhibitor gene sets based on pairwise genetic difference comparisons and on SNPs-based phylogenetic analyses of *C. shasta* genotypes. \((A)\) Chart shows relative frequency of conserved genes in pairwise comparisons between datasets, subtracting relative frequencies of all conserved genes. Negative values indicate that the subset of genes is less conserved than average. Positive values indicate that the genes are more conserved than average; \((B-C)\) SNP-based ML trees of transcriptomes of *Ceratonova shasta* genotypes using strictly curated datasets of \((B)\) cell migration genes \((N=164)\) and \((C)\) proteases and inhibitor genes \((N=41)\). Values at nodes indicate bootstrap support.

**Table 1.**- Sequencing, filtering and read mapping results from transcriptomes obtained during this study (NHP: neither host nor parasite).

| *C. shasta* genotype | IIR-reference \((RBT6)\) | IIR \((RBTC16)\) | IIR \((RBTJ7)\) | IIC | I |
|----------------------|-------------------------|----------------|----------------|-----|-------|
| SRA Run Acc. Num. Bioproject: PRJNA241036 | SRR6782113 | SRR1575205 | SRR1205836 | SRR1573049 | SRR1461768 |
| Tissue | Ascitic fluid | Intestine |
| Fish Host | Allopatric rainbow trout \((O. mykiss)\) | Sympatric coho salmon \((O. kisutch)\) | Sympatric Chinook salmon \((O. tshawytscha)\) |
| Fish Origin (Stock) | Roaring River Hatchery, OR | Iron Gate Fish Hatchery, CA |
| Location of exposure | Willamette River (WR), OR | Keno Eddy, Upper Klamath River (UKR), CA | Lower Klamath River (LKR), CA | Lower Klamath River (LKR), CA |
| Year | 2015 | 2014 | 2014 | 2014 |
| Illumina Sequencer | HiSeq 3000 | HiSeq 2000 |
| Total number of reads (paired) | 788,875,442 | 15,639,718 | 16,728,880 | 41,920,212 | 44,310,970 |
| Reads after quality filtering | 759,877,916 | 3,027,123 | 6,039,071 | 12,139,060 | 17,729,817 |
| Reads C. shasta origin | 452,498,284 | 1,829,473 | 1,968,164 | 1,552,718 | 787,739 |
| Reads fish host origin | 118,745,658 | 287,874 | 1,949,421 | 6,890,494 | 12,448,482 |
| Reads NHP | 182,816,267 | 900,957 | 2,112,865 | 3,681,236 | 4,487,947 |
| Reads match both C. shasta and host equally well | 5,817,707 | 8,819 | 8,621 | 14,612 | 5,649 |

Table 2.- Assembly and annotation statistics for de novo assemblies of reference transcriptome of IIR (RBT6) C. shasta from ascitic fluid: C. shasta only (more conservative) and C. shasta +NHP (less conservative)

| Host filtered reads into assembly | C. shasta only | C. shasta + NHP |
|----------------------------------|----------------|-----------------|
| **Assembled transcripts**        | 452,498,284    | 452,498,284 + 182,816,267 = 635,314,551 |
| Length min - max                 | 201 – 10,626   | 201 - 18,696    |
| Average length                   | 589            | 550             |
| N50                              | 871            | 810             |
| Size of assembly                 | 26.5 Mb        | 41.3 Mb         |
| Contigs after redundancy         | 18,253         | 28,503          |
| Gene annotation (UniProt)        | 8,460 (18.8%)  | 15,906 (21.2%)  |
| GO annotations (% total UniProt BLAST matches; % total assembly) | 5,229 (61.8%; 11.6%) | 11,219 (70.5%; 14.9%) |
| Taxon ID annotation (nr)         | 15,831 (35.2%) | 29,560 (39.4%)  |
| Taxon ID Cnidaria (% total Taxon ID annotations, % total assembly) | 1,530 (9.7%; 3.4%) | 2,126 (7.2%; 2.8%) |

**Filtration post assembly – transcripts**

| Best match to C. shasta | 43,395 (96.4%) | 48,824 (65%) |
| Neither                | 1,476 (3.3%)   | 14,635 (19.5%) |
| Best match to rainbow trout | 115 (0.3%) | 11,628 (15.5%) |
| Other taxa contamination (microorganisms) | 5,465 (12.1%) | 6,669 (8.9%) |
| **Final version assemblies** (read, contig and other taxa filtration) | 39,407 | 56,876 |
| Length min - max | 201 - 10,626 | 201 - 18,696 |
| Average length     | 569            | 579             |
| N50                 | 825            | 870             |
| Size of assembly    | 22.4 Mb        | 32.9 Mb         |
| Gene annotation (UniProt) | 7,074 (18%) | 10,679 (18.8%) |
| GO annotations (% total UniProt BLAST matches; % total assembly) | 4,525 (64%; 11.5%) | 7,280 (68.2%; 12.8%) |
|---------------------------------------------------------------|---------------------|---------------------|
| Taxon ID annotation (nr)                                      | 10,291 (26.1%)      | 15,503 (27.3%)      |
| Taxon ID Cnidaria (% total Taxon ID annotations, % total assembly) | 1,530 (14.9%; 3.9%) | 2,116 (13.6%; 3.7%) |
De novo Trinity assembly

Mapping reads

Illumina HiSeq 3000 Sequencing (1 full lane)

Quality control processing of raw data

Transcriptome library prep

O. mykiss genome

(Czardoa et al. 1995; Berthelot et al. 2014)

Host filtered C. shasta genome (purified myxospores)

blastn

O. mykiss contigs

discarded

O. mykiss best matching reads

NHP: neither host nor parasite matching reads

C. shasta best matching reads

Host filtering at read level

Host filtering at transcript level

De novo Trinity assembly

C. shasta + NHP assembly

C. shasta only assembly

blastx

Annotations (Uniprot, GO, Taxon ID)

Compare Contaminants

C. shasta + NHP

C. shasta only

C. shasta transcriptomes
A

IIC – LKR

Kudoa iwatai

Ceratonova shasta – IIC – LKR

Ceratonova shasta – I – LKR

Ceratonova shasta – IIR(RBTC16) – UKR

Ceratonova shasta – IIR(RBT7) – LKR

Ceratonova shasta – IIR(RBT6) – WR

Polypodium hydriforme

0.09

B

N=22,755 genes, 918 SNPs

IIR (RBTC16) – UKR

IIR (RBT7) – LKR

IIR (RBT6) – WR

IIC – LKR

Sympatric coho salmon

Sympatric Chinook salmon

Allopatric rainbow trout

0.6

C

N=593 genes, 235 SNPs

IIR (RBTC16) – UKR

IIR (RBT7) – LKR

IIR (RBT6) – WR

IIC – LKR

Sympatric coho salmon

Sympatric Chinook salmon

Allopatric rainbow trout

0.01

http://mc.manuscriptcentral.com/gbe

I – LKR
Allopatric rainbow trout  
Sympatric coho salmon  
Sympatric Chinook salmon

Cell migration genes  
Proteases & inhibitors genes

Genotype comparison

A Cell migration vs all genes  
Proteases/inhibitors vs all genes

IIC/IIR  
I/II  
II/II  
IIC/IIRs

Gene families

Cell migration vs all genes  
Proteases/inhibitors vs all genes

Relative frequency

Genotype comparison

I vs IIC  
I vs IIR (RBT6)  
IIC vs IIR (RBT6)  
IIR vs IIR (RBT6)

B Cell migration genes  
(N=164 genes), 40 SNPs

93  
93  
78  
0.2

Gene families

Cell migration vs all genes  
Proteases/inhibitors vs all genes

Relative frequency

Genotype comparison

I vs IIC  
I vs IIR (RBT6)  
IIC vs IIR (RBT6)  
IIR vs IIR (RBT6)

C Proteases & inhibitors genes  
(N=41 genes), 14 SNPs

63  
63  
29  
0.0009

Gene families

Cell migration vs all genes  
Proteases/inhibitors vs all genes

Relative frequency

Genotype comparison

I vs IIC  
I vs IIR (RBT6)  
IIC vs IIR (RBT6)  
IIR vs IIR (RBT6)