TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

HOST-PARASITE GENOMICS AND ECOLOGY: LINKING GENES AND TRANSCRIPTOMES TO DISEASE AND CONTEMPORARY SELECTION

Freed Ahmad

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AII OSA – TOM. 397 | BIOLOGICA – GEOGRAPHICA – GEOLOGICA | TURKU 2023

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

HOST-PARASITE GENOMICS AND ECOLOGY: LINKING GENES AND TRANSCRIPTOMES TO DISEASE AND CONTEMPORARY SELECTION

Freed Ahmad

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AII OSA – TOM. 397 | BIOLOGICA - GEOGRAPHICA - GEOLOGICA | TURKU 2023

University of Turku

Faculty of Science
Department of Biology
Doctoral programme in Biology, Geography and Geology (BGG)

Supervised by

Professor, Anti Vasemägi
Department of Aquatic Resources (SLU Aqua)
Swedish University of Agricultural Sciences
Sweden
Chair of Aquaculture, Institute of Veterinary Medicine and Animal Sciences, Estonian University of Life Sciences
Estonia

Associate Professor, Paul V. Debes
Department of Aquaculture and Fish Biology
Hólar University, Sauðárkrókur
Iceland

Reviewed by

Professor, Jouni Aspi
Ecology and Genetics Research Unit
University of Oulu
Finland

Professor, Phillip Watts
Department of Biological and Environmental Science
University of Jyväskylä
Finland

Opponent

Lecturer, Jason W. Holland
Fish Health and Aquaculture
School of Biological Sciences
University of Aberdeen
Scotland

ISBN 978-951-29-9232-4 (PRINT)
ISBN 978-951-29-9233-1 (PDF)
ISSN 0082-6979 (Print)
ISSN 2343-3183 (Online)
Painosalama, Turku, Finland 2023

أُطْلُبُوا الْعِلْمَ مِنَ الْمَهْدِ اِلَى اللَّحْد

"Seek knowledge, from the cradle to the grave."

Holy Prophet Muhammad SAW

UNIVERSITY OF TURKU
Faculty of Science
Department of Biology
Subject
FREED AHMAD: Host-parasite genomics and ecology: linking genes and transcriptomes to disease and contemporary selection
Doctoral Dissertation, 118 pp.
Doctoral Programme in Biology, Geography and Geology (BGG)
December 2022

ABSTRACT

Infectious diseases in natural populations are important areas of research in ecology and evolution as they describe how parasites influence the host fitness. A host may undergo adaptive evolution against the parasite by acquiring either resistance or tolerance through developing intricate biochemical and molecular defence strategies. However, the knowledge about the genes associated with these traits remains limited. Furthermore, the strength of ongoing natural selection on transcript abundance has not been studied directly, despite the fact that gene regulation has a major role in adaptive evolution. In this thesis, I applied genomic and transcriptomic approaches to study the host parasite system of anadromous brown trout (*Salmo trutta*) infected with a myxozoan parasite, *Tetracapsuloides bryosalmonae*. In salmonids, *T. bryosalmonae* causes an emerging temperature-dependent disease, proliferative kidney disease (PKD). The parasite infects the kidney and spleen of juvenile fish, and at elevated temperatures, causes strong inflammatory response, anaemia and kidney hypertrophy.

In this thesis, I performed one of the first association mapping attempts on parasite resistance and tolerance in brown trout and demonstrated the possibilities as well as limitations of association analysis in natural populations. As there was very limited genomic information available for *T. bryosalmonae*, I also generated an annotated assembly of the parasite transcriptome. Furthermore, by combining -omic approaches with genetic mark-recapture and classical regression-based selection analysis, I demonstrated the effect of temperature-driven parasite-induced contemporary natural selection on transcript abundance and co-regulated gene networks in this wild vertebrate species.

I identified several promising candidate genes involved in PKD resistance and severity in brown trout. I also characterized more than three thousand transcripts of *T. bryosalmonae*. Among these, I also identified four novel protein drug targets, which can help in curing the infected fish. I also showed that the myxozoan parasite induces massive cell proliferation in the fish host whose variation is associated with the survival selection on the co-regulated gene networks. The directional selection on the individual transcript abundances was weak, similar to the published selection estimates on phenotypic traits. Finally, I also discovered many transcripts exhibiting widespread signal of disruptive selection, related to host immune defence, host–pathogen interactions, cellular repair and maintenance.

Overall, my thesis showcases the power of integrating ecological and genomic perspectives to gain novel insights into the functional genomic basis of resistance against a parasite, health damage (i.e., anaemia) caused by the parasite and the ongoing associated natural selection in the wild. Altogether, my thesis combines multiple levels of biological complexities and represents a significant step forward towards understanding the molecular basis of *T. bryosalmonae* and PKD.

KEYWORDS: Proliferative kidney disease, PKD, host-parasite genomics, transcriptomics, natural selection, resistance, tolerance, parasite load, GWAS, transcriptome, de novo assembly, salmonids, ecology and evolution

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Biologian laitos
FREED AHMAD: Isäntä-loissuhteen genomiikka ja ekologia: geenien ja transkriptomien yhdistäminen sairauksiin ja valintaan
Väitöskirja, 118 s.
Biologian, maantieteen ja geologian tohtoriohjelma (BGG)
Toukokuu 2023

TIIVISTELMÄ

Tarttuvat taudit luonnonpopulaatioissa ovat tärkeitä ekologian ja evoluution tutkimuskohteita koska ne kuvaavat, kuinka loinen vaikuttaa isäntänsä kelpoisuuteen. Isännässä voi kehittyä evolutiivisia adaptaatioita loista vastaan joko resistenssin tai toleranssin kautta. Tällaisten piirteiden voidaan katsoa olevan isännässä kehittyneitä biokemiallisia ja molekyylisiä puolustusstrategioita loisia vastaan. Näihin piirteisiin liittyviä geenejä tunnetaan heikosti. Lisäksi luonnonvalinnan voimakkuutta transkription määrään ei ole tutkittu suoraan, huolimatta siitä, että geenisäätelyllä on tärkeä rooli adaptiivisessa evoluutiossa.

Tässä väitöskirjassa käytin genomisia ja transkriptomisia lähestymistapoja *Tetracapsuloides bryosalmonae* -loistartunnan saaneen anadromisen taimenen (*Salmo trutta*) isäntä-loisjärjestelmän tutkimiseen. Lohikaloissa *T. bryosalmonae* aiheuttaa lämpötilasta riippuvan sairauden, proliferatiivisen munuaissairauden (PKD). Tämä Myxozoa-pääjaksoon kuuluva loinen infektoi nuorien kalojen munuaiset ja pernan, ja aiheuttaa korkeissa lämpötiloissa voimakkaan tulehdusvasteen, anemiaa ja munuaisten liikakasvua. Tein ns. ”association mapping”-analyysin liittyen taimenen resistenssiin ja toleranssiin *T. bryosalmonae*-loista kohtaan, ja osoitin sekä assosiaatioanalyysin käytön mahdollisuudet että rajoitukset tutkittaessa luonnonpopulaatioita. Koska *T. bryosalmonaesta* oli saatavilla hyvin vähän genomista tietoa, loin myös annotoidun koosteen loisen transkriptomista.

Lisäksi, yhdistämällä kolme metodia: ns. ”-omics”-lähestymistavan, geneettisen merkintä-jälleenpyynnin ja klassiseen regressioon perustuvan valinta-analyysin, pystyin osoittamaan lämpötilasta riippuvaisen loisinfektion indusoiman luonnonvalinnan vaikutuksen transkription määrään ja yhteissäädeltyihin geeniverkostoihin tässä luonnossa elävässä selkärankais-lajissa.

Tunnistin useita lupaavia kandidaattigeenejä, jotka liittyvät PKD-resistenssiin ja taudin vaikeusasteeseen taimenessa. Kuvasin myös yli kolmetuhatta *T. bryosalmonaen* transkriptia. Näiden joukosta tunnistin myös neljä uutta proteiinilääke-kohdetta, joiden avulla voidaan parantaa tartunnan saaneita kaloja. Lisäksi osoitin, että tämä Myxozoan-pääjaksoon kuuluva loinen indusoi massiivista solujen lisääntymistä kalaisännässä, ja jonka vaihtelu liittyy selviytymisvalintaan yhteissäädellyissä geeniverkostoissa.

Yksittäisten transkriptien runsauteen kohdistuva suuntaava valinta oli heikkoa, ollen samantasoista kuin kirjallisuudessa esitetyt arviot fenotyyppisiin ominaisuuksiin kohdistuvasta valinnasta. Löysin myös monia transkripteja, joihin kohdistui hajoittavaa valintaa. Nämä liittyvät isännän immuunipuolustukseen, isännän ja patogeenin väliseen vuorovaikutukseen, solujen korjaamiseen ja ylläpitoon.

Väitöskirjani tuo hyvin esille ekologisten ja genomisten lähestymistapojen yhdistämisestä avautuvat uudet tutkimusmahdollisuudet ja näkökulmat loisen vastustuskyvyn genomiseen ja toiminnalliseen perustaan, loisen aiheuttamaan terveyshaittaan (eli anemiaan) ja siihen liittyvään luonnonvalintaan. Väitöskirjani yhdistää monia biologian tasoja ja on merkittävä askel kohti *T. bryosalmonaen* ja PKD:n molekyyliperustan ymmärtämistä.

ASIASANAT: Proliferatiivinen munuaissairaus, PKD, Isäntä-loisgenomiikka, transkriptomiikka, luonnonvalinta, resistenssi, toleranssi, loistaakka, GWAS, transkriptomi, de novo assembly, lohikalat, ekologia ja evoluutio

Table of Contents

Abbreviations

3D	Three dimensional
DAPC	Discriminant analysis of principal components
ddRAD seq	double digest restriction site-associated DNA (or dRAD) sequencing
DE	Differentially expressed or differentially expression
DNA	Deoxyribonucleic acid
BP	Base pairs
BUSCO	Benchmarking universal single-copy orthologs
FDR	False dicovery rate
GBS	Genotyping by sequencing
h^2	Heritability
HWE	Hardy–Weinberg equilibrium
KAAS	KEGG automatic annotation server
Kb	Kilo base pairs
KO	KEGG orthology
LD	Linkage disequilibrium
MAF	Minor allele frequency
MDS	Multi-dimensional scaling
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
PCR	Polymerase chain reaction
PE	Paired-end
PL	Parasite load (or RPL)
PPI	Protein-protein interaction
qPCR	Quantitative polymerase chain reaction
r	Pearson's correlation coefficients
r^2	Squared allele-frequency correlation
RAD seq	Restriction-site associated DNA sequencing
RF	Random Forest
RNA	Ribonucleic acid
RPL	Relative parasite load (or PL)
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction

s	Linear selection differential
SE	Single-end
SPRI	solid-phase reversible immobilization
Ss4R	Salmonid-specific fourth round (genome duplication)

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Ahmad, F., Debes, P. V., Palomar, G. and Vasemägi, A. Association mapping reveals candidate loci for resistance and anaemic response to an emerging temperature-driven parasitic disease in a wild salmonid fish. *Molecular Ecology*, 2018; 27(6): 1385-1401.

II Ahmad, F., Debes, P.V., Pukk, L., Kahar, S., Hartikainen, H., Gross, R. and Vasemägi, A. Know your enemy – transcriptome of myxozoan *Tetracapsuloides bryosalmonae* reveals potential drug targets against proliferative kidney disease in salmonids. *Parasitology*, 2021; 148(6): 726-739.

III Ahmad, F., Debes, P.V., Nousiainen, I., Kahar, S., Pukk, L., Gross, R., Ozerov, M. and Vasemägi, A. The strength and form of natural selection on transcript abundance in the wild. *Molecular Ecology*, 2021; 30(12): 2724-2737.

Author contributions to the original publications:

	I	II	III
Original idea	AV	**FA**, AV	AV
Data (sample) collection	PVD, AV	**FA**, AV, PVD, SK, LP	**FA**, AV, PVD, SK, LP
Lab work	**FA**, GP, PVD	PVD, AV	PVD, IN
Data analyses	**FA**, PVD, AV	**FA**	**FA**, AV, PVD, MO
Writing of the manuscript	**FA**, AV, PVD	**FA**, AV	AV, **FA**, PVD

FA= Freed Ahmad, AV= Anti Vasemägi, PVD= Paul Vincent Debes, GP= Gemma Palomar, IN= Ilkka Nousiainen, SK= Siim Kahar, LP= Lilian Pukk, RG= Riho Gross, MO= Mikhail Ozerov, HH= Hanna Hartikainen.

1 Introduction

1.1 Parasitism and evolution

Parasitism is a form of symbiotic association between two different species, which is beneficial to one organism (parasite) but harmful to the other (host). Parasites can be either endoparasites (e.g., malaria-causing *Plasmodium* spp.) or ectoparasites (e.g., lice *Pediculus* spp.), depending upon whether they live inside the host body or only anchor to the body surface. For a host, resource scavenging or severe tissue or even organ damage caused by the parasite are mostly detrimental, leading to compromise health and fitness of the host (Råberg 2014). This impairment might result in the development of typical clinical symptoms of a disease (i.e., pathogenicity) in the host. Thus, all pathogens (mostly microscopic bacteria and viruses) are parasites, but the converse is not always true. Some infections e.g., tuberculosis and malaria in humans, can even cost life to their hosts depending on a virulence of the parasite. The parasite virulence generally depends upon the life cycle of the parasite. For instance, some parasites are spread with the dead host while others tend to find a balance between within-host exploitation and transmission to other hosts to maximize their fitness (Gandon et al. 2001).

Over the course of evolution, both host and parasite may co-evolve and constantly acquire new tactics to overcome each other. For such "arms race" or antagonistic interaction, Van Valen (1973), while paying homage to Lewis Carroll, used the term "Red Queen's hypothesis". According to this hypothesis, competing species must change continuously to be in the same place, i.e., average fitness remains constant (Van Valen 1973). In order to limit the impact of the parasite, hosts may evolve to develop physiological barriers to evade the infection or develop complex biochemical and molecular strategies, such as resistance or tolerance. Resistance is broadly defined as the ability of the host to control the parasite burden (or pathogen load), whereas tolerance is an ability of the host to reduce harm caused by the given parasite amount (Råberg et al. 2007, Vander Wal et al. 2014). In a relatively simple evolutionary scenario, the fitness of a host is likely to be increased by acquiring resistance at an expense of the parasite. However, tolerance is not expected to have any drastic negative fitness-related consequences for the parasite (Kutzer and Armitage 2016, Little et al. 2010). Thus, resistance and tolerance are

likely to have different impact on the coevolution of host and parasite. Furthermore, the environment may also play an important role in modifying resistance and tolerance (Debes et al. 2017, Marcogliese 2008) as the fitness benefits and costs associated with these traits may change across different environmental settings (Råberg et al. 2009).

Parasites, on the other hand, have evolved strategies to evade the physiological barriers and immune defenses of hosts. Parasites are adapted to rely upon hosts, either fully or partially, for the nutrient acquisition, development, reproduction and dispersal. Such a reliance on hosts has often brought an overall simplification in body plan, and in special cases, a reduction in genome size (Poulin and Randhawa 2015). Parasites are considered as strong drivers of evolutionary change posing a serious threat to the survival of the hosts, while their number arguably, exceeds the number of free-living organisms (Eizaguirre and Lenz 2010).

1.2 Natural selection on quantitative phenotypes

Natural selection is a simple yet the most significant idea in evolutionary biology (Futuyma 2009). It is the primary mechanism by which biological diversity is produced through population divergence and speciation. Natural selection operates on the phenotypes ("targets of selection") of the individuals facing environmental challenges ("agents of selection" or "selective agents") within generations and brings about a non-random survival or reproductive success of individuals having different phenotypes (Kingsolver and Pfennig 2007). The agents of selection can be either biological (e.g., competitors, predators or parasites) or non-biological (e.g., weather, fire or drought). In order to cause evolutionary change in a population, i.e., selection responses have effects across generations, the phenotype under selection must possess heritable variation among the selected individuals with differential reproduction success. The differential reproduction success means that favoured individuals have more offspring, i.e., a higher fitness (Kingsolver and Pfennig 2007).

Three modes of selection are conceivable (Figure 1), when natural selection acts on quantitative traits (summarized in Kingsolver and Pfennig 2007). All three modes cause evolution by, more likely, removing individuals with lower fitness and favouring individuals with higher fitness. Briefly, in the case of directional selection (i.e., a linear fitness function) the fitness steadily increases (or decreases) with the value of the quantitative phenotype. For the heritable traits under positive directional selection, the population will evolve towards more extreme trait values of a certain direction, (i.e., either higher or lower). For example, Quinn et al. 2007, analysed the migration data (1969-2003) from two populations of sockeye salmon in Alaska and found that average migration date in both populations became earlier and was undergoing directional selection. In the case of stabilizing selection, the population

will evolve towards intermediate trait values. Finally, in the disruptive selection mode, the population will evolve towards extreme trait values in both directions, i.e., towards both higher and lower values.

Both stabilizing and disruptive selection can be described with nonlinear fitness function (Figure 1). The population under stabilizing selection will evolve to a shorter range of phenotype values (smaller phenotypic variance). On the other hand, the population under disruptive selection will evolve to a wider range of trait values (larger phenotypic variance) (Kingsolver and Pfennig 2007). A classic example of stabilizing selection on human birth weight, where intermediate birth weight is associated with increased survival in both males and females (Karn et al. 1951). On the other hand, bill size variation in black-bellied seedcracker (*Pyrenestes ostrinus*), is a textbook example of disruptive selection. Birds if the small morphs prefer soft seeds whereas large morphs eat hard-seeds. Whereas birds with intermediate phenotype are less efficient in eating both type of seeds (Smith 1993).

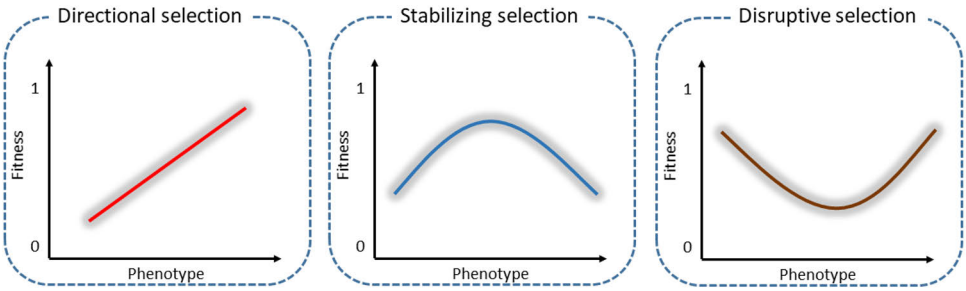

Figure 1. Three modes of natural selection defined through to the shape of the fitness function that illustrate the relationship between phenotype and the fitness (adapted from Kingsolver and Pfennig 2007).

As stated earlier, natural selection acts on the phenotypes that are, if heritable, encoded in the genotype. During the 1960s, population geneticists presumed that natural selection would have resulted in an optimal genotype for a specific environment and the disappearance of genetic variation (Bell et al. 2021). A lack of genetic variation must then reduce mean fitness of the population as it loses the ability to evolve to further environmental changes. However, this assumption was contradictory to the high genetic variation reported in several studies of that time (see Bell et al. 2021). To explain the presence of such genetic variability, Bruce Wallace introduced the concept of soft and hard selection (Wallace 1975). According to him, soft selection happens when the absolute fitness of an individual relies on the phenotypic composition of the population. Soft selection results from both density-dependent selection and frequency-dependent selection. On the other hand, hard

selection is dependent on the compatibility between the absolute trait value of an individual and its environment and does not depend upon the phenotypic makeup of the population (Wallace 1975). Furthermore, hard selection causes high mortality (more than background mortality) in order to remove the genotypes of less favoured individuals (Bell et al. 2021).

1.3 Proliferative kidney disease of salmonids

For this thesis, I studied a host-parasite system, which consists of the anadromous brown trout (*Salmo trutta*) as a host and its histozoic malacosporean parasite *Tetracapsuloides bryosalmonae*. *S. trutta* belongs to the genus *Salmo* of the family Salmonidae (commonly referred to as salmonids). In salmonids, the infection caused by *T. bryosalmonae* may result in proliferative kidney disease (PKD). PKD is an emerging and serious temperature-dependent disease of salmonids and can cause significant damage to both aquaculture and wild populations (Okamura et al. 2001). The main symptoms of PKD in infected fish are renal hyperplasia (kidney swollenness) and anaemia (Bettge et al. 2009, Clifton-Hadley et al. 1987). In the past two decades, *T. bryosalmonae* has been found responsible for mass mortalities in both freshwater fish farms and in wild salmonid populations (Feist 2002, Hedrick et al. 1993, Robbins 2016, Sterud et al. 2007, Wahli et al. 2002). Salmonid species affected by the parasite (both wild and in aquaculture) are given in the Table 1.

Table 1. Salmonids affected by the myxozoan parasite *T. bryosalmonae*

SUBFAMILY	GENUS	SPECIES (COMMON NAME)	REFERENCE
SALMONINAE	Salmo	*Salmo trutta* (brown trout)	Seagrave et al. (1981)
		S. salar (Atlantic salmon)	Ellis et al. (1982)
	Salvelinus	*Salvelinus fontinalis* (brook trout)	Plehn (1924)
		S. alpinus (Arctic charr)	Brown et al. (1991)
	Oncorhynchus	*Oncorhynchus mykiss* (rainbow trout)	Plehn (1924)
		O. tshawytscha (Chinook Salmon)	Hedrick et al. (1984)
		O. keta (chum salmon)	Gorgoglione et al. (2020)
		O. kisutch (coho salmon)	Hedrick et al. (1984)
		O. gorbuscha (pink salmon)	Braden et al. (2010)
		O. nerka (sockeye salmon)	Arkush and Hedrick (1990)
		O. clarkii (cutthroat trout)	Macconnell and Peterson (1992)
COREGONINAE	Coregonus	*Coregonus lavaretus* (European whitefish)	Sobociński et al. (2018)
	Prosopium	*Prosopium williamsoni* (mountain whitefish)	Hutchins et al. (2021)
THYMALLINAE	Thymallus	*Thymallus thymallus* (European grayling)	Clifton-Hadley et al. (1984)

T. bryosalmonae belongs to the minute subclass Malacosporea of Myxozoa (Giribet and Edgecombe 2020), has a complex life cycle (two stages) with freshwater bryozoans as the primary and some of the salmonid fishes as the secondary hosts (Figure 2, Okamura et al. 2011). The spores released from the bryozoans are called malacospores, while the spores released from fish are called fish-malacospores. *T. bryosalmonae* malacospores are mass-released from the bryozoans during the period of spring to early summer. They infect fish (juveniles) through gills and/or skin (Longshaw et al. 2002) and use the host vascular system to eventually move to internal organs (e.g., kidney and spleen). The sporogony (formation of spores) of *T. bryosalmonae* takes place in the lumen of the host kidney tubules where it transforms into a pseudoplasmodium. Using the pseudopodial extensions, the parasite remains attached with the epithelial cells of tubules. A fish-malacospore is matured inside each pseudoplasmodia and is comprised of two polar capsules, four valve cells, and one sporoplasm (Morris and Adams 2008). The bryozoans-infective fish-malacospores are finally released into the water via fish urine (Hedrick et al. 2004).

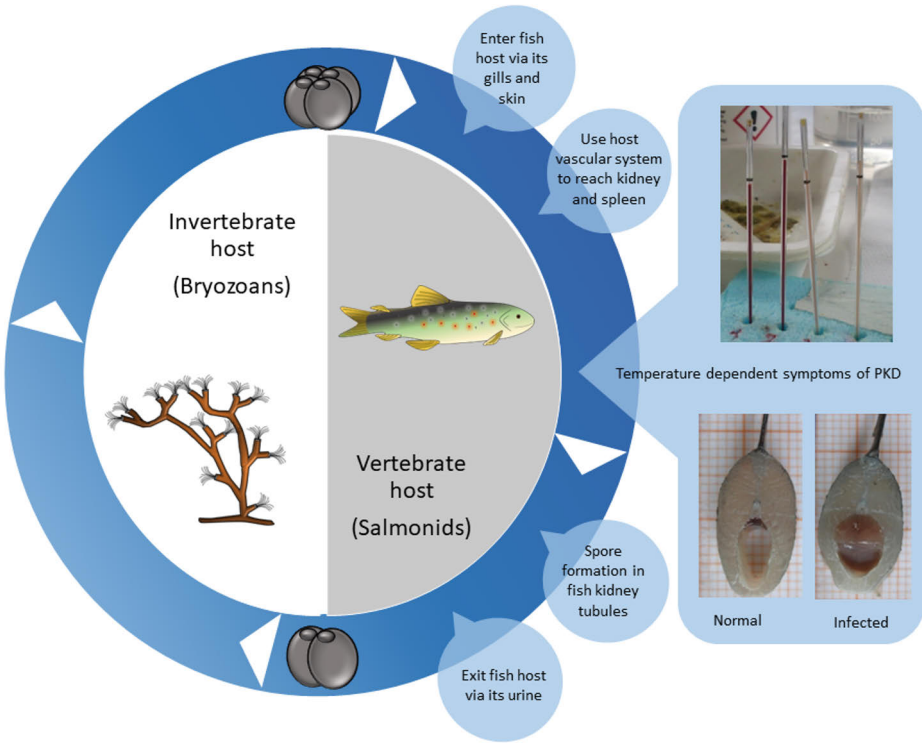

Figure 2. Simplified illustration of the life cycle of *Tetracapsuloides bryosalmonae* and its effect of brown trout.

T. bryosalmonae is widespread in Europe and North America (Hedrick et al. 1993, Dash and Vasemägi 2014, Debes et al. 2017, Mo and Jørgensen 2016, Okamura et al. 2011, Skovgaard and Buchmann 2012, Vasemägi et al. 2017, Gorgoglione et al. 2020). Interestingly, *T. bryosalmonae* exists as two different strains, each adapted to the native species in their respective continent. The North American strain forms spores in rainbow trout (Kent and Hedrick 1986, Hedrick et al. 2004) but the European strain does not produce mature spores in rainbow trout (Bucke et al. 1991). Thus, it has been suggested that the European strain is likely adapted to the genus *Salmo* and the North American to the genus *Oncorhynchus* (Bucke et al. 1991, Henderson and Okamura 2004, Morris et al. 1997). However, Arctic charr (genus *Salvelinus*; occurring in both Europe and North America) infected with the European strain, have also exhibited sporogonic stages of the parasite (Svavarsdóttir et al. 2021) indicating that more experimental work on wider range of species is needed to further explore this hypothesis. Nevertheless, rainbow trout infected with the European strain cannot pass the infection to bryozoan and acts as dead-end host. On the other hand, the infected brown trout can pass the infection to bryozoans and is a natural host for the European strain (Grabner and El-Matbouli 2008, Kumar et al. 2013).

The wild brown trout and *T. bryosalmonae* represent an appropriate system for studying the genetic basis of resistance and tolerance and contemporary natural selection on host gene expression. *T. bryosalmonae* not only shows a high prevalence (Hedrick et al. 1993) but also imposes a strong temperature-dependent effect on host physiology, performance (Bruneaux et al. 2016) and survival (Hedrick et al. 1993). Many methodological challenges associated with host-parasite systems (e.g., variation in host age, infection onset, co-infection or exposure avoidance (Bishop et al. 2012, Doeschl-Wilson et al. 2012, Graham et al. 2010), are minimal or absent in this system. Furthermore, like other salmonids, brown trout is a highly fecund organism, represent a promising model to further understand the role of genetic variation on contemporary evolutionary processes that are not only linked to host–parasite coevolution but, through the association between water temperature and disease severity, also to current climate change. Furthermore, *T. bryosalmonae* is expected to expand its geographic distribution and increase in virulence with the increasing temperatures (Tops et al. 2006, but see Lauringson et al. 2022). Under global warming with extreme events, PKD might cause hard selection on wild brown trout populations via high mortality, which otherwise are mostly exposed to soft selection. Thus, information gained through this host-parasite system is crucial to understand the complex interplay of host-parasite co-evolution, which also aids in understanding effects from current climate change.

1.3.1 Available genomic resources of both host and parasite

As genomic studies in general, also this thesis relied on the genomic resources and was confronted with specific methodological challenges related to the studied species. Specifically, no reference genome was available for either host or parasite at the time of the analysis of the three thesis chapters. In fact, among the salmonids, a high-quality (chromosome level) genome was only available of *Salmo salar,* a genetically close sister species to brown trout. In addition, salmonids diverged from their ancestors by going through a salmonid-specific fourth round (Ss4R) genome duplication event that occurred ~60–80 million years ago (Macqueen and Johnston 2014) and the analysis of the *S. salar* genome revealed that about 10% of the genome is still in a tetraploid state (Allendorf et al. 2015, Lien et al. 2016, Limborg et al. 2016), which complicates analyses relative to a diploid genome. This tetraploidy causes residual tetrasomic inheritance at the distal ends of some duplicated (homeologous) chromosomes (Limborg et al. 2016). Furthermore, the gene paralogues resulting from the genome duplication still need functional characterization.

Although a high economic and ecological significance of PKD has been recognised decades ago, the genomic resources for *T. bryosalmonae* remained limited to a few hundred sequences (e.g., Saulnier et al. 1996, Holland et al. 2011, Hartikainen et al. 2014, Hartikainen 2015, Carraro et al. 2018). Only recently, Faber et al. (2021) reported an intersect transcriptome of *T. bryosalmonae* consisting of *de novo* assembled transcripts common in both the farmed rainbow trout (*Oncorhynchus mykiss*) and a bryozoan host (*Fredericella sultana*). However, as the dead-end host *O. mykiss* was used to recover the *T. bryosalmonae* transcriptome, it is likely that it might not contain all the gene transcripts related to parasite sporogony, which is achieved only in the natural vertebrate host.

1.4 Genomics to host-parasite ecology

1.4.1 Genotype-phenotype association in the wild

Genome-wide association studies (GWAS) are performed to detect the associations of genotypes with a specific phenotype or disease. In GWAS, a large number of genetic variants (usually several hundred thousand), typically single nucleotide polymorphisms (SNPs), across many individual genomes (several hundreds to thousands) are tested to find those statistically associated with a phenotype (Uffelmann et al. 2021). Over recent years, GWAS have been increasingly used to detect allelic variants and corresponding genes that contribute towards resistance against infectious diseases in humans, livestock animals and commercially important

crops (Chapman and Hill 2012, Bishop and Woolliams 2014, Poland et al. 2011, Richards et al. 2017, Yu et al. 2011). Infectious diseases and related traits, such as resistance and tolerance in natural populations, are important areas of research in ecology and evolution as they describe how parasites influence the host fitness. However, association studies of infectious diseases in the wild populations have just started to appear in the past decade (Armstrong et al. 2018, Santure and Garant 2018). This advancement has become possible only after the development of cost-effective alternatives of DNA sequencing where, instead of the whole genome, only a fraction of it is sequenced.

By using the reduced representation genome sequencing techniques, such as restriction site-associated DNA (RAD) sequencing (Miller et al. 2007, Baird et al. 2008) or its variant double digest RAD (ddRAD or dRAD) sequencing (Peterson et al. 2012), it is possible to determine the genotype of thousands of SNPs without sequencing the entire genome. In RAD sequencing, genomic DNA is first digested with a restriction enzyme of choice and then the DNA fragments with overhanging ends are selected for sequencing through adapter ligation, pooling, shearing, size selection and PCR amplification. The use of barcodes (few base pairs) in the adapter sequence allows the pooling and sequencing of multiple samples in one lane (Baird et al. 2008) making this approach highly cost effective. In ddRAD sequencing, two restriction enzymes are used to digest the genomic DNA and only those fragments are selected for sequencing which have flanking sites for both enzymes (Peterson et al. 2012). The sequencing reads generated from the RAD sequencing can easily be used for the marker discovery using different bioinformatics tools. If the genome of the species of interest is available, then the reads are mapped on it and SNPs are called. In the absence of a reference genome, RAD loci are *de novo* assembled and reads are subsequently mapped to call the SNPs. Thus, the double RAD sequencing is a promising and cost-effective solution to genotype host samples from natural population and identify SNPs potentially associated with PKD resistance and tolerance traits.

1.4.2 Parasite transcriptome assembly

In order to develop the genomic resources for previously unexplored non-model yet ecologically important species, aiming at the transcriptome seems an intuitive approach for now. This is because, compared to the whole genome, the RNA sequencing costs are lower and downstream analysis is less computer intensive. Furthermore, in a dynamic host-parasite system, transcriptome sequencing can provide a clear snapshot of active genes of one species in response to the defences of the other. However, in case of a microscopic parasite like *T. bryosalmonae,* which uses multiple hosts to complete its lifecycle, isolation of parasite RNA is

very challenging. In such situations, dual-RNA sequencing can be particularly useful.

In dual RNA-sequencing, the RNA to be sequenced originates from both host and parasite i.e., infected tissue, and allows simultaneous transcriptome quantification for both species. However, a major challenge for the *de novo* transcriptome assembly of the parasite using dual-RNA sequencing data is the accurate *in silico* removal of host RNA (Schulze et al. 2016, Videvall et al. 2017). Recently, Alama-Bermejo et al. (2020) applied a two-step filtering approach to successfully remove the host RNA from the myxozoan *Ceratonova shasta* transcriptome. In the first step, they excluded the host reads by aligning reads to host and parasite reference genomes. During the second step (transcript-level filtering), Alama-Bermejo et al. (2020) searched the assembled transcripts (using BLASTN) against both host and parasite genomes to further exclude the potential host transcripts. They extended this search to exclude contaminant sequences (of bacterial, fungal or viral origins) using BLASTX (Alama-Bermejo et al. 2020). Thus, dual-RNA sequencing, coupled with two-step filtering, can facilitate accurate *in silico* elimination of host RNA sequences.

1.4.3 Selection on host transcriptome

Even though many studies indirectly deduce the contribution of different evolutionary forces in influencing gene transcription (Fraser et al. 2010, Gilad et al. 2006), our knowledge about how contemporary natural selection affects transcript abundance in the wild is very limited (Miller et al. 2011). This lack of knowledge is noteworthy as variation in transcript abundance is of fundamental importance to evolution (King and Wilson 1975, Emilsson et al. 2008, Fraser 2013, Fraser et al. 2010, Gilad et al. 2006, Miller et al. 2011, Price et al. 2022), linking molecular functions to phenotypes and thus to performance and fitness. Traditionally, strength and mode of selection is quantified using linear and quadratic regression between the fitness measure and standardised trait values. The regression of individual traits with fitness results in selection differentials (both linear and quadratic), whereas the multiple regression of all traits with fitness as predictor yields selection gradients (Lande and Arnold 1983). Thus, selection can be studied by combining traditional analyses of selection differentials and gradients with the high-throughput screening of ´molecular´ phenotypes at the mRNA level. Such use of intermediate molecular phenotypes has been greatly fruitful in medicine for discovering the mechanisms underlying complex diseases (e.g., Cheung and Spielman 2009). Thus, by integrating -omic approaches with genetic mark-recapture and classical regression-based selection analysis, the effect of parasite-induced and temperature-driven contemporary natural selection on

transcript abundance and co-regulated gene networks in a wild vertebrate host can be assessed.

Studying natural selection on the host transcriptome in the wild is challenging because regular RNA sequencing is costly for hundreds of individuals. In order to quantify transcript abundance in many individuals, I opted for a cost effective Lexogen QuantSeq 3′ mRNA sequencing approach. QuantSeq is an efficient protocol for generating strand-specific NGS libraries near the 3′ end of polyadenylated RNAs. Only one fragment per transcript is generated, directly connecting the number of reads mapping to a gene to its expression (Moll et al. 2014). Thus, compared to standard RNA sequencing, the QuantSeq approach can provide accurate estimates of transcript abundances with a fraction of the sequencing effort (Lohman et al. 2016).

1.5 Aims of the thesis

A better understanding of the genetic architecture conferring resistance (to parasites) and tolerance (to the disease parasites may cause) is essential to predict evolutionary dynamics and ecological interactions between hosts and parasites. Several quantitative genetic studies (both in wild and aquaculture settings) have identified moderate-to-high heritability for resistance to different infectious diseases in fishes, (Debes et al. 2017, Mazé-Guilmo et al. 2014, Ødegård et al. 2011), whereas obtaining precise heritability estimates for tolerance have remained a challenge due to the extremely large sample sizes required (Kause 2011, Debes et al. 2017). However, the underlying causative genetic variants controlling resistance, or tolerance, are less understood (Campbell et al. 2014, Moen et al. 2015). Additionally, detailed knowledge of the *T. bryosalmonae* transcriptome and gene expression may significantly contribute to decipher the underlying mechanisms of the parasite life cycle, host–parasite interactions, parasite virulence and, eventually, identify potential new drug targets (Cowell and Winzeler 2019). Finally, the variation in transcript abundance, like any other morphological or physiological traits, is assumed to be caused by natural selection through whole-animal performance (Arnold 1983). Therefore, it is valid to ask which transcribed genes and pathways are "visible" to selection in the studied host–parasite system and associate with survival.

This thesis applies genomic and transcriptomic sequencing to address several eco-evolutionary aspects between *S. trutta* and *T. bryosalmonae*, and thus of PKD. The overarching aims of this doctoral dissertation are:

1. To identify the genes and genomic regions responsible for parasite resistance and other PKD-related traits in brown trout.

2. To *de novo* assemble and subsequently characterize the transcriptome of *T. bryosalmonae* and, to identify potential drug targets against PKD.
3. To evaluate the occurrence and strength of contemporary natural selection at the transcriptome level and identify pathways, rather than individual genes, for generating new hypotheses related to the role of gene expressiong in host-parasite co-evolution.

2 Materials and Methods

2.1 Study area, field sampling and phenotyping

In Estonia, the PKD-causing parasite *T. bryosalmonae* is widespread among wild brown trout populations and many Baltic Sea coastal rivers have shown high prevalence of this parasite (Dash and Vasemägi 2014). For this thesis, parasite-infected juvenile *S. trutta* samples were collected from two such rivers, namely Altja and Mustoja (Table 2, Figure 3a-c). Both Altja (length = 17.5 km; catchment area = 46.1 km²) and Mustoja (length = 28 km; catchment area = 135 km²) are small rivers (mostly less than 1 m deep) and support a limited number of wild trout spawners. Both rivers support naturally reproducing anadromous *S. trutta* ("sea trout") populations that assume feeding migrations in the proximate Baltic Sea. In these rivers, *S. trutta* juveniles and small males are present throughout the year but large adults only enter for spawning in autumn. Both rivers have dams upstream (few kilometres) from the sampling locations, which inhibit entree to upstream areas. The presence of downstream beaver dams in the river Altja might also cause temporary blockage to returning spawners (Debes et al. 2017).

Table 2. Brown trout juvenile samples used in the thesis.

CHAPTER	RIVER	SAMPLING YEAR	NO OF SAMPLES	NGS METHOD	TISSUE FOR NGS
I	Altja	2014	275(255*)	Double RAD (SE)	Kidneys
II	Mustoja	2014	8	RNA (PE)	Kidneys
II	Preedi	2015	3	RNA (PE)	Kidneys
II	Vodja	2015	3	RNA (PE)	Kidneys
III	Altja	2015	14	RNA (PE)	Pelvic fins (7 July, 7 August)
III	Altja	2015	238	QuantSeq 3′mRNA (SE)	Pelvic fins

* Actual number of samples analysed for the association testing as some samples were excluded because of low throughput or missing phenotypes.

In September 2014, several hundred brown trout juveniles were sampled from two small areas of these rivers (**I**-**II**) during an extensive electrofishing procedure (Debes et al. 2017). The fish were euthanized by an overdose of buffered MS-222 (Sigma-Aldrich, St. Louis, MO), and their fork length (FL, in mm) and wet body mass (in g) were measured. Haematocrit (Hct; the ratio of red blood cell volume to total blood volume) was measured from the whole blood taken via tail ablation. Kidney swollenness (KS) was estimated from the photographs of fish cross sections. The DNA from the kidney of each fish was extracted using the salt extraction method (Aljanabi and Martinez 1997) and was later used for microsatellite-based genotyping (Debes et al. 2017) and double RAD sequencing (**I**, only from Altja). As resistance can be defined as the inverse of the amount of parasite on the affected host, the extracted DNA was also used for the detection and quantification of parasite abundance in the host kidney using quantitative PCR (qPCR) (detailed description in Debes et al. 2017). Briefly, we used qPCR on two DNA targets: a *T. bryosalmonae* 18S ribosomal DNA fragment (166-bp, GenBank accession U70623) and a conserved *S. salar* prefoldin subunit 6 (74-bp, GenBank accession BT049744.1) nuclear fragment. This quantification allowed standardization of the parasite DNA to the amount of salmonid DNA in the assay, and provided relative parasite load (RPL or PL (**II** and **III**)). Furthermore, the utilised *T. bryosalmonae* DNA fragment is from a multi-copy gene, whereas the salmonid target is a single-copy gene, which beneficially increases the assay sensitivity for the microscopic parasite. The qPCR method has been shown to be superior to the immunohistochemical method at low and high parasite numbers, at which the latter approach suffers from spatial heterogeneity and saturation effects (Bettge et al. 2009). Three technical replicates were run for both targets in each sample. FL and other PKD traits (Hct, KS and RPL) were significantly correlated with each other as shown in Figure 4 (Debes et al. 2017). For **Chapter I**, ~85% of the fish sampled from the river Altja were selected for single-end (SE) double RAD sequencing (Figure 5).

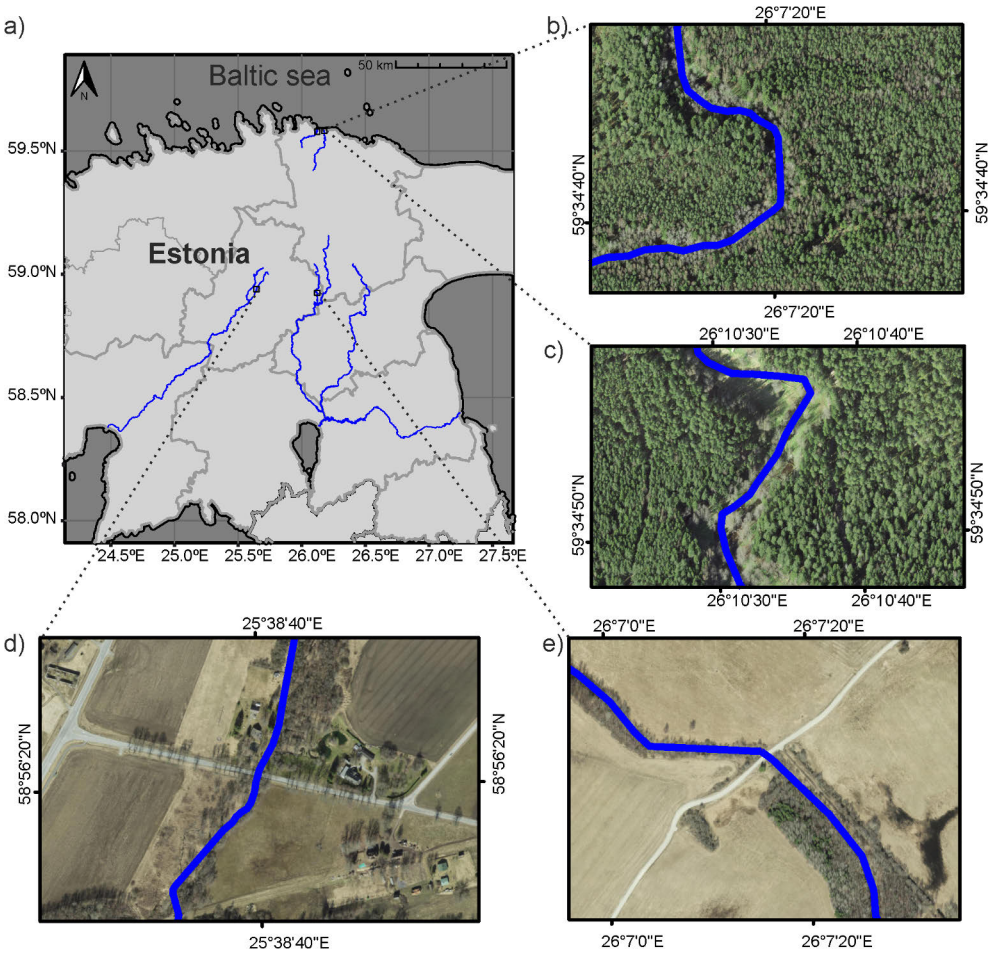

Figure 3. Location of the four Estonian rivers (a) namely, Altja (b), Mustoja (c), Vodja (d) and Preedi (e).

For **Chapter II**, eight infected samples from river Mustoja with moderate (n = 4) to severe (n = 4) parasite load were selected for paired-end (PE) RNA sequencing. The moderately infected fish also displayed lower kidney swelling as compared to severely infected fish (see supplementary figure 1 in **Chapter II**). The sampling of uninfected individuals (healthy controls in **II**) from the same river area was not conceivable because of the 100% prevalence of *T. bryosalmonae* (Dash and Vasemägi 2014, Debes et al. 2017). Therefore, six uninfected fish collected from two parasite-free Estonian rivers (Figure 3d & e), namely, Preedi (n = 3) and Vodja (n = 3) in September 2015 were also selected for PE RNA sequencing (**II**) (Figure 5)

For **Chapter III**, the same sampling area of the river Altja, divided into five sub areas, was electrofished on 30th of August 2015 and 278 brown trout juveniles were captured. This time only fork length (FL) and body mass of the live-captured fish were measured. Most importantly, small biopsies of the right pelvic fins (fin cut) were collected and instantly stored in liquid nitrogen at −80°C. These biopsies were used for the genetic mark–recapture analysis and 3′ RNA sequencing. The juveniles were released back into the original areas of capture after the fin clipping. Almost one month later (in September), we electrofished a total of 685 fish from the same 330-m stretch of initial capture and further up- and downstream areas, and also collected small biopsies (kidney) from every fish. The recaptured fish were identified based on having a fin cut, and matching genotypes of fish caught in August and September (see below). We extensively sampled the initial five capture areas thrice to allow for calculating the capture probability and total number of fish in the system using the depletion method (Zippin 1958), utilizing the fsa R-package (Ogle 2017). Out of 685 total fish captured in September, we euthanized 363 fish by an overdose of buffered MS-222 and measured the required phenotypes (e.g., FL, Hct, KS and PL), extracted DNA and called microsatellite-based genotypes (both fins and kidneys) as described earlier (Debes et al. 2017).

We identified the recaptured individuals by matching microsatellite genotypes between the fin (August) and the kidney tissue (September). These genetically recaptured samples were labelled as “observed” survivors in **III.** Interestingly, we found that observed survival was not dependent on fish size (FL) (Wilcoxon's rank sum test, p = .190; Welch's two-sample t test, n = 278, p = .216). Total RNA was extracted from the infected (from Mustoja) and non-infected kidney samples (**II**) and the pelvic fins (**III**) using NucleoSpin® RNA kit. For **Chapter III**, ~86% fish fins (observed survivors = 114, non-survivors = 124) collected in August 2015 were selected for Quantseq 3’ RNA sequencing (Figure 5). Additionally, two fin-clip mRNA pools both consisting of seven individuals were also subjected to conventional Illumina mRNA paired-end sequencing along with **Chapter II** samples.

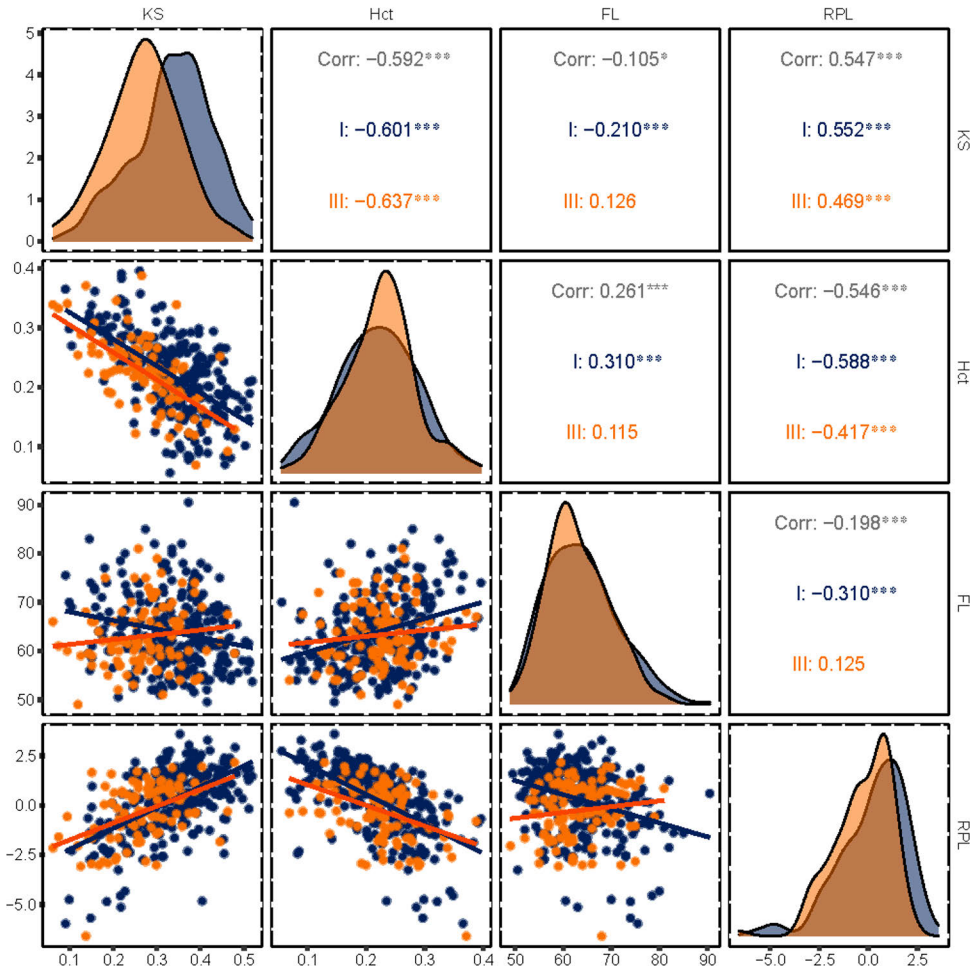

Figure 4. Distributions and correlations between the phenotypes (RPL, KS, Hct and FL) of the samples captured from the river Altja in 2014 (**I**; deep blue) and among the 2015 survivors (**III**; orange).

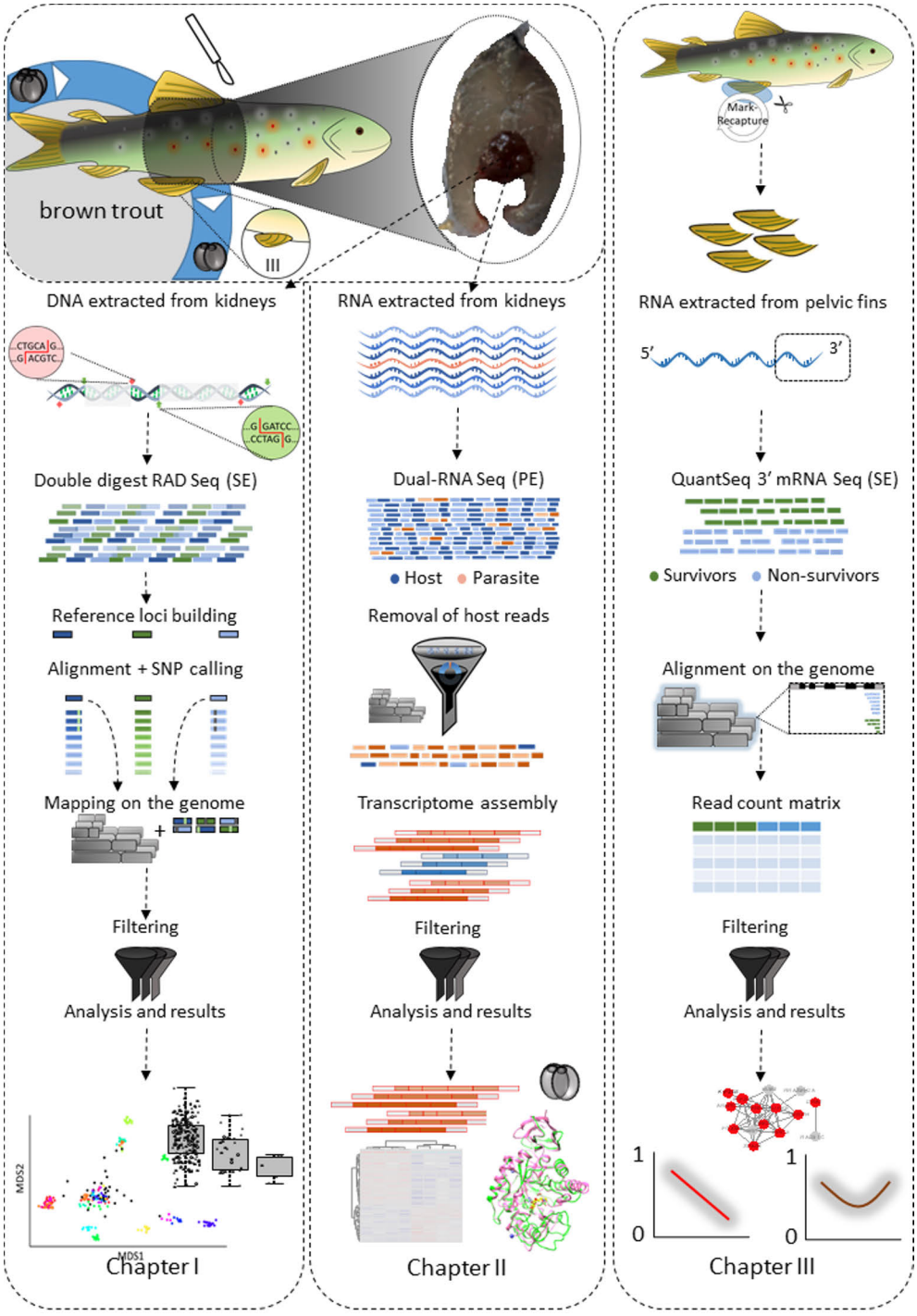

Figure 5. Graphical summary of all thesis chapters (**I-III**).

2.2 NGS library preparation and sequencing

The details of library preparation and sequencing can be found in the respective chapters. For all three chapters, the short read sequencing was performed at the Finnish Functional Genomics Centre, Turku Centre for Biotechnology, Turku, Finland. I deposited the raw sequence reads to NCBI SRA under the BioProject IDs PRJNA429104 (**I**), PRJNA668017 (**II**) and PRJNA517427 (**III**).

For **Chapter I,** three double-digest restriction-site associated DNA (ddRAD or dRAD) sequencing libraries (93 brown trout samples in each) were prepared using a slightly adjusted protocol (Elshire et al. 2011, Palomar et al. 2017). The DNA extracted from the fish kidneys, was first double digested with the restriction enzymes *PstI* and *BamHI* and the digested fragments were ligated with barcoded adapters using T4 DNA ligase. By utilizing this approach, only those fragments were selected for sequencing that had *PstI* restriction site (5'-CTGCAG-3') at 5'-end and *BamHI* (5'-GGATCC-3') restriction site at 3'-end. For each library, the adapter ligated DNA samples were then combined and purified using the QIAquick® PCR Purification Kit. After the purification, DNA fragments of different lengths (200, 300, 400, 500 and 600 bp) were selected using E-Gel® SizeSelect™ (Invitrogen) as in Pukk et al. (2015). Fragments from each size range were amplified by PCR and later purified using the QIAquick® PCR Purification Kit. Each library, first only 400 bp long fragments and later pooled (200-300 and 500-600 bp) fragments, was sequenced on a separate lane. The single-end (SE) sequencing was done using Illumina HiSeq 2500 and TruSeq version 2 Rapid sequencing chemistry.

For RNA sequencing, total RNA from the 14 kidney samples (**II**, dual RNA sequencing) and 14 pelvic fins samples (**III**) was used. The RNA from the fin tissues was first combined into two pools (July and August pools) prior to quality assessment. The quality and quantity of RNA were assessed using the Advanced Analytical Fragment Analyzer and Nanodrop ND-2000 spectrophotometer, respectively. Libraries were prepared according to Illumina TruSeq® Stranded mRNA Sample Preparation Guide (part # 15031047) starting from 100 ng of total RNA. In order to pool several samples for sequencing, a unique Illumina TruSeq indexing adapter was ligated to each sample during the adapter ligation step. The quality and quantity of the prepared libraries was evaluated with Advanced Analytical Fragment Analyzer and Qubit® Fluorometric Quantitation, Life Technologies, respectively. The libraries were pooled into a single pool which was then sequenced in three lanes. Paired-end (PE) sequencing was performed using Illumina HiSeq 2500 instrument and TruSeq v3 sequencing chemistry.

The quality of total RNA extracted from the pelvic fin tissue of 238 individuals was assessed using the Agilent 2100 Bioanalyzer. Three barcoded libraries (64, 91 and 96 individuals) were prepared using Lexogen QuantSeq 3′ mRNA-Seq Library Prep Kit FWD for Illumina according to the manufacturer's recommendations

(Lexogen). The qualities of barcoded libraries were assessed with the fragment analyzer (Advanced Analytical, AATI) using the High Sensitivity NGS Fragment Analysis Kit. The three pooled barcoded libraries were single-end sequenced using an Illumina HiSeq2500 in 14 lanes. For the first two pooled libraries, we generated 125-bp-long reads in two lanes. For the remaining 12 lanes, we generated 100-bp reads.

2.3 Pre-processing and NGS data analyses

The detailed description of each step can be found in the respective chapters. Briefly, I performed the quality assessment of sequenced reads (**I-III**) using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In **Chapter I**, I used fastq-multx (Aronesty 2011), fastx–Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and custom python scripts to perform de-multiplexing and other pre-processing tasks such as trimming and subsequent filtering of reads with ambiguous bases and low-quality scores. I also excluded some individual samples (n = 14) because of their low sequencing yield. Because the brown trout genome was not available at the time of data analysis, I generated reference sequences (each 88 bp) from a subset of quality-filtered reads (from 28 individuals) using clustering followed by alignment approach. Prior to clustering, I also excluded reads with complete restriction enzyme recognition sites (either of the *PstI* or *BamHI*) to avoid potential chimeric reads (Pukk et al. 2015). Then, I used CD-HIT-EST (Li and Godzik 2006) for three-round clustering and Bowtie2 (Langmead and Salzberg 2012) for mapping the reads back on the resulting sequences. By using information from the later alignment, I applied coverage filter (≥ 5 reads) to produce final reference sequences.

I subsequently aligned all pre-processed reads from all samples on these reference loci using Bowtie2 (Langmead and Salzberg 2012). I then called and retained biallelic SNPs using samtools and bcftools (Li et al. 2009). Next, using a custom python script, I obtained the allele counts of SNPs from the BAM files, and subsequently converted these counts into genotypes using criteria described in Hecht et al. (2013). Additionally, I changed the individual genotypes of the low coverage SNPs (<5 reads) to missing. Further, I aligned the SNP-containing reference sequences on the Atlantic salmon genome using Bowtie2 and recalculated the SNP positions of the mapped loci according to the genome. Prior to the data analysis using the genABEL package (Aulchenko et al. 2007), I also excluded potential erroneous SNPs by applying additional filters, such as location, quality and occurrence. The input for genABEL comprised of both mapped (on the Atlantic salmon genome) and unmapped SNPs. Finally, I further divided SNPs into three subsets based on their call rate, minor allele frequency (MAF), Hardy–Weinberg equilibrium (HWE) and

mapping on the Atlantic salmon chromosomes by using the genABEL package (Aulchenko et al. 2007). I used these subsets to separately perform familial relatedness, linkage disequilibrium (LD) and association analyses.

For RNA sequencing reads (**II** and **III**), I performed the adapter trimming and pre-processing using Trimmomatic (v. 0.35) (Bolger et al. 2014). During the pre-processing of QuantSeq data (**III)**, I also identified and cut longer runs of poly-As at the end of the reads and removed shorter sequences (<40 bp) using cutadapt (Martin 2011). Prior to the read mapping in **Chapter III**, I modified the reference genome of Atlantic salmon. For this alteration, I first mapped the quality filtered PE RNA sequence reads from fin pools on the genome using Hisat2 (Kim et al. 2015) and called the SNPs for the pooled data. Subsequently, I adjusted the reference by replacing the reference allele with alternative alleles. Furthermore, I obtained the trout fin-specific splice sites using stringtie (Pertea et al. 2015) and extract_splice_sites.py script available with Hisat2. I then used these splice sites as a guide during the alignment of QuantSeq 3′ mRNA reads to the adjusted reference genome using Hisat2. From the resulting alignments, I obtained the exon-level read counts for all samples using R package genomicalignments (Lawrence et al. 2013) and the Atlantic salmon genome annotation. As a next step, I aggregated the read counts from the different lanes, runs and replicates to individual sample counts (n = 238) using collapseReplicates function from the R package deseq2 (Love et al. 2014). Furthermore, I selected only protein-coding nuclear genes (> 10 average reads) for further analysis. For linear modelling, I transformed the raw read counts into quantile-normalized log2-counts per million (logCPM) using the voom function (Law et al. 2014) implemented in the R package limma (Ritchie et al. 2015). Finally, I corrected the batch effect (from the pooled libraries) by using the ComBat function in the R package sva (Leek et al. 2012).

2.4 Genotype-phenotype associations (I)

For assessing the statistical association between the genotype and four estimated phenotypes (**I**) (RPL, Hct, KS and FL), I used the FASTA (FAmily based Score Test Approximation) method implemented in the R package genABEL (Aulchenko et al. 2007). This method takes into account the familial structure by using a genomic kinship matrix, which I calculated using ~7K filtered SNPs. To control for variation of the focal traits caused by fish size variation, I used FL as a covariate for the phenotypes of RPL, Hct and KS. To incorporate the nonlinear relationships of the phenotypes of Hct or KS with that of RPL (i.e., tolerance), I used second-order polynomials of RPL as covariates for the analysis of these disease traits (Debes et al. 2017). Finally, I set Bonferroni threshold for considering SNPs to be genome wide significant i.e., a p-value < 0.05/number of SNPs.

Subsequently, I also performed Random Forest (RF) analysis to determine the possible polygenic architecture for the four focal traits. As the implementation of RF analysis did not allow missing values, I imputed the missing genotypes using TASSEL (Bradbury et al. 2007) and exported all genotypes as reference probabilities (values ranging from 0 to 1) rather than minor allele counts. I also performed principal component analysis (PCA) in TASSEL and used the first five PCs (explaining 22.1% variation) for correcting the spurious false positives caused by the familial relatedness. For the RF analysis, I used an R script developed by Brieuc et al. (2015) with slight adjustments (described in Hess et al. 2016). For consistency, I added the same covariates in the model as used in the association analyses. Furthermore, I also carried out an additional permutation analysis to evaluate the magnitude of variation of each trait explained by the random combinations of markers. From these permutations, I further calculated the p-values using pnorm function in R to show whether the observed outcome was higher than for the randomly permuted data sets.

Parallel to performing the association analysis, I also assessed the linkage disequilibrium (LD) and familial relatedness among the studied brown trout juveniles. The estimation of the LD pattern and spread was important because the strength of association analysis depends on the spread of LD between SNPs. For this purpose, I selected a subset of 2,834 SNPs that mapped only on Atlantic salmon chromosomes and had a relatively high MAF (≥ 0.1). I then estimated the squared allele-frequency correlation (r^2) values between adjacent SNPs in each chromosome using r2fast implemented in genABEL package. To further assess how far LD extends along the chromosome, I further measured the half-length of r^2, which is the position where r^2 is decrease up to 50% (Reich et al. 2001). For the familial relatedness assessment, I converted the respective SNPs subset into format acceptable by the software Colony (Jones and Wang 2010), which reconstructs parentage based on genetic variants. For this SNP subset, I used mean genotyping error rate of 0.066 and an initial allelic dropout rate of 0.0001. Using parental reconstruction in colony, I obtained the maximum-likelihood probability estimates of any two individuals to be either full- or half-siblings. I further visualized and compared the resulting sibship assignment with a multidimensional scaling plot created from the kinship matrix.

2.5 Host free transcriptome assembly (II)

In order to exclude reads derived from the host kidney transcriptome in **Chapter II**, I consecutively mapped the filtered reads on the Atlantic salmon (GCF_000233375.1) and the northern pike (*Esox lucius*, Esociformes) (GCF_000721915.2) genomes using Hisat2 (v. 2.0.1) (Kim et al. 2015). Of the reads

from the infected fish samples that were not mapped to either genome, I subsequently performed taxonomic classification using Kraken2 (Wood et al. 2019) with a custom database. Next, I selected those reads, that were not classified within the NCBI bony fish division, for the de novo parasite transcriptome assembly using Trinity (v. 2.5.1) assembler (Grabherr et al. 2011). This enabled clustering together those transcripts and isoforms that were up to 95% identical using CD-HIT-EST (Fu et al. 2012). On the resulting transcripts, I aligned the reads from both infected and uninfected samples using Bowtie2 and generated a read count matrix in R. At the same time, I also subjected the longest isoforms of these transcripts to a second round of classification using Kraken2 and BLASTN and BLASTX (Altschul et al. 1990) against NCBI non-redundant databases (only best hits). Then I obtained the transcripts of the parasite using filters for read coverage (≥ 40 reads), origin (≥ 90% of the mapped reads from the infected fish) and taxonomic classification (either assigned to GenBank's invertebrate division or no match was found using Kraken2, BLASTN and BLASTX). I evaluated the completeness of the putative parasite transcriptome using Benchmarking Universal Single-Copy Orthologs (BUSCO, v. 3.0) (Simão et al. 2015) against 978 Metazoan orthologs. I further compared the assembled transcripts against the recently assembled intersect (between bryozoan and rainbow trout) transcriptome (Faber et al. 2021) using reciprocal BLAST. Finally, I functionally annotated the parasite transcriptome using KEGG automatic annotation server (KAAS; Moriya et al. 2007) and Trinotate (Bryant et al. 2017).

2.6 Drug targets in *T. bryosalmonae* transcriptome (II)

To identify potential drug targets, I simultaneously searched the putative parasite transcripts against ChEMBL protein sequences (using TBLASTN) and the Atlantic salmon proteome (using BLASTX). In order to circumvent potential drug toxicity due to homology with host proteins, I excluded sequences that exhibited similarity with *S. salar* proteins. To increase sensitivity of the search, I only retained sequences with alignment lengths more than 100 amino acids with ChEMBL proteins. Furthermore, for homology modelling, I additionally narrowed down the focus to the sequences with conserved alignments at active or binding sites of matched drug targets. I then submitted the encoded protein sequences of the selected sequences to homology-based structure modelling in Phyre2 (intensive mode; http://www.sbg.bio.ic.ac.uk/phyre2) (Kelley et al. 2015). Finally, I used Chimera (Pettersen et al. 2004) to visualize the best-supported structures and superimposed them on the template structures.

2.7 Correction for survivors (III)

I performed random forest (RF) classification in iterations to identify the missing survivors based upon their transcriptome profiles. For this classification, I used the batch corrected logCPM matrix and survival status (response variable) as an input for the ranger (Wright and Ziegler 2017) R package. After each RF iteration, I omitted genes with permutation importance values less than zero to keep only informative genes for the next iteration. After the 64th iteration, all the genes (1270) in the input matrix have permutation importance values ≥ 0. RF analysis with these genes classified samples into survivors and non-survivors with an error rate of 16%. The high error rate resulted from the misclassification of 25 (10.5%) recaptured samples as non-survivors and may reflect the poor physiological status represented in transcript profiles of these individuals at the August capture.

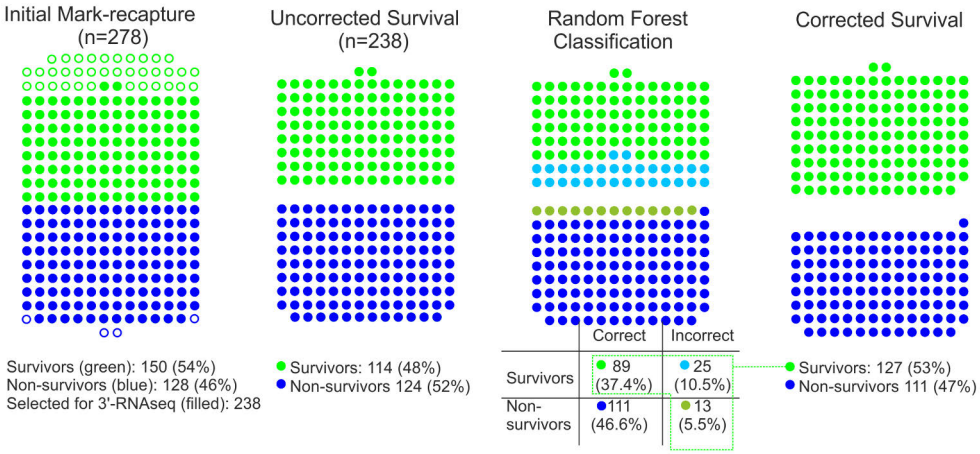

Figure 6. Random Forest classification of survivors and non-survivors. Each circle represents a brown trout sample captured in August 2015 and filled circles represents samples (n = 238) selected for the Quantseq 3' RNA sequencing (**III**).

As a key finding, RF misclassified 13 (5.5%) uncaptured samples as survivors (yellowish green circles in Figure 6). This was in line with the mark–recapture analysis, which also showed that a small number (~14) of survivors (i.e., fish marked in August) were not caught in September. Therefore, I considered the 13 uncaptured individuals as the missing survivors (classified as 'corrected survival'). I performed the subsequent analyses differential gene expression on both observed and corrected survival statuses. The main results stayed impervious irrespective of the survivors' classification (see below). For example, I found a considerable overlap (n = 171) among the top 416 genes based on p-values for the differential expression analyses

for both observed and corrected survival, both of which displayed highly significant enrichment for mitotic cell cycle genes.

2.8 Differential gene expression analysis (II & III)

In **Chapter II**, I performed gene-level differential expression (DE) analysis for the *de novo* assembled *T. bryosalmonae* transcripts. For this purpose, I first summed up the read counts of different isoforms of the same genes to get the gene-level read counts. I then used the R package DESeq2 to detect *T. bryosalmonae* genes DE between the moderately infected (low PL) and severely infected fish (high PL). As one of the main goals of **Chapter II** was to detect candidate DE genes that required further validation (e.g. by RT-qPCR), I used an explorative significance level (FDR < 0.1).

In **Chapter III**, the parasite load (PL) phenotype was measured in the observed survivors (September capture). As PL is a continuous phenotype, I used linear regression (using glm function in R) to describe its relationship with transcript abundance. In this approach I used each gene in the logCPM matrix (batch corrected) as a predictor against the phenotype (response variable) assuming the normal distribution for errors of both. I further used DESeq2 to identify differentially expressed genes for both corrected and uncorrected survival (**III**). In order to control the batch effect, I also included library IDs (I-III; factor term) as a covariate in the model. For uncorrected survival status, the p-value distribution was hill-shaped instead of uniform (as would be expected under an absence of differential expression). However, for the corrected survival, the distribution of p-values was approximately uniform but with some excess near zero, indicating the presence of differentially expressed genes. The main aim of **Chapter III** was to identify pathways rather than individual genes and generate new hypotheses, therefore, I used relaxed significance thresholds (unadjusted $p < .01$) for both PL and survival.

For both PL and corrected survival in **Chapter III**, I additionally performed the discriminant analysis of principal components (DAPC) on the corrected gene expression matrix using the adegenet package (Jombart 2008, Jombart et al. 2010) in R. Because the dapc function required categorical data, I first converted the PL values into three categories: low ($n = 24$; $PL < -1.27$), intermediate ($n = 39$; $PL \geq -1.27 \leq 0.58$) and high ($n = 48$; $PL > 0.58$) using the cut command available in R. Furthermore, I also subjected the differentially expressed genes for survival to the automatic network construction and module detection using the wgcna (Langfelder and Horvath 2008) package in R.

2.9 Natural selection on transcript abundance (III)

In order to quantify the strength and form of contemporary natural selection on transcript abundance, I computed standardized linear (s) and nonlinear (quadratic) selection differentials (λ) based on both corrected and uncorrected survival. I calculated the linear and nonlinear selection differentials for each of the 18,717 gene transcripts and for both corrected and uncorrected survival (binary status: non-survivor = 0, survivor = 1).

For these calculations with survival as the response, I used generalized linear models (glm function in R) with a logit-link function and binomial error distribution. For linear differentials, I mean-centred and variance-scaled (mean = 0, SD = 1) the expression of every transcript as a predictor. To estimate nonlinear differentials, I added a square of the scaled transcript-level as another predictor into the model. I subsequently estimated, the p-value for each selection differential using *t*-test that was based on logistic regression coefficients, and, in the same way, standard errors and model residual degrees of freedom.

Furthermore, I randomized the survival values and performed thousand permutations of selection differentials as described above. This generated a distribution of linear and nonlinear selection differentials under the absence of selection on transcript abundance. I also estimated distributional selection differentials (DSD) (Henshaw and Zemel 2016) to compare the strength of directional and nondirectional selection. By using this approach, I was able to divide total selection into selection on the trait mean (dD) and selection on the shape of the trait distribution (dN).

I further estimated the linear selection gradients for the DE genes using a recently developed approach based on the principal components (Chong et al. 2018). This approach can better deal with the multicollinearity among the predictors and seems intuitive when the number of predictors surpasses the number of observations (i.e., high-dimensionality). For this purpose, I first calculated the principal components from the correlation matrix of the standardized expression values of DE genes (from the batch corrected logCPM matrix). I then estimated the linear selection gradients for the first 55 PCs (explaining 76% of the variation) with the glm function in R (using logit-link function and binomial error distribution). I finally generated the selection gradients for individual genes by matrix multiplying the eigenvectors from the original 55 PCs with their estimated linear selection gradients. Using the same approach, I also calculated the selection gradients for FL and DE genes. I further calculated the standard errors (as in Chong et al. 2018), *t*-statistic (by dividing the gradients by their standard errors) and p-values (237 degrees of freedom). Finally, I used the p.adjust R-function to obtain FDR-corrected p-values.

2.10 Gene Ontology (GO) and protein–protein interaction network analysis (III)

I used the rentrez (Winter 2017) package in R to search Atlantic salmon complete gene names in NCBI and retrieved orthologue gene information (e.g., symbols, entrez IDs, and description) in humans (86.8%), zebrafish (3.6%) or other organisms (9.6%). I used both string-db (Szklarczyk et al. 2015) and gorilla (Eden et al. 2009) to perform the protein–protein interaction (PPI) and GO enrichment analyses. In string-db, I searched single lists (e.g., DEGs) of gene symbols against the human dataset. In order to incorporate more evidence-based interactions, I further set a minimum interaction score of 0.70 and disabled the text mining in the interaction sources. For the enrichment analysis I in gorilla, I also used the entire orthologue gene symbols as a background list.

3 Results and Discussion

In **Chapter I**, I obtained a total of 1.30 B single end reads (length = 101 bp) from the double digest RAD sequencing (Table 3) and after demultiplexing, I was able to assign 1.27 B (98%) of these reads to 279 individuals. I initially excluded 14 barcoded individuals because of the low throughput (< 190 K). After quality filtering, I retained a total of 1,21 B reads from 265 individuals (on average 4.6 M reads per sample; range = 0.19 M –13.39 M). By three-round clustering of 68.54 M quality-filtered reads from 28 individuals using CD-HIT-EST, I generated a total of 355.7 K double RAD loci. From these I used around 62.17 K as reference RAD loci for mapping the filtered reads from all individuals. I was able to map 93.5% of the reads from all the individuals. From this mapping, I initially obtained 19,837 SNPs. Among these I selected only biallelic SNPs (n = 19764) for further analysis. After applying multiple filtering steps on the biallelic SNPs, I created three data sets for the association testing (7,661 SNPs), familial relatedness (1,728 SNPs) and LD calculations (2,836 SNPs) for 255 individuals.

In **Chapter II**, I obtained a total of 506.2 M PE reads from the RNA sequencing of mRNA extracted from 14 kidney samples of both infected (n = 8) and uninfected (n = 6) fish. In **Chapter III**, I obtained a total of 2.21 B SE reads from the QuantSeq 3′ mRNA Sequencing of 238 individuals (median 8.9 M; range = 1.5–34.6 M). After quality filtering, I retained a total of 1.70 B reads (median = 6.8 M; range = 1.1 M – 27.1 M).

Table 3. Number of short reads produced in all chapters of this thesis.

CHAPTER	NGS METHOD	BIOPROJECT	TOTAL READS	QUALITY-FILTERED READS
I	Double RAD sequencing (SE)	PRJNA429104	1.29 B	1.21 B
II	RNAseq (PE) infected fish	PRJNA668017	243.6 M	220.7 M (+45.3 M unpaired)
	RNAseq (PE) uninfected fish		262.6 M	180.9 M (+46.2 M unpaired)
III	RNAseq (PE) fin mRNA pools	PRJNA517427	55.4 M	44.5 M (+9.1 M unpaired)
III	QuantSeq 3′mRNA (SE) of Pelvic fins		2.21 B	1.70 B

3.1 Error rate in double RAD sequencing-based genotypes

In **Chapter I**, I estimated the genotyping error rate by comparing the constancy of more than 9,000 genotype calls between two replicates for each of three samples. Among these samples, I found an average genotyping error rate of 6.6% (range: 3.7% to 11.1%) which is higher than the earlier reported error rate using RADseq (Mastretta-Yanes et al. 2015). Therefore, I further investigated the potential cause of this higher error rate. After pooling the allele counts from the two replicates of the same samples, I found that the majority (~85%) of inconsistent genotype were heterozygous. This finding indicated that the incorrect calling of heterozygous individuals as homozygotes was the main cause of high genotyping error. Furthermore, I also found that mean coverage of inconsistent genotype calls was lower in all samples as compared to the consistent calls. Interestingly, I noticed that the sample, exhibiting highest genotyping error rate, also had the biggest difference in heterozygosity estimates between the two replicates (0.20 vs. 0.28). The heterozygosity value of one particular replica was 0.280, which is highest among all samples indicating rare cross-contamination during library preparation (Jun et al. 2012).

3.2 Familial relatedness and linkage disequilibrium in brown trout

Based on a smaller subset of SNPs (n = 1,728), I observed on average seven individuals per full-sib family (range: 2 to 21 individuals) in **Chapter I**. Overall, I inferred 67 full-sib families, with 22 families having more than three siblings. In the MDS plot generated from the kinship matrix, I also observed that full- and half-sib samples positioned closer to each other. Furthermore, I also found an almost complete agreement between the pedigree constructed in **Chapter II** and pedigree built previously using 14 microsatellite markers. Taken together, these observations indicated the ~1,700 SNPs provide sufficient power for deducing familial relationships, and that the familial relationship estimates using two different marker types were indeed accurate (Debes et al. 2017). Based on another subset of SNPs (n = 2,836) present on Atlantic salmon chromosomes, I observed low levels of LD in brown trout with mean pairwise r^2 value of .043. I also found half-length of r^2 to be around 200 Kb, which means that 50% of LD decay (r^2 = .09) was reached at this distance. This observation leads to the conclusion that more SNPs (e.g., 100 K, rather than the 7.6 K) markers are required to cover the entire brown trout genome for the future association studies.

3.3 Association mapping reveals candidate SNPs for parasite load and disease traits

In **Chapter I**, I tested a total of 7,661 SNPs in 255 wild brown trout samples for the association with the four traits (RPL, Hct, KS and FL) and identified promising SNPs associating with RPL (inverse of resistance) and Hct (anaemic response to PKD). For all traits, I found the inflation factor (λ) estimates close to one, which reflected an efficient incorporation of familial relatedness in the models. Based on the Atlantic salmon genome, I found the promising SNP candidates to be located nearby or within genes (i.e., intergenic or intronic regions) that have previously been associated with various diseases in both fish and human. For example, the top RPL-associated SNP (reaching Bonferroni-corrected significance level) is an intergenic variant mapped in-between two protein-coding genes on the Atlantic salmon chromosome 10; situated between cerebellin 1 precursor, *cbln1* (LOC106560864), and NEDD4-binding protein 1, *n4bp1* (LOC106560863). Interestingly, the upregulation of *n4bp1* gene in grass carp (*Ctenopharyngodon idella*) inhibits the viral gene transcription in reovirus-infected kidney cells (Cai et al. 2014). Thus, *n4bp1* may also contribute in controlling infections (viral and other) in fish kidney. The second RPL-associated SNP was mapped in an intronic region of the prickle-like protein 2, *prickle2* (LOC106583026) gene on chromosome 22 and which has previously been associated with kidney development (McNeill 2009).

In the predecessor study of **Chapter I**, Debes et al. (2017) used a quantitative genetic framework in the brown trout populations from both Altja and Mustoja rivers to describe the relative contributions of genetic and nongenetic variation to resistance and tolerance to PKD. They reported a high heritability estimate for RPL ($h^2 = 0.44 \pm 0.14$) in the infected fish captured from Altja river. This high h^2 specifies that a substantial amount of RPL variation can be attributed to genetic effects which can either be by few SNPs with large effects or multiple SNPs with small effect (Visscher et al. 2008). Consistent with the quantitative genetic study, I also found two most significant SNPs to be associated with RPL. Furthermore, using Random Forest analysis, I found 15 SNPs that explained more residual variation than random permutations. This consistency advocates that a substantial amount of the variation in RPL is caused by genetic components which, in turn, is manifested by relatively few loci.

The topmost Hct- associated SNP was located on the chromosome 19 within an intron of the putative 5′-AMP-activated protein kinase subunit gamma 2 protein (LOC106579597) gene. In humans, the ortholog of this gene, *PRKAG2* (~70% protein level identity), has been found associated with haematocrit and other blood-related traits (Ganesh et al. 2009, Tragante et al. 2014). This finding makes *PRKAG2* not only an important candidate in brown trout but may even influence intraspecific variation of haematocrit in evolutionary distant species (Arendt and Reznick 2008) and, therefore, grasps attention for the future anaemia studies. Thus, even though this

SNP did not pass the Bonferroni corrected p-value threshold, the association of the respective gene with blood-related traits in human makes *PRKAG2* an appropriate candidate for Hct in brown trout. Finally, in contrast to RPL and Hct, I observed weak genotype associations for either KS or FL.

3.4 Multilocus associations using Random Forest

In **Chapter I**, using RF analysis, I found that around 30 to 60 SNPs could explain 25.2% to 41.4% of the residual variation for each trait. Surprisingly, when I permutated the trait values and then performed the RF analysis, ~37% of the residual variation for Hct and KS was purely explained by randomization. Thus, these permutations confirmed that even random combinations of genotype and phenotype could produce a substantial amount of variation explained by genotypes. However, for both traits, I also did not find any of the 13 SNP combinations that reached the significance level of 0.05. On the other hand, for RPL and FL, I found that the observed SNPs explained greater proportions of variation than their respective permutations, which indicated that the RF analysis discovered true multilocus associations for both traits. For example, I observed 15 SNPs that explained ~40% of the residual variation for RPL with a p-value of 5.2×10^{-7}. On contrary, 15 SNPs in the permuted data sets could explain as much of only 33.8% of residual variation. These results were consistent with the earlier quantitative genetic study, in which both RPL and FL displayed higher heritability than Hct and KS (Debes et al. 2017). Thus, through permutation analysis, I was able to assess the statistical significance to the RF findings and differentiate the true multilocus effects from the false positives. Taken together, through simple permutation of the phenotypes, I showed a successful strategy to evaluate the significance of the RF classifications. Finally, similar to some earlier studies (Hess et al. 2016, Holliday et al. 2012), I detected only a modest overlap between the top SNPs of association mapping and the RF analysis (33% for FL and 47% for RPL). Furthermore, with association mapping, I could not find any strong SNP candidates associating with FL. However, it is highly likely that with RF, I obtained a subset of SNPs with a small but true effect on FL. Thus, based on two different analytical procedures, the combination of two methods can deliver a more inclusive list of SNPs associated with the phenotype.

3.5 Host free *de novo* transcriptome assembly and functional annotation

A major challenge for the *de novo* assembly of transcriptome (**II**) using dual-RNA sequencing data was to correctly separate host and parasite RNA reads. By employing a multi-step filtering approach followed by *de novo* assembly, I was able to report a

partial transcriptome assembly of *T. bryosalmonae* along-with functional annotation. From the quality-filtered 266 M reads from the infected fish, I first separated 7.6 M (3.4%) PE reads that neither mapped on the fish reference genomes nor subsequently classified as bony fishes origin using Kraken2. From these reads, I generated a *de novo* assembly of 16,499 transcripts (12,676 genes) and subsequently merged ≥ 95% similar transcripts together into 1,474 transcripts (12,621 genes) using CD-HIT-EST. On the merged transcripts, I attained 83.73 and 33.24% overall average read alignment rates from both infected and uninfected samples, respectively. Subsequently, after applying read coverage and other taxonomic classification filters, I retained a total 3,427 transcripts belonging to 2,905 genes of *T. bryosalmonae*. Similar to the previous studies of myxozoans which revealed AT-richness in the genomes and transcriptomes (Yang et al. 2014, Foox et al. 2015, Yahalomi et al. 2020), the GC content in these transcripts was low (mean = 31.5%) and stayed within the range of other myxozoan genomes (23.6–37.5%) (Yahalomi et al. 2020). Finally, the raw read counts on the majority of these transcripts were positively correlated with PL (amount of parasite in the host kidney) i.e., more than 92% of transcripts had $r \geq 0.5$, indicated that the pipeline developed in chapter II successfully excluded the host sequences and acquired true RNA reads from *T. bryosalmonae*.

3.6 *T. bryosalmonae* transcriptome completeness and functional annotation

By using BUSCO, I found low completeness (26%) estimates with only 14.5% complete and 11.6% partial sequences of core single copy metazoan genes present in assembled *T. bryosalmonae* transcriptome. Compared to the recently published *T. bryosalmonae* transcriptome (Faber et al. 2021), that had twice higher completeness (48.3%), I found 58 additional BUSCO genes that were missing in the recently published *T. bryosalmonae* transcriptome (Faber et al. 2021). As Myxozoans are highly reduced metazoans with very small and diverged genomes, the estimation of completeness of *de novo* assemblies is not a trivial task (Chang et al. 2015). Furthermore, earlier myxozoan transcriptome and genome studies has also reported low completeness estimates using BUSCO (Hartigan et al. 2020; Yahalomi et al. 2020). In order to provide a better view for completeness estimates in myxozoans, I assessed BUSCO core single copy metazoan genes in all published myxozoan parasite datasets. Despite of variable completeness and coverage, I identified a set of 231 BUSCO genes that were present in all myxozoans. Based on this smaller set of BUSCO genes, I found higher completeness (56%) for the *T. bryosalmonae* transcriptome assembly. On the other hand, I also found that many of metazoan single-copy genes (23.7%) were not detected in any of the myxozoan datasets. It is very likely that some of these genes were lost during the myxozoan evolutionary transition from a free living to a parasitic

lifestyle as they have gone through a reduction in body plan and genome size (Chang et al. 2015). However, the alternative explanation that the unidentified metazoan single-copy genes are still present in myxozoans but were not detected using BUSCO because of highly divergent sequences, cannot be ruled out completely.

Using KAAS, I was able to annotate only 1,742 transcripts (1,509 genes) of *T. bryosalmonae* to KEGG orthologs (KO) genes, majority of which were enzymes (38%), membrane trafficking (14%), spliceosome (12%), exosome (10%) and chromosome and associated proteins (9%). The most abundant enzymes were protein kinases (14%) and peptidases (proteases) and inhibitors (11%). In the pathway searches, the majority of annotated transcripts belonged to metabolic pathways (ko01100; n = 262), spliceosome (ko03040; n = 141), ribosome (ko03010; n = 74), RNA transport (ko03013; n = 72), endocytosis (ko04144; n = 59), oxidative phosphorylation (ko00190; n = 54) and phagosome (ko04145; n = 54). Similar to myxozoan parasite *Thelohanellus kitauei* (Yang et al. 2014), I also identified *T. bryosalmonae* transcripts related to multicellular metazoan signalling pathways e.g., mTOR (40 transcripts) and Wnt (22 transcripts) signalling pathways. These enzymes and pathways provided valuable functional information of the *T. bryosalmonae* transcriptome. For example, peptidases (proteases), enzymes that hydrolyze proteins into tiny polypeptides or single amino acids, were the second most abundant enzymes (11%) found in the *T. bryosalmonae* transcripts. In many parasites, peptidases are known virulence factors and also considered as important antiphrastic drug targets (McKerrow et al. 2006, Pina-Vazquez et al. 2012, Siqueira-Neto et al. 2018). Recently, proteases have also gained considerable attention in myxozoan parasites because of their proposed role of nutrient digestion (Yang et al. 2014, Hartigan et al. 2020). In *T. bryosalmonae*, I found higher abundance of cysteine (32%) and metallo (27%) proteases while serine (16%), threonine (16%) and aspartic proteases (8%) had smaller proportions. Compared to other protease classes, I found higher expression of aspartic proteases despite their low abundance (5 genes). Finally, the protease composition of *T. bryosalmonae* followed the trend of other myxozoans such *as Sphaerospora molnari*, *C. shasta*, *Kudoa iwatai and Myxobolus cerebralis*.

Among identified pathways, I expected the endocytosis pathway to have a broader role in *T. bryosalmonae* because in the myxozoan *T. kitauei*, endocytosis is associated with nutrients acquisition (Yang et al. 2014). In *T. bryosalmonae*, endocytosis of the cellular material of the host phagocyte and from the primary cell to the enclosed secondary and tertiary cells has also been observed (Morris and Adams 2008). Furthermore, the primary cells of *T. bryosalmonae* engulfs other cells to form the internal cells (both secondary and tertiary) (Morris 2010). Thus, the same endocytosis pathway genes are likely to be involved in these engulfment processes.

Using protein translations with Trinotate, I was able to annotate 1,944 *T. bryosalmonae* genes to more than 25,000 GO terms, which I sub-categorized into

biological processes (BP, 45%), cellular component (CC, 31.6%) and molecular functions (MF, 23.5%). The most frequent GO terms in the subcategories were integral component of the membrane (CC), protein phosphorylation (BP) and ATP binding (MF).

I found a majority (81.5%) of the transcripts matching 1,615 contigs of the recently reported intersect transcriptome (from bryozoan and rainbow trout, (Faber et al. 2021) using an adjusted reciprocal BLASTN. Only 634 transcripts (528 genes), did not produce a hit within the published transcriptome and are unique to the assembly reported in **Chapter II**. In these unique transcripts, I found higher relative proportion of genes related to Ribosome (5.13%), Ribosome biogenesis (3%), Ubiquitin system (2%) and Peptidases and inhibitors (1.6%). Trinotate assigned GO terms comparison using WEGO 2 showed that these unique transcripts have higher ($P < 0.05$) percentage of genes related to endoplasmic reticulum membrane (CC, GO:0005789) and mitochondrial protein complex (CC, GO:0098798), structural molecule activity (MF, GO:0005198) and structural component of ribosome (MF, GO:0003735).

3.7 Isoforms in the *T. bryosalmonae* transcriptome

One fourth of the putative parasite transcripts (n = 929) were the isoforms of 407 genes and less than half of these were annotated to single KO genes (430 isoforms of 197 genes). Some of the *T. bryosalmonae* genes with ≥ 4 isoforms are; summarised in Table 4. I also detected multiple isoforms of the these*T. bryosalmonae* genes (except ETFDH) in the genome-guided transcriptome assembly of *K. iwatai* (assembled in **Chapter II**). The presence of multiple isoforms in the reference-guided assembly provides further evidence that these isoforms are not *de novo* assembly artefacts. Thus, it can be hypothesized that the functional diversity of proteins in Myxozoa is also expanded through different transcript isoforms of the same genes. However, this hypothesis requires further experimental validation for the functional characterization different isoforms of *T. bryosalmonae* genes.

Table 4. Some *T. bryosalmonae* genes with more than three isoforms.

GENE	GENE SYMBOL	# OF ISOFORMS
Cyclin-dependent kinase 5	CDK5	5
Annexin A7/11	ANXA7_11	4
Atpase family AAA domain-containing protein 3A/B	ATAD3A_B	4
Histone-binding protein RBBP4	RBBP4	4
Swi/snf-related matrix-associated actin-dependent regulator of chromatin subfamily a member 2/4	SMARCA2_4	4
Cyclin T	CCNT	4
Electron-transferring-flavoprotein dehydrogenase	ETFDH	4
V-type h + -transporting atpase subunit h	ATPEV1H	4

3.8 Anti-parasitic drug targets in the *T. bryosalmonae* transcriptome

By searching the *T. bryosalmonae* transcripts in the ChEMBL proteins and discarding hits with any homology within the salmon proteome, I identified six parasite genes (**Chapter II**) as potential drug targets. These include two Endoglycoceramidase (EGCase1 and EGCase2) genes, Carbonic anhydrase 2 (CA2), Legumain-like protease (LGMN) and two genes of Pancreatic lipase-related protein 2 (PLRP2s). Using the Phyre2, I further predicted the 3D structure models for the *T. bryosalmonae* EGCases and CA2 (>90% of residues) with high confidence (> 90%). Endoglycoceramidase is an enzyme that hydrolyses glycosphingolipids (acidic and neutral) into ceramides and oligosaccharides. In earlier studies, it has been found in two free-living cnidarians (jellyfish and hydra) and has displayed a distinctive role in the dietary pathway of glycosphingolipids catabolism in hydra (Horibata et al. 2000, Horibata et al. 2004). Therefore, even though the EGCases have not yet been reported in myxozoans, the existence of two copies in a reduced *T. bryosalmonae* genome suggests that this enzyme might also play an important role in the parasite lipid catabolism. The predicted three-dimensional structures of both EGCase1 and EGCase2 displayed a high overlap with the template (endoglycoceramidase ii from *Rhodococcus* sp.) structure. Furthermore, the six inhibitor binding site residues of the template were also conserved in *T. bryosalmonae* EGCase2 (one amino acid missing) and EGCase2 (missing three amino acids), and acquired the same orientation in the 3D models. Thus, the conserved residues of the binding site of *T. bryosalmonae* EGCases can also be tested for the inhibition with the cellobiose-like imidazole inhibitor-like structures.

CA2, on the other hand, a plasma membrane anchored alpha carbonic anhydrase, is an emerging anti-parasitic drug candidate for human diseases such as Chagas disease, leishmaniasis and schistosomiasis (Vermelho et al. 2017, Da'dara et al. 2019). Importantly, the CAs of the causative parasites of these diseases are also in vitro inhibited by sulphonamide, thiol and hydroxamate chemicals (Vermelho et al. 2017, Da'dara et al. 2019, Angeli et al. 2020). The superimposed structures of *T. bryosalmonae* CA2 and *S. mansoni* CA (Da'dara et al. 2019) also revealed the structural conservation of four out of five active site residues and all three zinc-binding histidines. Thus, it is also possible that the above-mentioned inhibitor molecules will also cause inhibition of CA2 of *T. bryosalmonae*.

LGMN, an asparaginyl endopeptidase, has also been considered as an important anti-parasitic drug target in human parasites such as *Trichomonas vaginalis*, *S. mansoni* and *Ixodes ricinus*. In these parasites, Aza-peptidyl Michael acceptors (Götz et al. 2008) and epoxide inhibitors inhibited the legumain activity (Ovat et al. 2009). In *T. bryosalmonae*, the last two drug target genes encode lipase enzyme PLRP2. This gene is also present in genomes of several free-living cnidarians,

including *Exaiptasia pallida*, *Stylophora pistillata* and *Nematostella vectensis*. Presence of PLRP2 in invertebrates and cnidarians suggest a potential role in nutrient acquisition and lipid metabolism. Interestingly, Orlistat, which is a lipase inhibitor and used for obesity treatment in humans, inhibit in vitro growth of the parasite *Giardia duodenalis* (Hahn et al. 2013). In the alignments of the protein sequences of LGMN and PLRP2s with their respective ChEMBL hits, the active- or binding site residues were missing. Therefore, in **Chapter II**, I did not predict the 3D structures for these drug targets.

3.9 Differentially expressed *T. bryosalmonae* genes

By comparing moderately and severely infected fish, I aimed to first identify parasite genes potentially linked with infection intensity (Råberg 2014). In **Chapter II**, I found 55 genes of the parasite that were differentially expressed (FDR < 0.1, 23 down- and 32 upregulated in severely infected fish) between severely and moderately infected fish host. Among the 55 genes, I could not assign functional annotation to 24 genes. Interestingly, the glutathione peroxidase (*gpx*) gene of the parasite was upregulated in the moderately infected fish. This *T. bryosalmonae* gene was expressed in all infected fish and possessed highest mean expression level among DE parasite genes. The antioxidant enzyme encoded in this gene safeguards the parasite from the harms of free radicals introduced by the host immune cells (Changklungmoa et al. 2018). Thus, the higher expression of *T. bryosalmonae gpx* in moderately infected fish is a confirmatory indicator that these hosts exerted stronger immune response in terms of free radicals.

More interestingly, I observed that the three polar capsule related genes (*TBNCOL-1*, *TBNCOL-2* and *TBNCOL-4*) had higher expression levels in moderately infected fish. Additionally, I also detected four interesting genes namely; Niemann-Pick C2 protein (*NPC2*), ammonium transporter Rh and two cathepsin L (*CTSL*s), among the upregulated genes in the moderately infected fish. In other myxozoans, the homologs of these genes are distinctively expressed in polar capsule (in *C. shasta*) or sporogonic stages (in *S. molnari*) (Piriatinskiy et al. 2017, Hartigan et al. 2020). In *T. bryosalmonae*, the polar capsules are present only on the spores when the parasite prepares to leave the brown trout host. Thus, the higher expression of these genes in low PL fish most likely indicates the relative abundance of final *T. bryosalmonae* sporognic stage. However, this hypothesis awaits confirmation through further experimental time-series analysis.

Among the 32 upregulated genes in the severely infected fish, I detected low-density lipoprotein receptor class A-like protein 3 (MT070967.1) gene with highest fold change. I also noticed four *T. bryosalmonae* retrotransposon genes (annotated

as Retrovirus-related Pol polyprotein from type-1 retrotransposable element R2; *pol*) with high expression in severely infected fish. I also observed the Talin (*TLN*) gene to be upregulated in severely infected fish. TLN helps in cytoskeletal adhesion to the extracellular matrix (Klapholz and Brown 2017). In *C. shasta* infected rainbow trout, upregulation of the *TLN* gene of the myxozoan parasite is linked with the change of a localized (intestine) to a proliferated infection (other organs) (Alama-Bermejo et al. 2019). Thus, *TLN* might contribute to proliferation and adhesion also in *T. bryosalmonae*.

3.10 Parasite load and survival are linked to the mitotic cell cycle in fin transcriptome

In the **Chapter III**, I analysed the expression of ~18 K genes of 238 brown trout samples from the pelvic fins. I first assessed the relationship between PL and the transcript abundance in fin measure a month earlier. Using linear regression, I detected 804 genes that correlated with PL at an unadjusted $p < .01$ (FDR < 0.19). This FDR indicates that 81% of (~650) these genes are true positives and reflects a genuine relationship between transcript abundance and PL. Using DAPC analysis (Jombart et al. 2010), I was also able visualize and confirm that the host transcriptome includes some genes that co-vary with the PL. Interestingly, for the genes that are positively correlated with PL, I observed enrichment of 59 GO terms (FDR < 0.05). The top three biological process GO terms were: cell cycle process (GO:0022402, 57 genes), mitotic cell cycle process (GO:1903047, 48 genes) and cell division (GO:0051301, 41 genes). I then extended the analysis towards survival (both corrected and observed). Using DESeq2, I identified 416 genes differentially expressed with survival at unadjusted $p < .01$ (FDR < 0.45). This FDR indicate that 55% of these (~229) genes reflect a genuine relationship between transcript abundance and survival. The PPI networks (string-db) for both PL and survival genes (PPI enrichment, $p < .001$) were also highly enriched for genes involved in the mitotic cell cycle (GO:0000278). Furthermore, genes common to both survival and PL included some well-recognized oncogenes and tumour suppressors such as *AURKB, BTG1, UBE2C, BIRC5 and EEF2K*. Using WGCNA, I was able to cluster differentially expressed survival genes into seven co-expression modules, two of which were also correlated with PL. Among these, the red module (27 genes) was highly enriched for the mitotic cell cycle GO term. This module included vital mitotic cell cycle regulatory genes like *AURKB, UBE2C, BIRC5 and CENPH*, which displayed high correlations with PL. Thus, the functional categorization of genes and correlated modules for both survival- and PL-associated transcripts reveals (III) that mitotic cell cycle is the performance trait that is "visible" to the natural selection in this fish–myxozoan system.

Earlier studies have reported that the mitosis arrest is a general response to stress (Burgess et al. 2014, Kassahn et al. 2009, Martín-Hernández et al. 2017). In **Chapter III**, the expression of key mitotic cell cycle genes (*AURKB*, *UBE2C* and *BIRC5*) was positively correlated with the parasite load. Therefore, it is implausible that the genes associated with survival are a mere reflection of stress response of the host caused by the infection.

Other possible explanations of observed associations between fin tissue transcriptome, PL and survival are: i) It might be a response to the penetration of *T. bryosalmonae* in its salmonid host through gills and skin (Longshaw et al. 2002) which may include fins. However, there is no evidence that the entry of *T. bryosalmonae* causes upregulation of cell-cycle activity in the mucosal tissues of the host. ii) In salmonids, the main PKD symptom is tumour-like proliferation of the kidney tissue that results in kidney swollenness (Bettge et al. 2009, Clifton-Hadley et al. 1987, Hedrick et al. 1993). The parasite also causes enlargement of the spleen and several studies indicated that PKD is a systemic disease that affects multiple organs and tissues (Bettge et al. 2009, Bruneaux et al. 2016, Clifton-Hadley et al. 1987, Hedrick et al. 1993, Longshaw et al. 2002, Okamura et al. 2011). Thus, the observed association might illustrate the serious physiological consequences of PKD on host at the whole-organismal level. iii) Kidney is the major organ responsible for haematopoiesis (formation of blood) in fish. PKD causes impairment of kidney, which results in anaemia. The pelvic fin also contains blood vessels along with bony rays, ligaments, nerve fibres and connective tissue cells. Thus, the potent links among cell-cycle-related host genes, PL and survival might also be a consequence of abnormal kidney function and defective blood homeostasis.

3.11 Linear selection is mostly weak on the individual transcripts

Using selection differential estimates (s) that quantify selection on relative fitness in the units of standard deviations of the trait (Lande and Arnold 1983), I was able to compare my estimates on transcription abundance with a larger collection of phenotypic selection estimates for survival for a variety of taxa (Siepielski et al. 2017). Compared to the 1,834 published estimates of s, the majority of the s values for individual transcripts were small with a median absolute value of 0.047. In terms of the strength of linear selection (which includes both direct and indirect selection components), both transcriptomic data and the published phenotypic selection estimates acquired similar frequency distributions with a large majority of s values being close to zero. However, compared to the majority of transcripts, the co-regulated gene modules for corrected survival showed larger s values. I also observed a similar pattern for the s estimates of uncorrected survival. Contrary to the

transcripts, a noticeable number of the published s values for phenotypic traits for a variety of taxa possessed higher “tails”, which might indicate either rare but very strong selection or bias due to low sample sizes. These results indicated that only a limited number of transcripts are probably affected by directional selection. Likewise, a recent work in rice also have shown that the directional selection is usually weak at microevolutionary times, and the strength of selection relies on environmental conditions (Groen et al. 2020). Nevertheless, the broader range of selection differentials (from −0.26 to 0.23) suggests that in the presence of heritable variation and absence of other constraints, selection might cause evolutionary changes in transcript abundances at shorter evolutionary timescales (Campbell-Staton et al. 2017, Donihue et al. 2020, Kingsolver et al. 2001, Kingsolver and Pfennig 2007).

Selection differentials are measured for each trait separately and quantify the total selection acting on the trait (both direct and indirect). To further evaluate the effects of individual traits on the relative fitness, multiple regression (i.e., estimating selection gradients) can be applied. However, estimating selection gradients becomes impractical for the many transcripts that are often highly correlated with each other. To overcome this restraint, I further calculated the linear selection gradients (β) for the 416 DE for the corrected survival using principal component scores (Chong et al. 2018). With this calculation, I was able to quantify the strength of direct selection on individual genes after removing indirect selection from other correlated transcripts. I identified 67 genes with significant β estimates (ranged from −0.47 to −0.15 and from 0.15 to 0.63, FDR < 0.05). These significant genes were also enriched for regulation of the cell cycle (GO:0051726, FDR = 0.031, n = 9). It is noteworthy that the body size (FL), which had no direct effect on survival, showed a significant selection gradient (β = 0.473; FDR = 6.0 × 10−4), when jointly analysed with 416 genes. However, the β estimates for 416 DE genes, either controlled or not controlled for fish size, were highly correlated ($r^2 = 0.950$) indicating that controlling for fish size has a limited effect on these estimates. These results, overall, indicate that direct selection acting on transcript abundances has the potential to cause considerable evolutionary changes at relatively short timescales.

3.12 Host transcriptome exhibiting signals of disruptive selection

I also compared the strength of linear vs. nonlinear selection using recently developed distributional selection differentials (DSD) (Henshaw and Zemel 2016). DSD showed that the linear component of selection was mostly higher than the nonlinear component, which in turn represents the selection on the shape of the trait distribution (mean dD = 0.053, dN = 0.031; signed test, $P = 9.4 \times 10^{-206}$). However,

the nonlinear differentials for 7,273 (40.8%) genes were higher than the linear selection differentials. The distribution of linear differentials and its permutations revealed that only a small proportion of transcripts showed a pattern consistent with the directional selection. On the other hand, comparing the distribution of estimated quadratic differentials (λ) versus the distribution of quadratic differentials from permutations suggested that the data set was highly enriched for transcripts influenced by disruptive selection. The distribution was moved towards the right side (disruptive selection) and the mean was significantly different from both zero and the mean value of the 658 phenotypic λ estimates for a range of taxa (Siepielski et al. 2017). These results suggest that the survival of trout was associated with the transcripts that showed both extremes, i.e., low or high abundance. This is an unexpected outcome because disruptive selection is believed to be uncommon in nature and conversely, if most populations are well adapted to their current environment, stabilizing selection must be a common norm in nature (Kingsolver et al. 2001, Kingsolver and Pfennig 2007). Nonetheless, disruptive selection may also be more widespread than previously thought, and could reflect density-dependent or frequency-dependent competition for resources (Kingsolver and Pfennig 2007). Thus, host transcript abundance might respond to parasite infection. We hypothesize that it may be more advantageous for a host to either show higher resistance (a resilient immune response to limit PL) or by exhibiting higher tolerance (controlling health damage instead of limiting PL). On the contrary, the intermediate response, i.e., partially controlling both parasite load and health damage, might not be the most beneficial strategy. This functional categorization of genes under disruptive selection provided further support to this hypothesis. The GO enrichment analysis of the genes with $\lambda > 0.2$ (n = 1,652) resulted in a wide range of enriched molecular processes. Some interesting GO terms were multi-organism process (GO:0051704), regulation of cell death (GO:0010941), iron ion homeostasis (GO:0055072), vesicle-mediated transport (GO:0016192) and neutrophil activation (GO:0042119). These functions might have reflected the intrinsic complexity of host–parasite interactions. Using WGCNA, I was able to cluster genes affected by disruptive selection ($\lambda > 0.2$) into six co-expression modules. For all these modules, survivors exhibited greater variance in transcript abundance compared to non-survivors, which is a strong indication of disruptive selection favouring extreme values (Levene's test, FDR < 2.0×10^{-4}). These modules were further enriched for more specific GO terms, some of which relate to processes that appear meaningful in response to parasite infection. For example, the brown module showed enrichment for the GO terms cellular response to cytokine stimulus and antigen processing and presentation of peptide antigen via MHC class II.

3.13 Limitations and implications

While linking host-parasite genomics with ecology, I encountered several common and study-specific limitations. The first limitation common in all chapters (**I-III**) was, at the time of data analyses, the unavailability of the brown trout reference genome. In the absence of the host genome, I used the reference genome of the congeneric Atlantic salmon. Aided by the Atlantic salmon genome in **Chapter I**, I was not only able to identify candidate genes for the association but also to characterize the extent and pattern of LD over the chromosomes (**I**). Even though brown trout and Atlantic salmon have different number of chromosomes (2n; *S. trutta*: 80, *S. salar*: 54–58), the LD patterns across shorter distances are expected to be compatible because of the high sequence similarity between the two species (Leitwein et al. 2017). Concurrently, as the main association analysis in **Chapter I** was not dependent on the exact genomic position of the SNPs, I also retained ~900 SNPs that failed to map on the reference genome because of too low similarity. However, the identified SNP candidates for RPL and Hct that were mapped on the Atlantic salmon chromosomes.

In **Chapter II**, the host reference genome was required to remove the host reads from the dual-RNA sequencing data through alignment and thereby to extract the parasite reads. Because of the unavailability of the host reference genome, I used two available reference genomes of fish species related to the study species (Atlantic salmon and northern pike), through which I was able to exclude a large majority of host-derived reads prior to the *de novo* assembly. However, it is also possible that the true parasitic sequences were also lost due to the short nucleotide homology with the fish genomes. In **Chapter III**, I also used the Atlantic salmon genome for read alignment and quantification of transcript abundance. However, In order to enhance the read mapping, I also modified the genome for brown trout specific differences. Most importantly, I first called SNPs from the PE fin RNA sequencing reads that were mapped on Atlantic salmon genome and subsequently replaced the reference alleles of the genome with the alleles called from the mapped reads. Furthermore, I also used brown trout-specific splicing information, which was also assembled from fin PE data, as a guide during the read alignment, respectively. These alterations reduced the otherwise observed mismatches during the alignment of QuantSeq SE reads.

Another host genome related limitation, which is also common in both **Chapter I** and **II**, arises because of the whole genome duplication in salmonid fishes (Ss4R) (Macqueen and Johnston 2014). It is estimated that at least 10% of the Atlantic salmon genome is still in a tetraploid state (Allendorf et al. 2015, Lien et al. 2016, Limborg et al. 2016) and a similar residual tetraploid genome state of *S. trutta* poses challenges during the genomic data analysis and interpretation. To counteract these challenges in **Chapter I**, I used multiple checks to exclude duplicated loci as the

method adapted for association analysis assumed Mendelian segregation of diploid loci. Thus, like many population-level genetic and association studies in salmonids, tetrasomically inherited genomic regions are also most likely underrepresented. Furthermore, the presence of multiple copies of candidate genes in both **Chapter I** and **III**, makes it difficult to connect a single gene to a specific phenotype or molecular function and necessitates further experimental validation.

The next limitation common in both **Chapter I** and **II** is low throughput from the NGS data. Whereas, a robust unravelling of a genotype–phenotype association requires a large number of SNPs, I was only able to generate a small number of genetic markers (7.6 K) to test for the disease association in **Chapter I**. Thus, it is possible that I only identified a subset of markers and thus candidate genes that associate with *T. bryosalmonae* resistance and severity of infection in brown trout. Furthermore, in **Chapter II**, I was able to generate only a small number of *T. bryosalmonae* transcripts (3,427) that corresponded to 26% BUSCO completeness for single-copy metazoan genes. Thus, further genomic and transcriptomic studies, incorporating rapid advances in technology, such as using single cell methods, are needed to extend the gaps posed by the two studies.

Another important limitation present in **Chapter I** is the use of limited number of brown trout samples (n = 255). This low sample size may be the reason why I was only able to detect relatively common SNPs for RPL with large effects. Specifically, the detection of strictly rare variants (i.e., MAF < 0.01) requires thousands of samples to be genotyped (Wang and Xu 2019). However, as the estimation of RPL in the kidney requires lethal fish sampling, it is impossible and unethical to sacrifice the whole year-class of wild fish in small streams to obtain more reliable assessment of rare variants. Related to lethal sampling, another limitation specific to **Chapter III** is the use of the fin transcriptome, which is not a primary target tissue for replication of the parasite, although parasites enter the host via skin and which includes the fin. The well-known target tissue for parasite replication is kidney, whose study-access is possible only through lethal sampling and makes it impossible to estimate survival thereafter. However, because of the systemic nature of PKD (i.e., affecting many organs or even the whole body), I was able to extract biologically meaningful information from the fin transcriptome with a minimal expected effect on fish survival (Gjerde and Refstie 1988).

4 Summary/Conclusions

In conclusion, by harnessing a polygamous mating system in wild stream-dwelling salmonid fish, my thesis exhibits the power of linking host parasite ecology with genomics. The results of **Chapter I** describe the genetic basis of resistance to *T. bryosalmonae* and the role of genetic and environmental variation on the severity of the temperature-dependent disease *T. bryosalmonae* induces. Because of the ecological and economic importance of PKD, additional validation of the here-identified candidate genes and detection of causative mutations represent an area of precedence for the future research. Altogether, my results directly facilitate follow-up research on the validation of SNPs associated with parasite resistance, *T. bryosalmonae* genome assembly and impact of environmental conditions on the strength and form of natural selection on transcript abundance. From a methodological point of view, **Chapter I** represents a special case where genotyping error rate has actually been measured and it highlights the importance of using several replicates to be able to reliably estimate the genotyping error rate. Thus, the results described in **Chapter I** not only open new research avenues towards understanding PKD, but also echo the earlier call for more thorough evaluation and transparent reporting of SNP calling errors in genotyping-by-sequencing data sets (Mastretta-Yanes et al. 2015) and therefore have a more general relevance to a reporting of omics data results.

In **Chapter II**, I established an *in-silico* pipeline to efficiently separate transcripts of *T. bryosalmonae* from the kidney tissue of a native fish host, followed by de novo transcriptome assembly. I anticipate that the transcriptome assembly of *T. bryosalmonae* represents a valuable general resource for future studies on genomic architecture and evolution of the myxozoans and will facilitate future research on the specific molecular mechanisms underpinning PKD. Furthermore, the here identified core set of 231 BUSCO genes can be utilized as an alternative to estimate the completeness of the assemblies in future myxozoan studies. **Chapter II** also illustrates the usefulness of parasite transcriptome as a utility for identifying drug targets against PKD in salmonids. The identified four novel protein drug targets can help in curing the infected fish via direct follow-up research targeted at finding

therapeutic solutions against PKD in especially farmed populations, but possibly also in wild populations.

Irrespective of the specific physiological mechanism, the inferences made in **Chapter III** adds to the increasing body of work showing that parasites influence the cellular machinery of hosts (Guo et al. 2016, Kassahn et al. 2009, Martín-Hernández et al. 2017). Until now, we lack the knowledge about the inflammatory, mitotic and immune processes across organs (Chevrier 2019) during the progression of PKD. Results from my thesis thus strongly suggest that analyses of the host transcriptomes from various other tissues (e.g., blood, kidney, spleen, skin and fin) or even the different cell types (using single cell RNA sequencing) of both naturally occurring and laboratory-infected fish, may contribute significantly to disentangle the molecular mechanisms of the host response and better understand the fitness consequences of transcript abundance. As gene expression variation has a strong environmental dependence, future research using transcriptome data from other tissues and from different years and populations is clearly needed in order to assess any fluctuations in selection patterns (e.g., hard selection by PKD-induced mortality) (Price et al. 2022).

Acknowledgemets

Coming to this point, the last part of my doctoral thesis, I wish to acknowledge some remarkable people who helped me enormously along this lengthy and bumpy road.

First, I would like to express my sincere gratitude to my supervisor, Prof. Anti Vasemägi for providing me this excellent opportunity. I will forever remain indebted of his complete support, guidance and patience throughout this work. Then, I would also like to thank my co-supervisor, AP. Paul V. Debes, whose guidance was always there when needed. Many thanks to Gemma Palomar, Katja Salminen, Meri Lindqvist, Lilian Pukk, Siim Kahar, Riho Gross and Mikhail Ozerov for their help in DNA/RNA library preparation, sample collection during field work and manuscript writing. Moreover, I would also like to thank both respected reviewers, Prof. Jouni Aspi and Prof. Phillip Watts, for their valuable comments and suggestions.

Most importantly. I would like to express my deepest gratitude to my wife, Tanzeela, as I could not complete this journey without her selfless support and unshakeable trust in me. Finally, I thank my parents from the core of my heart for their continuous encouragement and prayers.

At the end, I wish to express my respect to Prof. Jon E. Brommer and Docent Sari Järvi, for conducting helpful annual training days and providing excellent educational opportunities for BGG students. I also would like to acknowledge the Ella and Georg Ehrnrooth Foundation and the University of Turku Foundation for their financial support. Lastly, I am also grateful to CSC - IT Center for Science, Finland for providing the computational resources.

April 2023
Freed Ahmad

List of References

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

Alama-Bermejo, G., Meyer, E., Atkinson, S. D., Holzer, A. S., Wiśniewska, M. M., Kolísko, M., & Bartholomew, J. L. 2020. Transcriptome-Wide Comparisons and Virulence Gene Polymorphisms of Host-Associated Genotypes of the Cnidarian Parasite *Ceratonova shasta* in Salmonids. *Genome biology and evolution*, 12, 1258-1276. doi: 10.1093/gbe/evaa109.

Aljanabi, S. M., & Martinez, I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, 25(22), 4692–4693.

Allendorf, F. W., Bassham, S., Cresko, W. A., Limborg, M. T., Seeb, L. W., & Seeb, J. E. 2015. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity*, 106(3), 217-227. doi:10.1093/jhered/esv015

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410. doi: 10.1016/S0022-2836(05)80360-2.

Angeli, A., Pinteala, M., Maier, S. S., Simionescu, B. C., Da'dara, A. A., Skelly, P. J., & Supuran, C. T. 2020. Sulfonamide Inhibition Studies of an α-Carbonic Anhydrase from *Schistosoma mansoni*, a Platyhelminth Parasite Responsible for Schistosomiasis. *International Journal of Molecular Sciences*, 21, 1842. doi: 10.3390/ijms21051842.

Arendt, J., & Reznick, D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology & Evolution*, 23(1), 26-32. doi:10.1016/j.tree.2007.09.011

Arkush, K. D., & Hedrick, R. P. 1990. Experimental transmission of PKX, the causative agent of proliferative kidney disease, to three species of Pacific salmon. *Journal of Applied Ichthyology*, 6(4), 237–243. https://doi.org/10.1111/J.1439-0426.1990.TB00584.X

Armstrong, C., Richardson, D. S., Hipperson, H., Horsburgh, G. J., Küpper, C., Percival-Alwyn, L., Clark, M., Burke, T., & Spurgin, L. G. 2018. Genomic associations with bill length and disease reveal drift and selection across island bird populations. *Evolution Letters*, 2(1), 22–36. https://doi.org/10.1002/EVL3.38

Arnold, S. J. 1983. Morphology, performance and fitness. *American Zoologist*, 23(2), 347-361. doi: 10.1093/icb/23.2.347

Aronesty, E. 2011. ea-utils: "Command-line tools for processing biological sequencing data"; https://github.com/ExpressionAnalysis/ea-utils

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10), 1294-1296. doi:10.1093/bioinformatics/btm108

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE*, 3(10), e3376. https://doi.org/10.1371/JOURNAL.PONE.0003376

Bell, D. A., Kovach, R. P., Robinson, Z. L., Whiteley, A. R., & Reed, T. E. 2021. The ecological causes and consequences of hard and soft selection. *Ecology Letters*, 24(7), 1505–1521. https://doi.org/10.1111/ELE.13754

Bettge, K., Segner, H., Burki, R., Schmidt-Posthaus, H., & Wahli, T. 2009. Proliferative kidney disease (PKD) of rainbow trout: temperature- and time-related changes of *Tetracapsuloides bryosalmonae* DNA in the kidney. *Parasitology*, 136, 615-625. doi:10.1017/S0031182009005800

Bishop, S. C., & Woolliams, J. A. 2014. Genomics and disease resistance studies in livestock. *Livestock Science*, 166, 190-198. doi:10.1016/j.livsci.2014.04.034

Bishop, S. C., Doeschl-Wilson, A., & Woolliams, J. A. 2012. Uses and implications of field disease data for livestock genomic and genetics studies. *Frontiers in Genetics*, 3 doi: 10.3389/fgene.2012.00114.

Bolger, A. M., Lohse, M., & Usadel, B. 2014. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633–2635. https://doi.org/10.1093/BIOINFORMATICS/BTM308

Braden, L. M., Prosperi-Porta, G., Kim, E., & Jones, S. R. M. 2010. *Tetracapsuloides bryosalmonae* in spawning pink salmon, *Oncorhynchus gorbuscha* (Walbaum), in the Quinsam River, British Columbia, Canada. *Journal of Fish Diseases*, 33(7), 617–621. https://doi.org/10.1111/J.1365-2761.2010.01145.X

Brieuc, M. S., Ono, K., Drinan, D. P., & Naish, K. A. 2015. Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology*, 24(11), 2729-2746. doi:10.1111/mec.13211

Brown, J. A., Thonney, J. P., Holwell, D., & Wilson, W. R. 1991. A comparison of the susceptibility of *Salvelinus alpinus* and *Salmo salar* ouananiche to proliferative kidney disease. *Aquaculture*, 96(1), 1–6. https://doi.org/10.1016/0044-8486(91)90134-S

Bruneaux, M., Visse, M., Gross, R., Pukk, L., Saks, L., & Vasemägi, A. 2016. Parasite infection and decreased thermal tolerance: impact of proliferative kidney disease on a wild salmonid fish in the context of climate change. *Functional Ecology*, 31, 216-226. doi: 10.1111/1365-2435.12701

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T. H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Peshkin, L., Tabin, C. J., Regev, A., Haas, B. J., & Whited, J. L. 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, 18, 762-776. doi: 10.1016/j.celrep.2016.12.063.

Bucke, D., Feist, S. W., & Clifton-Hadley, R. S. 1991. The occurrence of proliferative kidney disease (PKD) in cultured and wild fish: further investigations. *Journal of Fish Diseases*, 14(5), 583–588. https://doi.org/10.1111/J.1365-2761.1991.TB00614.X

Burgess, A., Rasouli, M., & Rogers, S. 2014. Stressing mitosis to death. *Frontiers in Oncology*, 4 140. doi: 10.3389/fonc.2014.00140

Cai, J., Yang, L., Wang, B., Huang, Y., Tang, J., Lu, Y., ... Jian, J. 2014. Identification of a novel N4BP1-like gene from grass carp (*Ctenopharyngodon idella*) in response to GCRV infection. *Fish & Shellfish Immunology*, 36(1), 223-228. doi:10.1016/j.fsi.2013.11.003

Campbell, N. R., LaPatra, S. E., Overturf, K., Towner, R., & Narum, S. R. 2014. Association mapping of disease resistance traits in rainbow trout using restriction site associated DNA sequencing. *G3: Genes, Genomes, Genetics*, 4(12), 2473–2481. http://doi.org/10.1534/g3.114.014621

Campbell-Staton, S. C., Cheviron, Z. A., Rochette, N., Catchen, J., Losos, J. B., & Edwards, S. V. 2017. Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard. *Science (American Association for the Advancement of Science)*, 357(6350), 495-498. doi: 10.1126/science.aam5512.

Carraro, L., Hartikainen, H., Jokela, J., Bertuzzo, E., & Rinaldo, A. 2018. Estimating species distribution and abundance in river networks using environmental DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 11724-11729. doi: 10.1073/pnas.1813843115.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 14912-14917. doi: 10.1073/pnas.1511468112.

Changklungmoa, N., Chaithirayanon, K., Cheukamud, W., Chaiwichien, A., Osotprasit, S., Samrit, T., Sobhon, P., & Kueakhai, P. 2018. Expression and characterization of glutathione peroxidase of the liver fluke, *Fasciola gigantica*. *Parasitology Research*, 117, 3487-3495. doi: 10.1007/s00436-018-6046-9.

Chapman, S. J., & Hill, A. V. S. 2012. Human genetic susceptibility to infectious disease. *Nature Reviews Genetics*, 13, 175-188. doi:10.1038/nrg3114

Cheung, V. G., & Spielman, R. S. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews Genetics*, 10:9, 10(9), 595–604. https://doi.org/10.1038/nrg2630

Chevrier, N. 2019. Decoding the body language of immunity: Tackling the immune system at the organism level. *Current Opinion in Systems Biology*, 18, 19-26. doi: 10.1016/j.coisb.2019.10.010

Chong, V. K., Fung, H. F., & Stinchcombe, J. R. 2018. A note on measuring natural selection on principal component scores. *Evolution Letters*, 2(4), 272-280. doi: 10.1002/evl3.63

Clifton-Hadley, R. S., Bucke, D., & Richards, R. H. 1984. Proliferative kidney disease of salmonid fish: a review. *Journal of Fish Diseases*, 7(5), 363–377. https://doi.org/10.1111/J.1365-2761.1984.TB01201.X

Clifton-Hadley, R. S., & Alderman, D. J. 1987. The effects of malachite green upon proliferative kidney disease. *Journal of Fish Diseases*, 10, 101-107. doi: 10.1111/j.1365-2761.1987.tb00725.x.

Cowell, A. N., & Winzeler, E. A. 2019. Advances in omics-based methods to identify novel targets for malaria and other parasitic protozoan infections. *Genome Medicine*, 11, 1-17. doi: 10.1186/s13073-019-0673-3.

Da'dara, A. A., Angeli, A., Ferraroni, M., Supuran, C. T., Skelly, P. J. 2019. Crystal structure and chemical inhibition of essential schistosome host-interactive virulence factor carbonic anhydrase SmCA. *Communications Biology*, 2, 1-11. doi: 10.1038/s42003-019-0578-0.

Dash, M., & Vasemägi, A. 2014. Proliferative kidney disease (PKD) agent *Tetracapsuloides bryosalmonae* in brown trout populations in Estonia. *Diseases of Aquatic Organisms*, 109, 139-148. doi: 10.3354/dao02731

Debes, P. V., Gross, R., & Vasemägi, A. 2017. Quantitative genetic variation in, and environmental effects on, pathogen resistance and temperature-dependent disease severity in a wild trout. *American Naturalist*, 190(2), 244-265.

Doeschl-Wilson, A., Bishop, S. C., Kyriazakis, I., & Villanueva, B. 2012. Novel methods for quantifying individual host response to infectious pathogens for genetic analyses. *Frontiers in Genetics*, 3, 266. doi: 10.3389/fgene.2012.00266

Donihue, C. M., Kowaleski, A. M., Losos, J. B., Algar, A. C., Baeckens, S., Buchkowski, R. W., … Herrel, A. 2020. Hurricane effects on neotropical lizards span geographic and phylogenetic scales. *Proceedings of the National Academy of Sciences*, 117(19), 10429-10434. doi: 10.1073/pnas.2000801117.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. 2009. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1), doi: 10.1186/1471-2105-10-48

Eizaguirre, C., & Lenz, T. L. 2010. Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. *Journal of Fish Biology*, 77(9), 2023–2047. https://doi.org/10.1111/J.1095-8649.2010.02819.X

Ellis, A. E., McVicar, A. H., & Munro, A. L. S. 1982. A preliminary report on the epidemiology of proliferative kidney disease in brown trout (*Salmo trutta*) and Atlantic salmon parr (*S. salar*) in Scotland. *Bulletin of the European Association of Fish Pathologists*, 2(1), 13–15.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5), e19379. doi:10.1371/journal.pone.0019379

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., ... Stefansson, K. 2008. Genetics of gene expression and its effect on disease. *Nature*, 452(7186), 423-428. doi: 10.1038/nature06758

Faber, M., Shaw, S., Yoon, S., de Paiva Alves, E., Wang, B., Qi, Z., Okamura, B., Hartikainen, H., Secombes, C. J., & Holland, J. W. 2021. Comparative transcriptomics and host-specific parasite gene expression profiles inform on drivers of proliferative kidney disease. *Scientific Reports*, 11:1, 11(1), 1–17. https://doi.org/10.1038/s41598-020-77881-7

Feist, S. W., Peeler, E. J., Gardiner, R., Smith, E., & Longshaw, M. 2002. Proliferative kidney disease and renal myxosporidiosis in juvenile salmonids from rivers in England and Wales. *Journal of Fish Diseases*, 25(8), 451-458. doi:10.1046/j.1365-2761.2002.00361.x

Foox, 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(1) doi:10.1186/s12864-015-2039-6.

Fraser, H. B. 2013. Gene expression drives local adaptation in humans. *Genome research*, 23(7), 1089-1096. doi: 10.1101/gr.152710.112

Fraser, H. B., Moses, A. M., & Schadt, E. E. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences*, 107(7), 2977-2982. doi: 10.1073/pnas.0912245107

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150-3152. doi: 10.1093/bioinformatics/bts565.

Futuyma, D. J. 2009. Evolution. Sinauer Associates.

Gandon, S., Jansen, V. A. A., & van Baalen, M. 2001. Host life history and the evolution of parasite virulence. *Evolution*, 55(5), 1056–1062. https://doi.org/10.1111/J.0014-3820.2001.TB00622.X

Ganesh, S. K., Zakai, N. A., van Rooij, F. J. A., Soranzo, N., Smith, A. V., Nalls, M. A., … Lin, J. P. 2009. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nature Genetics*, 41(11), 1191-1198. doi:10.1038/ng.466

Gilad, Y., Oshlack, A., & Rifkin, S. A. 2006. Natural selection on gene expression. *Trends in Genetics*, 22(8), 456-461

Giribet G., & Edgecombe, G. D. 2020. The Invertebrate Tree of Life. Princeton University Press, Princeton, USA

Gjerde, B., & Refstie, T. 1988. The effect of fin-clipping on growth rate, survival and sexual maturity of rainbow trout. *Aquaculture*, 73(1-4), 383-389. doi: 10.1016/0044-8486(88)90071-3

Gorgoglione, B., Bailey, C., & Ferguson, J. A. 2020. Proliferative kidney disease in Alaskan salmonids with evidence that pathogenic myxozoans may be emerging north. *International Journal for Parasitology*, 50(10–11), 797–807. https://doi.org/10.1016/J.IJPARA.2020.03.010

Götz, M. G., James, K. E., Hansell, E., Dvořák, J., Seshaadri, A., Sojka, D., Kopáček, P., McKerrow, J. H., Caffrey, C. R., & Powers, J. C. 2008. Aza-peptidyl Michael acceptors. A new class of potent and selective inhibitors of asparaginyl endopeptidases (legumains) from evolutionarily diverse pathogens. *Journal of Medicinal Chemistry*, 51, 2816-2832. doi: 10.1021/jm701311r.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644-652. doi: 10.1038/nbt.1883.

Grabner, D. S., & El-Matbouli, M. 2008. Transmission of *Tetracapsuloides bryosalmonae* (Myxozoa: Malacosporea) to *Fredericella sultana* (Bryozoa: Phylactolaemata) by various fish species. *Diseases of Aquatic Organisms*, 79(2), 133–139. https://doi.org/10.3354/DAO01894

Graham, A. L., Shuker, D. M., Pollitt, L. C., Auld, Stuart K. J. R., Wilson, A. J., & Little, T. J. 2010. Fitness consequences of immune responses: Strengthening the empirical framework for ecoimmunology. *Functional Ecology*, 25(1), 5-17. doi: 10.1111/j.1365-2435.2010.01777.x

Groen, S. C., Ćalić, I., Joly-Lopez, Z., Platts, A. E., Choi, J. Y., Natividad, M., ... Purugganan, M. D. 2020. The strength and pattern of natural selection on gene expression in rice. *Nature*, 578(7796), 572-576. doi: 10.1038/s41586-020-1997-2

Guo, Z., González, J. F., Hernandez, J. N., McNeilly, T. N., Corripio-Miyar, Y., Frew, D., ... Li, R. W. 2016. Possible mechanisms of host resistance to *Haemonchus contortus* infection in sheep breeds native to the Canary Islands. *Scientific Reports*, 6(1), 26200. doi: 10.1038/srep26200

Hahn, J., Seeber, F., Kolodziej, H., Ignatius, R., Laue, M., Aebischer, T., & Klotz, C. 2013. High sensitivity of *Giardia duodenalis* to tetrahydrolipstatin (orlistat) in vitro. *PloS one*, 8(8), e71597. doi: 10.1371/journal.pone.0071597.

Hartigan, A., Kosakyan, A., Pecková, H., Eszterbauer, E., & Holzer, A. S. 2020. Transcriptome of *Sphaerospora molnari* (Cnidaria, Myxosporea) blood stages provides proteolytic arsenal as potential therapeutic targets against sphaerosporosis in common carp. *BMC genomics*, 21, 1-404. Doi: 10.1186/s12864-020-6705-y.

Hartikainen, H., Filippenko, D., Okamura, B., & Vasemägi, A. 2015. First microsatellite loci of the myxozoan parasite *Tetracapsuloides bryosalmonae*, the causative agent of proliferative kidney disease. *Diseases of Aquatic Organisms*, 113, 85-88 doi: 10.3354/dao02833

Hartikainen, H., Gruhl, A., & Okamura, B. 2014. Diversification and repeated morphological transitions in endoparasitic cnidarians (Myxozoa: Malacosporea). *Molecular Phylogenetics and Evolution*, 76(1), 261-269. doi: 10.1016/j.ympev.2014.03.010

Hecht, B. C. Campbell, N. R. Holecek, D. E. & Narum, S. R. 2013. Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Molecular Ecology*, 22, 3061-3076. doi:10.1111/mec.12082

Hedrick, R. P., Kent, M. L., Rosemark, R., & Manzer, D. 1984. Proliferative Kidney Disease (PKD) In Pacific salmon and steelhead trout. *Journal of the World Mariculture Society*, 15(1–4), 318–325. https://doi.org/10.1111/J.1749-7345.1984.TB00166.X

Hedrick, R. P., MacConnell, E. & de Kinkelin, P. 1993. Proliferative kidney disease of salmonid fish. *Annual Review of Fish Diseases,* 3, 277-290. doi:10.1016/0959-8030(93)90039-E

Hedrick, R. P., Baxa, D. V., De Kinkelin, P., & Okamura, B. 2004. Malacosporean-like spores in urine of rainbow trout react with antibody and DNA probes to *Tetracapsuloides bryosalmonae*. *Parasitology Research*, 92(1), 81-88. doi: 10.1007/s00436-003-0986-3.

Henderson, M., & Okamura, B. 2004. The phylogeography of salmonid proliferative kidney disease in Europe and North America. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1549), 1729–1736. https://doi.org/10.1098/RSPB.2004.2677

Henshaw, J. M., & Zemel, Y. 2016. A unified measure of linear and nonlinear selection on quantitative traits. *Methods in Ecology and Evolution*, 8(5), 604-614. doi: 10.1111/2041-210x.12685

Hess, J. E., Zendt, J. S., Matala, A. R., & Narum S. R. 2016. Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. *Proceedings of the Royal Society B: Biological Sciences*, 283, 20153064. doi:10.1098/rspb.2015.3064

Holland, J. W., Okamura, B., Hartikainen, H., & Secombes, C. J. 2011. A novel minicollagen gene links cnidarians and myxozoans. *Proceedings of the Royal Society B: Biological Sciences*, 278, 546-553. Royal Society doi: 10.1098/rspb.2010.1301.

Holliday, J. A., Wang, T., & Aitken, S. 2012. Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using Random Forest. *G3: Genes, Genomes, Genetics*, 2(9), 1085-1093. doi:10.1534/g3.112.002733

Horibata, Y., Okino, N., Ichinose, S., Omori, A., & Ito, M 2000. Purification, characterization, and cDNA cloning of a novel acidic endoglycoceramidase from the jellyfish, *Cyanea nozakii*. *Journal of Biological Chemistry*, 275, 31297-31304. doi: 10.1074/jbc.M003575200.

Horibata, Y., Sakaguchi, K., Okino, N., Iida, H., Inagaki, M., Fujisawa, T., Hama, Y., & Ito, M. 2004. Unique catabolic pathway of glycosphingolipids in a hydrozoan, *Hydra magnipapillata*, involving endoglycoceramidase. *Journal of Biological Chemistry*, 279, 33379-33389. doi: 10.1074/jbc.M401460200.

Hutchins, P. R., Sepulveda, A. J., Hartikainen, H., Staigmiller, K. D., Opitz, S. T., Yamamoto, R. M., … Okamura, B. 2021. (2021). Exploration of the 2016 Yellowstone River fish kill and proliferative kidney disease in wild fish populations. *Ecosphere*, 12(3), e03436. https://doi.org/10.1002/ECS2.3436

Jombart, T. 2008. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403-1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., & Balloux, F. 2010. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1), 94. doi: 10.1186/1471-2156-11-94

Jones, O. R., & Wang, J. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10, 551-555. doi:10.1111/j.1755-0998.2009.02787.x

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., … Kang, H. M. 2012. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *American Journal of Human Genetics,* 91(5), 839–848. http://doi.org/10.1016/j.ajhg.2012.09.004

Karn, M. N., & L. S. Penrose. 1951. Birth weight and gestation time in relation to maternal age, parity and infant survival. Annals of Eugenics, 16, 147–164. doi: 10.1111/J.1469-1809.1951.TB02469.X.

Kassahn, K. S., Crozier, R. H., Pörtner, H. O., & Caley, M. J. 2009. Animal performance and stress: Responses and tolerance limits at different levels of biological organisation. *Biological Reviews*, 84(2), 277-292. doi: 10.1111/j.1469-185X.2008.00073.x

Kause, A. 2011. Genetic analysis of tolerance to infections using random regressions: a simulation study. *Genetics Research*, 93(4), 291–302. https://doi.org/10.1017/S0016672311000176

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10, 845-858. doi: 10.1038/nprot.2015.053.

Kent, M. L., & Hedrick, R. P. 1986. Development of the PKX myxosporean in rainbow trout *Salmo gairdneri*. *Diseases of Aquatic Organisms*, 1:169-182

Kim, D., Langmead, B., & Salzberg, S. L., 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, 12, 357-360. doi: 10.1038/nmeth.3317.

King, M., & Wilson, A. 1975. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184), 107–116. https://doi.org/10.1126/science.1090005

Kingsolver, J. G., & Pfennig, D. W. 2007. Patterns and power of phenotypic selection in nature. *Bioscience*, 57(7), 561-572. doi: 10.1641/B570706

Kingsolver, J. G., Hoekstra, H. E., Hoekstra, J. M., Berrigan, D., Vignieri, S. N., Hill, C. E., ... Beerli, P. 2001. The strength of phenotypic selection in natural populations. *The American Naturalist*, 157(3), 245-261. doi: 10.1086/319193

Klapholz, B., & Brown, N. H. 2017. Talin - The master of integrin adhesions. *Journal of Cell Science*, 130, 2435-2446. doi: 10.1242/jcs.190991.

Kumar, G., Abd-Elfattah, A., Saleh, M., & El-Matbouli, M. 2013. Fate of *Tetracapsuloides bryosalmonae* (Myxozoa) after infection of brown trout *Salmo trutta* and rainbow trout *Oncorhynchus mykiss*. *Diseases of Aquatic Organisms*, 107(1), 9–18. https://doi.org/10.3354/DAO02665

Kutzer, M. A. M., & Armitage, S. A. O. 2016. Maximising fitness in the face of parasites: a review of host tolerance. *Zoology*, 119(4), 281-289. doi:10.1016/j.zool.2016.05.011

Lande, R., & Arnold, S. J. 1983. The measurement of selection on correlated characters. *Evolution*, 37(6), 1210. doi: 10.2307/2408842

Langfelder, P., & Horvath, S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), doi: 10.1186/1471-2105-9-559

Langmead, B., & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357-359. doi: 10.1038/nmeth.1923.

Lauringson, M., Ozerov, M. Y., Lopez, M. E., Wennevik, V., Niemelä, E., Vorontsova, T. Y., & Vasemägi, A. 2022. Distribution and prevalence of the myxozoan parasite *Tetracapsuloides bryosalmonae* in northernmost Europe: analysis of three salmonid species. *Diseases of Aquatic Organisms*, 151, 37–49. https://doi.org/10.3354/DAO03688

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. 2014. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), R29. doi: 10.1186/gb-2014-15-2-r29

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., ... Carey, V. J. 2013. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8), e1003118. doi: 10.1371/journal.pcbi.1003118

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6), 882-883. doi: 10.1093/bioinformatics/bts034

Leitwein, M., Guinand, B., Pouzadoux, J., Desmarais, E., Berrebi, P., & Gagnaire, P. 2017. A dense brown trout (*Salmo trutta*) linkage map reveals recent chromosomal rearrangements in the Salmo genus and the impact of selection on linked neutral diversity. *G3: Genes, Genomes, Genetics*, 7(4), 1365-1376. doi:10.1534/g3.116.038497

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, W. & Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658–1659. doi: 10.1093/bioinformatics/btl158

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

Limborg, M. T., Seeb, L. W., & Seeb, J. E. 2016. Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular Ecology*, 25(10), 2117-2129. doi:10.1111/mec.13601

Limborg, M. T., Seeb, L. W., & Seeb, J. E. 2016. Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular Ecology*, 25(10), 2117–2129. https://doi.org/10.1111/MEC.13601

Little, T. J., Shuker, D. M., Colegrave, N., Day, T., & Graham, A. L. 2010. The coevolution of virulence: tolerance in perspective. *PLoS Pathogens*, 6(9), e1001006. doi:10.1371/journal.ppat.1001006

Lohman, B. K., Weber, J. N., & Bolnick, D. I. 2016. Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. *Molecular Ecology Resources*, 16(6), 1315-1321. doi: 10.1111/1755-0998.12529

Longshaw, M., Le Deuff, R. M., Harris, A. F., & Feist, S. W. 2002. Development of proliferative kidney disease in rainbow trout, *Oncorhynchus mykiss* (Walbaum), following short-term exposure to *Tetracapsula bryosalmonae* infected bryozoans. *Journal of Fish Diseases*, 25, 443-449. doi: 10.1046/j.1365-2761.2002.00353.x.

Love, M. I., Huber, W., & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15,. doi: 10.1186/s13059-014-0550-8.

Macconnell E., & Peterson J. E. 1992. Proliferative Kidney Disease in Feral Cutthroat Trout from a Remote Montana Reservoir: A First Case. *Journal of Aquatic Animal Health*, 4(3), 182–187.

Macqueen, D. J., & Johnston, I. A. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778), 20132881. http://doi.org/10.1098/rspb.2013.2881

Marcogliese, D. J. 2008. The impact of climate change on the parasites and infectious diseases of aquatic animals. *Revue Scientifique Et Technique*, 27(2), 467-84.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 10. doi: 10.14806/ej.17.1.200

Martín-Hernández, R., Higes, M., Sagastume, S., Juarranz, Á, Dias-Almeida, J., Budge, G. E., ... Boonham, N. 2017. Microsporidia infection impacts the host cell's cycle and reduces host cell apoptosis. *PloS One*, 12(2), e0170183. doi: 10.1371/journal.pone.0170183

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D. and Emerson, B. C. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15: 28–41. https://doi.org/10.1111/1755-0998.12291

Mazé-Guilmo, E., Loot, G., Páez, D. J., Lefèvre, T., & Blanchet, S. 2014. Heritable variation in host tolerance and resistance inferred from a wild host–parasite system. *Proceedings of the Royal Society B: Biological Sciences*, 281, 20132567. doi:10.1098/rspb.2013.2567

McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P., & Sajid, M. 2006. Proteases in parasitic diseases. *Annual review of pathology*, 1, 497-536. doi: 10.1146/annurev.pathol.1.110304.100151.

McNeill, H. 2009. Planar cell polarity and the kidney. *Journal of the American Society of Nephrology*, 20(10), 2104-2111. doi:10.1681/ASN.2008111173

Miller, K. M., Li, S., Kaukinen, K. H., Ginther, N., Hammill, E., Curtis, J. M. R., ... Farrell, A. P. 2011. Genomic signatures predict migration and spawning failure in wild Canadian salmon. *Science*, 331(6014), 214-217. doi: 10.1126/science.1196901

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2), 240. https://doi.org/10.1101/GR.5681207

Mo, T. A., & Jørgensen, A. 2016. A survey of the distribution of the PKD-parasite *Tetracapsuloides bryosalmonae* (Cnidaria: Myxozoa: Malacosporea) in salmonids in Norwegian rivers - additional information gleaned from formerly collected fish. *Journal of Fish Diseases*, 40(5), 621-627. doi:10.1111/jfd.12542

Moen, T., Torgersen, J., Santi, N., Davidson, W. S., Baranski, M., Ødegård, J., … Lien, S. 2015. Epithelial cadherin determines resistance to infectious pancreatic necrosis virus in Atlantic salmon. *Genetics*, 200(4), 1313-1326. doi:10.1534/genetics.115.175406

Moll, P., Ante, M., Seitz, A., & Reda, T. 2014. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nature Methods*, 11(12), i-iii. doi: 10.1038/nmeth.f.376

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35, W182-W185. doi: 10.1093/nar/gkm321.

Morris, D. J., Adams, A., & Richards, R. H. 1997. Studies of the PKX Parasite in Rainbow Trout via Immunohistochemistry and Immunogold Electron Microscopy. *Journal of Aquatic Animal Health*, 9(4), 265–272. https://doi.org/10.1577/1548-8667(1997)009<0265:SOTPPI>2.3.CO;2

Morris, D. J. 2010. Cell formation by myxozoan species is not explained by dogma. *Proceedings of the Royal Society B: Biological Sciences*, 277, 2565-2570. doi: 10.1098/rspb.2010.0282.

Morris, D. J., & Adams, A. 2008. Sporogony of Tetracapsuloides bryosalmonae in the brown trout Salmo trutta and the role of the tertiary cell during the vertebrate phase of myxozoan life cycles. *Parasitology*, 135, 1075-1092. doi: 10.1017/S0031182008004605.

Ødegård, J., Baranski, M., Gjerde, B., & Gjedrem, T. 2011. Methodology for genetic evaluation of disease resistance in aquaculture species: challenges future prospects. *Aquaculture Research*, 42, 103-114. doi:10.1111/j.1365-2109.2010.02669.x

Ogle, D. H.,. 2017. FSA: Fisheries stock analysis. R package 0.8.17.

Okamura, B., Anderson, C. L., Longshaw, M., Feist, S. W., & Canning, E. U. 2001. Patterns of occurrence and 18s rdna sequence variation of pkx (*Tetracapsula bryosalmonae*), the causative agent of salmonid proliferative kidney disease. *Journal of Parasitology*, 87, 379-385. doi: 10.1645/0022-3395(2001)087[0379:pooars]2.0.co;2.

Okamura, B., Hartikainen, H., Schmidt-Posthaus, H., & Wahli, T. 2011. Life cycle complexity, environmental change and the emerging status of salmonid proliferative kidney disease. i, 56, 735-753. doi:10.1111/j.1365-2427.2010.02465.x

Ovat, A., Muindi, F., Fagan, C., Brouner, M., Hansell, E., Dvořák, J., … Powers, J. C. 2009. Aza-peptidyl michael acceptor and epoxide inhibitors - Potent and selective inhibitors of *Schistosoma mansoni* and *Ixodes ricinus* legumains (asparaginyl endopeptidases). *Journal of Medicinal Chemistry*, 52, 7192-7210. doi: 10.1021/jm900849h.

Palomar, G., Ahmad, F., Vasemägi, A., Matsuba, C., Nicieza, A. G., & Cano, J. M. 2017. Comparative high-density linkage mapping reveals conserved genome structure but variation in levels of heterochiasmy and location of recombination cold spots in the common frog. *G3: Genes, Genomes*, Genetics, 7(2), 637–645. http://doi.org/10.1534/g3.116.036459

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T., Mendell, J. T., & Salzberg, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, 33, 290-295. doi: 10.1038/nbt.3122.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. 2012. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, 7(5), e37135. https://doi.org/10.1371/JOURNAL.PONE.0037135

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. 2004. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612. https://doi.org/10.1002/JCC.20084

Pina-Vazquez, C, Reyes-Lopez, M, Ortiz-Estrada, G, de la Garza, M and Serrano-Luna, J 2012. Host-Parasite Interaction: Parasite-Derived and -Induced Proteases That Degrade Human Extracellular Matrix. Journal of parasitology research 2012, 748206-24. doi: 10.1155/2012/748206.

Piriatinskiy, G, Atkinson, SD, Park, S, Morgenstern, D, Brekhman, V, Yossifon, G, Bartholomew, JL and Lotan, T 2017. Functional and proteomic analysis of Ceratonova shasta (Cnidaria: Myxozoa) polar capsules reveals adaptations to parasitism. Scientific Reports 7, 1-10. doi: 10.1038/s41598-017-09955-y.

Plehn, M. 1924. Praktikum der Fischkrankheiten. Schweizerbart´sche Verlagsbuchhandlung G.m.b.H.

Poland, J. A., Bradbury, P. J., Buckler, E. S., & Nelson, R. J. 2011. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences*, 108(17), 6893-6898. doi:10.1073/pnas.1010894108

Poulin, R., & Randhawa, H. S. (2015) 2015. Evolution of parasitism along convergent lines: from ecology to genomics. *Parasitology*, 142(S1), S6–S15. https://doi.org/10.1017/S0031182013001674

Price, P. D., Palmer Droguett, D. H., Taylor, J. A., Kim, D. W., Place, E. S., Rogers, T. F., … Wright, A. E. 2022. Detecting signatures of selection on gene expression. *Nature Ecology & Evolution*, 6(7), 1035–1045. https://doi.org/10.1038/s41559-022-01761-8

Pukk, L., Ahmad, F., Hasan, S., Kisand, V., Gross, R., & Vasemägi, A. 2015. Less is more: extreme genome complexity reduction with ddRAD using Ion Torrent semiconductor technology. *Molecular Ecology Resources*, 15, 1145-1152. doi:10.1111/1755-0998.12392.

Quinn, T. P., Hodgson, S., Flynn, L., Hilborn, R., & Rogers, D. E. 2007. Directional selection by fisheries and the timing of sockeye salmon (*Oncorhynchus nerka*) migrations. Ecological applications : a publication of the Ecological Society of America, 17(3), 731–739. doi: 10.1890/06-0771.

Råberg, L. 2014. How to live with the enemy: understanding tolerance to parasites. *PLoS Biology*, 12(11), e1001989. doi:10.1371/journal.pbio.1001989

Råberg, L., Graham, A. L., & Read, A. F. 2009. Decomposing health: tolerance and resistance to parasites in animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1513), 37-49. doi:10.1098/rstb.2008.0184

Råberg, L., Sim, D., & Read, A. F. 2007. Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. *Science*, 318(5851), 812-814. doi:10.1126/science.1148526

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., & Sabeti, P. C. Richter, D. J., ... Lander, E. S. 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834), 199-204. doi:10.1038/35075590

Richards, J. K., Friesen, T. L., & Brueggeman, R. S. 2017. Association mapping utilizing diverse barley lines reveals net form net blotch seedling resistance/susceptibility loci. *Theoretical and Applied Genetics*, 130(5), 915-927. doi:10.1007/s00122-017-2860-1

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47. doi: 10.1093/nar/gkv007

Robbins, J. 2016. Tiny Invader, Deadly to Fish, Shuts Down a River in Montana. Retrieved December 19, 2022, from https://www.nytimes.com/2016/08/24/us/tiny-parasite-invader-deadly-to-fish-shuts-down-yellowstone-river-in-montana.html

Santure, A. W., & Garant, D. 2018. Wild GWAS—association mapping in natural populations. *Molecular Ecology Resources*, 18(4), 729–738. https://doi.org/10.1111/1755-0998.12901

Saulnier, D., Brémont, M., & De Kinkelin, P. 1996. Cloning, sequencing and expression of a cDNA encoding an antigen from the Myxosporean parasite causing the proliferative kidney disease of salmonid fish. *Molecular and Biochemical Parasitology*, 83, 153-161. doi: 10.1016/S0166-6851(96)02761-2.

Schulze, S., Schleicher, J., Guthke, R., & Linde, J. 2016. How to Predict Molecular Interactions between Species? *Frontiers in Microbiology*, 7, 442. doi: 10.3389/fmicb.2016.00442.

Seagrave, C. P., Bucke, D., Hudson, E. B., & Mcgregor, D. 1981. A survey of the prevalence and distribution of proliferative kidney disease (PKD) in England and Wales. *Journal of Fish Diseases*, 4(5), 437–439. https://doi.org/10.1111/J.1365-2761.1981.TB01155.X

Siepielski, A. M., Morrissey, M. B., Buoro, M., Carlson, S. M., Caruso, C. M., Clegg, S. M., ... MacColl, A. D. C. 2017. Precipitation drives global variation in natural selection. *Science*, 355(6328), 959-962. doi: 10.1126/science.aag2773

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210-3212. doi: 10.1093/bioinformatics/btv351.

Siqueira-Neto, J. L., Debnath, A., McCall, L., Bernatchez, J. A., Ndao, M., Reed, S. L., & Rosenthal, P. J. 2018. Cysteine proteases in protozoan parasites. *PLoS neglected tropical diseases*, 12, e0006512. doi: 10.1371/journal.pntd.0006512.

Skovgaard, A., & Buchmann, K. 2012. Tetracapsuloides bryosalmonae and PKD in juvenile wild salmonids in Denmark. *Diseases of Aquatic Organisms*, 101(1), 33-42. doi: 10.3354/dao02502.

Smith, T. B. 1993. Disruptive selection and the genetic basis of bill size polymorphism in the African Finch Pyrenestes". Nature 1993 363:6430 363: 618–620. doi: 10.1038/363618a0.

Sobociński, B., Huusko, A., & Vasemägi, A. 2018. First record of Tetracapsuloides bryosalmonae (Myxozoa; Malacosporea) in European whitefish (*Coregonus lavaretus*). *Bulletin of the European Association of Fish Pathologists*, 38(4), 171–176.

Sterud, E., Forseth, T., Ugedal, O., Poppe, T., Jørgensen, A., Bruheim, T., ... Mo, T. 2007. Severe mortality in wild Atlantic salmon *Salmo salar* due to proliferative kidney disease (PKD) caused by *Tetracapsuloides bryosalmonae* (Myxozoa). *Diseases of Aquatic Organisms*, 77, 191-198. doi:10.3354/dao01846

Svavarsdóttir, F. R., Freeman, M. A., Antonsson, P., Árnason, F., & Kristmundsson, Á. 2021. The presence of sporogonic stages of Tetracapsuloides bryosalmonae in Icelandic salmonids detected using in situ hybridisation. *Folia Parasitologica*, 68, 2021.020. https://doi.org/10.14411/FP.2021.020

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... von Mering, C. 2015. STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1), D447-D452. doi: 10.1093/nar/gku1003

Tops, S., Lockwood, W., & Okamura, B. 2006. Temperature-driven proliferation of *Tetracapsuloides bryosalmonae* in bryozoan hosts portends salmonid declines. *Diseases of Aquatic Organisms*, 70(3), 227-236. doi: 10.3354/dao070227

Tragante, V., Barnes, M. R., Ganesh, S. K., Lanktree, M. B., Guo, W., Franceschini, N., ... Keating, B. J. 2014. Gene-centric meta-analysis in 87,736 individuals of European ancestry identifies multiple blood-pressure-related loci. *American Journal of Human Genetics*, 94(3), 349-360. doi:10.1016/j.ajhg.2013.12.016

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9

Van Valen, L. 1973. A new evolutionary law. *Evolutionary Theory*, 1, 1–30.

Vander Wal, E., Garant, D., Calmé, S., Chapman, C. A., Festa-Bianchet, M., Millien, V., Rioux-Paquette, S., & Pelletier, F. 2014. Applying evolutionary concepts to wildlife disease ecology and management. *Evolutionary Applications*, 7(7), 856-868. doi:10.1111/eva.12168

Vasemägi, A., Nousiainen, I., Saura, A., Vähä, J. P., Valjus, J., & Huusko, A. 2017. First record of proliferative kidney disease agent *Tetracapsuloides bryosalmonae* in wild brown trout and European grayling in Finland. *Diseases of Aquatic Organisms*, 125(1), 73-78. doi:10.3354/dao03126

Vermelho, A. B., Capaci, G. R., Rodrigues, I. A., Cardoso, V. S., Mazotto, A. M., & Supuran, C. T. 2017. Carbonic anhydrases from Trypanosoma and Leishmania as anti-protozoan drug targets. *Bioorganic and Medicinal Chemistry*, 25, 1543-1555. doi: 10.1016/j.bmc.2017.01.034.

Videvall, E., Cornwallis, C. K., Ahrén, D., Palinauskas, V., Valkiūnas, G., & Hellgren, O. 2017. The transcriptome of the avian malaria parasite *Plasmodium ashfordi* displays host-specific gene expression. *Molecular Ecology*, 26, 2939-2958. doi: 10.1111/mec.14085.

Visscher, P. M., Hill, W. G., & Wray, N. R. 2008. Heritability in the genomics era – concepts and misconceptions. *Nature Review Genetics*, 9(4), 255-266. doi:10.1038/nrg2322

Wahli, T., Knuesel, R., Bernet, D., Segner, H., Pugovkin, D., Burkhardt-Holm, P., Escher, M., & Schmidt-Posthaus, H. 2002. Proliferative kidney disease in Switzerland: current state of knowledge. *Journal of Fish Diseases*. 25, 491-500. doi: 10.1046/j.1365-2761.2002.00401.x.

Wallace, B. 1975. Hard and Soft Selection Revisited. *Evolution*, 29(3), 465. https://doi.org/10.2307/2407259

Wang, M., & Xu, S. 2019. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity*, 123(3), 287–306. https://doi.org/10.1038/s41437-019-0205-3

Winter, D. J. 2017. Rentrez: An R package for the NCBI eUtils API. *The R Journal*, 9(2), 520. Doi: 10.32614/rj-2017-058.

Wood, D. E., Lu, J., & Langmead, B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20, 257. doi: 10.1186/s13059-019-1891-0.

Wright, M. N., & Ziegler, A. 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), doi: 10.18637/jss.v077.i01

Yahalomi, D., Atkinson, S. D., Neuhof, M., Chang, E. S., Philippe, H., Cartwright, P., Bartholomew, J. L., & Huchon, D. 2020. A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome. *Proceedings of the National Academy of Sciences*, 117, 5358-5363. doi: 10.1073/pnas.1909907117.

Yang, Y., Xiong, J., Zhou, Z., Huo, F., Miao, W., Ran, C., Liu, Y., Zhang, J., Feng, J., Wang, M., Wang, M., Wang, L., & Yao, B. 2014. The genome of the myxosporean *Thelohanellus kitauei* shows adaptation sto nutrient acquisition within its fish host. *Genome Biology and Evolution*, 6, 3182-3198. doi: 10.1093/gbe/evu247.

Yu, L., Lorenz, A., Rutkoski, J., Singh, R. P., Bhavani, S., Huerta-Espino, J., & Sorrells, M. E. 2011. Association mapping and gene–gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. *Theoretical and Applied Genetics*, 123(8), 1257-1268. doi:10.1007/s00122-011-1664-y

Zippin, C. 1958. The removal method of population estimation. *The Journal of Wildlife Management*, 22(1), 82. doi: 10.2307/3797301

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2023

ISBN 978-951-29-9232-4 (PRINT)
ISBN 978-951-29-9233-1 (PDF)
ISSN 0082-6979 (Print)
ISSN 2343-3183 (Online)