Development of Genomic Resources and Tools for Precision Farming of Pikeperch through High-Throughput Sequencing and Computational Genomics

DISSERTATION

for the degree of

Doctor of Engineering (Dr.-Ing.)

Doctoral field: Bioinformatics

Submitted in the

Faculty of Agricultural and Environmental Sciences

at the University of Rostock, Germany

Universität Rostock

Traditio et Innovatio

by

Julien A. Nguinkal

born in Ngaoundéré, Cameroon

Rostock, December 2021 ©

https://doi.org/10.18453/rosdok_id00004127

Dissertation Referees

Reviewer #1 : **Prof. Dr. Tom Goldammer**
Dept. of Molecular Biology and Genetics of Fish
Faculty of Agricultural and Environmental Sciences
University of Rostock, Germany

Reviewer #2: **Prof. Dr. Olaf Wolkenhauer**
Dept. of Systems Biology and Bioinformatics
Faculty of Computer Science and Electrical Engineering
University of Rostock, Germany

Reviewer #3: **Prof. Dr. Stefan Simm**
Institute of Bioinformatics
University Medicine Greifswald
University of Greifswald, Germany

Defense Date: December 2022

La seule manière de faire du bon travail, c'est d'aimer ce que vous faites.

–Steve Jobs

Acknowledgements

Many thanks to Prof. Tom Goldammer for providing me with the great opportunity and funding to engage in the pikeperch genome project and pursue this work, as well as for the general academic supervision of this PhD project. I am also very grateful for the motivating scientific discussions we had during the last few years and for the freedom to develop my own ideas and work independently on them.

I'd like to express my gratitude to Prof. Olaf Wolkenhauer for accepting me as an external PhD student and integrating me into his lab, the Systems Biology and Bioinformatics department (SBI). I have learned a lot from the scientific discussions and group meetings, as well as from informal discussions and seminars I had the opportunity to attend. Thanks to my SBI fellow colleagues, who provided me with helpful support and constructive input when I felt stuck in my research.

Special thanks to my FBN colleagues and co-authors, including Ronald, Alex, Marieke, Nadine, Lidia, Fabio, Dörte, and Frieder, for the unprecedented collegial atmosphere and continuous support throughout the years, including scientific, technical, and private or administrative matters. I cannot forget to acknowledge Ingrid Hennings, Luisa Falkenthal and Brigitte Schöpel (FBN, Dummerstorf), for their technical assistance in molecular biology analyses.

This work would have nearly been impossible without the gracious support of the German Network for Bioinformatics Infrastructure (de.NBI), which provided me with a dedicated instance on their cloud infrastructure. I was then able to perform resource-intensive tasks such as genome assembly or read mapping easily. Therefore, I would like to thank the whole de.NBI team for the great technical support.

My family and friends were a crucial support during this hard journey. I'd like to thank my uncle Franklin in particular for his unwavering support of my education over the years. He inspired me to study bioinformatics, as it was just becoming a hot research topic in the late 2000s. I dedicate this achievement to my beloved mum.

Sincere thanks to all of you!

Abstract

Computational genomics develops and leverages genomic tools and resources to address fundamental questions, e.g., in biomedical, livestock, or aquaculture research. Aquaculture genomics, for instance, addresses crucial questions, such as maker-assisted selection, development of species-specific breeding programs, and stock management to ensure efficient and sustainable production. Genome and transcriptome sequencing, assembly and annotation, as well as genotyping of genetic variants have been dramatically enhanced in the last decade due to marked improvements in computational methods and emerging next-generation sequencing technologies. Pikeperch (*Sander lucioperca*) is a fresh and brackish water percid fish natively inhabiting the northern Eurasian hemisphere. It has recently gained a high commercial relevance and is emerging as a promising candidate for intensive inland aquaculture. However, the successful positioning of pikeperch in large-scale aquaculture requires a comprehensive understanding of its genome structure and organization, and the identification of critical genes that can be used as genetic parameters for its optimal domestication and smart breeding.

This thesis provides the first genomic tools and resources to enhance pikeperch's innovative farming, optimal domestication, and adaption into modern intensive aquaculture systems. These primarily include an annotated high-quality chromosome-level assembly, a reference transcriptome, along the gene expression atlas based on multiple tissues and individuals. The about 900 Mb genome assembly with 24 chromosomes was generated combining second and third-generation high-throughput sequencing methods, high-density SNP-based genetic linkage maps, and by integrating different bioinformatics approaches. In addition, the transcriptomics data were used to refine and improve the annotation of the predicted ~24,000 protein-coding genes, and to analyze the expression and evolution of crucial stress and development genes associated with fish performance and welfare. In a subsequent application of these genomics tools, the pikeperch genome was utilized as a reference for comparative genomics analyses among *Percidae* species, including positive selection, gene duplication, and phylogenetic analyses. Finally, population genetics analyses in a cohort of domesticated individuals were performed to establish the landscape of genetic variations of pikeperch including structural variants (SVs), short tandem repeats (STRs) and single-nucleotide polymorphism (SNP).

The reported genomic tools and resources provide groundbreaking data for genomic-based breeding studies targeting phenotypic and production traits in pikeperch aquaculture. Both genomic and transcriptomic data are valuable resources to investigate molecular hallmarks for phenotypic characteristics of this species. Moreover, these findings lay the foundation for addressing critical issues in genomics-informed pikeperch farming.

Zusammenfassung

Computergestützte Genomik entwickelt und nutzt genomische Werkzeuge und Ressourcen, um grundlegende Fragen zu beantworten, z. B. in der biomedizinischen, Nutztier- oder Aquakulturforschung. Die Aquakultur-Genomik befasst sich beispielsweise mit entscheidenden Fragen wie Markergestützte Selektion, Entwicklung artspezifischer Zuchtprogramme und Bestandsmanagement, um eine effiziente und nachhaltige Produktion zu gewährleisten. In den letzten Jahren wurden Genom-und Transkriptomsequenzierung, Assemblierung und Annotation sowie Genotypisierung genetischer Varianten aufgrund deutlicher Verbesserungen der Rechenmethoden und der aufkommenden Sequenzierungstechnologie der nächsten Generation erheblich verbessert. Zander (*Sander lucioperca*) ist ein Süßwasserfisch, der ursprünglich die nördliche eurasische Hemisphäre bewohnt. Es hat in letzter Zeit eine hohe kommerzielle Bedeutung erlangt und entwickelt sich als vielversprechender Kandidat für die Diversifizierung der Binnenland-Aquakultur in Europa. Die erfolgreiche Positionierung des Zanders in einer wirtschaftlich ertragreichen Aquakultur erfordert jedoch das umfassende Verständnis seiner Genomstruktur und -organisation sowie grundlegende Erkentnisse über kritische Gene, die als genetische Parameters für seine optimale Domestikation und Züchtung genutzt werden können. Doch diese wichtigen Erkenntnisse und grundlegenden genomischen Daten sind bisher für diese Fischart trotz ihrer hohen kommerziellen Relevanz nicht verfügbar.

Diese Arbeit liefert die ersten genomischen Werkzeuge und Ressourcen, um die innovative Zucht des Zanders, seine optimale Domestikation und Anpassung an moderne intensive Aquakultursysteme zu untersuchen. Dazu gehören in erster Linie eine hochwertige Assemblierung seines Genoms auf Chromosomenebene, ein Referenztranskriptom sowie ein Genexpressionsatlas, der auf mehreren Geweben und Individuen basiert. Das etwa 900 Mb große Genom mit insgesamt 24 Chromosomen wurde durch Kombination von Hochdurchsatz-Sequenzierungsmethoden der zweiten und dritten Generation, hochdichten SNP-basierten genetischen Kopplungskarten und verschiedenen bioinformatischen Methoden und Ansätzen generiert. Die Expression und Evolution von entscheidenden Stress- und Entwicklungsgenen, die mit der Leistungsfähigkeit und dem Wohlergehen von Fischen in verbunden sind, wurden ermilttelt. In einer nachfolgenden Anwendung dieser genomischen Werkzeuge wurde das Zandergenom als Referenz für vergleichende Genomanalysen in Barschartigen verwendet, um positive Selektion, Genduplikation und phylogenetische

Orthologie-Inferenz zu bestimmen. Schließlich wurden populationsgenetische Analysen in einer Kohorte domestizierter Individuen (N=394) durchgeführt, um eine Landkarte genetischer Variationen beim Zander zu erstellen, einschließlich struktureller Variationen (SVs), *Short Tandem Repeats* (STRs) und Single-Nukleotid-Polymorphismen (SNP).

Die hier vorgestellten genomischen Werkzeuge und Ressourcen liefern erste grundlegende Daten und Ansatzpunkte für genombasierte Züchtungsstudien, die auf phänotypische und leistungsbezogene Merkmale in der Zanderaquakultur abzielen. Beide Referenzgenom- und Transkriptomsequenzen bieten wertvolle Ressourcen, um molekulare Kennzeichen für wichtige phänotypische Eigenschaften dieser Spezies zu untersuchen. Außerdem sind diese Ergebnisse hilfreich, um kritische Probleme in der genombasierten Aquakultur von Zander anzugehen sowie die Effizienz der Marker-basierten Selektion, Krankheitsresistenz und andere kommerzielle Merkmale zu erforschen und zu verbessern.

Contents

List of Figures

List of Tables

List of Abbreviations

BUSCO	Benchmarking Universal Single-Copy Orthologs
CCS	Circular Consensus Sequence
CLR	Circular Long Reads
DGB	De Bruijn Graph
DNA	Deoxyribonucleic Acid
FAIR	Findability, Accessibility, Interoperability, and Reusability
GWAS	Genome-wide Association Studies
HGP	Human Genome Project
HiFi	High Fidelity reads
HTS	High-throughput Sequencing
LD	Linkage Disequilibrium
LINEs	Long Interspersed Nuclear Elements
LTR	Long Terminal Repeat
Mb	Mega base pairs
MITEs	Miniature Inverted-repeat Transposable Elements
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
ONT	Oxford Nanopore Technologies
PMC	PubMed Central
RNA	Ribonucleic Acid
SD	Segmental Duplications
SDs	Segmental Duplications
SINEs	Short Interspersed Nuclear Elements
SMART	Selection with Markers and Advanced Reproductive Technologies
SMRT	Single Molecule Real-time Sequencing
SNPs	Single Nucleotide Polymorphism
STRs	Short Tandem Repeats
SVs	Structural Variants
TE	Transposable Elements
TPM	Transcripts Per Million

1. General Introduction

A genome is the complete set of nucleotides (i.e., A, C, G, and T), the building blocks that make up all chromosomes of an organism or species. Genomics aims to characterize and analyze the structure and function of entire genomes employing computational and statistical approaches commonly termed bioinformatics. Nowadays, sequencing the complete genome of any organism is made possible through substantial progress in sequencing technology and bioinformatics. Here, I will introduce the general context of this thesis, highlight its significance, and outline a few significant developments in the field of genome sequencing and assembly for enhancing aquaculture research.

1.1. Background and Motivation

Genome sequencing and analysis technologies are developed to study genomes' structure, expression, and function to bring humans medical, economic, or nutritional benefits. Deciphering the genome sequence of an organism produces large amounts of genomic data. Bioinformatic methods are essential to mine, analyze, and transform these big genomic data into meaningful and usable insights for different applied life science branches. The modern genomics and bioinformatics era started in the early 2000s with the initial release of the first human draft genome sequence by the Human Genome Project (HGP) Consortium [1]. Since then, life science research has witnessed tremendous developments in deoxyribonucleic acid (DNA) high-throughput sequencing (HTS) and assembly algorithms. These developments have substantially contributed over the years to more affordable sequencing costs per megabase (Mb) and the emergence of more accurate and efficient assembly algorithms and computational approaches (Figure 1.1). For example, long reads sequencing (i.e., third-generation sequencing) has dramatically changed the way genomes are sequenced and assembled today [2, 3], by tackling major shortcomings and drawbacks of short reads technology [4]. Technological advances and rapidly decreasing sequencing costs have enhanced genome research in all taxa, including aquaculture and fishery species.

A decade ago, the daunting tasks of whole-genome sequencing and assembly of vertebrate genomes were conducted mainly by consortiums and required hundreds of thousands of dollars and several years to complete. Today, sequencing and assembling a midsize vertebrate genome (e.g., fish) can be achieved in the scope of a regular Ph.D. project with no more than a five-digit budget. Like biomedical science, agriculture, or livestock genomics, aquaculture genomics is also leveraged with these advancements. Hence, the advanced genome sequencing technologies and assembly algorithms offer an unprecedented opportunity to decipher and characterize the genome of any economically relevant fish species at affordable costs and in a reasonable time.

Although the first teleost fish genome, Fugu (*Takifugu rupriens*) [5], was sequenced in 2002, the first chromosome-scale assembly of an aquaculture species leveraging HTS technology was only achieved in late 2011 with the publication of the complete genome sequence of Atlantic cod (*Gadus morhua)* [6]. The amount of published chromosome-scale fish genomes deposited in public data repositories has experienced a sharp growth from only five genomes in 2011 to a total of 307 as of the time of this writing[*] (Figure 1.1A). Among these high-quality fish genome assemblies, farmed species alone account for nearly 88% (N=270), attesting to the ongoing breakthrough in aquaculture genomics.

Aquaculture's efficient production and profitability basically depend upon harnessing species-specific genomic resources [7]. The majority of published genomes of farmed fish are at chromosome levels with comprehensive annotation of genes and functional pathways, that have practical relevance to aquaculture production. It could be the basic understanding of the genetic makeup, evolutionary life history, performance and production traits, or thorough analysis of genetic variations associated with environmental factors. Hence, the insights gained from computational genomics are susceptible to improve production efficiency, sustainability, and to ensure food security. Some of the most important applications of computational genomics in diverse aquaculture species include the identification of SNPs and small insertion-deletions (Indels) in linkage disequilibrium (LD) with genetic loci associated with skin color, body mass, cold tolerance, and growth rate in common carp (*Cyprinus carpio*) [8, 9, 10]. In addition, harnessing genomic resources has allowed to identify genetic markers linked with adaptation to salinity in European seabass (*Dicentrarchus labrax*) [10, 11], to predict sex-related genes for marker-assisted selective breeding in the flatfish turbot (*Scophthalmus maximus*)[12], and to elucidate disease resistance traits in salmonid species [13, 14, 15, 16]. As last illustration, sequencing and analyzing the grass carp (*Ctenopharyngodon idella*) genome and transcriptome provided valuable support to discover the genes responsible for the vegetarian adaptation and other commercial traits in that herbivorous fish [17].

[*] NCBI Genomes, 30th July 2021.

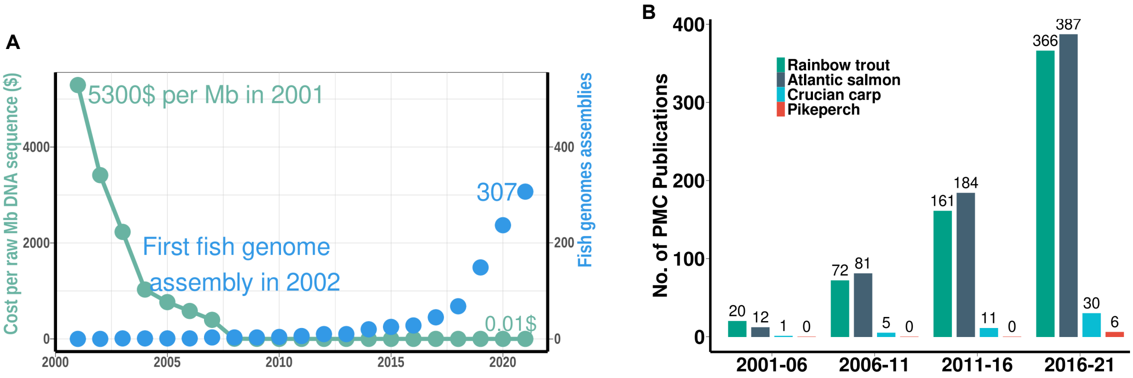

Figure 1.1.: Trend of sequencing costs and number of omics related publications. (**A**), depicts the dramatic decrease of sequencing costs per Mb DNA sequence and the sharp growth of published high-quality fish genomes assembly since 2002. Sequencing cost data were obtained from National Institute of Health [18]. (**B**), Cumulative number of PubMed Central (PMC) referenced publications since 2001 reporting omics studies in exemplary aquaculture fish species compared to pikeperch.

Pikeperch (*Sander lucioperca* L., 1758, taxonomy-ID: 283035), a member of the *Percidae* family, is a fresh and brackish water piscivorous fish species, that has recently gained high commercial and ecological significance in Europe. The international consortium conducting the research project DIVERSIFY[†] has recently identified pikeperch as one of the six species with the highest potential for inland aquaculture diversification in Europe [19, 20, 21]. Further European projects such as LUCIOPERCA and LUCIOPERCIMPROVE have demonstrated the bioeconomic potential of pikeperch intensive rearing in recirculating aquaculture systems (RAS), which are modern and ecologically viable alternatives to ponds [22, 23, 24]. Thanks to its flesh quality, which features low fat content (1-2%), highly assimilable proteins, and delicate flavour without small intramuscular bones [24], the aquaculture industry is facing a growing demand of this highly valued perch-like fish (13€/kg on average, FAO European Price Report 2021) in the international markets. Its relatively rapid growth, the resilience to disease and handling stress in captive environments also make pikeperch an attractive candidate for large-scale aquaculture[25, 26]. In the decade 2007–2017, the global capture production of pikeperch increased from 17,891 to 20,481 tonnes, while the global inland aquaculture production doubled in the same time, from 627 to 1418 tons [27], attesting the growing demand for this species.

However, despite this growing commercial significance, a number of production issues remain unresolved—what still hampers an efficient and sustainable production of pikeperch. These include cannibalism, low larval survival, high incidence of deformities, impaired and heterogeneous growth [28, 29, 30, 31]. In addition, we have currently no understanding on the genetic variability and markers as a framework to study performance, welfare

[†] European project aiming to explore the biological and socio-economic potential of new/emerging candidate fish species for the expansion of the European aquaculture industry.

and production traits in pikeperch. Moreover, the lack of a reference genome and transcriptome sequences as fundamental resources for genetic studies have been hindering our better understanding of the genomic structure, and organization of this obligate piscivore species, as well as the investigation of key genetic features at the basis of its adaptation to environmental factors including stressors and pathogens. As a candidate aquaculture species, it is essential to have genomic tools available to map complex traits such as body mass, hypoxia and stress tolerance, sexual maturation time, fitness, viability, and fecundity of the fish, among many others. Furthermore, indispensable tools for any farmed species, including SNPs chips and sex-determination markers for selective breeding are completely unavailable for this species. Customized use of genomic tools can be applied at each stage of the domestication and selective breeding process to inform about optimal hatchery parameters [32]. Addressing these gaps and unmet needs is still a pressing research question.

Although international efforts have been initiated in recent years to develop initial genomic resources for pikeperch [25, 26, 33, 34, 35], the scientific literature on multi-omics studies on this species is still sparse. There are only six publications (N=6 abstracts and full articles) reporting multi-omics (i.e., genomics, proteomics, metabolomics) studiess in pikeperch, as recorded by PubMed Central (PMC), July 13, 2021– which is negligible compared to successful aquaculture species such as rainbow trout (N=366), Atlantic salmon (N=387), or crucian carp (N=30) (Figure 1.1B).

Taken all together, valuable genomic tools and resources are needed to enhance studies that will contribute to the successful positioning of pikeperch in the aquarfarming sector. Whole genome and transcriptome sequencing, assembly, and annotation of pikeperch will provide important clues for stable and optimal rearing conditions, smart management of broodstocks, and genome-based customized breeding programs.

1.2. Pikeperch Biology

The *Percidae* family is a diverse and economically important group of mostly freshwater fishes that comprises 11 genera and about 275 identified species [21]. Some of these species are valuable candidates for inland aquaculture, while other play important role in recreational fishery. The native range of *Sander lucioperca* essentially spans streams and seas of mainland Europe including the Caspian, Black, Aral and Baltic Sea drainages (Kazakhstan, Azerbajan, Hungary, Czechia, Germany, Finnland), where they inhabit brackish waters. Meanwhile, pikeperch has been anthropogenically introduced into most regions in Europe, Northern America, Asia [36, 37], and in few Maghreb countries, making it the Percid species with the largest geographic expanse [38] (Figure 1.3). Although pikeperch are characterized as cool to warm water fish species, their high level of phenotypic plasticity enable them to adapt to different environmental conditions throughout their geographical ranges [21]. Actually, *Percidae* are temperate mesotherms, that is, they are capable to tolerate a large range of temperature distribution [39]. This makes them highly plastic species able to thrive in environments ranging from small streams to large lakes and bays. As such, the mechanisms controlling important traits including growth, reproduction, recruitment, and mortality likely vary both spatially and temporally across different species, populations, and environments [21].

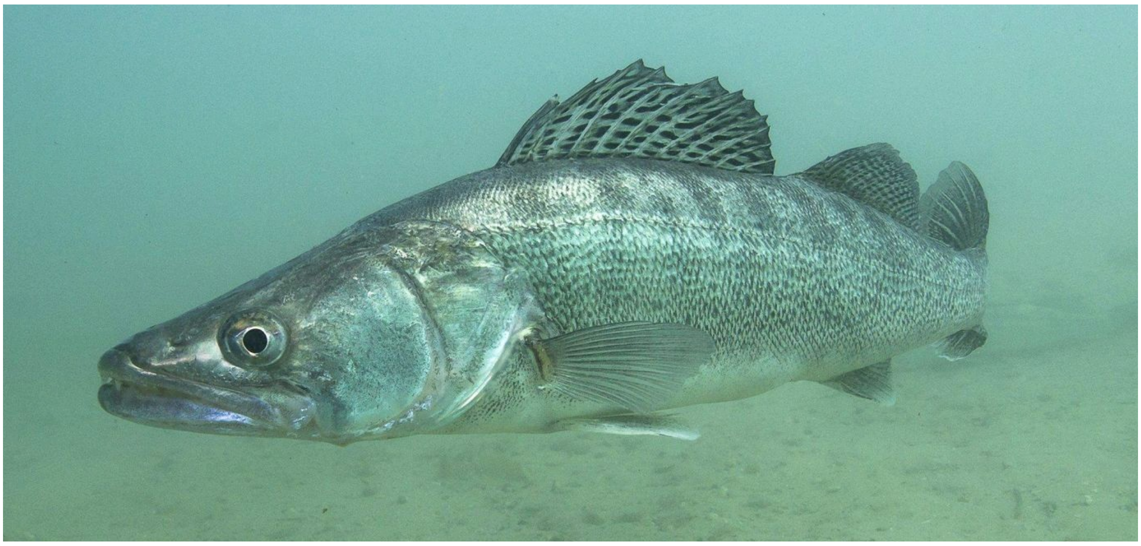

Figure 1.2.: The pikeperch (*S. lucioperca*) also known as zander. © Fotolia/Adobe Stock.

The size of the pikeperch haploid genome was estimated to 1.14 pg (i.e., 1114 Mb) utilizing cytometric methods [40]. A diploid number of 48 (2n=48) chromosomes was reported for this species [41, 42]. Previous studies have also reported a XY/XX heterogametic sex chromosome system in the *Percidae* fish family [43]. Genetic analyses based on microsatelites markers have indicated low levels of genetic diversity, and high rates of heterozygosity and inbreeding in domesticated broodstocks and natural populations of pikeperch [44, 45]. These findings are of concern towards the future establishment of optimal genetic breeding programs for sustainable domestication of pikeperch, because high inbreeding can negatively impact the growth rate and other production traits [46, 47]. Besides, intra-cohort cannibalism in all life stages [48] is one of the major issues while rearing pikeperch. Population genetic studies could shed light on the molecular markers associated with juvenile cannibalism and predation avoidance. However, essential genomics tools and resources including reference transcriptome or genome sequences to conduct genome-wide association studies (GWAS), and understand the mechanisms underlying these traits are currently lacking.

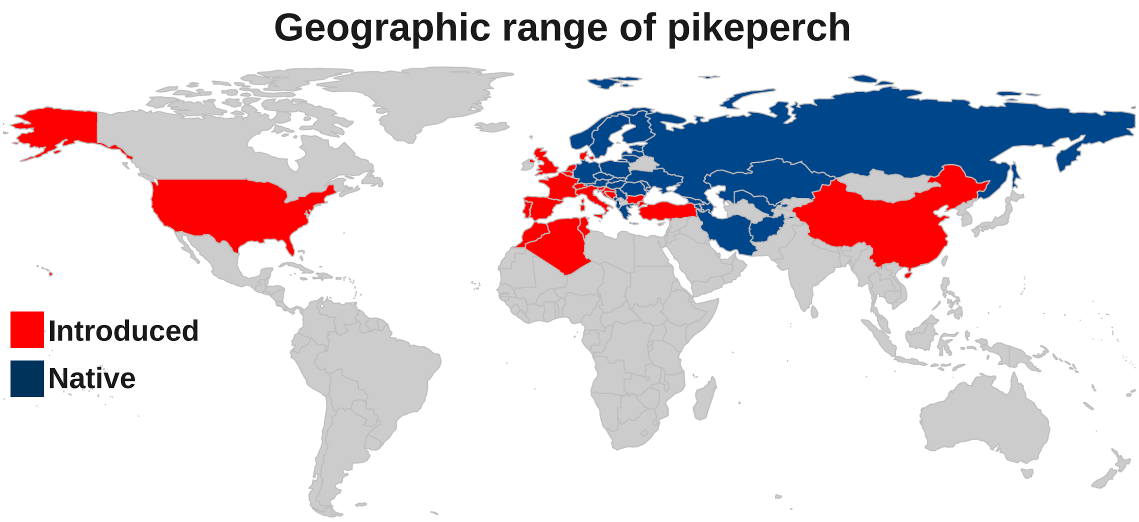

Figure 1.3.: Geographic range of pikeperch (*Sander luicioperca*). Pikeperch is native to continental Europe including Germany, Finland, Poland, Serbia, Hungary, Letvia, Kazakhstan, Iran, Norway, Finland, Azerbaijan, etc. (blue countries). Currently, pikeperch has anthropogenically been introduced into many regions/countries around the world such as China, Turkey, Algeria, Morocco, Tunisia, USA, Bulgaria, Croatia, Cyprus, Sweden, Afghanistan, Denmark, France, Italy, The Netherlands, Portugal, Slovenia, Spain, the United Kingdom, Russia, Belgium the USA, etc. (red countries). Extensive data on the geographical expanse of pikeperch are available in this online reference: https://www.cabi.org/isc/datasheet/65338

1.3. Objectives and Significance of this Thesis

Objectives

The overall aim of this Ph.D. research project is to develop state-of-the-art genomic resources and provide genetic tools as a starting point towards understating systems biology, evolutionary life history, and adaptive diversity of the emerging aquaculture species pikeperch. Specifically, the research objectives are at:

i Building high-quality annotated reference genome and transriptome assemblies, to serve as a backbone resource for future genetic and genomics studies in pikeperch;

ii Utilizing the built genome and transcriptome to establish a genome-wide catalog of gene expression and co-expression atlas, and to capture global expression and recent positive selection patterns in pikeperch;

iii Identifying and quantifying genomic diversity and variability of captive pikeperch broodstocks in experimental RAS, to allow precise and targeted selection to improve aquaculture performance;

iv Conducting whole genome comparative analyses with related percids species to identify core genomic elements that cast light on their phylogentic and evolutionary history;

v Establishing a catalog of putative genome-wide candidate markers associated with key biological and commercial traits in pikeperch, in order to help decision how future genetic breeding programs should be established for sustainable and optimal domestication.

To that end, I leveraged different HTS technologies to deeply sequence the pikeperch genome and transcriptome. For de novo assembly and subsequent computational genomics analyses, I setup and deployed analysis pipelines by creatively combining several command line bioinfomatics tools/softwares and workflows with minimal customization, as well as by using high-performance computing and scripting in R and Python programming languages. To ensure the reproducibility of the data outputs, the analysis pipelines and workflows I have deployed have been documented and containerized in reproducible environments including Conda, Singularity or Nextflow.

Significance

Developing species-specific genomic resources (e.g., high-quality genome and transcriptome assemblies) is a quantitative and qualitative foundation for future indept and targeted genome research. Therefore, the resources produced in this work aim to build a framework for aquaculturists to enhance their production through SMART Breeding (**S**election with **M**arkers and **A**dvanced **R**eproductive **T**echnologies). Moreover, the data and tools generated in the scope of this thesis along with the insights gained will serve as fundamental resources to a broader community of genome scientists enabling the investigation of causal links between genome and phenome, and to understand the genomic architecture of bioeconomical traits, thereby significantly enhancing the species-specific breeding and production systems of pikeperch. Following the FAIR principles (**F**indability; **A**ccessibility; **I**nteroperability and **R**eusability), whole genomic data including genome and transcriptome sequences, gene expression and genotyping data have been structured and deposited in public omics-data repositories for immediate use by the scientific community. Finally, and maybe most impactful, this thesis is a modest contribution to the *Fish10K* project [49], a large-scale genome sequencing effort aiming to sample, sequence, assemble, and analyze genomes of 10,000 representative fish species within 10 years, till 2030.

1.4. Outline of the Dissertation

This dissertation is structured into five separate, though built on one another parts. Namely, theoretical background on methods deployed in this thesis (1); whole genome sequencing and analysis (2); Multitissues transcriptome assembly and analysis (3); population-wide genotyping of genetic variations (4), and comparative phylogenomics analyses (5).

Part I provides readers detailed background of existing state-of-the-art methods and approaches to genomes sequencing, assembly, annotation and analysis.

Part II comprehensively presents the newly assembled pikeperch genome, including functional and structural annotation. Here, I additionally quantified further genomic features such as repetitive elements, noncoding RNA (ncRNA), and recent genome duplication events.

In **Part III**, I performed data integration of the newly built genome with multitissue bulk RNA-Seq to generate a reference transcriptome and gene expression atlas of pikeperch. Moreover, I used these transcriptomic data to perform tissue-specific expression and co-expression analysis.

In **Part IV**, I investigated genetic variations in a sample of farmed pikeperch population. In particular, diffident types of genetic variants are genotyped and annotated at genome-scale and population-level.

In **Part V**, I interrogated the evolutionary history of pikeperch genome with a series of comparative phylogenetics analyses.

Part I.

Fundamental Concepts and Methods in Genome Sequencing, *de novo* Assembly and Annotation

Introduction

The genome is like an information storage medium. It harbors and encodes the functions of all observable physical or physiological traits (phenome) of an organism. The genome contains structural and functional information, which is transferred and expressed in phenotypic traits through complex biological processes (Figure 1.4). Two of these processes are transcription and translation, by which segments of DNA are written into other biomolecules such as RNA and proteins, respectively. Proteins can further get involved in metabolic processes with other biological molecules.

However, the genetic information stored in DNA is essentially abstract and encoded by four organic compounds (nucleotides) including adenine (A), cytosine (C), guanine (G), and thymine (T). Genome sequencing and assembly aim to reliably read, decipher, and mine this abstract information using molecular genetics and diverse biological technologies. Shotgun sequencing is a molecular genetics technique for reading the DNA sequence of an organism's genome. This method involves breaking up the genome randomly into libraries of smaller pieces that are individually sequenced (i.e., read sequentially) to obtain sequencing reads. A computer program called assembler looks for overlaps between these reads and places them in the correct order and orientation to reconstitute the initial genome. The raw assembly, i.e., the sequence of nucleotides alone, is meaningless. Therefore, the subsequent step is to annotate the assembled genome, to mine its structural and functional features of scientific relevance.

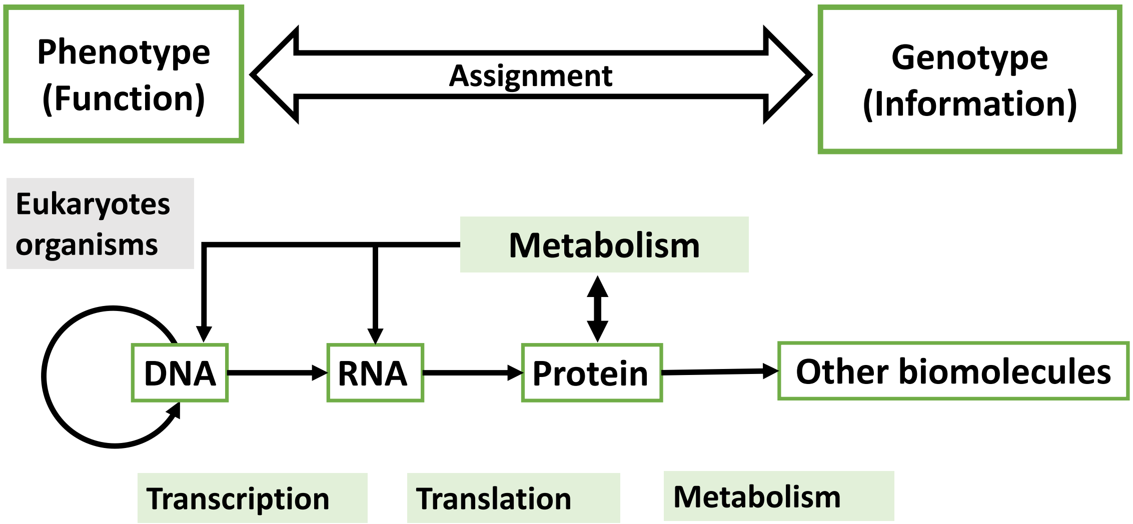

Figure 1.4.: The flow of genetic information. The information preserved in DNA flows into RNA via transcription and ultimately to proteins via translation. Processes like reverse transcription and replication also represent mechanisms for propagating genetic information in different forms in eukaryote organisms. Proteins are the functional units of the genome inveled in enzymatic reactions and processes with other proteins and biomolecules. Proteins are also inveled in metabolic processes with other molecules in the cells.

This section (**Part** I) will provide a comprehensive introduction to fundamental concepts and experimental approaches in the generation, integration, analysis, and interpretation of experimental genomics data. It will be covering the fundamental concepts of genome sequencing and analysis as well as computational and statistical methods used in this thesis. A diverse range of genomics and bioinformatics concepts, including modern DNA sequencing technologies, genome assembly, computational genome annotation, and analysis, as well as data validation and dissemination, will be explored. Readers who are more or less familiar with genomics and bioinformatics will gain a basic technical and theoretical understanding of the main methods and approaches used in the different experiments and analyses throughout this thesis.

2. Massively Parallel DNA Sequencing

Next-generation sequencing (NGS) is a generic term referring to different massively parallel genome sequencing technologies offering ultra-high throughput, higher scalability, yield, and speed, compared to the traditional dideoxynucleotide sequencing technologies based on the Sanger chain termination method. NGS includes two major paradigms: short-read sequencing and long-read sequencing. Short read NGS approaches are highly accurate (error rate $< 0.05\%$) [50, 51], cost-effective, and are well suited for applications such as high-resolution population-level genotyping and discovery of SNPs and short indels associated with quantitative or commercial traits.

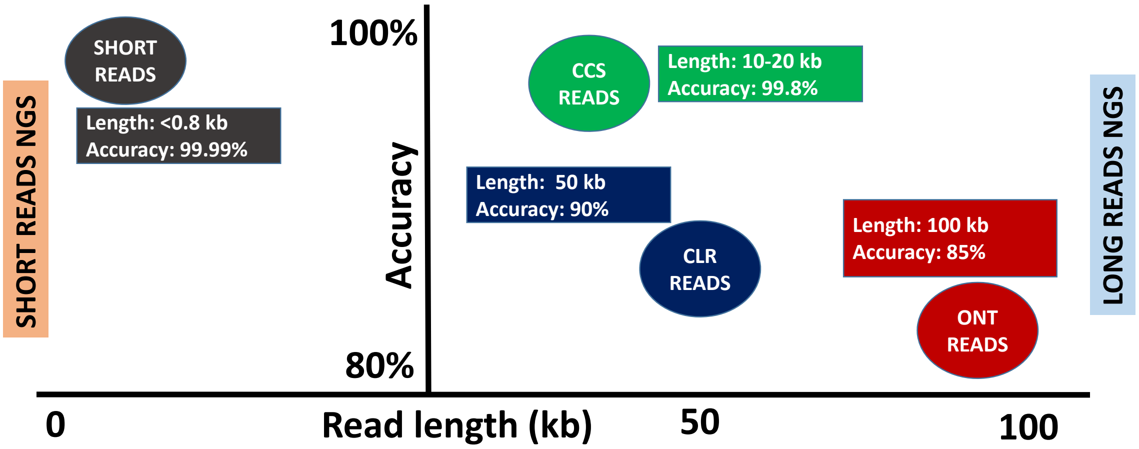

Figure 2.1.: NGS reads length and accuracy. The reads length and accuracy of the two main NGS technologies are shown here. Short reads NGS such as from Illumina usually do not exceed 800 bp and yield an average base-level accuracy of 99.99%. Long reads NGS include circular consensus sequence (CCS) also known as HiFi reads, and circular long reads (CLR), both from Pacific Bioscience. Oxford nanopore technology (ONT) reads are much longer, but also with higher average error rate.

In genome assembly, highly accurate short reads (SRs) are harnessed to correct erroneous long reads (LRs), assist scaffolding, and polish assemblies. However, due to their short length, they are unable to resolve long eukaryotic repeats and segmental duplications (SDs). *De novo* genome assembly with SRs leads to highly fragmented assemblies with a substantial amount of short contigs. Hence, SRs alone are not adequate to produce high-quality assemblies of complex eukaryotes' genome. By contrast, although less accurate, long read assembly produces much longer contigs that span most of the problematic regions of the genome. The average error rate of LRs ranges from 1% (e.g., HiFi reads) to 15%

(e.g., ONT reads), depending on the sequencing platform and chemistry used (Figure 2.1). Though, they have been excellent for applications such as *de novo* assembly and genome finishing, full-length RNA-Seq, and identification of structural variants (SVs) and complex genomic rearrangements. Today, LRs sequencing is the technology of choice in genome assembly projects of eukaryote organisms and many other animal genomics applications.

2.1. Short Reads NGS Technology

Illumina is currently the leading short read sequencing platform. The Illumina platform uses sequencing by synthesis (SBS) in a workflow described as following (Figure 2.2):

1. Since entire DNA cannot be read at once, the first step is to break up the DNA into more manageable fragments of around 300 to 800 bp, called inserts. Adapter sequences (short artificial DNA sequence) are then attached to these DNA fragments to prepare sequencing libraries.

2. Libraries are loaded into the a flow cell and fragments are hybridized to the flow cell surface. Each bound fragment is amplified into clonal cluster through bridge amplification.

3. DNA polymerase, connector primers and four dNTPs labelled with base-specific fluorescent dyes are added to the reaction system. The 3'-OH of these dNTPs are protected by chemical methods, which ensures that only one base will be added at a time during the sequencing process. The flow cell is imaged and the emission of each cluster is recorded. The incorporated base (A, T, C or G) is uniquely and precisely identified by its emission wavelength and intensity. The cycles are repeated k times or theoretically, until the polymerase lifetime, to create reads of length k bases.

The library inserts can be sequenced either from both ends (3'-end and 5'-end) or from solely one end—resulting in paired-end (PE) and single-end (SE) reads, respectively. PE sequencing generate high-quality and mappable reads that enable the detection of repeats, genomic rearrangements, as well as gene fusions and novel transcripts.

Depending on the library preparation protocol, we distinguish between two kinds of PE-reads: short-insert paired-end reads (SIPERs) and long-insert paired-end reads (LIPERs). The latter one is also known as mate-pair (MP) reads. The substantial difference between the two variants is the insert length. SIPERs are 200-800 bp long, while LIPERs can be longer up to 10 kbp. Besides Illumina, other cutting-edge short-reads sequencing technologies such as linked-reads sequencing are commercialized by 10x Genomics.

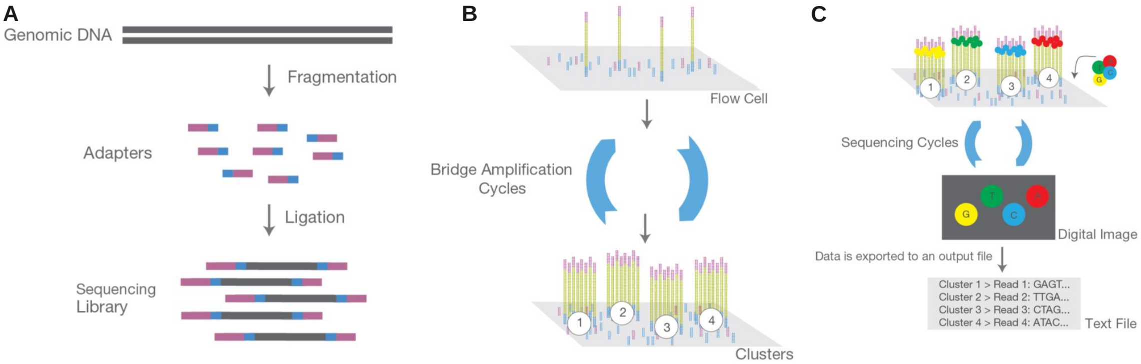
A
Genomic DNA
Fragmentation
Adapters
Ligation
Sequencing Library
B
Flow Cell
Bridge Amplification Cycles
1
2
3
4
Clusters
C
1
2
3
4
T
A
C
G
Sequencing Cycles
T
A
G
C
Digital Image
Data is exported to an output file
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
Text File

Figure 2.2.: Illumina DNA sequencing workflow. Illumina NGS includes three main steps: Library preparation (**A**), cluster generation (**B**), and sequencing by synthesis (**C**). Adapted and minimally edited from Illumina technical handbook: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

2.2. Long Reads NGS Technology

There are currently two competing technologies that dominate the long-read sequencing space, namely nanopore sequencing commercialized by Oxford Nanopore Technologies [52], and single-molecule real-time sequencing (SMRT) by PacBio [2]. Also termed third-generation sequencing, both technologies offer real-time, long-read, direct, and large-scale sequencing of DNA or RNA.

Nanopore Sequencing

ONT allows nucleic acids to be sequenced without PCR amplification or chemical labeling of the sample. A nanopore is simply a nanometer-size hole found on specific transmembrane cellular proteins. The sequencing principle relies on detecting electrical current changes induced by the passing DNA stand (one base at a time or in a stretch of k bases) through the nanopore channel immersed in a conducting fluid with voltage applied on it. Each passing nucleotide induces a specific shift in the current value. The resulting signal is then decoded to infer the specific DNA or RNA sequenced read. With ONT sequencing, there is no limit to read length since a single molecule can (theoretically) be sequenced end-to-end, achieving ultra-long reads that can reach a length of hundreds of kilobases. The longest read recorded so far was over four Mb.

A major limitation of nanopore sequencing is its high base calling error rate, which, despite recent improvements to nanopore chemistry and base caller tools, still ranges between 5% and 15% [53]. The most recurrent errors are homopolymers, recording the same signal for different nucleotides in a row. However, this limitation can be mitigated through error correction with highly accurate SRs from Illumina.

PacBio Sequencing

In contrast to its counterpart (ONT), PacBio SMRT sequencing requires a library preparation step including DNA fragmentation and size selection to the optimal length determined by the sequencing protocol and the ligation of specific adapters to fragments of the DNA be to sequenced. The circularized DNA templates to be sequenced are immobilized in a zero-mode waveguide on a sequencing device. Next, four nucleotides are added, each labeled with a different fluorescent dye. When the DNA polymerase incorporates a specific nucleotide, the emitted light signal is measured in real-time, the corresponding base is called, and the fluorophore is cleaved off. This process can be repeated until the polymerase dies and breaks the read. Most SMRT reads obtained from newer PacBio sequencing platforms such as Sequel II are larger than 10 kb, while some reads surpass 100 kb.

Here also, high base calling error rate used to be a notorious limitation. However, an improved sequencing protocol, HiFi sequencing, has recently been released [54]. HiFi sequencing has dramatically improved the average read quality to $>$ 99.99% (Q40; 1 error in 10 kb) by leveraging that sequencing errors tend to be random. Therefore, if the same region is sequenced several times, the resulting consensus accuracy of the circular consensus sequence is optimized. Hence, HiFi protocol enables us to get the benefits of short reads and traditional long reads in one easy-to-use technology, making it to the new gold standard for reference-quality *de novo* assembly.

3. *De novo* Genome Assembly

Genome assembling problem is usually illustrated as solving a large jigsaw puzzle with million of pieces. Hence, assembling sequence reads to the initial genome is similar to assembling millions of puzzle pieces into the reference picture. Here, some pieces might be dirty or missing (e.g., sequencing bias or errors), other might be ambiguous and highly similar (e.g., repetitive sequences in the genome). Thus, the assembly complexity increases with the number of reads, the genome size and the repeat content. One can build an assembly using a backbone sequence as reference by read alignments. This approach is termed reference-guided or comparative assembly. De novo assembly refers to building a genome from 'scratch', with no a priori knowledge of the correct sequence or order of the genomic fragments (reads).

3.1. Definitions and Notations

Let $\sum = \{A, C, G, T\}$ the DNA alphabet, and $R = \{r_1, r_2, ..., r_m\}$ the set of m strings (e.g reads) obtained by sequencing the target genome S of length $|S| = g$. If $S = uvw$ for possibly empty sequences $u, v, w \in \sum^*$, we call u a prefix, v a substring (subsequence), and w a suffix of S. A super-sequence (superstring) is a large sequence containing every sequence of a string set as substrings, e.g., S is a superstring of all reads in R. A k-mer of the read r_i is a substring of length k, $k < |r_i|$. More generally, a sequence of length L will have $(L - k + 1)$ k-mers, and n^k total possible k-mers, where $n = |\sum|$ (e.g., four in the case of DNA alphabet).

A contig is a contiguous (gap-free) sequence assembled from a set of overlapping reads. Assembly scaffolds are obtained by chaining contigs together utilizing additional information about the relative position and orientation of the contigs in the genome. Gaps of variable number of 'N' letters are introduced between contigs in the scaffolds during a computational process called scaffolding.

A de Bruijn graph (DBG) of order k built on a set of reads R is a directed graph DG_k representing the overlaps between reads in R. Depending on the way of expressing the nodes and edges, DBGs are classified into two types: Hamiltonian and Eulerian *DBGs* [55]. In the former type, the k-mers are nodes and edges are pairwise overlaps between consecutive reads in R, whereas in the latter, edges represent k-mers and nodes are (k-1)-mers [56] (Figure 3.1). An overlap graph (OG), on the other hand, is a directed graph where each sequence is a node and the overlap between suffix of r_i and prefix of r_{i+1} (consecutive reads) are edges joining the two nodes (reads) involved. OGs differ from Hamiltonian DBG in that nodes are reads (strings) in OGs, whereas they are k-mers in Hamiltonian DBGs.

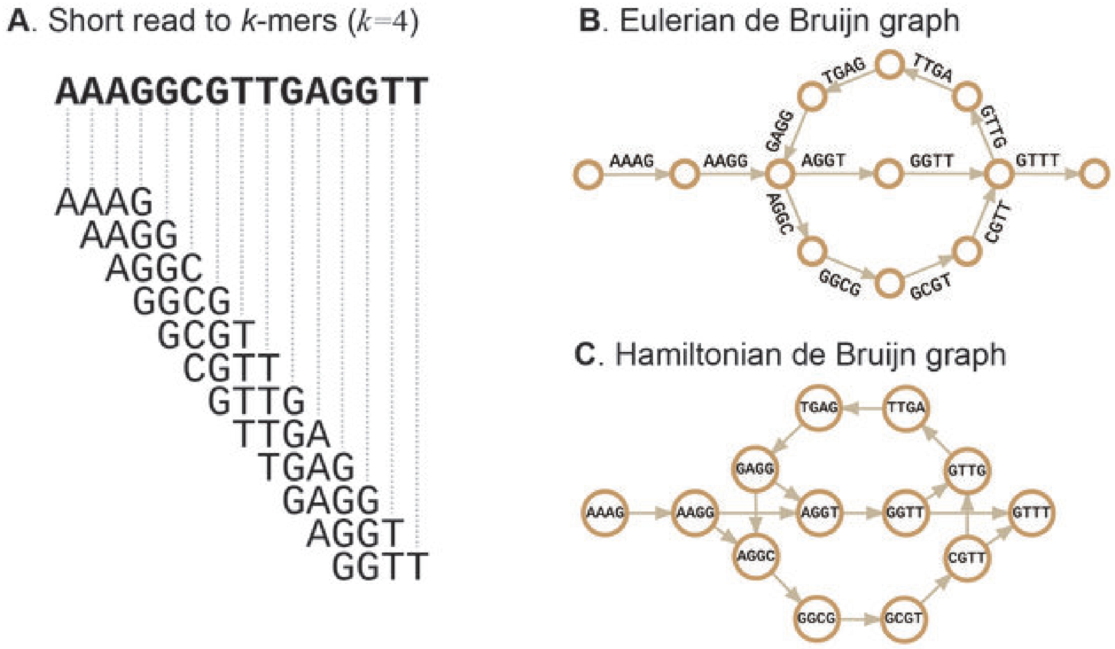

Figure 3.1.: The two types of de Bruijn graphs. A genomic read split in 4-mers (**A**). Eulerian de Bruijn graph, where the k-mers (substrings) are the egdes (**B**), whereas they are nodes in Hamiltonian de Bruijn grpahs (**C**). Reprinted from Sohn and Nam (2016) [57]. © Jang-il Sohn

3.2. *De novo* Assembly Problem

In the strict theoretical perspective, the *de novo* assembly problem is stated as follows: Given a set of strings (reads) $r \in R$, find a minimum length string $S \in \sum^*$ such that every $r \in R$ is a substring of S. The *de novo* assembly problem is a shortest common superstring problem (SCS), which is a combinatorial optimization problem known to be NP-complete [58]. The SCS problem ultimately aims to find the shortest possible string that contains every string in a given set as substrings. Exact solutions are intractable, but most assemblers implement approximation algorithms. For the case of Hamiltonian DBGs and OGs, solving the SCS problem is to find a Hamiltonian path, that is, a walk in the directed graph that visits each vertex exactly once. For Eulerian DBG, the SCS problem is reduced to finding an Eulerian path in the directed graph, a trail that visits every edge of the graph exactly once. The complexity scales polynomial with the number of vertices. The SCS assumes no errors and known orientation of the reads. Moreover, if entire reads do not span long repeats, then the target common superstring (genome) will not be the shortest possible, i.e., the assembly algorithm will not produce the correct result. However, real-world assemblers give up on unresolvable repeats and use a tractable algorithm to assemble the resolvable portions of the genome. That is why real-world assemblies are usually fragmented, consisting of hundreds to thousands of contigs.

3.3. Bioinformatic Approaches in Genome Assembly

There are two computational strategies for genome assembly, which depend on the type of input reads: De Bruijn graph (1) and overlap-layout-consensus (OLC) approaches [56, 59]. Since DBGs are particularly sensitive to sequencing errors, highly accurate SRs are well suited for algorithmic solutions finding Eulerian trails. On the other hand, LRs, which are more prone to errors, are well adapted for the OLC approach (finding Hamiltonian trails). OLC approaches leverage the fact that OGs can tolerate a moderate amount of errors (mismatches and indels) in the overlap alignment, in contrast to DBGs which usually require exact matches in the overlaps (Figure 3.2).

De Bruijn Graph (DBG) Assemblers

DBG-based genome assemblers are widely applied to SRs. A whole panel of short-read assembly softwares have been developed for small eukaryote and bacterial genomes as well as for metagenome assembly. Notable examples include SPAdes, DISCOVAR, Megahit, SOAPdenovo2, and Platanus. The conceptual approach of all these DBG assemblers is to first build a k-mer set from genomic reads. Second, for a given k-mer, construct its prefix and suffix, and connect two k-mers (nodes) with an directed edge, if they completely overlap except for one nucleotide at each end. Finally, look for an Hamiltonian path that represents the candidate genome. This path will have minimal length (minimality requirement for the superstring) because per definition, a Hamiltonian trail travels each k-mer (node) exactly once. DBG approach does not require all vs. all overlaps alignment, and consequently, does not need to store the reads and their overlaps in the memory, which makes assembly of small genomes very efficient. However, for larger genomes, massive amount of memory resources might be necessary to contain all k-mers and their DBG.

Overlap-Layout-Consensus (OLC) Assemblers

OLC assembly tools leverage the length of LRs to build genomes *de novo*. Some popular OLC assemblers include Flye [60], Hifiasm [61], Canu [62], Falcon [63], and Shasta [64]. The algorithmic approach of OLC assemblers builds on three main stages to capture the global relationship between the input reads [65]. The first stage is to construct an OG that captures all read connections by exhaustively computing all-vs-all overlaps (O) among the reads. The overlaps are only taken into consideration between reads if they share the same k-mer. Therefore, the choice of the *k*-mer length and the minimum overlap alignment score are critical parameters of this step. Then, the exact layout (L) of the OG containing all reads and overlaps is determined by simplifying the graph in that the errors are corrected where possible, and transitively inferrable connections are pruned. Dynamic programming (DP) is used in a computationally and memory intensive process to find the optimal overlap alignments. Finally, the consensus sequence is inferred by tracing the Hamiltonian path and collapsing paths until a breakpoint is reached, e.g, an unresolved repetitive region of the genome that is longer than the read. Each collapsed path generates a contig.

In practice, both DBG and OLC assembly approaches handle ambiguities or unresolvable repeats by essentially leaving them out. Unresolved repeats break the assembly graph into different contigs. Hence, genome resolution will increase with the read length. Moreover, the advantages of SRs and LRs can be combined to improve the completeness and correctness of an assembly, making it a hybrid assembly method. The hybrid sequencing and assembly approach is the optimal strategy for obtaining high-quality data in *de novo* assembly. The MaSuRCA assembler [66] is an example of such computational solutions that combine the benefits of DBG and OLC assembly approaches to produce more accurate and highly contiguous assemblies. However, it should be noted that no assembler is a silver bullet to reconstruct high-quality genomes. A combination of different kinds of sequencing data, assembly approaches, and iterative parameters tweaking and manual curation are usually needed to obtain reference-quality assemblies.

3.4. Chromosome-scale Scaffolding

The raw assembly output is a large collection of fragmented contigs, hampering downstream analyses such as functional annotation and comparative genomics, which heavily rely on assembly of high contiguity. Thus, it is crucial to anchor the contigs into chromosome-level pseudomolecules and in the correct orientation, relative distance, and order. Unfortunately, NGS data alone are currently unable to achieve chromosome-level scaffolding. Additional technologies harnessing long-range chromosomal interactions are required to achieve chromosome-scaffolding.

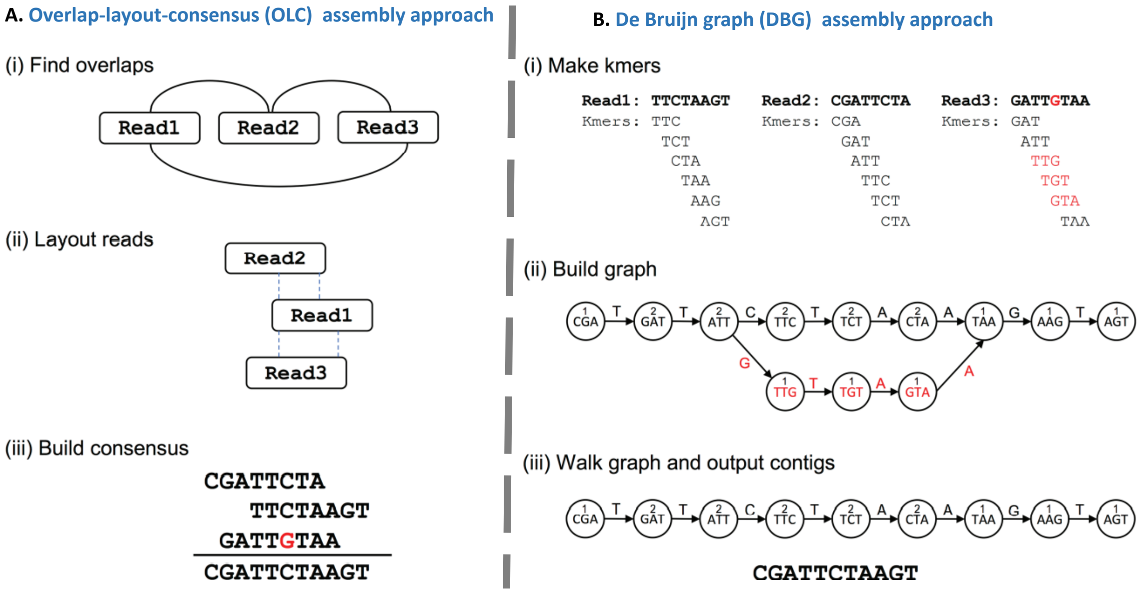
A. Overlap-layout-consensus (OLC) assembly approach
(i) Find overlaps
Read1
Read2
Read3
(ii) Layout reads
Read2
Read1
Read3
(iii) Build consensus
CGATTCTA
TTCTAAGT
GATTGTAA
CGATTCTAAGT
B. De Bruijn graph (DBG) assembly approach
(i) Make kmers
Read1: TTCTAAGT
Kmers: TTC TCT CTA TAA AAG ACT
Read2: CGATTCTA
Kmers: CGA GAT ATT TTC TCT CTA
Read3: GATTGTAA
Kmers: GAT ATT TTG TGT GTA TAA
(ii) Build graph
CGA T GAT T ATT C TTC T TCT A CTA A TAA G AAG T AGT
G TTG T TGT A GTA A
(iii) Walk graph and output contigs
CGA T GAT T ATT C TTC T TCT A CTA A TAA G AAG T AGT
CGATTCTAAGT

Figure 3.2.: Simplified overview on OLC and DBG assembly approaches. Illustration of the Overlap-Layout-Consensus assembly paradigm (**A**), and de Bruijn graph paradigm (**B**).

Incorporation of Hi-C Data

Hi-C sequencing is a high-throughput chromosome conformation capture technique to analyze the 3D genome organization and map higher-order chromosome folding and topological associated domains [67, 68]. Hi-C quantifies all interactions between genomic loci that are closely located in the 3-D space but may be separated by millions of base pairs in the linear genome. Genomic segments that are spatially close along the DNA sequence are preferentially ligated to one another. These fragments are used to prepare paired-end libraries sequenced using NGS technology. Paired-end sequencing allows the retrieval of sequence information from each end of ligated fragments. This long-range information is used in assembly scaffolding to retrieve, anchor, and order contigs that originated from the same chromosome.

Incorporation of Genetic Maps Data

A genetic map shows the relative locations of genes and other genetic markers on a linkage group (LG), usually mapping the entire chromosome. Genetic maps build on the idea of linkage, which suggests that the closer two genes/markers are to each other on the chromosome, the greater the probability that they will be inherited together. The inheritance patterns allow to find the relative locations of markers along the chromosome [69]. High-density linkage maps for example, are used as a chromosomal framework to anchor *de novo* assemblies, by orienting and ordering contigs/scaffolds into chromosome-scale sequences.

3.5. Assembly Assessment and Validation, Genome Finishing

Assembly assessment and validation is a crucial step of any *de novo* assembly workflow since it can inform potential users on the levels of structural and functional accuracy, and completeness of the data, as well as on their inherent limitations. There are different methods and metrics to validate and gauge the quality of genome assemblies, summarized in Table 3.1. These quality metrics gauge both the completeness and contiguity of an assembly and provide confidence in using the data for downstream analyses.

Assembly Contiguity

Assembly contiguity describes how contiguous an assembly is, using length metrics such as contig/scaffold N50, NG50, and L50. Contig N50 is the shortest contig's size (in bp) that needs to be included in the assembly for covering 50% of the assembly size. At the same time, L50 represents the number of contigs whose sum length makes 50% of the assembly size. The NG50 metric is the same as N50 except that the statistic considers 50% of the known or estimated genome size instead of the assembly size, typically shorter than the known genome size. The Vertebrate Genome Project (VGP) standard suggests that high-quality genomes should have a minimum contig and scaffold N50 length of 1Mb and 10Mb, respectively [70, 71].

Genes and Repeat Space Completeness

A highly accurate and complete assembly should span the whole known repeat and genes space of the sequenced species or orthologs from different species in the same lineage. BUSCO (Benchmarking Universal Single-Copy Orthologs) provides robust metrics for quantitative and qualitative assessment of genome assembly, gene set, and transcriptome completeness based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs [72]. LTR Assembly Index (LAI) is a standardized metric to evaluate the assembly in repeats space completeness. A complete and structurally accurate assembly should resolve must of repeats and genes ($> 98\%$) [71].

Genome Finishing

Genome finishing is the final step in genome assembly workflows prior to submission into public genome databases. Genome finishing is achieved by filling the remaining gaps in the chromosomes and performing intensive manual curation to fix errors (misassembly) and ambiguities. The final completed assembly is often referred to as ”closed” or ”finished”.

Table 3.1.: Overview of the assessment methods and quality standards of high-quality genome assemblies.

Quality	**Metrics**	**Methods**
Contiguity	Contig N50 $\geq$ 1 Mbp Scaffold N50 $\geq$ 10 Mbp	Quast
Structural accuracy	False duplications, Reliable blocks Genomic reads mapping rate (>95%)	Merqury, BWA samtools
Base accuracy	QV >50, k-mer completeness	MaSuRCA, Merqury
Haplotype phasing	No. of phased blocks (NG50)	Merqury
Gene space completeness	BUSCO (> 90%) RNA-Seq reads mappability (>95%)	Merqury
Repeat space completeness	LAI	GenomeQC
Functional completeness	Functional databases mapping(>90%) RNA-seq reads mappability (>95%)	Merqury, BLAST
Gene metrics	No. of genes models, CDS content Average exon/intron length distribution exons count per genes, etc.	GenomeQC, Quast

***Note**: In addition to these metrics, the Vertebrate Genome Project (VGP) committee proposes additional quantitative assembly standards for reference-quality genome assemblies. These include:* < 5% *false duplications;* > 90% *kmer completeness;* > 90% *sequence assigned to candidate chromosomal sequences;* > 90% *transcripts from the same organism mappable. QV: consensus assembly quality value. LAI: LTR asembly index.*

4. Bioinformatic Approaches to Genome Annotation

The raw assembly sequence is nearly meaningless without information on its structural features and functions. Genome sequences become only meaningful and practically useful when they are annotated with genes in particular [73]. Genome annotation is the subsequent step after *de novo* assembly that can provide this information [74]. Annotation is the process of identifying and describing structural and functional elements along the genome sequence.

4.1. Structural Genome Annotation

Structural annotation of genomes refers to the process of predicting coding and noncoding gene models and all other structural elements of the genome, such as repetitive elements, genes, and segmental duplication events. The prediction of protein-coding genes is generally performed combining *ab initio* and evidence-driven computational methods.

In *ab initio* gene finding, the intrinsic features of the DNA sequence are scanned for gene-specific signals. These can be the promoter and regulatory sequences, transcription factor binding sites (TFBS), CpG islands (regions with a high frequency of CpG sites), GC-content (region of high GC density), and k-mer statistics. To that end, different gene finders trained to recognize and discriminate these gene-associated patterns apply complex statistical and probabilistic models such as hidden Markov models (HMMs) or machine learning techniques like support vector machines and neural networks. Some popular and sound performing *ab initio* gene finders include Augustus, GLIMMER, GeneMark, GENSCAN, and SNAP. In contrast, evidence-driven or homology-based gene prediction methods rely on external evidence provided by RNA-Seq, CDSs, or protein sequences of closely related species to predict gene models [74]. The available well-curated protein databases such as NCBI nonredundant protein, RefSeq, and Uniprot provide strong support for homology-based gene prediction. In addition, transcript and protein sequences from the same species or evolutionarily related species provide valuable evidence and pieces of information to be exploited in the structural gene annotation. When developing structural genome annotation pipelines in practice, both intrinsic and extrinsic-based methods and tools should be integrated with different kinds of evidence to achieve high-confident and reliable gene model predictions. However, errors (false predictions) propagation from extrinsic evidence remains a real caveat in computational genome annotation that should be considered when using the data.

4.2. Functional Genome Annotation

The functional annotation aims to assign putative names and biologically relevant information to predicted structural features of the genome by homology searches. The underlying idea is that a high degree of sequence similarity implies a high likelihood of homology, hence a high probability of functional convergence. For instance, the observed structural similarities between proteins suggest that these proteins perform identical or similar functions. The function of genes can then be computationally inferred based on their sequence similarity to gene sequences in public repositories, including but not limited to Uniprot, Pfam, PANTHER, or eggNOG. Caution should be taken because sequences may be similar by chance without strictly being homologous. Therefore, it is better to perform different searches and merge results into a consensus annotation. Obtaining the best functional annotation in any case will still require intensive manual curation and experimental validation through *in vivo* or *in-vitro* assays.

5. Genomic Data Sharing, Dissemination, and Usability

A significant challenge in whole-genome sequencing projects is to make the amount of post-genomic datasets and resources accessible and reusable by the scientific community. Some challenges in sharing and disseminating genomic data include the large data volume (big omics data), the complexity of data structure, formats, nomenclature, the metadata description, and the choice of appropriate repositories to deposit these data. Datasets and resources, including genome and transcriptome assemblies, sequencing reads, genotyping data, codes, and workflows generated during this thesis, have been made accessible in public data repositories following the FAIR standards where possible. I will briefly introduce some of these principles here and the used omics data repositories.

5.1. FAIR Genomics Data Principles and Standards

FAIR data principles are simple guidelines to ensure that researchers can find, reuse, and interpret scientific data together with the tools and workflows that led to these data.

Making Data Findable

The produced (omics)-data and metadata should be discoverable, uniquely identifiable and locatable by means of standardized identification mechanisms such persistent and unique identifiers (e.g., Digital Object Identifiers) [75]. Here, community adopted naming and conventions should be followed and version numbers of the data clearly provided.

Making Data Openly Accessible

The findabilty of omics data is ensured by making data openly available in public genomic repositories as well as documenting the methods and softwares needed to access the data.

Making Data Interoperable

To make genomic datasets interoperable, it is essential to share them in standard file formats and ontologies compliant with open software applications, to enable smooth data exchange and reuse between researchers.

Making Data Reusable

To guarantee data reuse, it is important to clarify any licences or restrictions on the data. Moreover, data should be released with a clear and accessible data usage license, when applicable.

5.2. Public Genomic Data Repositories

Genomic data generated in this project have been submitted in the following public omics databases, mostly hosted and maintained by the National Center for Biotechnology Information (NCBI):

- **Sequence Read Archive (SRA)** is a bioinformatics database that provides a public repository for DNA high-throughput sequencing data, such as short reads and long reads data. SRA stores raw reads along with associated alignment information to enhance reproducibility and facilitate new discoveries through data analysis [76].

- **NCBI reference sequences (RefSeq)** is a curated non-redundant sequence database that organizes information on genomes, including transcripts, proteins, maps, chromosomes, and annotations [77].

- **The Single Nucleotide Polymorphism Database (dbSNP)** is public resource for genetic variation, in particular SNPs genotyping data.

- **The European Variation Archive (EVA)** is an open-access database of all types of genetic variation data from all species. EVA archives genotyping and variant data including, but not limited to, SNPs, SVs, Indels, short tandem repeats (STRs), and transposable elements (TE).

Most of these databases provide tools and application programming interface (API), for fully automated and programmatic upload and retrieval of genomic (meta)data and information. How to submit, retrieve, and access omics data in these public repositories is beyond the scope of the thesis.

Pics/IntroGenerall/Assembly_illustation.pdf

Part II.

The Pikeperch Reference Genome, Fundamental Resource for the Development of Customized and Smart Breeding Programs

The analyses and the results presented in this section of the thesis have already been published in peer-reviewed scientific journals. Hence, **Part II** is **substantially** based on the following published manuscripts:

- **Nguinkal, J.A.**; Brunner, R.M.; Verleih, M.; Rebl, A.; de Los Ríos-Pérez, L.; Schäfer, N.;Hadlich, F.; Stüeken, M.; Wittenburg, D.; Goldammer, T. **The First Highly Contiguous Genome Assembly of Pikeperch (Sander lucioperca), an Emerging Aquaculture Species in Europe**. Genes (Basel) **2019**,10(9):708.

  Contribution: *I designed and performed the bioinformtics analyses including pipelines development, data visualization and validation. I additionally wrote the manuscript.*

- De Los Ríos-Pérez, L.; **Nguinkal, J.A.**; Verleih, M.; Rebl, A.; Brunner, R.M.; Klosa, J.; Schäfer, N.; Stüeken, M.; Goldammer, T.; Wittenburg, D. **An ultra-high density SNP-based linkage map for enhancing the pikeperch (*Sander lucioperca*) genome assembly to chromosome-scale**. Sci Rep **2020**, 10(1):22335.

  Contribution: *I performed the bioinformtics analyses, visualization and validation. I contributed substantially in writing parts (sections) of the manuscript.*

6. Introduction

Dedicated breeding programs have substantially optimized rearing conditions and enhanced the productivity of a multitude of aquaculture species [78, 79, 80, 81, 82], showing that, smart and tailored breeding programs are crucial to support further growth of the sustainable aquaculture sector. The development of species-specific breeding programs in aquafarming is based on economic goals of the industry and relies on the comprehensive understanding of the genomic architecture of commercial traits, and the elucidation of genetically informed breeding parameters [83]. As a promising European aquaculture species, the elucidation of the pikeperch genome along with its genes and other functional features should provide an essential resource and fundamental meta(data) towards the development of customized, smart breeding approaches.

This chapter reports the first nearly complete high-quality reference genome sequence of pikeperch along with a comprehensive annotation and analysis of its functional features. This first ever developed genomic resource for pikeperch should pave the way for the investigation of genotypes associated with key production traits, analysing family and population structures of domesticated animals, as well as to aid the identification of genetic-informed indicators for monitoring the fish growth, health and welfare.

The assembly and annotation were constructed using different methods integrated into customized assembly and annotation pipelines. Therefore, I used long-read sequencing by PacBio and took advantage of accurate Illumina short reads to improve the base-level quality of the assembly and gene prediction reliability. The chromosome-level assembly was achieved by integrating an ultra-high-density SNP-based linkage map into the scaffolding process. I subsequently applied different approaches to evaluating the assembly, including reads alignment statistics, gene space statistics, and comparative alignments with other teleosts, which has revealed this genome assembly, to my knowledge, as the most complete in the *Percidae* fish family.

7. Materials and Methods

Whole genome sequencing is a modular project that involves a variety of methods, software tools and hardware requirements at each step, throughout data generation to data analysis and dissemination (Figure 7.1).

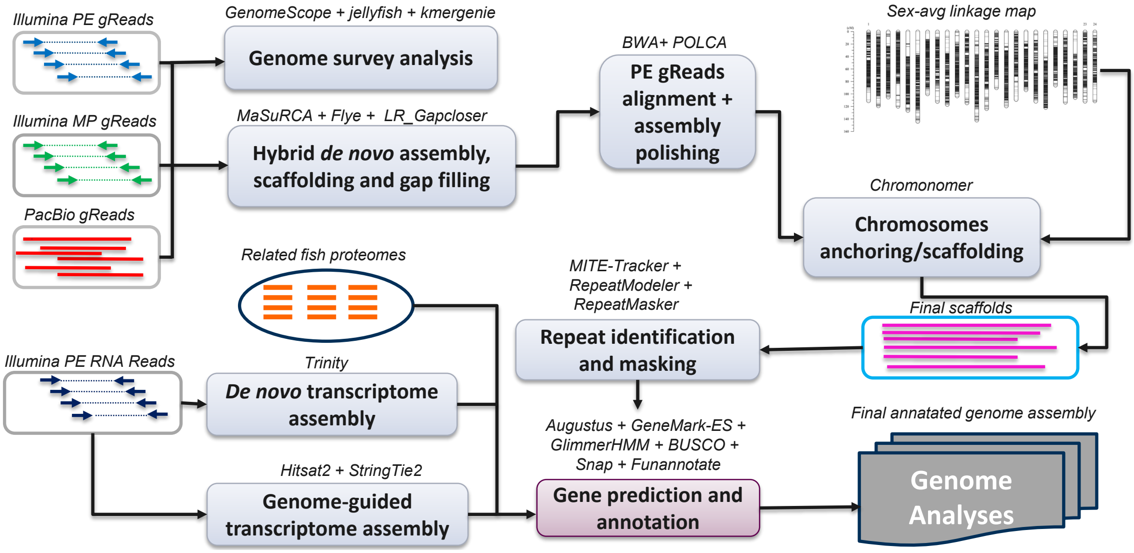
Illumina PE gReads
Illumina MP gReads
PacBio gReads
GenomeScope + jellyfish + kmergenie
Genome survey analysis
MaSuRCA + Flye + LR_Gapcloser
Hybrid de novo assembly, scaffolding and gap filling
BWA+ POLCA
PE gReads alignment + assembly polishing
Sex-avg linkage map
Chromonomer
Chromosomes anchoring/scaffolding
Related fish proteomes
MITE-Tracker + RepeatModeler + RepeatMasker
Repeat identification and masking
Final scaffolds
Illumina PE RNA Reads
Trinity
De novo transcriptome assembly
Hitsat2 + StringTie2
Genome-guided transcriptome assembly
Augustus + GeneMark-ES + GlimmerHMM + BUSCO + Snap + Funannotate
Gene prediction and annotation
Final annatated genome assembly
Genome Analyses

Figure 7.1.: Modular workflow including generation of input data, assembly and annotation pipelines and post-assembly analyses.

7.1. Hardware Platforms

Data generated in sequencing projects is ordinarily large, and genome assembly and annotation require specific hardware resources depending on the size of the input data. The following hardware platforms and resources were used here: Small automation scripts and analyses were performed on a Linux-based desktop computer equipped with eight cores and 32 GB of RAM. The sequencing data and annotation tasks were run in-house on the FBN compute server, requiring up to a 2 TB storage peak and 500 GB RAM for certain computations. More computationally intensive pipelines and tasks such as running the MaSuRCA assembly toolkit were deployed on a dedicated virtual machine on the de.NBI* OpenStack cloud system with up to 1.5 TB storage and 256 GB of peak memory.

*de.NBI is the German Network for Bioinformatics Infrastructure, https://www.denbi.de/cloud

7.2. Software and Application Tools

Most software tools used in the analyses performed in this thesis are community-developed and maintained open-source bioinformatics softwares. The different tools will be referenced where mentioned for the first time. The task-specific analysis pipelines were developed by automating workflows involving multiple softwares/tools, by means of bash and python scripting. Most software tools were installed, packaged and deployed in Anaconda [84] and Mamba [85] environments, or using containers like Singularity [86] and Dockers [87]. Downstream genomics analyses were performed mostly using Bioconductor packages in R. Data wrangling, integration and visualization were primarily performed in the R tidyverse ecosystem, and secondarily with the panda and numpy libraries.

7.3. Genome Survey Analysis

Every genome project is different because distinctive properties of the genome influence the decision on the type and amount of data needed, the suitable methods and techniques to use for analyses, as well as the overall budget [74]. Genome survey analysis aims to provide a preliminary understanding of the genomic characteristics before large-scale genome sequencing for the species of interest. Prior to sequencing, I investigated the properties of the pikeperch genome, including expected genome size, repeat content, heterogeneity, ploidy level, GC-content, and sex-determination system (SDS). I was not able to clarify all these properties. However, a summary of some of the putative or expected characteristics of the pikeperch genome to be sequenced is reported in Table 7.1.

Table 7.1.: Summary of pikeperch genome survey based on literature and previous work.

Genome Properties	Prospective Information	References
Genome size	C-value (pg): 1.14 ⇔ 1114 Mb	[40]
Ploidy level	Diploid ($2n = 48$)	[41, 42]
SDS	XY heterogametic	[43]
Genome sequence available	No	—
Transcriptome sequence available	No	—
Heterozygosity rate	??	—
Repeat content	??	—
GC content	??	—

7.4. High-throughput sequencing - Experimental Design

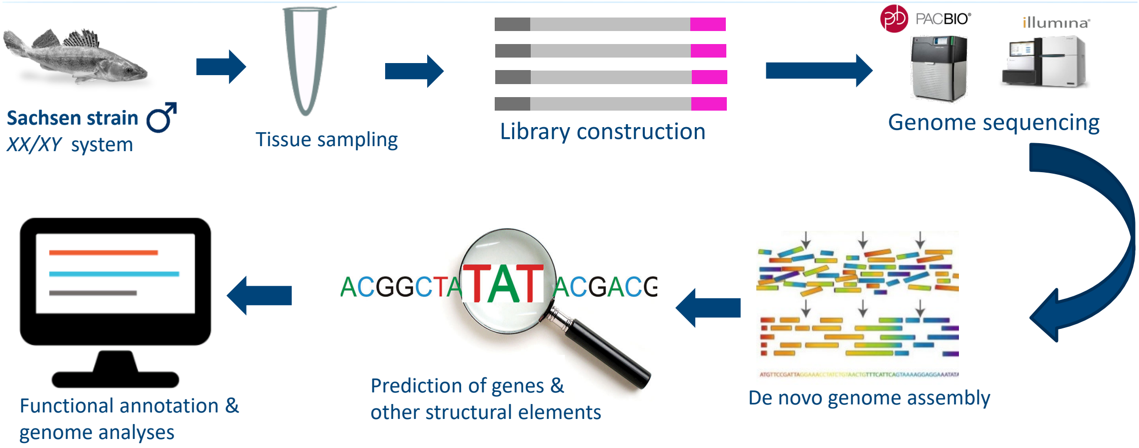

Figure 7.2.: Overview of NGS project's experimental design and workflow. A male Sachsen strain of pikeperch (fish) was sequenced with NGS technologies including PacBio and Illumina, followed by *de novo* assembly and annotation.

Sample Collection, Library Preparation

The sample tissues were obtained from a single adult male*S. lucioperca*, collected at the state's aquaculture facilities in Hohen Wangelin, Germany. Genomic DNA was extracted from liver, muscle, and spleen tissues that had been previously isolated and stored in liquid nitrogen. All DNA samples were pooled for the library's preparation. For whole-genome sequencing, multiple types of libraries were used. One short insert (paired-end, 470 bp) shotgun library was prepared using Illumina's TruSeq DNA PCR-free library preparation kit. In addition, two size-selected mate-pair libraries with 2–8 kb and 2–10 kb long inserts were prepared following the Nextera mate-pair library preparation protocol. To overcome the limitations of short reads for the assembly of complex eukaryote genomes, 20 kb large-insert PacBio libraries were also prepared according to the guidelines for preparing the SMRTbell template for sequencing on the PacBio Sequel System.

Whole Genome Sequencing, Quality Control

The size-selected 20 kb DNA libraries were pooled and sequenced in 10 single-molecule real-time sequencing (SMRT) Cells on the PacBio Sequel II systems according to the SMRT® sequencing guide. In total, 66 Gb of raw data, accounting for 6.4 million polymerase reads, were generated. Polymerase reads were trimmed using SMRT Link v6.0.0 to obtain 5.2 million high-quality subreads (Table 8.1). Additionally, one paired-end and two mate-pair libraries were sequenced on the Illumina HiSeq X Ten platform. The average length of a short insert was 470 bp, while long inserts ranged from 2 to 10 kb.

To assess the overall quality of sequencing data, FastQC (version 0.11.8) [88] was applied. Trimmomatic (version 0.38) was used to trim adapters, filter low quality reads (poorer quality $< Q28$), and discard reads that were less than 40 bp. Since the mate-pair reads contained overrepresented sequences (about 0.1–1.2%), which probably originated from TruSeq adapter contamination, I iteratively removed them using fastp (version 0.19.3) [89].

7.5. Genome Characteristics Estimation — k-mers Analysis

Extensive knowledge of basic genome properties such as genome size, repeat content, and heterozygosity rate supports the decision for an appropriate assembly strategy and adequate parameter tuning. K-mer profiles analysis is an efficient, assembly-independent approach to estimate these genome characteristics prior to assembly. I generated k-mer profiles from high-quality genomic paired-end reads using the program Jellyfish (version 2.2.10) [90], and a custom R script to estimate the genome size of *S. lucioperca*. As applied in previous studies [91, 92, 93, 94], the genome size G was calculated with the following formulas: $N = M * L/(L - k + 1)$ and $G = T/N$, where N is the mean PE-reads coverage, M the mean k-mer depth, L the mean read length, k the k-mer size, and T the total number of base pairs. To evaluate the robustness of this method, I applied different k-mer lengths, with $k \in \{17, 19, 21, 31\}$. The estimated genome size ranged from 1006.86 Mb (k = 17) to 1024.35 Mb ($k = 31$), depending on the k-mer length (Table 7.2).

The genome size estimate of 1014.28 Mb ($k = 19$) was considered to be more reliable, as a k-mer of 19 is long enough to yield fairly specific genomic sequences, but also short enough to give sufficient data (Figure 7.3). Low coverage (< 50) 19-mers with high frequency are putative erroneous k-mer, whereas deep coverage (> 450) k-mer with low frequency more likely originate from repetitive genomic sequences. The frequency of 19-mers follows a bimodal distribution with two distinct main peaks, α (heterozygous 19-mers) and β (homozygous 19-mers), indicating a heterozygous genome [95]. However, using the GenomeScope R-script [93], the heterozygosity rate, which is proportional to the α/β ratio, was estimated to be 0.14% (14 SNPs per 10 kb). The k-mers located in single-copy regions of the genome will appear uniquely in the k-mers profile and will fit the non-stationary portion of the k-mer histogram.

In the analysis summarized in Figure 7.3, these are 19-mers with depths between 150 and 450. Hence, the total length of unique genomic regions (i.e. single copy portion) is estimated by the area spanned by unique k-mers divided by the depth of the maximal k-mer frequency (here β peak) [96]. Based on the 19-mer histogram in Figure 7.3, the single copy portion in the pikeperch genome is estimated to approximately 55% and obtained with the following formula:

$$SC = \sum_{c=150}^{450} c \cdot freq_c / B$$

where SC is the single copy size (in bp), c is the k-mer depth, $freq_c$ the corresponding frequency and B the depth of the main peak β. Consequently, we expect repeated sequences including duplicated genes, interspersed and tandem repeats to account for about 45% of the pikeperch genome (Table 7.2).

Table 7.2.: Summary of genome characteristics based on k-mer analysis.

	Mean k-mer depth (bp)	**Estimated heterozygosity (%)**	**Estimated genome size (Mb)**	**Single copy size (Mb)**	**Single copy portion (%)**
$k = 17$	367	0.106	1006.86	533.41	53
$k = 19$	358	0.117	1014.28	552.45	54.5
$k = 21$	350	0.108	1024.35	562.85	55
$k = 31$	318	0.12	1039	645.33	62

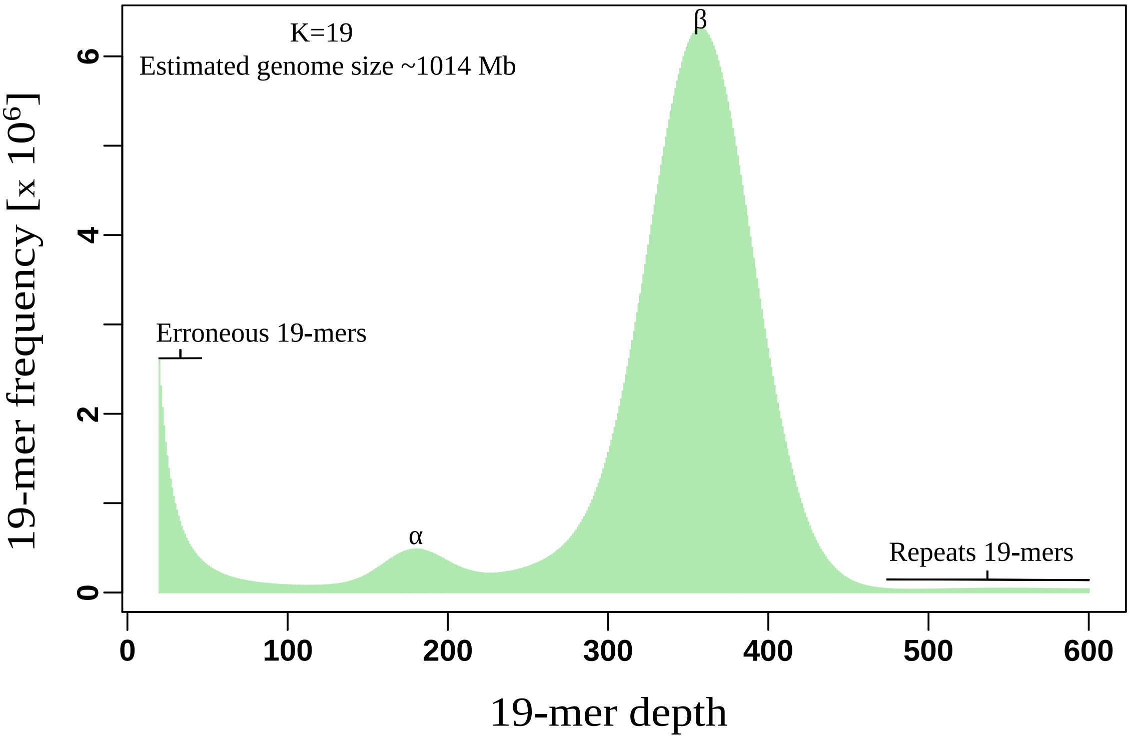

Figure 7.3.: Histogram of 19-mers distribution. K-mer histogram depicting estimated characteristics of *Sander lucioperca* genome based on 19-mer analysis. The horizontal axis represents the 19-mer depth, and the vertical their corresponding frequency. α is the heterozygous and β the homozygous peak. Low coverage (< 50) 19-mers are putative erroneous sequences, whereas deep coverage (> 450) 19-mer indicate repetitive genomic sequences.

7.6. *De novo* Assembly, Scaffolding

Hybrid and Modular *de novo* Assembly

A hybrid assembly approach was adopted to generate the primary contigs using the MaSuRCA Genome Assembly and Analysis Toolkit vers.3.4.02 and by integrating PacBio LRs with Illumina SIPERS and LIPERs data (Table 8.1). The MaSuRCA assembler combines the benefits of DBGs and OLC assembly approaches to build high-confident hybrid assemblies. Flye (version 2.3.7) [97] was used to generate preliminary contigs with an optimized k-mer size of 19. Flye is a fast and accurate de novo assembler for long error-prone and noisy reads using an A-Bruijn graph to find preliminary inaccurate contigs. The inaccurate contigs are transformed into a repeats graph, which can tolerate a higher noise level than DBGs. The long reads are then iteratively mapped back to the repeat graph to accurately resolve repeats and polish the contigs to a final assembly of high nucleotide-level quality.

Chromosome-scale Scaffolding with Genetic Maps

To anchor the initial assembly into the chromosomal framework, flanking sequences of the SNP loci (i.e., 100 bp upstream and downstream from the SNP position) of the sex-average genetic map generated in Part III were extracted and aligned to the newly assembled genome with BWA [98]. In total, 723,360 SNPs markers were uniquely mapped to 706 contigs which were ordered and integrated into 24 pseudomolecules using the software Chromonomer v1.11 [99] (Figure 7.1). The genome was polished in three iterations with POLCA polishing module [66] using short paired-end reads.

7.7. Assembly Quality Assessment and Validation

To evaluate the quality of the assembled genome, I analyzed gene space and k-mer completeness with BUSCO and Merqury [100], respectively. BUSCO provides quantitative measures for the assessment of assembly completeness in regard to the expected gene content, while the k-mer completeness informs how the assembly accurately represents the input reads, with a value of $> 97\%$ being indicative of a highly confident assembly. The structural accuracy of the assembly was additionally validated by mapping genomic paired-end reads obtained by re-sequencing of 394 conspecific individuals and gauging the rate of concordantly mapped reads. Additionally, misassemblies and base-level accuracy were evaluated.

7.8. Genome Structural and Functional Annotation

7.8.1. Genome Structural Annotation

Structural feature were annotated using homology to closely related species as well as intrinsic and extrinsic methods. Structural elements include different types of repetitive DNA sequences, protein coding genes, and diverse classes of ncRNA.

Repetitive Elements Annotation

RepeatModeler version 2.0.1 [101] was used to identify transposable elements (TE) in the genome assembly. To specifically identify miniature inverted-repeat transposable elements (MITE), the open source software MITE-Tracker was applied. Subsequently, the outputs from RepeatModeler and MITE-Tracker [102] were combined with FishTE database (`http://www.fishtedb.org/`) and RepBase [103] into a non-redundant custom repeat library for repeatMasker (`http://repeatmasker.org`), that was used to classify repetitive elements and estimate their distribution in the pikeperch genome. Finally, the landscape of segmental duplications (SDs), which are low copy repeats in the genome, was characterized using the Segmental Duplication Evaluation Framework (SEDEF) [104]. SDs < 5 Kb and with an identity score $< 90\%$ were discarded.

Prediction of Gene Models

To computationally annotate putative protein-coding genes in the pikeperch genome, I combined ab initio and homology-based methods along with RNA-Seq evidence in a customized annotation pipeline. In homology-based gene prediction, I obtained homologous protein sequences from closely related percid species, including walleye (*Sander vitreus*), yellow perch (*Perca flavescens*) [105], European perch (*Perca fluviatilis*), Arkansas darter (*Etheostoma cragini*), and orangethroat darter (*Etheostoma spectabile*) [106]. TBLASTN vers.2.5.0 [107] was used to align these homologous protein sequences to the reference pikeperch genome with an e-value cutoff of 1e-6, thereby retaining only the top scoring alignment for each protein with a minimum identity of 80%. Exonerate vers.2.4.0 [108] was then used to map these top-scoring proteins to the repeats-masked *Sander lucioperca* genome in order to predict putative gene models.

In the *ab initio* approach, I combined different intrinsic gene finders including AUGUSTUS [109], GENSCAN [110], GlimmerHMM [111], SNAP [112], and GeneMark [113] to predict gene structures on the repeat-masked genome. While Augustus was trained with randomly selected full-length protein-coding genes as predicted by Exonerate, GENSCAN was run with human genome parameters. The transcript-based gene prediction was performed using RNA-Seq data of a conspecific individual, whose paired-end reads were obtained from the Sequence Read Archive (SRA), (Accession-Nr: SRR2871497).

These reads were mapped to our pikeperch genome using HISAT2 vers.2.1 [114], a splice-aware aligner, to detect splice junctions. Cufflinks (version 2.2.1) [115] was subsequently used to assemble transcripts based on HISAT2 alignments and to build gene structure hints. The gene models prediction from the three methods was integrated using Evidence-Modeler [116], to build a consensus, non-redundant *Sander lucioperca* gene set. Ultimately, the resulting gene set was filtered to remove genes that had no start and/or stop codon, or had an in-frame stop codon, or had a coding sequence (CDS) shorter than 150 nt. Genes with significant open reading frame (ORF) homology to TE sequences were also discarded.

Moreover, I predicted four main classes of non-coding RNAs, which play important roles in different cellular processes. Transfer RNA (tRNAs) were predicted using tRNAscan-SE vers.2.0 with eukaryote parameters [117]. Eukaryotic ribosomal RNAs (rRNAs) were annotated utilizing the software package RNAmmer vers.1.2 [118], and putative micro RNAs (miRNAs) were predicted by homology to the known mature miRNAs sequences available in the miRBASE database [119], by using the miRDeep2 pipeline [120].

7.8.2. Genome Functional Annotation

Functional annotations of predicted protein-coding genes were carried out based on several functional databases, including Swissprot, Tr-EMBL, NCBI-NR, KEGG, and eggNOG. I also used InterProScan (version 5) [121] to map protein domains in the InterPro database, which includes CATH-Gene3D, CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMAT, SUPERFAMILY, and TIGRFAMs. Lastly, the high-scoring functional annotations in each database were retained as the final consensus functional annotation results.

8. Results

8.1. Genome Sequencing

Whole genome shotgun (WGS) strategy was used to produce 412.8 Gb (350X genome coverage), 74.2 Gb (66X genome coverage) and 71.4 Gb (63X genome coverage), corresponding to data yielded by Illumina paired-end, 2-8 kb and 2-10 kb mate-pairs libraries, respectively. In addition, 66 Gb (60X genome coverage) data were generated with size selected 20 kb PacBio libraries. The mean read length for Illumina data was 150 bp. The PacBio read data had a mean read and N50 length of 12.7 kb and 16.4 kb, respectively (Table 8.1). The paired-end reads were primarily used for genome properties estimation, to assess and improve the base-level quality of the assembly. The estimations based on k-mer analysis have shown that the pikeperch genome is as large as 1014 Mb, which is consistent with the previous estimate of 1114 Mb, based on cytometric methods [40]. The k-mer analysis also revealed that, one should theoretically expect about 45% of repetitive DNA sequences, since the single copy portion in the pikeperch genome was roughly estimated to 55%.

Table 8.1.: Summary statistics of generated whole genome sequencing data of pikeperch

Platform	Library Type	Insert Sizes	Total No. of Reads	Total No. of bp (Gb)	Mean Coverage	Mean read Length (bp)
Hiseq X Ten	Paired-end	470 bp	2,761,296,894	412.8	~350X	150
Hiseq X Ten	Mate-pair	2-8 kb	491,748,132	74.2	~66X	150
Hiseq X Ten	Mate-pair	2-10 kb	473,090,686	71.4	~63X	150
PacBio Sequel	—	20 kb	5,258,946	66.5	~51X	10,891

8.2. A Highly Accurate Chromosome-scale Assembly

The draft genome preliminary consisted of 1602 contigs with an N50 length of 6.6 Mb and spanned 900,47 Mb in total. In particular, 75.8% of the draft genome is covered by 207 contigs larger than 1 Mb, and only 3.9% of the genome is spanned by contigs shorter than 100 kb. The integrated final chromosome-scale assembly yielded 336 scaffolds with an N50 of 41.06 Mb. The 24 largest scaffolds represented the putative 24 pikeperch chromosomes, and covered 896.48 Mb (99.47%) of the total assembly size (901 Mb). Three hundred twenty-two (322) short contigs/scaffolds covering 4.74 Mb (0.53% of assembly size) were not placed into pseudomolecules. This chromosome-level assembly also yielded five contigs > 17 Mb spanning full chromosome arms, indicating high assembly contiguity. The average base-level accuracy (QV) was 99.9996 (Q50, i.e., one error in 100 kb), corresponding to the defined VGP (VGP-2020) accuracy standards [100]. Approximately 99.80% of genomic PE-reads were obtained by resequencing of ca. 400 conspecific individuals have been confidently aligned to the chromosome-level assembly with concordant read mapping at a rate of 97.50%. This high mapping rate of external reads of conspecific individuals is a solid and extrinsic cross-validation of the high structural accuracy of this assembly (Figure 8.1). Moreover, from a total of 4584 actinopterygians core genes, BUSCO assessment recovered 96.27% as full-length single-copy (with 2.45% being duplicated), 1.94% as fragmented and only 1.79% were missing. Similar rates were also obtained with Vertebrates single-copy orthologs, indicating that most genic regions were accurately assembled and that the gene space spanned by this assembly is nearly complete (Table 8.2). Collectively, these assessment scores show that this assembly is the most contiguous and complete genome published so far in the *Percidae* fish family (Figure 8.1, Table 8.3).

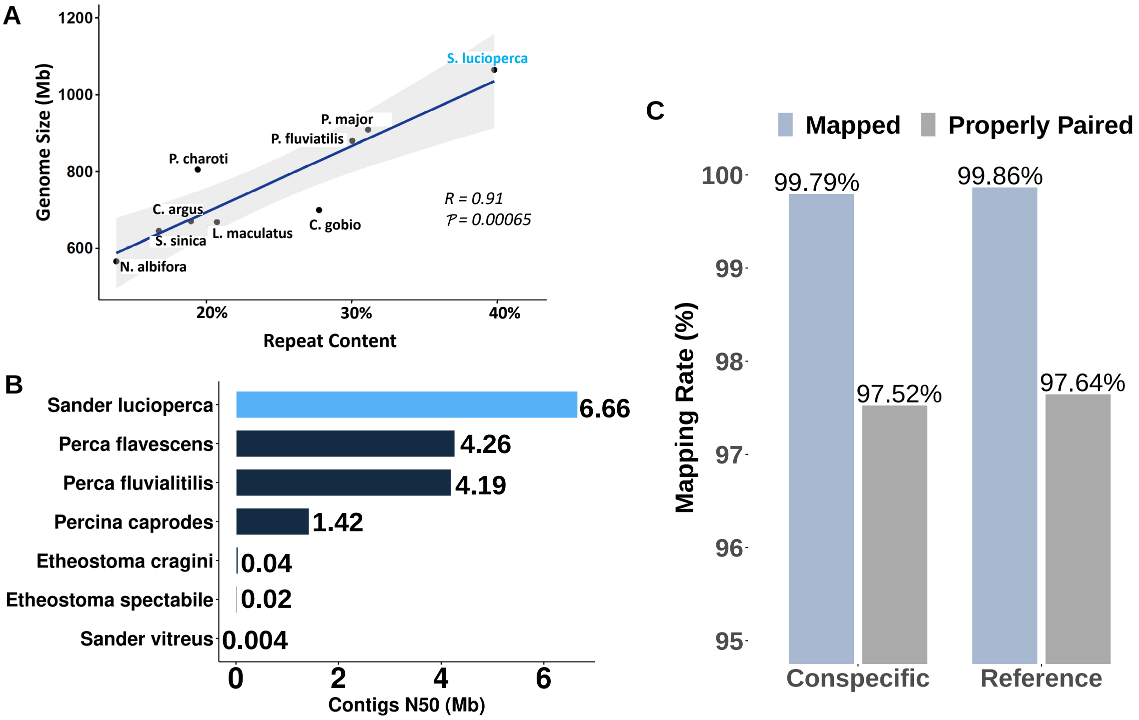

Figure 8.1.: Assembly assessment statistics. (A): Strong positive correlation between repeat content and the genome sizes in recently published genomes of Perciformes fish species. R is the Pearson's correlation coefficient and p the associated p-value. **(B)**: The contiguity (as contigs N50 metric) of the pikeperch assembly compared to all currently published genome assemblies in the *Percidae* fish family. **(C)**: Mapping rates of genomic paired-end reads from resequening of 394 pikeperch individuals (conspecific), compared with PE-reads of the same fish used for *de novo* assembly (reference).

Table 8.2.: Summary statistics of BUSCO analysis for *Sander lucioperca* genome assembly

	Actinopterygii		**Vertebrata**	
Categories	#Genes	Percentage (%)	#Genes	Percentage (%)
Complete single-copy	4413	96.27	2523	97.56
Complete duplicated	112	2.45	26	1.01
Fragmented	89	1.94	40	1.54
Missing	82	1.79	23	0.89

Table 8.3.: Comparison of *Percidae* genome assemblies currently available in GenBank/NCBI

Percid Species	**Total Size (Mb)**	**Gaps Size (%)**	**Contigs N50 (Mb)**	**Unplaced Size (Mb)**	**Completeness BUSCO (%)**
S. luciopera	901.23	**0.02**	**6.66**	**4.74**	**96.27**
S. vitreus	782.90	—	0.004	—	78.87
E. spectabile	854.80	0.47	0.026	148	94.10
E. cragini	643.07	0.51	0.045	14.81	92.67
P. fluviatilis	951.34	0.03	4.19	9.58	89.00
P. flavescens	924.95	0.04	4.26	10.65	93.50
P. caprodes	1011.96	—	1.42	—	85.23

8.3. Architecture and Organization of the Pikeperch Genome

Repeat Content

In total, repetitive sequences accounted for ca. 39% of the assembled genome and spanned 334 Mb, which is in range with the repeat content reported in other Percidae fish [105]. With more than 250 Mb (27.76% of assembly size), DNA transposons and retroelements were the most abundant type of repeats found in the pikeperch genome. In particular, long interspersed nuclear elements (LINEs), long terminal repeat (LTR) elements, and hobo-Activator occupied 10.16%, 3.22% and 4.94%, respectively, of the assembled genome (Figure 8.2A, Table 8.4). More than 46% of the base pairs are contained in SD regions which are uniformly distributed across the genome with the exception of chromosomes 4, 5 and 8 which have a higher SD density per Mb (Figure 8.3). Overall, 2050 MITEs families spanning 22.7 Mb were also predicted genome-wide. I investigated the correlation between repeat content and the genome size of the most contiguous assemblies of Perciformes species, which have recently been published, assuming that a strong positive correlation might support a coherent prediction of *S. lucioperca* repeats. As it was expected based on prior knowledge, the repeat content was strongly correlated with genome sizes (Pearson $R = 0.91$, $p = 0.00065$), whereby pikeperch had the largest genome length and repeat content among the compared Perciformes (Figure 8.1).

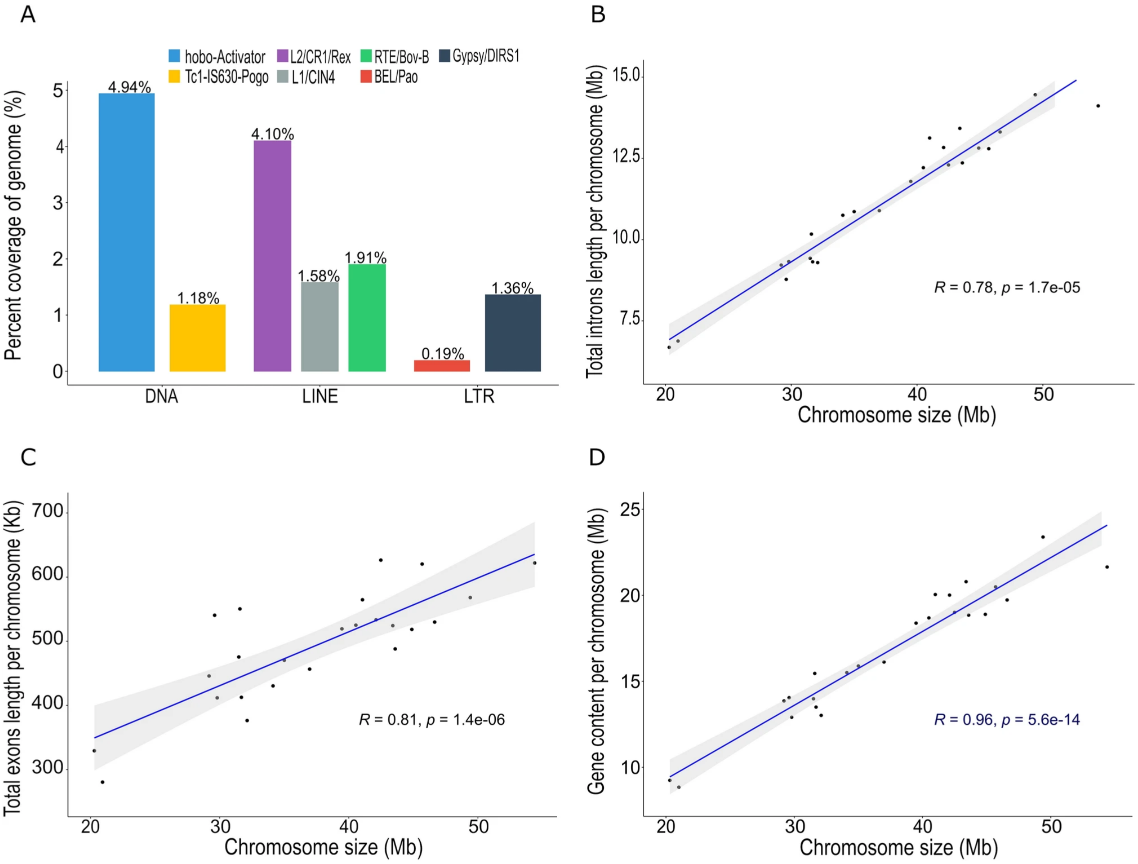

Figure 8.2.: Summary of TE elements and genomic features. (**A**):The percentage coverage of the most abundant families of transposable elements in pikeperch. LINE: long interspersed nuclear elements; LTR: long terminal repeat. Correlation between (**B**) otal introns length, (**C**) total exons length, and (**D**) gene content per chromosome and chromosome size (Mb).

Gene Models

Like in most sequenced percid species, the pikeperch genome comprises 24 diploid chromosomes, as revealed by this assembly. The obtained consensus gene models included a total of 33,456 high-quality gene models with an average coding sequence (CDS) length of 1451 bp. The genes possibly express up to 56,557 different proteins (Table 8.5). On average, each *S. lucioperca* gene comprises 7.8 exons, each with an average length of 156 bp. About 82% of the 278,346 exonic sequences were shorter than 200 bp, while introns showed an average length of 2276 bp, with only 2% of them being longer than 10 Kb. The total lengths of intronic and exonic DNA on each chromosome were significantly correlated to the chromosome size with correlation coefficients of $R = 0.78$ and $R = 0.81$, respectively (Figure 8.2B,C). Consequently, the gene content per chromosome was also in strong dependence to the chromosome size, with a correlation coefficient of $R = 0.96$ (Figure 8.2D). All together, the distribution of CDS length, introns length and exons number is comparable with other percid genomes [105]. The 24 chromosomes were sorted by physical size in bp, from largest to smallest and named accordingly (Table 8.5, Figure 8.4). Given a genome-wide average of 40 genes per Mb, the chromosomes 21 and 23 displayed the highest and lowest gene density with 52 and 34 genes per Mb, respectively. Additionally, we observed a putative nucleolus organizer region (NOR) on chromosome 7, which had already been observed in previous cytogenetics analysis on pikeperch [41].

8.4. Functional Annotation of Genomic Features

Homology and structure-based approaches were used for functional annotation of protein-coding genes. I was able to find 31,234 genes (93.36% of protein-coding genes) with at least one significant hit in the different functional databases queried for homology-based annotation (Table 8.4). Over 9277 (36%) genes were predicted to be involved in cellular processes and signaling functions, 3742 (14.62%) were associated with information storage and processing, and 3095 (12.10%) were associated with metabolism processes. A total of 7926 (30%) genes were poorly characterized or had unknown functions. A comprehensive annotation report is available in Appendix 22.4. Non-coding genes included 2345 transfer RNA (tRNA), 160 ribosomal RNA (rRNA) and 145 microRNA (miRNA) (Table 8.4).

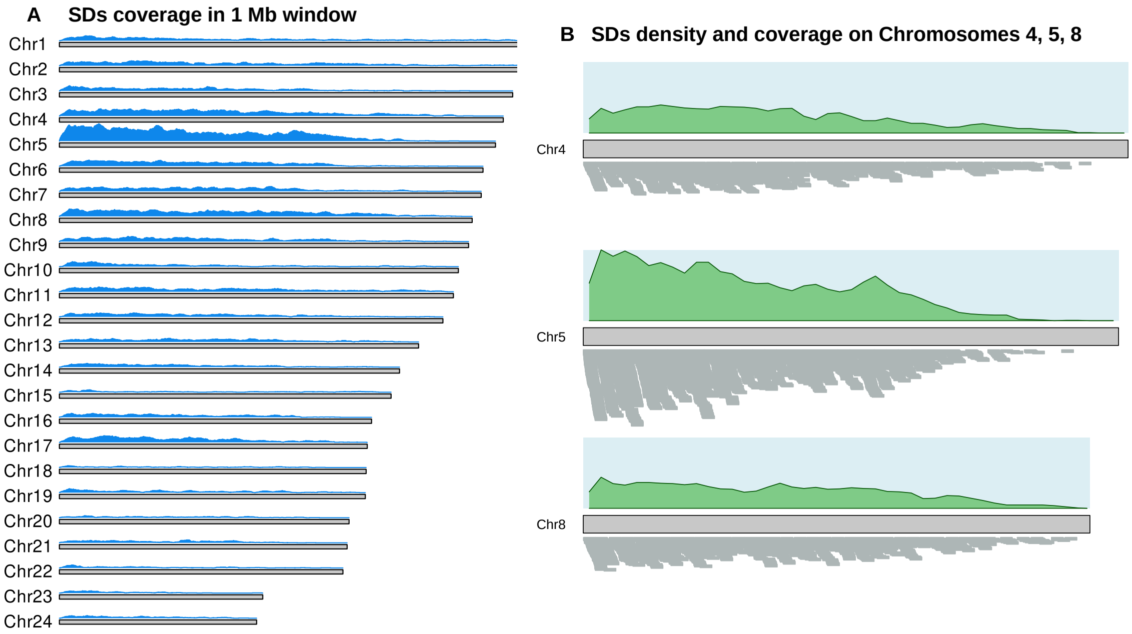

Figure 8.3.: Genome-wide SDs coverage in the pikeperch genome. (**A**): Genome-wide SDs coverage in 1 Mb window across pikeperch chromosomes (**B**): SDs coverage and density across chromosomes 4, 5, and 8.

Table 8.4.: Summary of the different assembly levels along with the landscape of structural genomic features and functional annotation of protein-coding genes.

	Chromosome-scale	**Draft assembly**
ASSEMBLY METRICS		
Total assembly size (bp)	901,221,791	900,477,756
Number of contigs	870	1602
Contig N50 length (bp)	6,668,792	2,995,800
Number of scaffolds	336	1313
Scaffold N50 length (bp)	41,060,379	4,929,547
Longest scaffold size (bp)	54,393,628	19,065,786
Scaffold L50	10	52
Base-level accuracy	99.9996 (QV50)	99.998 (QV40)
Σ Scaffolds >10 Mb (% of assembly size)	99.47	26.60
Σ Unplaced scaffolds (% of assembly size)	0.53	-
GC-content (%)	41.00	40.91
REPETITIVE DNA ANNOTATION		
DNA	27.76%	25.10%)
LINEs	10.16%	3.92%
SINEs	1.02%	0.69%
LTRs	3.22%	1.94%
SDs	46.30%	38.79%
SSRs	3.20%	3.78%
GENES ANNOTATION		
Number of genes	36,010	24,278
Number of protein-coding genes	33,456	21,249
Mean gene length (bp)	10,697	10,961
Mean CDS length (bp)	1451	1313
Mean exon count per CDS	7.80	6.70
Functional annotation	31,234 (93.36%)	18,536 (87.23%)
Mean intron length (bp)	276	1696
Mean exon length (bp)	156	196
% of genome covered by exons	3.82	3.11
Number of tRNA	2345	2313
Numer of rRNA	160	180
Number of miRNA	145	166

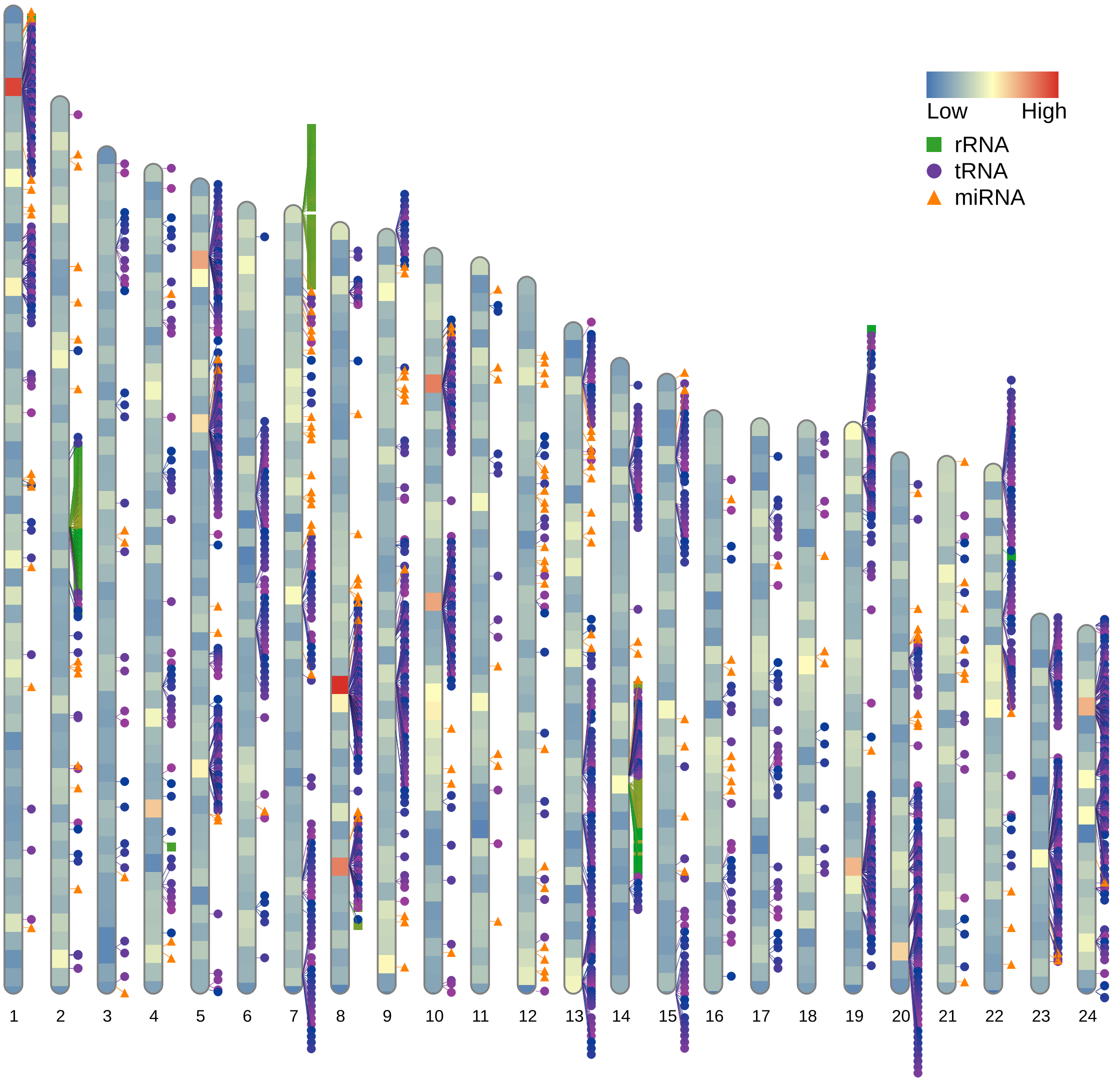

Figure 8.4.: Genome structure of pikeperch along with its gene density. (Gene density on each pikeperch chromosome ordered by length and distribution of non-coding RNA loci including miRNA (orange triangle), tRNA (purple circle) and rRNA (green square). The colour code within each chromosome represents the gene density from low (blue) to high (red) in a window of 1 Mb.

Table 8.5.: Summary of gene annotation statistics on chromosomes ordered by size with corresponding LG. LG: linkage group, Mb: Megabase.

ChrNr	LG	No. of Markers	No. of Contigs	No. of Genes	No. of Proteins	Physical Size (Mb)	Density (genes/Mb)
1	15	33,239	33	2095	3212	54.39	38.52
2	4	38,017	26	2071	3136	49.41	41.92
3	1	38,493	34	1598	2346	46.65	34.25
4	2	40,980	24	1846	2967	45.68	40.41
5	6	33,387	48	1675	2286	44.88	37.32
6	12	35,005	31	1722	2807	43.59	39.50
7	3	41,147	24	1692	2861	43.41	38.98
8	10	29,918	35	1739	2678	42.48	40.94
9	5	40,404	33	1793	3002	42.11	42.58
10	23	27,205	25	1777	2777	41.06	43.28
11	11	32,750	29	1668	2508	40.55	41.14
12	18	24,823	41	1697	2813	39.47	43.00
13	19	28,731	24	1329	2446	36.97	35.95
14	9	24,299	29	1260	2229	35.01	35.99
15	7	28,542	24	1249	2137	34.13	36.59
16	24	18,151	38	1263	1792	32.13	39.31
17	13	28,606	58	1288	2010	31.68	40.65
18	17	33,245	27	1383	2089	31.57	43.81
19	21	23,452	26	1288	1970	31.48	40.91
20	22	21,440	14	1202	1585	29.81	40.32
21	8	31,401	19	1531	2283	29.61	51.70
22	14	25,569	31	1254	2202	29.18	42.98
23	16	20,259	19	708	927	20.93	33.83
24	20	24,297	14	864	1390	20.30	42.57
Total	-	**723,360**	**706**	**35,992**	**56,557**	**896.48**	-
Average	-	*30,140*	*29.42*	*1500*	*2356*	*37.35*	*40.27*

9. Discussion and Conclusion

Discussion

In this study, I have used modern NGS technologies to sequence, assemble, and annotate the pikeperch's first chromosome-level genome, a promising species for aquaculture diversification in Europe. The pikeperch genome assembly was produced entirely with PacBio and Illumina sequencing data and integrated SNP-based genetic linkage map (LM) information. This approach has been used in several assembly projects of fish genomes, including Atlantic salmon [122], mandarin fish [123], or yellow perch [124]. However, chromosome-level assemblies harnessing LM data are becoming outdated because Hi-C technology offers a higher resolution and map of the chromosome architecture. Hi-C data can provide long-range linkage information across different length scales and spanning tens of megabases. Hence, this method has been established today as the standard approach for achieving chromosome-scale assembly [125]. Nevertheless, the scaffolding of high-quality long read assemblies assisted with LM data is still able to produce very competitive *de novo* genome assemblies, and in some cases, even outperforms Hi-C based scaffolding, as demonstrated in this work (Table 8.3, Figure 8.1).

The quality and accuracy of the reported genome assembly were assessed by state-of-the-art approaches such as the completeness of lineage-specific single-copy orthologs, estimating the mapping rate of genomic reads, or comparing the assembly and annotation metrics with closely related species. This assembly has outstanding contiguity metrics compared to recently reported assemblies of *Perciformes* fishes with even less complex genomes than pikeperch. In addition, the unprecedented rate of concordantly mapped paired reads is indicative of a highly contiguous and structurally accurate genome assembly. Nevertheless, further improvements on assembly structure and annotation are still required. The reported assembly still contains gaps, missamblies[*], and missing core genes at some places. Genome assembly has traditionally been an iterative process until full completion. For example, the complete gapless and telomere-to-telomere reference assembly of the human genome was achieved only this year (2021) by Nurk et al. [126], that is, exactly 20 years after the first draft genome sequence was released.

[*] missassembly is an assembly error where the same contig is mapped to different locations of a reference genome, suggesting that this contig was built by nonconsecutive genomic regions.

Conclusion

In summary, the genome assembly and sequencing data I have reported here are the most awaited genomic resources to pave the way for genomic studies such as genotyping by sequencing, genetic selection, and biodiversity on pikeperch. Such studies will provide an impetus for the industrial production of this species. The structural and functional annotations of genes provide the first overview of pikeperch's gene content and structure. These data represent the first step towards making deep sequencing data resources for this commercial fish species available in public HTS sequence databases. It will also serve as a basis and framework for the development of resources like population-scale genotyping and transcriptomics data in the following sections (Part III, IV) of this thesis.

Part III.

The Reference Transcriptome and Multi-tissue Expression Atlas – Powerful Tool for Functional Genomics in Pikeperch

The analyses and the results presented in the present Part of the thesis have already been published in a peer-reviewed scientific journal. Hence, **Part** III is **substantially** based on the following published manuscript:

- **Nguinkal, J.A.**; Verleih, M.; de los Ríos-Pérez, L.; Brunner, R.M.; Sahm, A.; Bej, S.; Rebl, A.; Goldammer, T. **Comprehensive Characterization of Multitissue Expression Landscape, Co-Expression Networks and Positive Selection in Pikeperch.** Cells **2021**, 10, 2289. https://doi.org/10.3390/cells10092289

  Contribution: *I designed the experiments and performed the bioinformtics analyses including data processing, visualization and validation. I additionally wrote the manuscript.*

10. Introduction

In the preceding section (Part II), the reference genome sequence of pikeperch was deciphered, and its global architecture was characterized. These data are fundamental genomic resources to understand organism biology. However, the functional features of the genome are better captured and understood through expressed genes, which are the basic functional units of genomic DNA [127]. Transcriptomics analyzes the complete set of RNA transcripts profiles produced by the genome, under specific circumstances, time points, or in a specific tissue or cell type. Transcriptome analysis is the state-of-the-art approach to gain a broader overview of cellular processes and interpret the functional elements of genomes, such as functional genes that correlate with crucial traits in aquaculture, including adaptation, growth, and disease resistance. Therefore, providing an annotated transcriptome sequence of pikeperch along with the landscape of gene expression will serve as another critical genomic resource and powerful tool for examining the relationship between the genotype and phenotypic traits with economical relevance [128]. It will additionally help to understand the genetic architecture of these traits comprehensively.

In this section of my thesis work (Part III), I used the reference genome developed in Part II as a framework along with deep RNA-Sequencing of ten vital tissues collected in eight pikeperch animals to build a high-confident and annotated trancriptome assembly and expression atlas, to characterize the tissue-specificity of genes expression and co-expression network modules. Pathway enrichment and protein-protein interaction network analyses were performed to characterize the unique biological functions of tissue-specific genes and co-expression modules.

11. Materials and Methods

11.1. Experimental Design, Multi-Tissues RNA-Sequencing

Tissue samples were collected from eight adult pikeperch individuals (3 males, 5 females) in the Experimental Aquaculture Facility of the Research Institute for Farm Animal Biology (FBN) in Dummerstorf/Germany. Prior to tissue collection, fish were euthanized by immersion for 15 min in an overdose of 2-phenoxyethanol (50 mg/L) followed by a bleed cut in the head as well as cutting of the spinal cord posterior to the head. For each individual, different tissues including gonads (testis or ovary),mliver, spleen, muscle, gills, brain, head kidney, skin, and heart were sampled and snap-frozen in liquid nitrogen. They were ultimately transferred to a −80 °C freezer until required for RNA extraction. In total, eight individuals were euthanized and 72 tissue samples were obtained. These samples were separately homogenized in 1 mL TRIzol reagent (Invitrogen, Darmstadt, Germany). Following phenol-chloroform extraction, the obtained RNA was purified using the RNeasy Mini Kit (Quiagen, Hilden, Germany) according to the manufacturer's protocol. Extracted RNA was quantified using the NanoDrop (Thermo Scientific™ NanoDrop 2000) and its integrity was assessed by electrophoretic profiling with Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Subsequently, the purified mRNA from the same tissue type were pooled. With exception of the gonads sample which were sex homogeneous, the other pools were sex heterogeneous in the ratio of 3 males vs. 5 females (Table 11.1).

11.2. *De novo* Transcriptome Assembly, Functional Annotation

Despite the availability of diverse tools and documented pipelines for the purpose of transcriptome assemblies, *de novo* transcriptome assembly from short PE-reads remains an extremely challenging task, which requires, like genome assembly, a customized use of different strategy to achieve the most optimal assembly. To achieve the most biologically meaningful and representative set of *S. lucioperca* transcripts, I setup an analysis pipeline combining different assembly strategies (Figure 11.1). *De novo* assembly algorithms included Trinity (version 2.8.1) [129], and rnaSPAdes (version 3.14.1) [130]. Trinity assembly was performed by pooling reads of all tissues and setting strand-specific parameters, whilst rnaSPAdes assembly was iteratively built with k-mer sizes of 27, 33, 55, 77, and 99. Additionally, I generated a genome-guided assembly with StringTie2 (version 2.1.2) [131].

Table 11.1.: Experimental design of multitissues RNA-Seq samples. F: female; M: male

Pools for RNA-Seq Libraries	Pooled Samples				Sex-specificity of samples
Pool SL01 He	S02	S03	S04	S73	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL02 HK	S06	S07	S08	S74	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL03 Mu	S10	S11	S12	S75	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL04 Li	S14	S15	S16	S76	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL05 Sk	S18	S19	S20	S77	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL06 Gi	S22	S23	S24	S78	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL07 Br	S26	S27	S79	/	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL08 Sp	S30	S31	S32	S80	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL09 Go ♂	S34	S35	S69	S72	Homogeneous $3 \times M$
Pool SL10 He	S37	S38	S39	S40	Heterogeneous $(3 \times M + 5 \times F\mathrm{f})$
Pool SL11 HK	/	S42	S43	S44	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL12 Mu	S45	S46	S47	/	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL13 Li	S49	S50	S51	S52	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL14 Sk	S53	S54	S55	S56	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL15 Gi	S57	S58	S59	S60	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL16 Br	S61	S62	S63	S64	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL17 Sp	S65	S66	S67	S68	Heterogeneous $(3 \times M + 5 \times F)$
Pool SL18 Go ♀	S36	S70	S81	/	Homogeneous $3 \times F$

Briefly, the combined RNA-Seq reads of all tissues were aligned to the latest pikeperch reference genome built in Part II using HISAT2 (version 2.2.0) [132]. Reads alignments were then assembled with StringTie2 setting the $' - merge'$ option to obtain a non-redundant set of transcripts across all tissue samples. Subsequently, raw assemblies were piled into a meta-assembly and large redundant transcripts were clustered using cd-hit-est (version 4.6.1) [133] with an identity threshold of 98%. Note that the clustering threshold of 98% used here, is slightly lower than the 100% suggested by Nakasugi et al. [134]. This approach was chosen because the pooled libraries of our tissue samples represent individuals with different genotypes. A too stringent identity threshold (e.g., 100%), would result in too many transcript variants remaining unclustered, and thus make the redundancy removal suboptimal. Finally, I used EvidentialGene tr2aacds pipeline [135] to collate the overassemblies into a less redundant and high confident transcript set, thereby maximizing the diversity and completeness of the final transcriptome assembly.

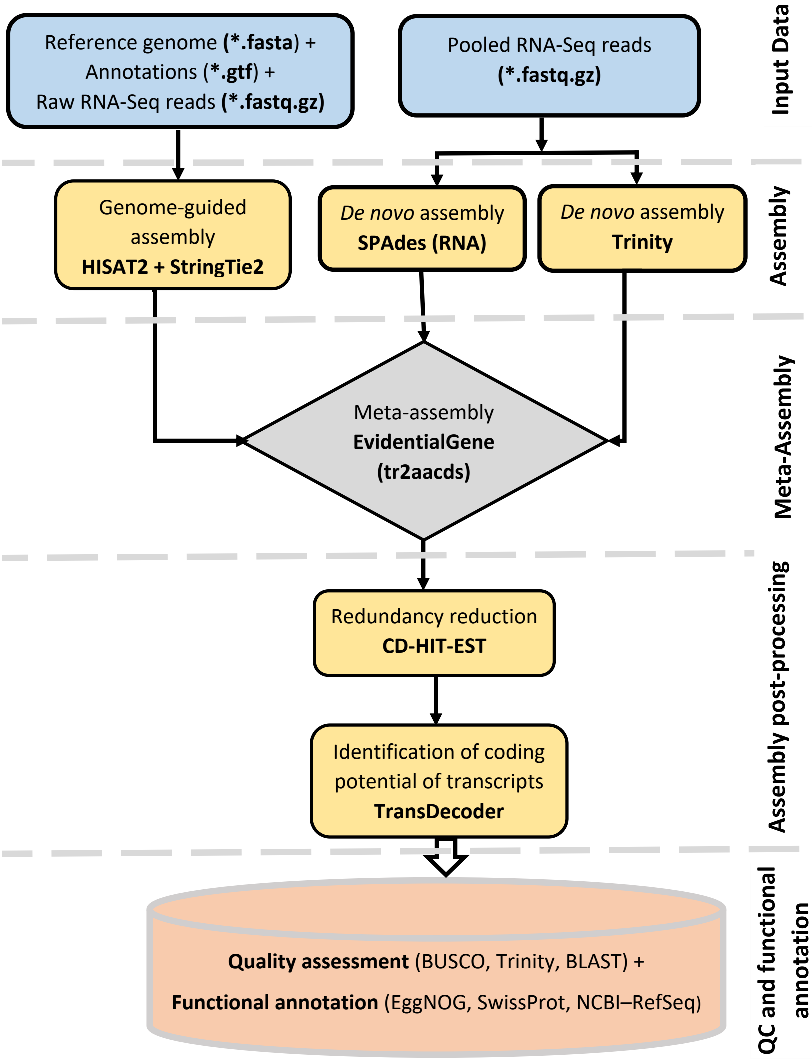

Figure 11.1.: Workflow of the different steps carried out in the RNA-Seq assembly pipeline. The diagram shows the main steps and bioinformatics tools of the RNA-Seq assembly pipeline.

The resulted multitissue transcriptome was functionally annotated using the eggNOG mapper (version 5.0) [136] as well as through homology search against protein sequence databases, including SwissProt and nonredundant RefSeq (NCBI) proteins. Moreover, I performed protein domain identification and functional sites mapping with InterProScan (version 5) [121]. The quality of our newly built transcriptome assembly was gauged using multiple strategies and quality metrics. Assembled transcripts were aligned with minimap2-splice option (version 2.21) [137] to the pikeperch reference genome to produce a GTF annotation file which was then compared to the pikeperch reference annotation (GenBank: GCF_008315115.2) with gffcompare (version 0.10.4) [138]. I also used BUSCO to explore the assembly completeness regarding the conserved actinopterygians (*Actinopterygii* dataset) single-copy orthologs. To assess the RNA-Seq reads representation of the assemblies, merged RNA-Seq reads were mapped to the assembly using Hisat2, and the mapping statistics were estimated. Finally, I queried (BLAST) the assembled transcripts against different protein databases, including UniProt and NCBI RefSeq nonredundant proteins, and performed full-length transcript analysis using a utility Perl script provided in Trinity.

11.3. Quantification of Tissues Expression Profiles

The trimmed and filtered RNA-Seq reads from each sample were individually mapped to the pikeperch reference genome (Part II) using STAR (version 2.7.5a) [139], in two-pass mode. Abundance levels of transcripts were estimated using TPMCalculator [140], a one-step software to quantify mRNA expression abundance directly from RNA-Seq alignments. TPMCalculator reported the expression matrix including transcripts per million (TPM) values and raw read counts for each gene across all samples. We removed genes with mean expression over all tissues ≤ 1 TPM, as well as those tagged as noncoding RNA (ncRNA).

11.4. Tissue Specificity Index, Differential Expression Analysis

Spatial transcriptomics (e.g., multi-tissue) aims to comprehensively characterize the gene expression landscape in an organism to explore these genes' function and adaptive evolution ultimately. Tissue-specific genes have significantly enhanced expression levels in a given tissue relative to the baseline expression in all other tissues. Tissue-specific expression profiling is crucial in elucidating the development, the complexity, and evolutionary history of an organism at the systemic level. Furthermore, the classification of genes concerning their expression patterns across organs or tissue types is important for a deeper understanding of the molecular mechanisms of tissue activity and function, to discover key regulatory features, and to shed light on the correlated phenotypic and functional evolution of tissues [141, 142, 143]. To characterize the gene expression atlas and the landscape of tissue-specific expression (TSE) in pikeperch, I made use of standard differential expression (DE) analysis statistics and the index of tissue specificity (τ). To that end, the per sample average (arithmetic mean) expression values (TPM) were used to calculate the index of tissue specificity (τ) for all pikeperch protein-coding genes. Following the approach described in Yanai et al. [144] and Mank et al. [145], I calculated tissue specificity (τ) for protein-coding genes with the formula:

$$\tau = \frac{\sum_{i=1}^{N}(1 - \frac{TPM_{g,i}}{TPM_{g,max}})}{N - 1}$$

where N is the total number of tissues examined, $TPM_{g,i}$ is the expression of a gene g in tissue i, and $TPM_{g,max}$ is the maximal expression level detected for a given gene g over the examined N tissues. Tau (τ) index has been demonstrated to be the most suitable metric to measure gene tissue specificity [146]. It varies between 0 and 1, where highly tissue-specific transcripts have values approaching 1 ($\tau > 0.85$), and broadly expressed transcripts, e.g., from housekeeping genes, have a tissue-specificity index approaching 0 ($\tau < 0.3$) [143, 146, 147]. In addition, I analysed differential expression (DE) between tissue samples in a "one vs. all" design, utilizing the likelihood ratio testing (LRT) under the generalized linear model (GLM) framework in the package edgeR [148].

I then iteratively detected genes whose expression levels change by a significant amount between the two groups—namely X and non-X, where X is one of the ten tissue samples. Instead of considering only statistical significance like in standard DE tests, we applied a combination of fold-change ($\log_2 FC > 3$) and p-value ($p < 0.05$) cut-offs to deem genes as differentially expressed (DEGs) between the respective comparison groups ($X\ vs.\ non - X$). Lastly, I performed tissue enrichment analysis using the *teGeneRetrieval* function in the Bioconductor package TissueEnrich [149]. This package applies the algorithm from the Human Protein Atlas (HPA) [150] on expression data (normalized counts) and classifies genes into different categories based on their expression levels across the tissues. More details about the tissueEnrich algorithm can be found in Jain et. al [149].

11.5. Tissue-specific Co-expression and PPI Networks Analysis

To assess the tissue-specificity of genes in the context of functional modules, I examined whether tissue-specific expression patterns are also reflected in tissue-specific co-expression and network modules. To that end, I used CEMiTool [151], a Bioconductor package to identify differential co-expression modules. Tissue-specific co-expression modules were defined as a subset of tissue-specific genes which show relatively high co-expression in one tissue, while having consistently lower co-expression in all other tissues. Tissue-specific protein–protein interaction (PPI) networks were predicted as follows: we first constructed global (i.e., based on all genes) interactions by exhaustively mapping pikeperch proteins to PPI networks supported by experimental evidences using STRING (version 11). Only proteins with a minimum interaction score of more than 0.7 were kept in the PPI network. Given a pikeperch tissue, a subnetwork in the global network is labeled specific to that tissue if the interacting proteins in that subnetwork are differentially co-expressed and tissue-specific. Hence, a PPI network is specific to a given tissue, if it is induced by tissue-specific proteins, whose coding genes are additionally co-expressed in that tissue.

12. Results

12.1. RNA-Seq, Assembly and Functional Annotation

In this work, transcriptome profiles of ten different pikeperch tissues were first analyzed by integrating state-of-the-art RNA-Seq methods. Messenger RNA-Sequencing yielded between 32.5 and 49 million PE-reads per library, with an average of nearly 38 million reads per library (Table 12.1). About 92% of the raw reads were ultimately retained after QC. Then, different sets of transcriptome assemblies, including Trinity, rnaSPAdes, and Stringtie2 were built. A summary of the assembly statistics and their characteristics including functional annotations are reported in Table 12.2 and Figure 12.1. The number of transcripts greatly varies among assemblies.

De novo assembly with Trinity and rnaSPAdes substantially yielded a more significant number of contigs, with 438,462 and 295,387 contigs, respectively. The reference-guided assembly with StringTie2 yielded 79,936 contigs in total. As expected, merging all assemblies with EvidentialGene substantially reduced the contig count to a number of 56,302 contigs, which is entirely consistent with the total number of proteins (N = 56,557) annotated in the pikeperch reference genome (Chapter 8), indicating that our multi-issue-assembly nearly spans the whole pikeperch proteome. Overall, the meta-transcriptome outperformed the separate assemblies (Trinity, rnaSPAdes and Hisat2+StringTie2) concerning BUSCO completeness and protein functional database records.

Interestingly, mapping this transcriptome assembly to the pikeperch reference genome showed that all reference loci were recovered (100%), and approx. 94% of reference introns were accurately captured, while only 4% of reference exons were missed by this transcriptome assembly, validating its high accuracy. Regarding assembly metrics and transcript coverage, Hisat2+StringTie2 yielded the best results (Table 12.2, Figure 12.1). In particular, nearly 89% of transcripts were assembled in full length (Table 12.2). This result is in line with previous studies [152], where reference-guided assemblies tended to produce longer and more full-length transcripts compared to reference-free approaches.

Table 12.1.: Summary statistics of paired-end RNA-Seq reads yielded from 18 libraries of ten different pikeperch tissues using Illumina NovaSeq 6000 System.

Libraries	**No. of Raw Reads**	**Q30 Raw Reads (%)**	**No. of Clean Reads (%)**	**Q30 Clean Reads (%)**
Heart-1	33,908,652	93.98	30,984,679	97.42
Heart-2	35,031,697	94.16	33,180,437	97.49
Head kidney-1	33,002,587	93.95	30,015,644	97.51
Head kidney-2	39,938,681	94.16	36,623,609	97.50
Muscle-1	35,047,416	94.74	32,410,797	97.52
Muscle-2	39,426,896	94.32	36,506,458	97.53
Liver-1	32,566,471	94.20	30,209,964	97.42
Liver-2	35,071,990	94.11	32,007,540	97.65
Brain-1	40,422,234	93.58	37,106,005	97.40
Brain-2	35,567,608	93.46	36,458,085	97.41
Skin-1	37,989,173	94.21	35,172,646	97.42
Skin-2	40,633,032	94.34	37,687,356	97.47
Gills-1	38,586,131	93.97	35,630,505	97.42
Gills-2	39,427,046	94.51	36,458,085	97.41
Spleen-1	45,127,155	94.00	41,790,579	97.36
Spleen-2	33,329,198	94.27	30,504,153	97.44
Ovary	37,553,020	94.38	34,742,133	97.51
Testis	48,694,199	93.90	44,903,971	97.37
Average	**37,851,288**	94.12	**35,132,924**	97.45
Total	**681,323,186**	—	**632,392,646**	—

12.2. Gene Expression Atlas, Tissue-specific Expression Patterns

Spatial transcriptomics (e.g., multitissue) aims to characterize the landscape of gene expression in an organism and to explore the function and adaptive expression of these genes. Tissue-specific genes are those with significantly enhanced expression levels in a given tissue, relative to the baseline expression in all other tissues. Tissue-specific expression profiling is crucial in elucidating the development, the complexity, and evolutionary history of an organism at the systemic level. Furthermore, tissue-specific expression patterns across organs or tissue types are important for a deeper understanding of the molecular mechanisms of tissue activity and function, to discover key regulatory features, and to shed light on the correlated phenotypic and functional evolution of tissues [141, 142, 143].

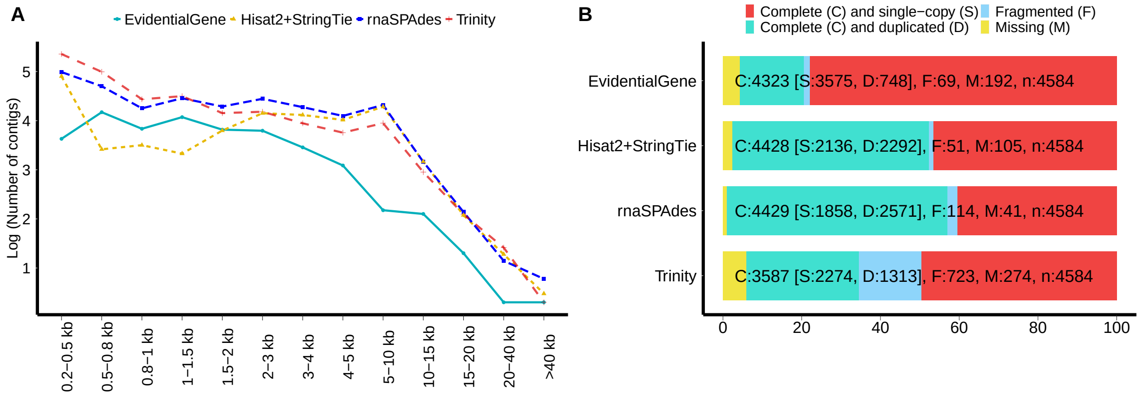

Figure 12.1.: Contigs length distribution and BUSCO score of the different assemblies. Contigs length (scaled to log10) of the different transcriptome assemblies (**A**). BUSCO completeness for each assembly, showing the proportion (%) of complete (C) and single-copy orthologs (S), complete and duplicated (D) orthologs, missing (M) and fragmented (F) orthologs. Transcripts were queried against the *Actinopterygii* gene set (N = 4584) (**B**).

Table 12.2.: Summary of pikeperch transcriptome assembly and assessment. Bold **values** are comparatively, the best values in term of quality metrics.

	Trinity	**rnaSPAdes**	**Hisat2 + StringTie2**	**EvidentialGene**
Number of contigs	438,462	295,387	79,936	**56,302**
Cumulative contigs length (Mb)	399.28	502.46	299.28	**85.73**
Mean contigs length (bp)	910.65	1701.05	**3744.49**	1522.81
N50 contigs length (bp)	1340	3436	**4934**	1977
Largest contig (bp)	70,079	**80,089**	78,909	79,815
σ contigs >1 Kb (%)	57.11	83.16	**98.31**	80.51
% of FL transcripts	60.57	72.84	**89.52**	86.73
% of transcripts with ORFs	76.73	80.53	**88.84**	85.07
% of BUSCO complete	80.27	96.58	96.62	**96.87**
% of transcripts with NCBI NR hits	72.83	78.04	86.27	**88.35**
% of transcripts with Swiss-Prot hits	55.76	60.23	75.86	**78.57**
Mapping rate RNA-Seq reads (%)	83.92	84.75	**90.86**	88.15

This analysis showed that most of the expressed protein-coding genes in pikeperch were detected in the testis (N = 22,097), brain (N = 19,481), and gills (N = 17,417), While muscle expressed the least genes (N = 10,529). A mean number of 15,820 genes were detected per tissue. Since cDNA libraries were constructed with equal amounts of cDNA from each tissue, the differences in the number of detected protein-coding genes suggest genuine biological variations. Genes with detected expression signals (N = 19,542) were binned into four categories, based on their expression levels and tissue-specificity index (Table 12.3).

Table 12.3.: Classification of protein-coding genes based on transcript expression levels and index of tissue specificity (τ) across 10 pikeperch tissues.

Category	No. of Genes	Fraction of Detected Genes (%)
Tissue-Specific	2930	15
Group-Enriched	3809	19.5
Expressed-in-All	5810	29.8
Mixed	6970	35.7
Total detected	19,541	100

Mixed-Expressed Genes

The largest group of genes (35.7%) encompassed 6970 genes in the category termed "*Mixed*", which includes detected genes that could not be assigned to any of "Tissue-Specific", "Group-Enriched" or "Expressed-In-All" categories (Figure 12.2). This class features the lowest expression variance ($\sigma = 102.3$) and the least average expression ($\overline{x} = 20.2$), suggesting a lower within-group expression variability. Moreover, the τ index in this category is more dispersed ($\sigma = 0.13$) compared to the three other categories ($\sigma < 0.06$). This is coherent with my definition of *Mixed-Expressed* genes, which are highly enriched in a subset of tissues while being broadly expressed at moderate or lower levels in the others. Thus, this explains the stretched distribution of the tissue specificity index (Figure 12.2A). GO enrichment analysis revealed that 3,586 genes (51.44% of "*Mixed*") contained significant ($FDR < 0.05$) enrichment for 78 GO-terms (20 GO:MF, 44 GO:BP, 14 GO:CC), five KEGG and two Reactome (REAC) pathways. However, most of these genes ($> 80\%$) were associated with only two significant GO terms, namely GO:0005515 (MF:Protein binding, FDR $< 10^{-9}$) and GO:0003824 (MF:Catalytic activity, FDR $< 10^{-4}$). Transcription factor genes (TF) of which 2733 have been identified in fish species and reported in the Animal Transcription Factor Database (AnimalTFDB3.0) [153], were mostly (12% of all TF genes in fish) found in this category.

Expressed-In-All Genes

The second largest class (29.8%) consists of 5810 genes ubiquitously expressed in all tissues, termed "*Expressed-In-All*". These gene products are needed in all cells and tissues for the maintenance of essential cellular functions. Functional enrichment of these genes include primarily, ribosomal and spliceosomal proteins involved in protein biosynthesis and metabolism, RNA processing and transport, as well as proteins responsible for the structural integrity and stability of the cells (Figure 12.3). The average expression levels ($\overline{x} = 65.8$) in this group is significantly higher (p-value $< 10^{-16}$) than in "*Mixed-Expressed*" and "*Group-Enriched* genes", but still lower than in "*Tissue-Specific*" genes

($\overline{x} = 102.9$), suggesting that these genes are relatively upregulated in all analysed pikeperch tissues (Figure 12.2B). The top 5 most abundant "*Expressed-In-All*" genes include known housekeeping genes such as *EEF1A1* (Elongation factor 1-alpha 1), *ACTB2* (Beta-actin), *RPS2* (40S ribosomal protein S2), *RPL7A* (60S ribosomal protein L7a) and *RPL4* (60S ribosomal protein L4).

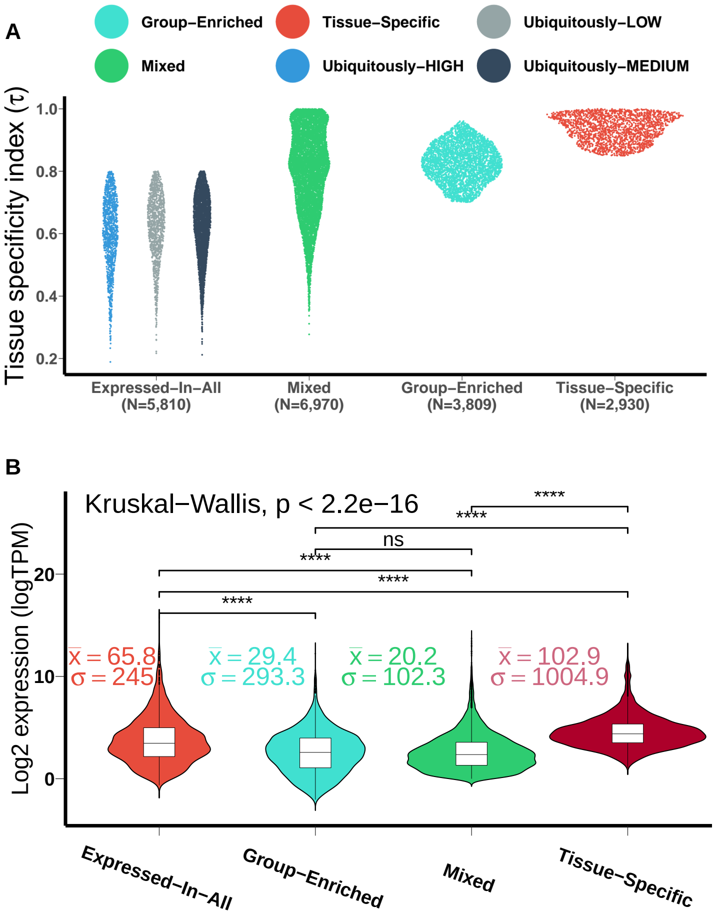

Figure 12.2.: Classification of protein-coding genes detected in pikeperch. (**A**), Sinaplot showing the distribution of the tissue specificity index τ in each genes class. (**B**), Violin plot showing the distribution of genes expression levels within each class. Statistical significance *ns*: Not significant; ****: Extremely significant ($p < 0.0001$).

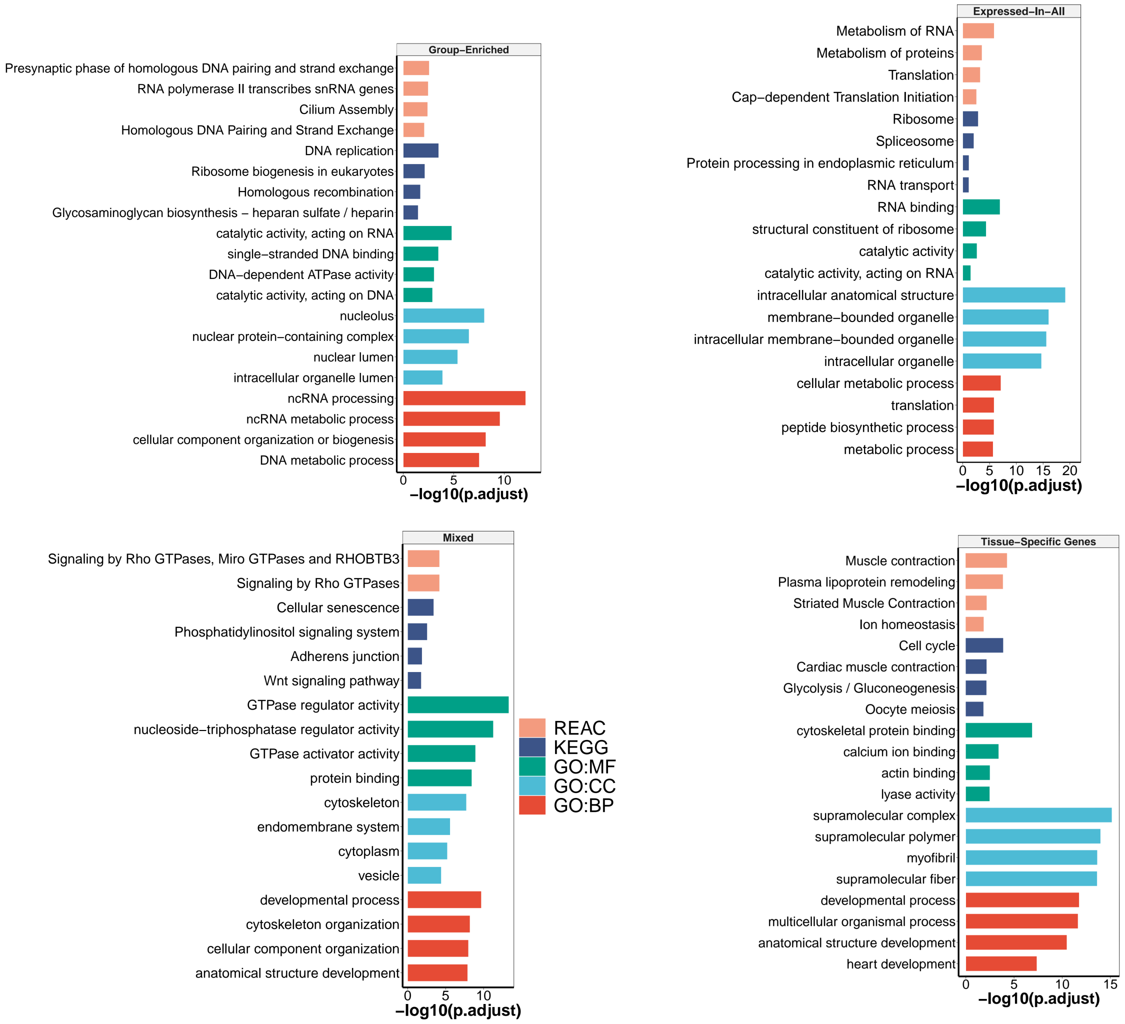

Figure 12.3.: Genes functional annotation. Enrichment of Gene Ontology (GO) terms, KEGG and Reactome (REAC) pathways for each genes class. The top 4 significant (FDR ¡ 0.05) GO terms/functional pathways are depicted here, including REAC (Reactome Pathways), KEGG (KEGG Pathways), GO:BP (Biological Process), GO:CC (Cellular Component), and GO:MF (Molecular Function).

Group-Enriched Genes

The third category contains 3809 genes (19.5%), termed "*Group-Enriched*". Group-Enriched genes are non-housekeeping genes with enhanced expression in a limited number of 2–7 tissues and with an index of tissue specificity $\tau > 0.5$. They are often involved in coordinated biological processes in different tissues/organs, and thus highly enriched in those tissues. I obtained 36 sets of *Group-Enriched* genes, comprising two (12 sets) to five (one set) different tissues. The groups {brain; testis} (N = 1760) and {ovary; testis} (N = 1458) are the pairs sharing most of the Group-Enriched genes (Figure 12.4). GO overrepresentation analysis indicated that genes enriched in the group {brain; testis} are predominantly involved in plasma membrane bounded cell projection organization (GO:0120036; FDR $< 10^{-16}$), and small conductance calcium-activated potassium channel activity (GO:0016286; FDR $< 10^{-3}$). Genes in the group {ovary; testis} (gonads) are primarily involved in cellular nitrogen compound metabolism (GO:0034641; FDR = 0) and in nucleotide and nucleic acid metabolic process (GO:0006139; FDR = 0). Genes enriched in the triplet {brain; testis; ovary} did not have any overrepresented GO terms. Global functional analyses indicated that ncRNA processing (GO:0034660) was the most significant biological process of all Group-Enriched genes (Figure 12.3).

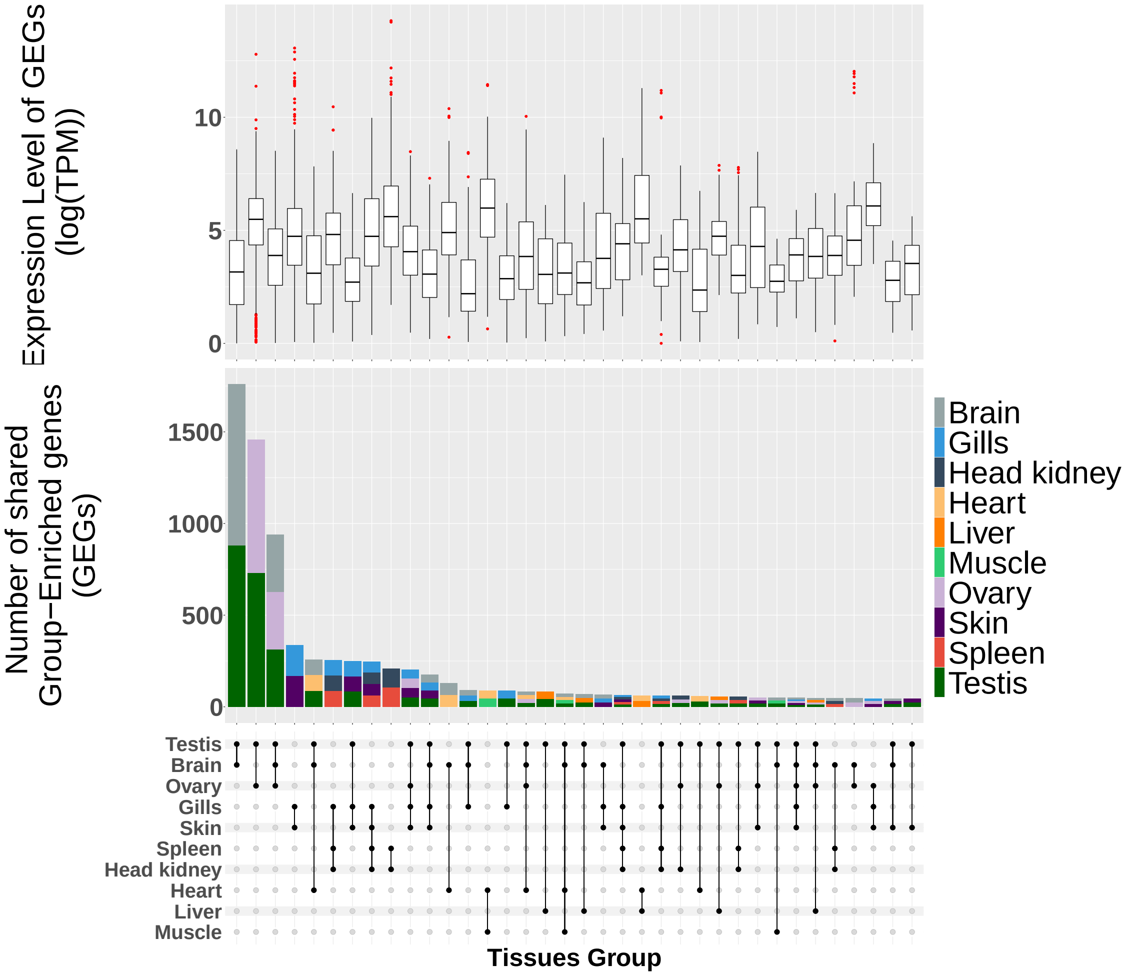

Figure 12.4.: Shared genes between tissues along with their expressions levels. Upset plot depicting different gene sets and the number of shared Group-Enriched genes in each set. The box-and-whisker plot summarizes the expression levels of all genes contained in each intersection set. Red dots beyond whiskers represent genes displaying outlier expression within the group.

Tissue-Specific Genes

The last category termed "*Tissue-Specific*" (N = 2930) constitutes about 15% of all detected protein-coding genes in pikeperch (Table 12.3). These are genes with an index of tissue-specificity $\tau > 0.85$ and at least five-fold higher expression in one tissue compared to all other tissues. Ovary (N = 563) and testis (N = 379) had the largest numbers of tissue-specific genes detected in our analysis, while the head kidney (N = 109) had the least (Figure 12.5D). GO enrichment analyses indicated that the most significant biological process in the ovary is cell cycle process (GO:0022402), reproductive process (GO:0022414) in the testis, homeostasis (GO:0007599) in liver, nervous system development (GO:0007399) in brain, developmental process (GO:0032502) in gills, humoral immune response (GO:0006959) in spleen, muscle structure development (GO:0061061) in muscle, circulatory system development (GO:0072359) in heart, and hematopoietic stem cell migration (GO:0035701) in head kidney. Tissue-specific genes in the skin did not show any significantly overrepresented terms. Moreover, KEGG pathway analysis was performed to identify which pathways were significantly enriched with tissue-specific genes. A total of nine significantly enriched pathways were identified, whereas cell cycle, biosynthesis of antibiotics, glycolysis/gluconeogenesis and biosynthesis of amino acids showed the strongest KEGG enrichment signal across multiple tissues. For example, seven tissue types were involved in the biosynthesis of amino acids with at least two genes (Supplementary Figure S1). As expected by construction, the logarithmized expression fold-change for a gene in a given tissue compared to all others tissues is positively correlated with the index of tissue specificity (τ), confirming that tissue-specific genes are significantly upregulated only in a particular tissue (Figure 12.5C).

12.3. Co-Expression Modules — Hubs and PPI Networks

To gain insights into the pikeperch interactome with the aim to detect hubs and co-expression networks containing tissue-specific genes, I conducted genome-wide co-expression network analysis. Overall, seven differentially co-expressed modules (M) displaying correlated expression were identified. They contain 55 to 14 genes, and involve 211 genes in total. The largest modules consisted of 55 (M1) and 53 (M2) genes. These were specifically upregulated in liver and muscle tissues, respectively (Table 12.4). By integrating interactome information with co-expression modules, I identified potential hub genes specific to each module (Figure 12.6A). GSEA highlighted which modules were induced or repressed in the different tissues (Figure 12.6B). Finally, I performed overrepresentation analysis (ORA) to determine which biological functions are associated with the identified modules. For instance, the glycolysis/gluconeogenesis pathway is overrepresented in module M4, which is enriched by muscle-specific genes (Supplementary file S1).

To identify network-based protein functional modules that are significantly associated with different tissues, I integrated tissue-specific co-expression modules with genome-wide PPI networks in pikeperch. Overall, four tissue-specific PPI networks (TSN) associated with four tissues were predicted including skin, liver, muscle and heart (Figure 12.7). These TSN

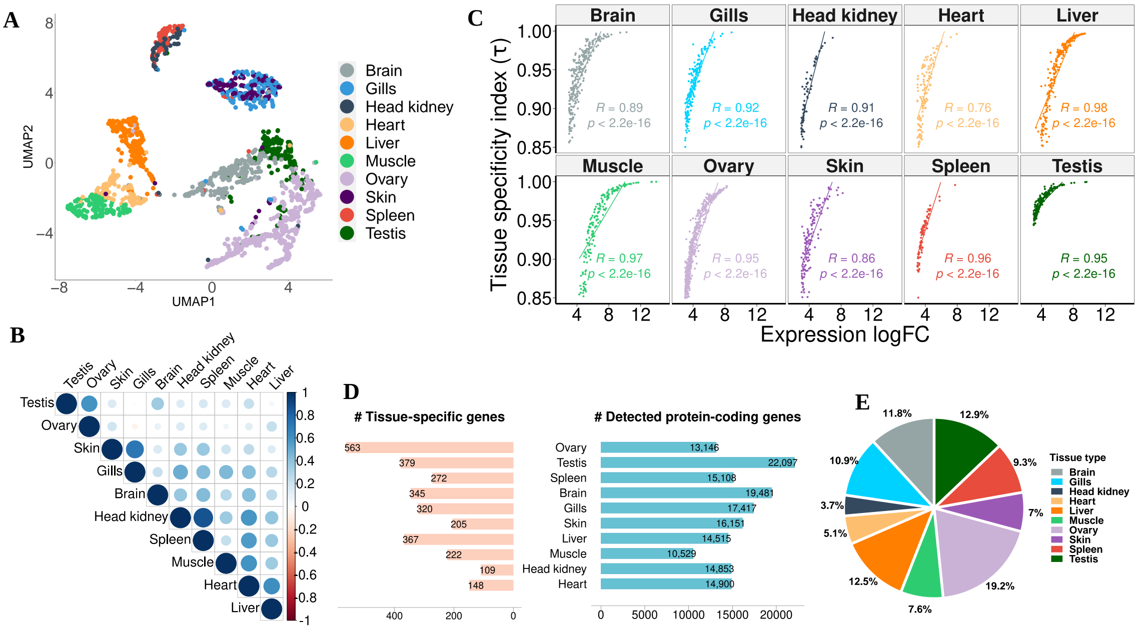

Figure 12.5.: Summary statistics on tissue-specific genes. (**A**), Uniform Manifold Approximation and Projection (UMAP) clustering of 2930 tissue-specific genes in pikeperch based on their expression levels (TPM), where clusters represent genes with similar or correlated expression. (**B**), Correlation heatmap between tissues, based on their specific transcriptome profile. (**C**), Spearman correlation between expression fold-change for each tissue vs. all, and the index of tissue specificity (τ). R is the spearman correlation coefficient and p the corresponding p-value. (**D**), Number of detected and tissue-specific genes in each tissues. (**E**), Percentage distribution of tissue-specific genes across tissues.

Table 12.4.: Tissue-specific (differential) co-expression modules with their hub genes.

Module	No. Genes	Tisssue-Specific Upregulation	Hubs (Gene Symbol)
M1	55	Liver	*C3, AFP4, C1QTNF3*
M2	53	Muscle	*PYGM, TNNT3A, TRIM21, MYLPFA*
M3	27	Ovary, Testis	*SERPINA12, ALOX12B, LOC116046623*
M4	21	Skin	*RPS7, RPS3A, RPL5, RPL13A, RPL7A*
M5	19	Head kidney, Spleen	*HBZ, NPRL3, AQP8A, HBB2*
M6	15	Gills, Skin	*LOC116046623, ZG16B, MPO*
M7	14	Heart	*TNNT2A, MYBPC3, TNNC1A, TNNI1, TPM4A*

involve between tree genes (skin, heart) and 13 genes (muscle). GO analysis identified no significantly enriched biological process in these functional modules. However, these genes mostly describe biological processes specific to these tissues.

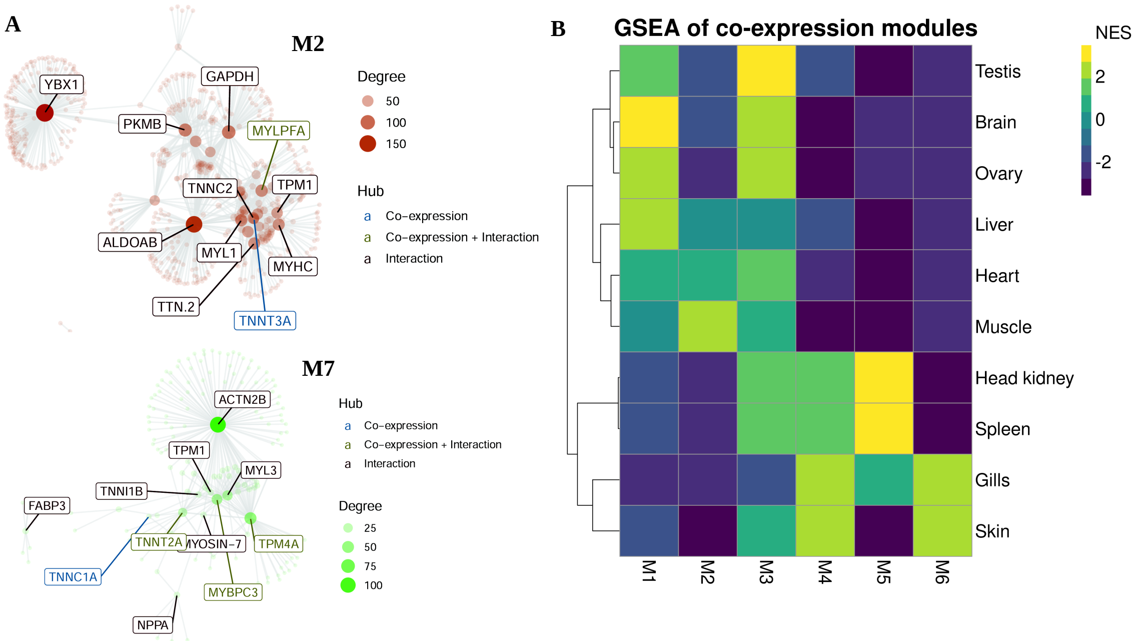

Figure 12.6.: Gene co-expression network analysis. (**A**), Gene co-expression networks of modules M2 (upregulated in muscle tissues) and M7 (upregulated in heart tissues). The top hubs (i.e., genes with highest connectivity) are labelled and coloured based on their source: if only present in the co-expression module predicted by CEMiTool, they are coloured blue; if additionally present in PPI networks, they are coloured green; if exclusively present in PPI network and not in co-expression a network, they are coloured red. The size of the node is proportional to its degree. (**B**), Gene Set Enrichment Analyses (GSA) showing the modules activity on each tissues type. NES is the normalized enrichment score. Exhaustive figures for co-expression modules are available in Supplementary file S2.

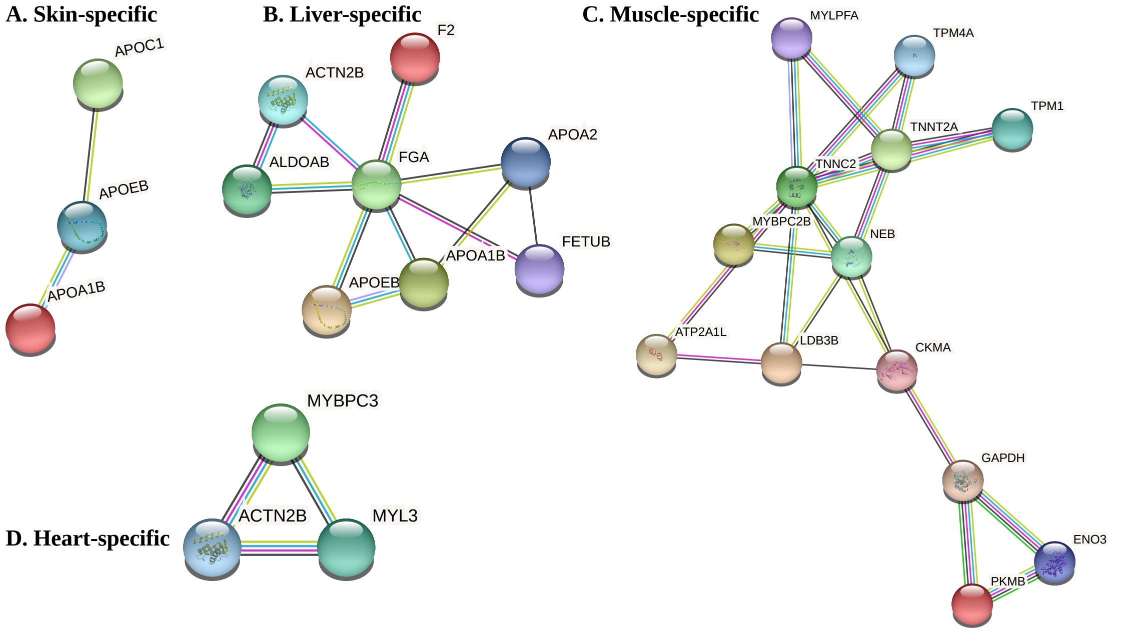

Figure 12.7.: Predicted tissues-specific PPI networks. Tissue specific protein–protein interaction networks predicted in four pikeperch tissues including skin, liver, muscle and heart.

13. Discussion and Conclusion

Discussion

Pikeperch is an emerging inland aquaculture species in Europe. For successful positioning of this species in the European aquaculture industry, genomics insights can be harnessed in all stages of its domestication to understand its adaption biology, optimize breeding programs and improve commercial traits [32]. Hence, the comprehensive transcriptomics data presented here provide a key molecular resource for in-depth informing on developmental, evolutionary, and behavioural questions throughout the domestication process of the pikeperch.

The quality assessment of this new pikeperch transcriptome using BUSCO and various metrics suggest that a wide range of full-length transcripts were resolved since nearly 95% of single-copy orthologs in ray-finned fish (*Actinopterygii*) were covered by this transcriptome assembly. Moreover, the merged assembly (EvidentialGenes) displayed the best contiguity and mappability metrics compared to the other assemblies generated with Trinity or rnaSPAdes assemblers (Figure 12.1). It has been demonstrated in similar studies that combining transcriptome assemblies from multiple assemblers or assembly approaches yields significantly better and optimized results compared to assemblies built with a single assembler [134, 154, 155].

Tissue-specific gene expression is a well-known biological phenomenon by which the genome expresses differentiated transcriptomes among tissues and cell types. Therefore, tissue-specific protein-coding transcripts can explain the difference in the composition and complexity of the transcriptomes of different tissues, as well as provide clues to detecting key pathways and physiological and regulatory processes unique in a tissue [143]. My analysis using the tissue specificity index (τ) with RNA-Seq expression profiles of tissues from 10 vital pikeperch organs allowed us to establish the first catalog of tissue-specific genes and capture their specific metabolic process. In this dataset, testis and brain tissues had the most complex transcriptomes. In contrast, ovary and testis tissues featured the highest number of tissue-specific protein-coding genes, accounting for 19% and 13% of all tissue-specific genes, respectively (Figure 12.5E). This trend is comparable with previous studies across different taxa, including domesticated animals. Different studies in pigs [156], salmons [157], and crucian carp [158], or well studied models such as rats [159] and mice [160]—birds [143] and even on higher-order mammals such as humans [150], have consistently shown that brain and gonad tissues express the most tissue-specific transcripts. These suggest a conserved tissue-specific expression pattern across main vertebrate taxa and lineages.

Among tissue-specific genes, I identified 151 transcription factors (TFs) validated in diverse fish species, 15 immune-related genes (IRG), and three hypoxia-related genes (HRG), suggesting that these important classes of genes are less likely to be uniquely expressed in a specific tissue. In contrast, these genes were similarly expressed in all tissues (Expressed-In-All), or moderately expressed in a subset of tissues (Mixed). More than one-third (35%) of detected TFs, IRG, and HRG, were either classified as "Expressed-In-All" or "Mixed", while only 9% were "Tissue-Specific". However, this is an expected observation, in that transcription factors, for example, are more likely to be ubiquitously expressed, as they are regulatory proteins acting as housekeeping genes in different tissue types. In addition, TFs identified as tissue-specific in our data trigger the expression of genes involved in highly specialized organ-limited functions. For instance, the *GATA* transcription factors family including *GATA4*, *GAT5*, and *GATA6*, which are known to play a key role in cardiac development and cardiomyocyte gene expression, was specifically expressed in the heart tissues of pikeperch. Similarly, SOX32 (SRY-box transcription factor 32) and HSF5 (Heat Shock Factor 5), which are known TFs playing an essential role in spermatogenesis in Zebrafish [161] and other fish species [162], were testis-specific in pikeperch. Lastly, we want to highlight three hypoxia-related genes including the hypoxia-inducible factor prolyl hydroxylase 2 (EGLN1), ceruloplasmin (CP), and solute carrier family 2 (SLC2A2), which have been identified as tissue-specific (in heart and liver, respectively). These transcription factors are known to regulate the expression of hypoxia-responsive genes [163, 164, 165].

Conclusion

In this experiment, I first reported a multi-tissue reference transcriptome from 10 pikeperch vital tissues along with a comprehensive landscape of tissue-specific expression and co-expression networks for classifying protein-coding genes regarding to their unique expression pattern across tissues. Next, I characterized the specific tissue function by identifying functional pathways and biological processes associated with tissue-specific genes and network modules. These transcriptomics resources will be useful for functional genomics analyses, including validating genetic markers and understanding the roles of specific metabolic cycles in different tissues. This knowledge will then lay a framework for investigating pikeperch's important production and domestication traits. Ultimately, the transcriptome dataset will complement this aquaculture species' information in public data repositories.

Part IV.

Genetic Diversity—Landscape of Genetic Variations in Domesticated Pikeperch Broodstock

The analyses and the results presented in the present section of the thesis have already been published or are being considered for submission in a scientific journal. Hence, **Part IV** is **partly** based on the following manuscripts :

- De Los Ríos-Pérez, L.; **Nguinkal, J.A.**; Verleih, M.; Rebl, A.; Brunner, R.M.; Klosa, J.; Schäfer, N.; Stüeken, M.; Goldammer, T.; Wittenburg, D. **An ultra-high density SNP-based linkage map for enhancing the pikeperch (*Sanderlucioperca*) genome assembly to chromosome-scale**. Sci Rep **2020**, 10, 22335.

  Contribution: *I performed the bioinformtics analyses, data visualization and validation. I also contributed in writing parts (sections) of the manuscript.*

- **Nguinkal, J.A.**; Adzigbii, L.; Brunner, R.M.; Goldammer, T. **Identification of High-Confidence Structural and STRs Variants in a Domesticated Pikeperch Broodstock Population** (Unpublished)

  Contribution: *Manuscript in preparation for first submission into a journal.*

14. Introduction

One of the grand challenges in aquaculture genomics is the lack of efficient genomic tools to inform genomic biodiversity and adaptation of farmed animals. Genomic variations are the basis of polymorphisms between the genomes of individuals within a population. Several types of genetic variants can be found at the genome level: deletion of insertions (indels) of one or more nucleotides (i); single base substitution at various position in the genome (ii); structural genomic rearrangement of at least 50 bp encompassing deletions, duplications, insertion , inversion and translations (ii); and microsatellites often referred to as STRs (iv). These variations do not only influence the expression of genes but also trigger advantageous or deleterious phenotypic changes.

This section (Part IV) of the thesis aims to provide catalogues of genetic variants in a domesticated pikeperch population including seven families and 375 progeny as basis resources and instrument for understanding how genetic variation influences the adaptation, health and well-being of animals in the domestication continuum. I restricted the variants detection here to SNPs (i.e. non-somatic single nucleotide substitution), STRs, and SVs, which are by far, the most common and widespread types of genomic variations with relevance in aquaculture genomics [166]. The genomic and transcriptomic resources developed in **Parts** II and III will serve here as backbone to perform the variants identification and analysis using different bioinformatic software tools.

15. Materials and Methods

The modular genotypoing pipeline includes reads mapping, variant calling, post-genoptyping, data filtering and optimization as well as annotation and downstream analyses.

Broodstock Population, DNA Sequencing

Adults pikeperch individuals (N=18, 11 males and 7 females) were used to generate seven matings representing seven families. A total of 375 progeny were obtained from the seven matings. Genomic DNA from the 18 broodstock and 375 progeny was isolated using DNeasy Blood and Tissue Kit (Qiagen) and following the manufacturer's protocol. Whole genome PE150 (150 bp paired-end) sequencing was performed for each individual on Illumina NovaSeq 6000 platform with an average coverage of 30×.

Reads Preprocessing, Reference Mapping

Raw PE150 genomic reads were preprocessed to remove adapters and overepresented sequences using fastp. Clean genomic paired reads were mapped to the reference genome built in **Part** II with BWA-MEM, and by including read group informations that are important metadata information in the processing of mapping and genotyping data.

15.1. Population-wide SNPs Genotyping

Note: *This specific experiment including the development of the SNPs genotyping workflow and and downstream analysis was mainly performed by my colleague* ***Lidia Perez****. I have specifically contributed in the bioinformatics analysis and data visualization and interpretation. (see Perez et.al., 2020 for detailed background and detailed methodology)* [167].

Briefly, SNPs variants were genotyped following the the Genome Analysis Toolkit v4.0 (GATK) pipeline [168]. High-confidence variants were obtained using the following filtering settings: QualByDepth (QD) < 10, Quality (QUAL) < 30, StrandOddsRatio (SOR) >3, FisherStrand (FS)> 60, RMSMappingQuality (MQ) < 40, MappingQualityRankSumTest (MQRankSum) < -12.5 and ReadPosRankSumTest (ReadPosRankSum)< -8.

15.2. Population-wide STRs Genotyping

STRs are highly repetitive genomic sequences of tandemly repeated DNA motifs of usually 2-7 bp, e.g., $(AT)_n$, $(AGACT)_n$, or $(TGCCA)_n$, where $n \geq 2$. They are prevalent in all fish genomes, and of particular interest in aquaculture because of their high mutation rate and genome coverage. To call STRs genotypes in the broodstock of 394 pikeperch individuals, I used HipSTR (version 0.5) [169], a genotyping tool that takes aligned reads (indexed bam files) and a reference set of STRs panel as input and computes the maximum likelihood diploid genotypes for each STR allele (length variation). To obtain phased STRs genotypes, I integrated the phased SNPs data obtained in Section 15.1. A phased STR genotype call is determined by a STR-containing reads that overlap a sample's heterozygous SNP. High-confident genotype calls were filtered using the filter_vcf.py script in the HipSTR package as following: min-call-qual≥ 0.9, max-call-flank-indel ≤ 0.15, max-call-stutter ≤ 0.15, min-call-allele-bias ≥ -2, min-call-strand-bias ≥ -2. Furthermore, STRs with call rate $< 80\%$ with expected heterozygosity $< 99\%$ were discarded to retain only highly polymorphic STRs genotypes.

15.3. Population-scale Mapping of Structural Variants

SVs are sequence variations greater than 50 pb [170]. Deletions (DEL), duplications (DUP), inversion (INV), and translocations (TRANS) are the most common types of SVs. They are beside SNPs and STRs, another important source of genetic variations that can impact gene expression and numerous phenotypes in farmed animals [171, 172]. To detect SVs genotypes, I used the smoove wrapper (https://github.com/brentp/smoove) to deploy the genotyping pipeline which includes existing tools such as lumpy, svtyper, svtools, mosdepth, bcftools and duphold. To reduce false positive calls, read pairs overlapping gaps, low complexity and high coverage ($> 200\times$) genomic regions were also discarded from the SV detection. Population-level SVs detection includes the following steps: SVs were initially identified for each sample in parallel, and then SVtools (version 0.5.1) [173] was utilized to combine all SVs calls across samples. SVtyper (version 0.7.1) [174] and Duphold (version 0.2.1) [175] were used to genoptype and annotate SVs with read coverage, respectively. High-confidence SVs calls were obtained after applying the following filtering criteria: SVs with size less than 50 bp were excluded using BCFtools v1.9 (i); SVs overlapping with the extreme high-read coverage ($\geq 200\times$) regions or gaps of the reference genome were also discarded using bedtools (ii); based on duphold annotation, deletions with DHFFC (fold change for the variant depth relative to flanking regions) > 0.7, duplications with DHFFC < 1.3 and inversions with DHFFC > 1.3 or DHFFC < 0.7 in at least one sample were excluded using BCFtools (iii); SVs variants with allele frequencies (AF) $\leq 1\%$ or $\geq 99\%$ were ultimately filtered out (iv).

16. Results

16.1. High-density SNPs Landscape of the Pikeperch Genome

A total of 1,387,169 high confident and informative SNPs variants were obtained after filtering, with an average of 1107 SNPs per million base pairs (i.e., one SNP per kb). A chromosome-wide SNPs density analysis (Table 16.1) has revealed that the highest number (6.12%) of SNPs were detected on chromosome 2. However, the highest density per Mb (1348 SNPs per Mb) was obtained on chromosome 24, the shortest (20 Mb) in the pikeperch genome. Based on their alternative allele frequency (AAF), the variants were binned into three classes, including rare (AAF< 5%), low frequency (AAF< 10%), and common (AAF >= 10%) variants. AAF denotes the percentage of reads supporting any base other than the reference found at that variant's locus. That is the relative frequency of an allele in the population not found in the reference genome. Nearly 90% of all informative SNPs were categorized as ”common” (Figure 16.1C). However, SNPs were only the second most abundant type of variants in terms of allele frequency, behind STRs, which were by far the most frequent type of genomic variants in the population (Kruskall-Wallis-Test, $p < 10^{-6}$) (Figure 16.2). Moreover, the substitutions G>A and C>T were the most common SNP genotypes in the analyzed pikeperch cohort, making in total more than 30% of all high-quality SNPs genotypes (Figure 16.1B). Genome-wide annotation found out that only 9% of these SNPs were located on exons (exonic), while most of them were either intronic (36%) or intergenic (55%).

16.2. A Population-level Catalog of STR Variants

A reference panel of 189,649 high quality STRs of *S. lucioperca* was used for population-wide STRs genotyping. In total, I identified 119,401 polymorphic STRs (mean heterozygosity: 0.44) in the analyzed samples after filtering to keep high-confident genotypes (Table 16.1). On average, there are 10,694 STR loci per sample, and 92% of loci are shared by at least 5 samples. Di-nucleotide motifs are by far the most abundant (89.86%) repeat motifs in the genome, while septa-nucleotide (7 bp) repeat motifs occur in less than 1% of all polymorphic STRs (Figure 16.3). I used Beagle (version 4.0) [176] to impute STRs to SNP genotype data generated in section 15.1. This imputation showed that 80% of the STRs variants are associated with phased SNPs with a genome-wide imputation accuracy

Table 16.1.: Number of polymorphic genetic variants including SNPs, STRs and SVs predicted in a broodstock population of 394 individuals along with the corresponding per mega base pairs (Mb) density.

CHR	Size (Mb)	SNPs	Mb Density	STRs	Mb Density	SVs	Mb Density
Chr1	54.43	54,338	998.86	7762	142.68	1145	21.05
Chr2	49.40	60,738	1229.51	6609	133.79	995	20.14
Chr3	46.65	57,479	1232.13	6292	134.88	1047	22.44
Chr4	45.68	55,388	1212.52	6817	149.23	1022	22.37
Chr5	44.88	50,563	1126.63	5440	121.21	743	16.56
Chr6	43.62	50,561	1159.66	5588	128.17	1022	23.44
Chr7	43.40	52,629	1212.65	5635	129.84	758	17.47
Chr8	42.48	52,108	1226.65	5395	127	853	20.08
Chr9	42.11	50,925	1209.33	5314	126.19	961	22.82
Chr10	41.05	41,011	1000.27	5691	138.8	854	20.83
Chr11	40.54	42,901	1058.24	5594	137.99	713	17.59
Chr12	39.46	33,634	852.36	5593	141.74	888	22.5
Chr13	36.96	32,928	890.91	4225	114.31	834	22.56
Chr14	35.08	34,228	977.94	4441	126.89	833	23.8
Chr15	34.13	38,897	1139.67	5038	147.61	642	18.81
Chr16	32.12	27,858	867.31	4283	133.34	788	24.53
Chr17	31.68	39,895	1259.31	3692	116.54	1087	34.31
Chr18	31.56	33,851	1072.59	4370	138.47	824	26.11
Chr19	31.48	28,745	913.12	4131	131.23	799	25.38
Chr20	29.80	31,481	1056.41	4689	157.35	644	21.61
Chr21	29.61	34,917	1179.23	3820	129.01	738	24.92
Chr22	29.17	33,940	1163.52	3697	126.74	663	22.73
Chr23	20.93	25,931	1238.94	2505	119.68	510	24.37
Chr24	20.30	27,367	1348.13	2780	136.95	544	26.8
Total	**896.47**	**992,313**	**1106.91**	**119,401**	**133.19**	**19,907**	**22.21**

of 95%. Since the family structure of this pikeperch population was known beforehand, I used this structure to detect potential *de novo* mutations for 1143 STR loci, where a progeny had an STR allele not observed in either of the parents. Overall, 8586 STR loci overlap coding regions, 70% of which contain either dimeric, trimeric, or tetrameric repeat motifs.

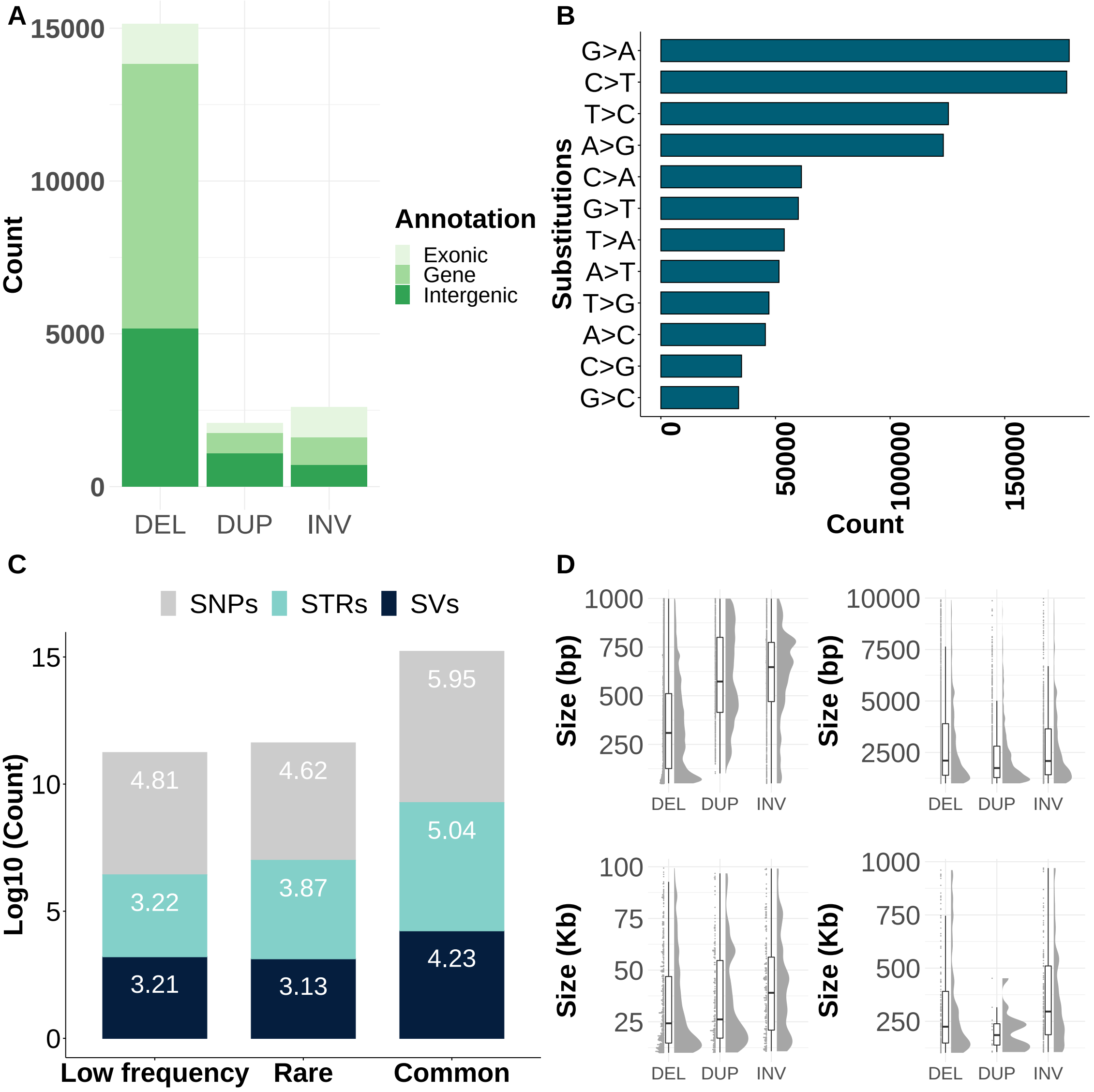

Figure 16.1.: Statistics on genetic variants in pikeperch genome. Annotation of structural variants (SVs) as exonic, intronic or genic (**A**). Most abundant types of base substitutions (SNPs) in the pikeperch genome (**B**). Classification of the three types of variants based on alternative allele frequency (AAF) into three classes including rare (AAF$<$ 5%), low frequency (5% $\leqslant$ AAF$<$ 10%) and common (AAF $\geqslant$ 10%) variants (**C**). Boxplots and distributions showing lengths distribution of SVs in four windows size (50-1000 bp; 1000-10,000 bp; 10-100 kb; and $>$ 100 kb (**D**).

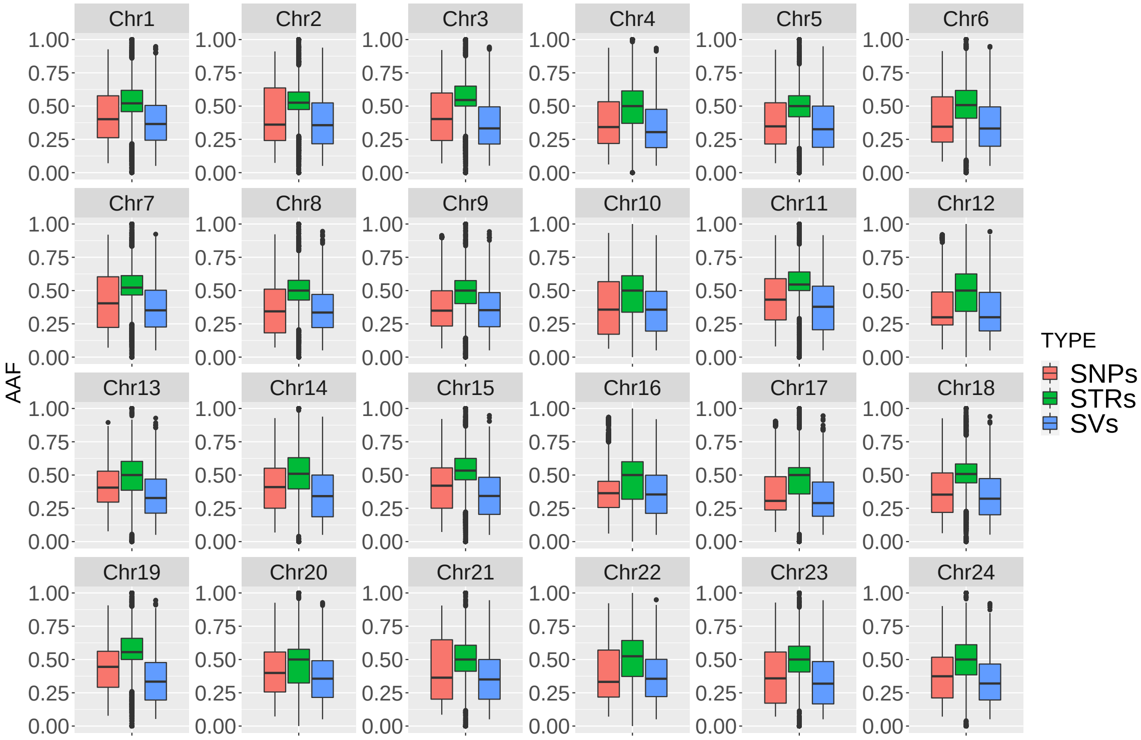

Figure 16.2.: Alternative allele frequency of the most common genomic variants in pikeperch. Boxplots summarizing the distribution of AAF (alternative allele frequency) on each pikeperch chromosome. SNPs, STRs and SVs genotypes were called in a domesticated pikeperch population of 394 individuals. AAF denotes the percentage of reads supporting any base, other than the reference, that is found at that variant's locus. That is, the relative frequency of an allele in the population, that is not found in the reference genome.

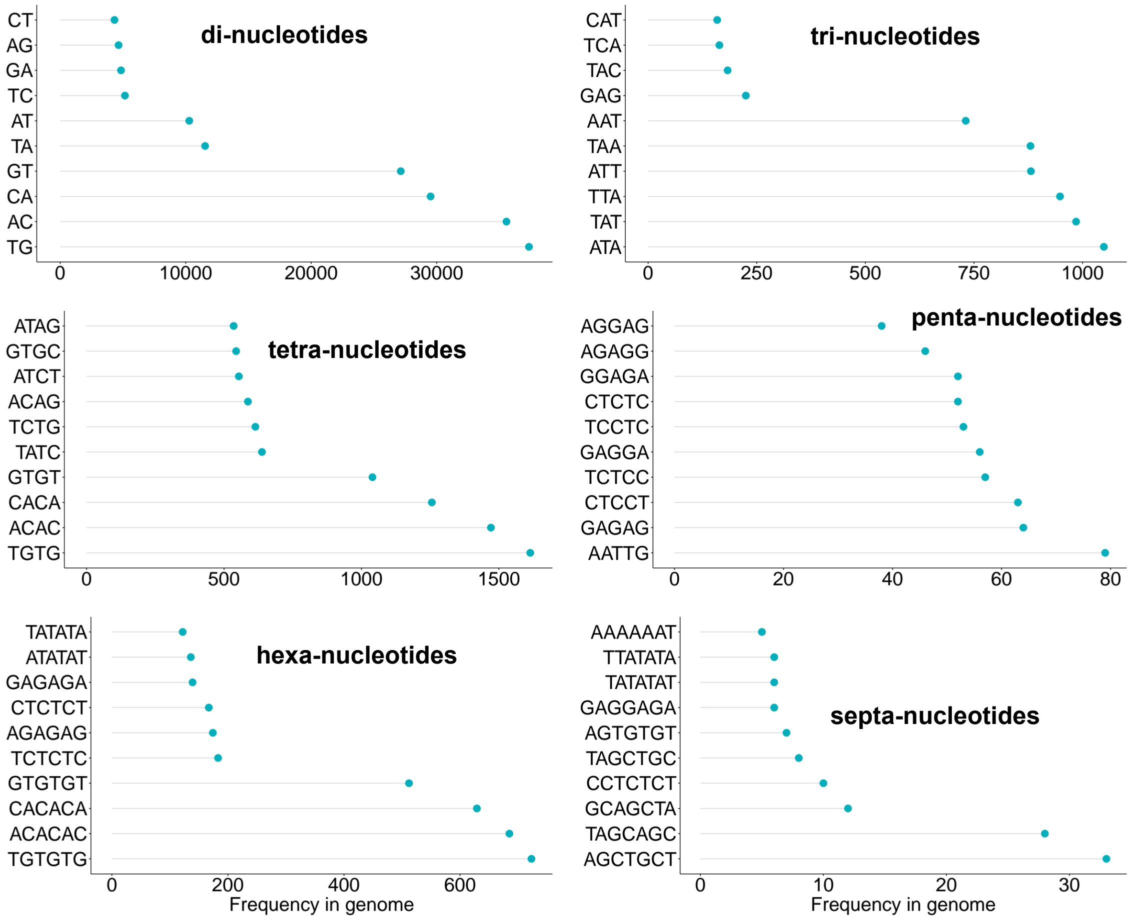

Figure 16.3.: Most abundant STR motifs in pikeperch genome. Summary of the top 10 most abundant di, tri, tetra, penta, hexa and septa STR repeat motifs in the pikeperch genome.

16.3. The Structural Variants Landscape in 394 Broodstock Individuals

After rigorous filtering as described in the methods section, 19,907 high-confidence SV calls including 15,147 deletions (DEL); 2601 inversions (INV); and 2145 duplications (DUP) were obtained across all 394 individuals (Table 16.1). On average, there are 22 SVs genotypes per Mb, affecting 551 million base pairs in total. The number samples (individuals) sharing high-confidence SVs ranged from 4 to 390 with a median of 171 samples associated with an SV event. The SV sizes ranged from 100 bp to 3 million bp (median: 656 pb, mean: 50.8 kb) and 95% of the SVs are shorter than 5 kb. The SV size distributions show that inversions span longer distances than deletions and insertions (Figure 16.1D). As expected and observed in other aquaculture species [171, 172], SVs are less common with a significant lower allele frequency than SNPs and STRs in the analyzed pikeperch sample population (Figure 16.2).

The annotation of high-confidence SV calls was performed using SnpEff [177]. Most deletions span both exons and introns (genic). In contrast, duplications mostly span intergenic regions, and inversions are predominantly located on exonic sequences. Overall, 45% of all SVs overlap at least with one protein-coding gene (Figure 16.1A). Some long deletions even span multiple genes. For example, a 55 kb DEL on chromosome 1 (Chr1) identified in nine samples, affects three genes including *NFKB1*, *SQSTM1*, and *MGAT4B*, which are known to play key roles in immune response in teleosts. (Figure 16.4).

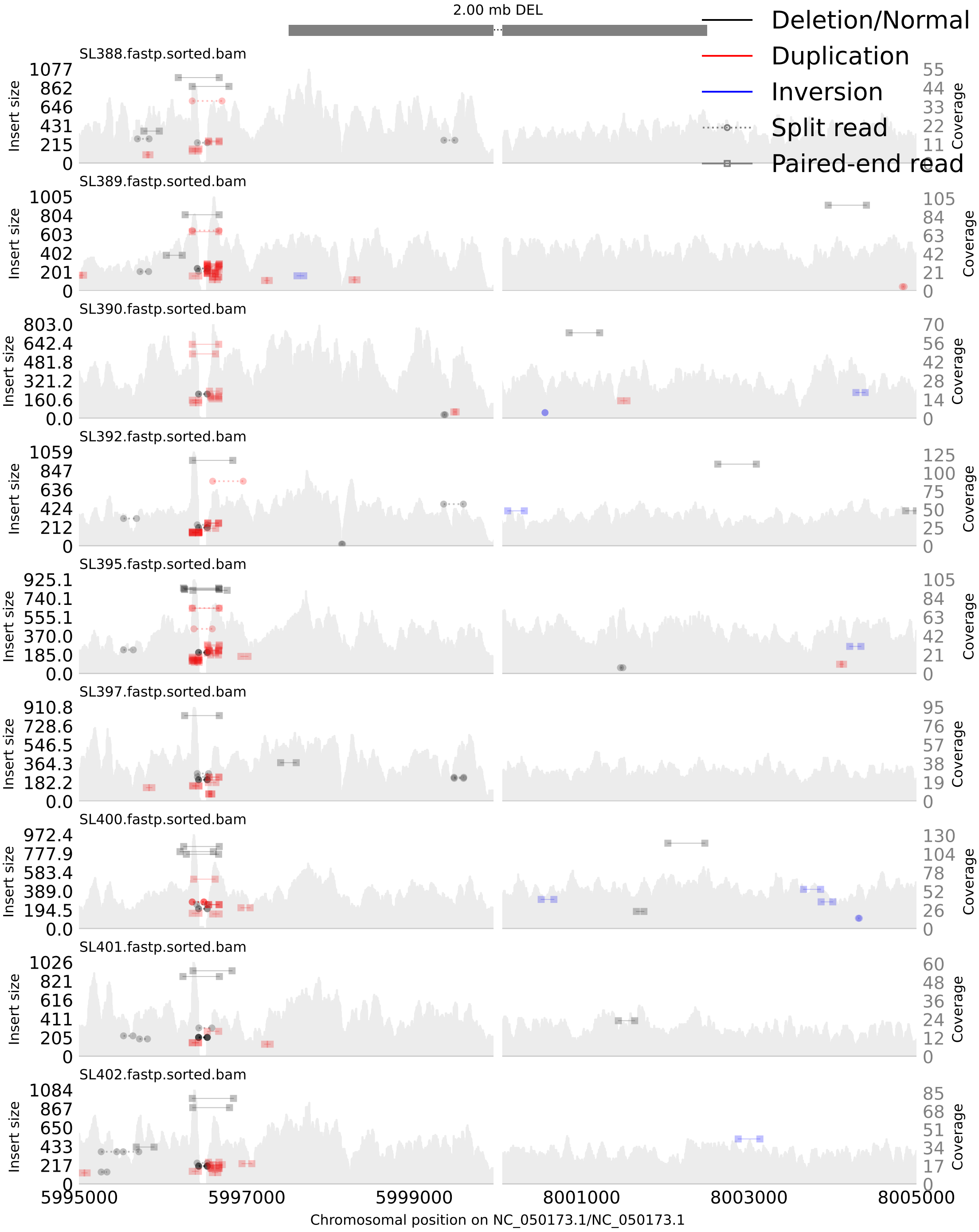

Figure 16.4.: A two Mb genomic region on chromosome 1 (RefSeq: NC_050173.1) displaying annotations of SV genotypes. Region on chromosome 1 (RefSeq: NC_050173.1) displaying SVs annotation including deletions, duplications and inversion predicted in nine samples. There is 55 kb deletion around 7 Mb affecting three genes, including *NFKB1*, *SQSTM1*, and *MGAT4B* on Chr1. The paired-end reads coverage as well as the inserts size are depicted for each sample. Red lines denote duplications, blue lines inversions and blank regions with zero coverage are deletions. The left y-axis display the reads insert sizes for each sample and the right y-axis the co responding reds coverage.

17. Discussion and Conclusion

Discussion

Genomic variants such as SNPs, STRs, or SVs constitute a significant genetic and phenotypic variation source, still largely unexplored in pikeperch. This study aimed to fill that gap by providing the first catalog of these variants and characterizing their landscape in a cohort of domesticated pikeperch. Using short-read NGS data, high-confident SVs, STRs, and SNPs variants were comprehensively identified, quantified, and annotated. This initial study on a cohort of broodstock individuals is of particular interest because domesticated fishes appear to accumulate more variants than their wild relatives, probably due to increased selection pressure in captive environments [171]. Hence, genetic variations offer unprecedented opportunities for advancing aquaculture production.

Future applications of this resource include, for example, gene expression analysis to quantify the impact of different variant classes, genome-wide analyses to confirm or refute their causative effects, and ultimately, their integration into GWAS fine-mapping to identify candidate QTLs [83, 171, 172]. STRs, for example, are efficient tools for genotyping and mapping specific traits or following the flow of genetic material in a population [169]. STRs have been successfully applied in diverse aquaculture species to derive causative variants [178, 179]. The predictive power of SVs and SNPs variants has shown promising results in estimating breeding values of farmed fishes [171, 180]. This clearly demonstrates the great potential of genetic variations to increase genetic gains for specific traits of interest using genomics insights. However, this study is limited to a purely descriptive characterization of genetic variations without further functional investigations of putative causative effects on phenotypic traits. In addition, the analyzed cohort comprising seven broodstock families is not representative and randomized enough for powerful and robust inferences on genetic diversity in pikeperch. Large-scale studies including hundreds of domesticated and wild animals and individuals from different geographical locations and environments would give a broader picture of pikeperch's genetic variation landscape. The results reported here merely serve as groundbreaking findings and initial work and framework to such large-scale analyses.

Conclusion

In summary, this part of my dissertation has provided a descriptive catalog of the most common genetic variants in pikeperch including SNPs, STRs, and SVs, which are a valuable genetic resource and tool for studying complex traits in pikeperch aquaculture. The landscape of SVs, SNPs and STRs reported provides a novel resource for future studies towards the elucidation of the genetic architecture of economical, welfare, and adaptation traits in pikeperch. They will serve as first layer of genomics information to investigate and understand the possible function and effects of every class of variants on the pikeperch genome.

Part V.

Comparative Genomics Analyses of Pikeperch in the *Percidae* Fish Family

The analyses and the results presented in this Part have already been partially published in a peer-reviewed scientific journal. Hence, **Part V** is **partially** based on the following published manuscript:

- **Nguinkal, J.A.**; Verleih, M.; de los Ríos-Pérez, L.; Brunner, R.M.; Sahm, A.; Bej, S.; Rebl, A.; Goldammer, T. **Comprehensive Characterization of Multitissue Expression Landscape, Co-Expression Networks and Positive Selection in Pikeperch.** Cells **2021**, 10, 2289. https://doi.org/10.3390/cells10092289

  Contribution: *I designed the experiments and performed the bioinformtics analyses including data processing, visualization and validation. I additionally wrote the manuscript.*

18. Introduction

Comparative genome analysis is a powerful tool in genomics studies, especially for poorly studied and understood species, which facilitates the transferring of genomic information and insights from well-studied species to newly or less investigated species within a given taxonomic lineage [181]. Comparative genomics allows a better understanding of genome arrangement during evolution and enhances the discovery of orthology among species of interest. For poorly studied aquaculture species like pikeperch, comparative genome analyses should inform the evolutionary conservation of genes and their functions, recent or ongoing selective signatures, and better characterization of its genome organization.

Traditionally, phylogenetic analyses were performed using morphological data and mitochondrial DNA only. However, with significant advances in analytical bioinformatics methods and the democratization of NGS technologies, phylogenomics analyses are now broadly performed at genome-scale. Phylogenomics analysis uses genome-wide data such as protein sequences or molecular markers to find evolutionary ties in a group of genetically related organisms (i), clarify relationships between ancestral sequences and their descendants (ii), and estimate divergence time between groups of organisms sharing a common ancestor (iii).

This section (Part V), will first focus on orthology analysis and inference between percid fishes, then elucidate the recent phylogeny of this family. Finally, positive selection will be investigated to capture fast-evolving genes among percids.

19. Materials and Methods

19.1. Orthologs Analysis and Inference

To investigate the phylogenetic history of *S. lucioperca*, genome-scale orthologous groups (OGs) from six percid species, namely, pikeperch, walleye, yellow perch, European perch, Arkansas darter, and orangethroat darter, were constructed to facilitate phylogenetic analyses. Orthologs are genes descended from a single gene in two species' last common ancestor (LCA). For this analysis, the Asian sea bass, (*Lates calcarifer*) and channel bull blenny (*Cottoperca gobio*), were used as outgroup species.

The longest CDS for each gene locus of these species were aligned in an all vs. all fashion using BLASTP with an e-value threshold of 1e-5. In order to reduce the redundancy caused by alternative splicing variations, only the longest CDS of gene models at each gene locus were retained in the analysis. Coding sequences shorter than 50 amino acids were also excluded. BLASTP alignments were fed to the OrthoFinder pipeline (version 2) [182], which applied the Markov Cluster Algorithm (MCL) to cluster BLASTP alignments into 38,222 orthogroups (families). Subsequently, a phylogenetic tree of all six species, including outgroups, was built based on the predicted 3980 one-to-one single-copy orthologous gene clusters. For each single-copy cluster (i.e., family), and each species, single-copy orthologs were concatenated into a supergene, and multiple sequence alignments (MSA) were generated using mafft (version 7) [183]. The rooted species tree was inferred from the generated MSA using approximately maximum-likelihood methods implemented in FastTree (version 2.1) [184].

To elucidate the evolutionary timescale of pikeperch in the *Percidae* lineage, the divergence time among the different species was estimated using the PATHd8 algorithm [185]. The input of PATHd8 consists of a tree file in Newick format, and an arbitrary number of age constraints of nodes, specified either as fixed age, minimum age, or maximum age. The output file includes a tree in Newick format as well as a list of estimated node ages (in million years), their mean path lengths, and their estimated substitution rates. The relative age was calibrated by using a reference node from fossil data (as recorded in the TimeTree database - http://www.timetree.org/) between sea bass and yellow perch on the one hand, and between channel bull blenny and Arkansas darter on the other hand.

19.2. Positive Selection Analysis in *Percidae*

Spatial transcriptomics (e.g., multitissues) and genome sequencing can explore the adaptive evolution of genes, i.e., genes under positive selection (GUPS), and provide a foundation for lineage-specific evolutionary and fitness features among species. Positive selection analysis can provide clues to the genes and mechanisms that drive adaptation to different environments, including external stimuli temperature and pathogens. Another interesting question is whether the expression of genes evolving under positive selection in pikeperch is correlated with tissue specificity. In addition, we want to clarify whether GUPS act as hubs (key players) in specific cellular and biological processes.

Here, I have investigated putative GUPS in pikeperch and related *Percidae* to examine how recent natural selection might be associated with tissue specificity (Section 12.2), and to interrogate how it might have shaped the phenotypic and physiological diversification in the *Percidae* branch. To that end, I performed a genome-wide analysis of positive selection in six representative species in the *Percidae* family. Briefly, I obtained coding sequences of six percids species, including pikeperch, walleye, yellow perch, European perch, Arkansas darter, and orangethroat darter. Based on these single-copy orthologs, positive selection was scanned using PosiGene pipeline [186], which makes use of CODEML program in the PAML package to conduct branch-site tests of positive selection. The same outgroup species in section 19.1 were used here. Candidate genes under positive selection were those with a false discovery rate (FDR) < 0.1. Finally, I explored the relationship between positive selection, tissue specificity, and gene expression levels in pikeperch.

20. Results

20.1. Orthology Inference — Phylogenetics Analysis of *Percidae*

To generate the phylogenetic relationship of pikeperch and find its relative evolutionary position in the *Percidae* clade, a rooted phylogenetic species tree was inferred based on 1:1 single single-copy orthologs between the analyzed species (see methods). A total of 38,222 orthogroups (gene families) were predicted, of which 3980 (10.41%) are 1:1 single-copy. Moreover, 627 gene families are pikeperch specific (Table 20.1). Among the compared species, pikeperch (SLUC) had the largest number of shared gene families (N=21,776). In contrast, its closest relative walleye (SVIT) had the smallest number (N=18,632) of shared OGs. All other species were involved in around 20,000 OGs (Table 20.1).

Phylogenetic analysis using 1:1 single-copy orthologs between these species suggested that the pikeperch and walleye have probably recently diverged as species from their last common ancestor (LCA), i.e., *Luciopercinae.* The relative divergence time of pikeperch dates around 3.5 million years ago (MYA), which is relatively recent for a species. The global picture shows that Sander spp. are evolutionary closer to *Perca spp.* (yellow perch and European perch) than *Etheosomatidae* (darters), since *Luciopercinae* and *Percinae* build a monophyletic clade with an estimated divergence time from their LCA around 27 MYA (Figure 20.1A). Darters clustered together in one branch, which is also consistent with the species taxonomy of percids.

As expected, the two Sander species included in this analysis shared the maximum number of orthologs when comparing pikeperch to all species pairwise. In particular, pikeperch and walleye (SVIT) shared the most many-to-one orthologs. Many-to-one orthologs are multicopy orthologs in one species with only one copy in the other species. They have 747 one-to-one orthologs in common. Only the European perch (PFLU) shared fewer 1:1 orthologs (N=280) with pikeperch (Figure 20.1B). Walleye (SVIT), however, experienced much fewer (N=1931) duplications than pikeperch (SLUC).

Additionally, species-specific gene duplication events were identified by examining each node of each orthogroup gene tree using the orthofinder pipeline. Species-specific duplication events occur in the branch leading to the corresponding species in the species tree (e.g., duplications on a terminal branch of the species tree). Pikeperch (SLUC) experienced the largest number of gene duplication events, with nearly 32% (N=9362) of its genes involved in at least one duplication event. The European perch (PFLU) had the fewest number with only 467 (2% of all genes) duplication events. With more than 7000 duplications involving 28% of its genes, the Arkansas darter also had a relatively high rate of gene duplication (Table 20.2).

Table 20.1.: Species-specific statistics on orthogroups analysis in the *Percidae* lineage. OGs: Orthogroups, ECRA: *Etheostoma cragini*, ESPE: *Etheostoma spectabile*, PFLA: *Perca pflavescens*, PFLU: *Perca pfluviatilis*, SLUC: *Sander lucioperca*, SVIT: *Sander vitreus*.

	ECRA	ESPE	PFLA	PFLU	SLUC	SVIT
No. of proteins (CDS)	41,178	41,686	38,939	24,326	50,557	34,698
Proteins in OGs	40,273	40,682	37,875	23,193	50,094	28,381
% of proteins in OGs	97.80	97.62	97.26	95.34	99.08	81.80
No. of OGs containing species	20,690	20,076	20,861	20,002	21,776	18,632
% of OGs containing species	85.00	82.50	85.70	82.21	89.53	76.64
No. of species-specific OGs	61	106	172	47	171	875
No. of proteins in species-specific OGs	216	383	514	171	627	3119

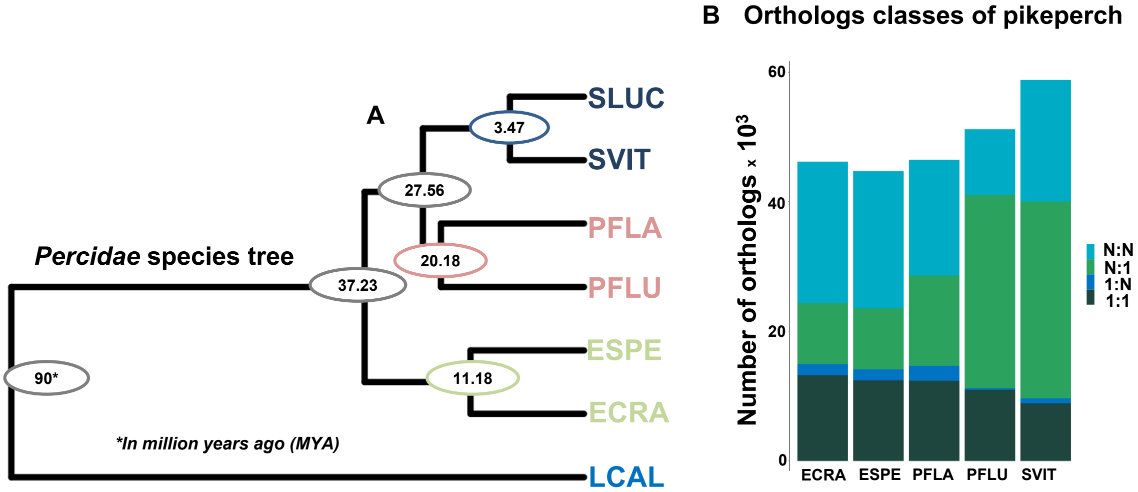

Figure 20.1.: ***Percidae* phylogenetic species tree.** (**A**), species tree of representative percids species based on single copy orthologs. Internal node labels indicate the divergence time in million years from their last common ancestor (LCA). ECRA: *Etheostoma cragini*, ESPE: *Etheostoma spectabile*, PFLA: *Perca pflavescens*, PFLU: *Perca pfluviatilis*, SLUC: *Sander lucioperca*, SVIT: *Sander vitreus*, LCAL: *Lates calcarifer* (as outgroup). (**B**) Number of SLUC ortholgs in different orthologous genes classes compared to other five percid species including N:N (many-to-may), N:1 (many-to-one), 1:N (one-to-many) and 1:1 (one-to-one) orthologs.

Table 20.2.: Species-specific statistics on gene duplication events in the *Percidae* lineage. ECRA: *Etheostoma cragini*, ESPE: *Etheostoma spectabile*, PFLA: *Perca pflavescens*, PFLU: *Perca pfluviatilis*, SLUC: *Sander lucioperca*, SVIT: *Sander vitreus*.

Species	**No. of gene duplication events**	**% of all genes**
ECRA	7061	29.42
ESPE	7193	27.66
PFLA	6875	24.55
PFLU	467	2.12
SLUC	9362	32.37
SVIT	1931	8.10

20.2. Insights into Positive Selection Signatures in Percids

Six representative percid species were analyzed for candidate GUPS (see Methods in 19.2). Overall, 43, 63, 137, 154, 152, and 124 putative candidate genes under selection pressure were detected in *S. lucioperca, S. vitreus*, *P. flavescens, P. fluviatilis*, *E. spectabile*, and *E. cragini*, respectively (Table 20.3). Only two tissue-specific genes, *SLC13A2* and *VWA1* were found to be under positive selection. Although the expression levels of GUPS in *Sander lucioperca* did not significantly vary across tissues (One-way ANOVA, $F < 1$), they were markedly expressed in higher levels in some tissues, such as head kidney, spleen, and gills (Figure 20.2B). Relative to tissue-specific genes (TS), GUPS were less tissue-specific (Kruskal–Wallis-Test, p-value < 0.0001). Although GUPS showed a higher tissue specificity index than genes not under positive selection (non-GUPS) (Kruskal–Wallis-Test, p-value < 0.01), and their expression levels were not significantly different (Figure 20.2A).

GO enrichment analysis of GUPS in *S.lucioperca*, *S. vitreus*, and *E. spectabile* revealed no significantly overrepresentated terms. Though, several GUPS in *Sander lucioperca* were associated with metabolic process, regulation of cellular process and response to stimulus. On the other hand, GUPS in *P. fluviatilis, P. flavescens* and *E. cragini* were significantly (FDR $<$ 0.05) enriched with immune-related biological processes, including regulation of immune system process (GO:0002682), regulation of defense response (GO:0031347), myeloid leukocyte activation (GO:0002274), neutrophil degranulation (GO:0043312), leukocyte mediated immunity (GO:0002274) or neutrophil activation (GO:0042119) (Supplementary file S3). A broader overview of the functional terms associated with GUPS is shown on the treeMap in Figure 20.3, representing clusters of GO terms based on their context similarity. Each rectangle in the treeMap represents a cluster of GO terms associated with genes under positive selection. The size of rectangles reflects the significance of the cluster (i.e., the number of GO terms in the cluster). Closely related GO terms are clustered together in a supercluster of the same colour (Figure 20.3).

Table 20.3.: Statistics on lineage-specific positive selection in the representative *Percidae* species.

Branch	No of. CDS	No. of GUPS	Mean ω (d_N/d_S)	Avg No. of Sites
Sander lucioperca	56,899	43	5.11	6.63
Sander vitreus	34,187	63	4.08	9.16
Perca flavescens	43,150	137	3.41	8.41
Perca fluviatilis	50,212	154	5.97	7.80
Etheostoma spectabile	45,699	152	4.07	9.10
Etheostoma cragini	45,199	124	3.24	9.22

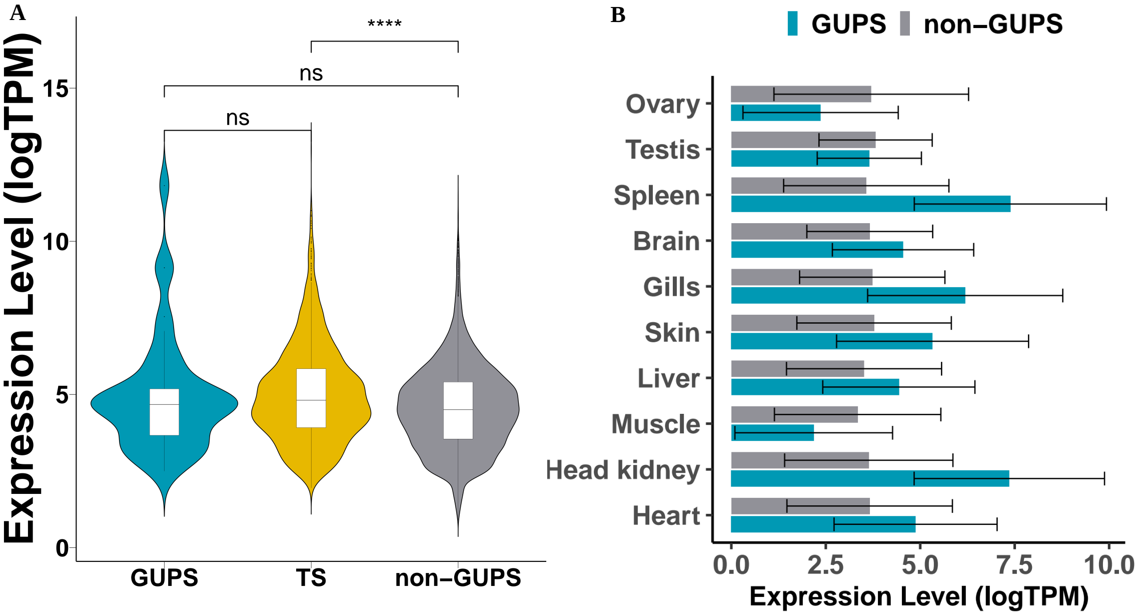

Figure 20.2.: Gene expression levels of GUPS vs. Tissue-specific genes. Violin plot comparing log transformed expression levels of GUPS, TS (Tissue-specific) and non-GUPS (all genes not under positive selection) (**A**). The expression levels between GUPS and non-GUPS are not significantly different (Kruskal–Wallis-Test). Log transformed mean expression of GUPS and non-GUPS in each tissue type (**B**). Statistical significance *ns* : Not significant; ****: Extremely significant ($p < 0.0001$).

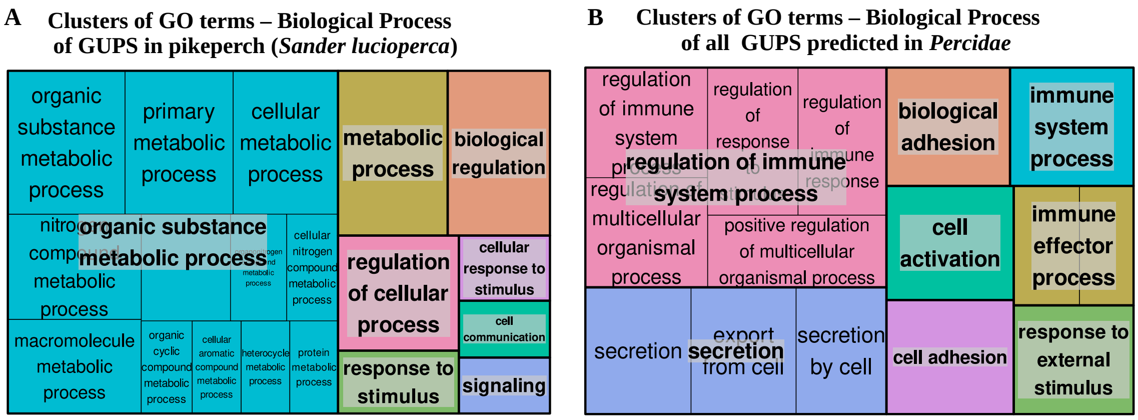

Figure 20.3.: Treemap clusters of GO-terms based on their semantic similarity. TreeMaps depicting GO terms (Biological Process) clusters GUPS in pikeperch (**A**) and in Percidae (all analysed species) (**B**), respectively. Each rectangle represents a cluster of related GO terms. The sizes of rectangle reflect the significance of the cluster (# of GO terms included in the cluster). Closely related GO terms are grouped together in a super-cluster of the same colour.

21. Discussion and Conclusion

Discussion

Comparative genomics is an approach to understanding and quantifying the contribution of natural selection to the molecular evolution of biological species. Given the divergent temperature habitats and morphological traits of *Percidae* species, this genome-wide orthology and positive selection analysis aimed to discover the genes and mechanisms that drive adaptation to different temperature environments on the one hand and to shed light on their phylogenetic history. Comparisons among representative *Percidae* genomes allowed us to identify fast-evolving genes in different species. These insights are crucial to investigate how molecular evolution relates to adaptation and phenotypic evolution in this fish family.

Functional annotation found no significantly overrepresented biological process among the putative positively selected genes in pikeperch. However, a substantial number (2/3) of pikeperch genes were associated with organic substances, metabolic process and cellular response to stimuli (Figure 20.3A). GO terms associated with GUPS in *Percidae* mainly included the regulation of the immune system and cellular secretion (Figure 20.3B). This result suggests that most of these genes are implicated in the process of immune regulation and activation [187]. Subsequent functional analyses would broadly characterize their adaption features in the life history of percids fish.

In an attempt to associate tissue specificity with fast-evolving genes in pikeperch, the expression levels of GUPS predicted in the pikeperch-specific lineage were analyzed across the pikeperch tissues reported in Part III. Unexpectedly, only two GUPS were found tissue-specific (liver-specific), and no significant correlation between positive selection and tissue-specificity was established. Moreover, most of the detected GUPS (27/32) were classified as "Mixed" (19) or "Expressed-In-All" (8) and tended to be expressed at lower levels relative to tissue-specific genes. This result strengthens the previous hypothesis that genes under natural selection are more likely to be expressed at moderate or lower levels [188, 189]. On the other hand, highly tissue-specific genes are significantly expressed at higher levels. Thus, we can hypothesize that high tissue specificity in *Percidae* might release some genes from selection pressure.

The phylogenomics analysis based on gene orthologs has confirmed the known evolutionary taxonomic classification of percid species by perfectly clustering each genus in the expected species tree. I also expected *Perca* species to be evolutionary closer to *Sander* spp. than *Etheostoma* spp. Interestingly, these results also reveal pikeperch as a relatively new species, as the genus Sander diverged only around four million years ago from the LCA of *Luciopercianae*. This result is consistent with previous studies dating the origin of the European Sander spp. to the Pliocene and Pleistocene, approx. 2.4 - 5.4 million years ago [190]. Moreover, the estimated divergence time of 20.18 MYA of the genus Perca including PFLU and PFLA, is also supported by previous findings, which suggested an estimated origin 19.8 million years ago during the early Miocene Epoch [191]. The same reference dated the fossil ancestor of *Sander* and *Perca* genera around 25-33 MYA, which is in the same range as my estimate of 27.18 MYA in this study. Collectively, these phylogenetic analysis results can be considered confident and accurate since they have strong support in independent studies, which were based on mitochondrial DNA rather than on whole-genome orthologs like in this work. Therefore, these results are more robust and significant since they involved thousands of single-copy genes across species.

The classification of orthologous gene families showed that the number of single-copy genes in Sander species (SVIT and SLUC) was lower than in the other species, suggesting a higher level of gene duplication in pikeperch, for example. This assumption is substantiated by a large number of many-to-one orthologs (30415 OGs) and duplication events (ca. 9000) identified in SLUC . Pikeperch has, namely, experienced the most gene duplication events among the analyzed percids species. I assume that this has also contributed to the higher repeat content (39%) of pikeperch predicted on the reference genome in Part II. SLUC has the highest rate of repetitive DNA among percids. Since the walleye (SVIT) genome (the closest relative of SLUC) used in this analysis was highly fragmented and not at the chromosome level, we cannot confidently interpret the lowest rate (2%) of gene duplication observed in its genome. However, we speculate that most gene duplication events could not be detected because of too many partial and missing genes in the walleye genome.

Conclusion

In summary, I have analyzed the phylogenetic relationship between selected species representing three subfamilies in the *Percidae* lineage based on genome-wide orthologs analysis. Comparative genome analyses provided new insights into recently duplicated genes in the species-specific branch. They informed about differences in terms of orthologs content between *Percidae.* In addition, molecular phylogenetic dating allowed to reconstruct the evolutionary history of percid species with confident estimates of their relative divergence time.

22. General Discussion, Perspectives and Conclusions

Fishes are the most diverse vertebrate lineage with about 34,000 living fish species recorded by FishBase (https://www.fishbase.de/), of which only 270 have their reference genome sequenced or other genomic resources and information available at different levels [10, 49]. The picture is even more dramatic for farmed fish species, where more than 90% have no genetically informed breeding programs for selective and precision farming [192]. At the start of this PhD project, pikeperch did not have any genome data and resources available to enhance its innovative farming, optimal domestication, and adaption into intensive aquaculture systems.

The primary purpose of this thesis was to fill these essentials, however unmet research needs by developing state-of-the-art genomic data, genetic tools, and resources as a landmark scientific contribution in the genome-based aquaculture research of pikeperch. In this chapter, the critical contributions will be discussed in a general context and its concrete implications for applied and translational aquaculture genomics research will be exemplified. Results and insights will be put into a general context. Their potential implications for aquaculture genomics will be discussed. Readers should please refer to the specific '**Results and Discussion**' sections of each Part for discussion on the methods developed and applied, as well as for the interpretation of the specific results.

22.1. General Context and Main Contributions of this Work

The development of genomic tools and resources and their application has enabled the emergence of genome-informed technologies to improve the rearing, conservation, and sustainability of aquaculture species. These various genomic resources promote basic research such as comparative genomics, association studies, and genome-wide detection of selection signatures and support applied and translational research in aquaculture and fisheries.

This thesis has contributed to that hot genomics research question by leveraging modern high-throughput sequencing and computational genomics and bioinformatics methods to develop and share fundamental, but necessary species-specific resources for the community and to support genome-based aquaculture research. Starting from a single animal, I have elucidated its genome organization and architecture with a comprehensive annotation of its structural elements. This genome of competitive quality helped in the subsequent transcriptomic analyses to develop and validate a genome-wide catalog of gene expression, co-expression networks, and functional modules across multiple tissues. Both genomic and transcriptomic resources and data sets have built a framework for resequencing and genotyping various genetic variants such SNVs, STRs, and SVs and for comparative genomics analyses. Gained insights and generated resources have been validated and made publicly available in peer-reviewed publications and public genomic databases for the scientific community.

This work does not claim nor aim to provide ready to use solutions for the research challenges in aquaculture genomics, since the data presented here only provide groundbreaking resources and instruments, which are indispensable for more targeted and applied large-scale investigations. However, this thesis demonstrates effectively how—like in transnational biomedical research—basic science in genomics and bioinformatics can provide useful data and basic insights for translation research in livestock and aquaculture. In particular, the developed resources and tools serve as raw materials for advanced studies towards the development of customized and SMART farming approaches as well as for translational research in the aquaculture of this species. The subsequent challenges are how to efficiently harness these genomic data and transform these basic findings into optimal aquaculture practices, such as smart stock management, precision and selective breeding.

22.1.1 Contribution: Genomic Resources for SMART Farming

Genome and transcriptome sequencing is the initial step in a whole genome project, since it generates reference genome and transcriptome data that is useful for downstream analyses and applications. In this work, I have leveraged different NGS technologies including SRs and LRs sequencing to produce the first genomic and transcriptomic sequencing data of the pikeperch.

Whole Genome and Transcriptome Sequencing

Even though decreased sequencing costs have made genome projects feasible for small labs—what was a few years ago only possible by the combined efforts of multiple research groups, whole-genome sequencing remains a challenging task for eukaryote species. In particular, fish genomes underwent at least three whole-genome duplications [193], resulting in a substantial accumulation of repetitive genomic sequences, which complicates sequencing and assembly for major fish species. For instance, well-established European farmed fish such as Atlantic cod or Atlantic salmon got their whole genome completely sequenced only in 2011 and 2016, respectively. This was achieved through international efforts involving scientists from different labs around the world, including Norway, Germany, Chile, Canada, China, and USA [6, 122]. With a genome size of nearly 1000 Mb, the sequencing of the pikeperch genome was challenged by low complexity regions, usually within repeats, which made up roughly 40% of the genome. However, combining LRs and SRs NGS technologies has mitigated this challenge, such that a high rate ($> 99\%$) of genome coverage was achieved, hence facilitating subsequent assembly and annotation.

The raw NGS data generated here are indispensable for developing genome and transcriptome assembly and analyzing genome-wide gene expression and design of genetic (bio)-markers for molecular analyses. Much more, its usefulness substantially relies on the efficient and adequate annotation of genomes, which in turn depends on the availability of several genomic resources and tools, including reference genome and transcriptome sequences, characterization of genetic variation, and noncoding features of the genome. The raw sequencing data developed here are available in public sequence databases. They can be retrieved and used by scientists for comparative genome mapping, design of primers for molecular analyses, and targeted screening of the pikeperch genome.

Fundamental Genomics Resources

Genomic and transcriptomic data of competitive quality have been developed in this thesis by integrative NGS data analysis and harnessing different computational and statistical methods in customized analysis pipelines (see Methods section in Parts II–V).
Building a high-confident genome sequence is a key prerequisite towards the development of specie-specific, customized breeding approaches in aquaculture, including the identification of genome-informed indicators and parameters for monitoring fish's health, growth, welfare, and adaption to environmental stressors [83]. Transcriptomics resources complement genomics information by providing a solid basis to develop molecular tools for, e.g., diagnosis and the recording and quantifying stress and performance parameters. Transcriptomic data additionally improve our thorough investigation and assessment of commercial traits. Transcriptomics ultimately aids our understanding of the adaptation, development, and metabolism parameters, which can inform aquaculturists in making customized adjustments to environmental conditions, such as hypoxia, temperature, nutrition, or salinity levels.

Genomics resources such as reference genomes have been developed for major comestible fishes of commercial relevance [7, 10], where they have proven to be instrumental in addressing many challenges of sustainable fisheries and aquaculture through the use of SMART breeding approaches. Since there are no SMART breeding programs nor genome-based intelligent monitoring systems for pikeperch farming so far, the resources developed here represent a fundamental contribution that will lay a ground framework for research towards innovative and optimal animal rearing, broodstock management, and breeding schemes. Moreover, with the genome and transcriptome annotation, the identified key genomic features, measuring characteristics like i.e. animal health, resilience, environmental impact, and reproduction parameters, are considerably facilitated. In sum, the comprehensive and informative genome, and transcriptome assembly, and functional annotation presented here could empower a standardized assessment and monitoring of both ecologically and production-relevant traits.

Resource for Gene Expression and Genetic Diversity

As part of my thesis work, I have comprehensively characterized the spatial landscape of gene expression and co-expression network modules in pikeperch (Part III), and provided an end-to-end workflow to predict, genotype, validate and annotate high-confident genetic variants including SNVs, SVs, and STRs in domesticated pikeperch broodstock (Part IV). Spatial, i.e., multitissues gene expression atlas and catalogs of genetic variations are other necessary resources used in fish genomics and biology for the characterization of family structures [171, 172], to perform GWAS [194], correlate important commercial traits with genetic variants, and to measure QTLs [195]. The genetic variation landscape of a species is a crucial resources to investigate genetic plasticity and diversity in that species, as demonstrated in some commercial fish species like salmonids [196, 197] or European seabass [198].

The expression and co-expression atlas in pikeperch was characterized for a panel of vital tissues sampled from animals of both sexes. This allowed quantifying the divergence and correlation between the transcriptomes of the different tissues, but also the characterization of sex-specific and tissue-specific gene regulation and evolution of transcript complexity. These insights are of scientific importance in aquaculture research in that they contribute to a deeper understanding of the molecular mechanisms of tissue activity and function, help to discover key regulatory features, and shed light on the correlated phenotypic and functional evolution of tissues, which in turn—provide a basic understanding of the key production and life-history traits of the animal [199]. Besides, the panels of genetic variants and their distribution in a domesticated pikeperch population have determined family structures and detected signatures of ongoing positive selective. It is well known that variants such as SVs and microsatellites constitute a significant source of genetic and phenotypic variations with potential impact on fitness, resilience, and adaptation [171, 172]. Integrating spatial gene expression and co-expression networks with the catalog of genetic variants and both transcriptomic and genomic data has provided raw ingredients to analyze genetic biodiversity, and investigate omics-informed farming practices for this percid fish. Collectively, the development of sequencing data, genome and transcriptome sequences, gene expression, and genetic variation data are instrumental in addressing most of the basic challenges hampering sustainable fisheries and aquaculture production of pikeperch.

22.1.2 Contribution: Genomics Tools towards Translational Aquaculture

Data such as gene expression profiles, SNV markers, structural variants, STRs variants generated in this thesis could also serve as genetic tools for applied aquaculture research, including genotype-phenotype correlation, marker-assisted selection, QTLs mapping, sex and straits discrimination, and genome editing technologies.

Tools for Customized Genomic Selection

Genomic selection (GS) is a technique used to estimate individualized breeding parameters using a large number of markers distributed across the genome [200]. The idea of GS is to directly predict productions characteristics in terms of breeding values (BVs) solely from high-density genomic markers, without integrating phenotypes information. It uses a test population with phenotypic data along with a training set of individuals who have been genotyped and phenotyped to develop mathematical and machine learning predictive models to determine the optimal BVs [200]. The new tools presented here aim to accelerate the identification and design of specific breeding parameters through targeted sequencing or resequencing of core genes affected by markers like SNPs, STRs, or SVs. Gene expression data can then be used to quantify these qualitative parameters. In European carp, for example, only 20 molecular markers were used to develop a practical selective breeding approach in a pool of over three million individuals [201], which demonstrates the vast opportunities that offer genetic tools for marker-assisted selection.

Tools for Genetic and Phenotypic Correlation

Genome-wide catalogs of genome variations, high-throughput genotyping, and expression datasets are an instrument for GWAS in a population of interest. A great challenge in the industrial application of aquaculture research is to quickly and efficiently measure performance and production traits in different broodstocks [7]. However, these traits often underlie hidden causative genomic signatures, revealed through reference genome and transcriptome sequences and annotations, genome-wide polymorphic markers, efficient genotyping, and high-density linkage maps. Hence, the data developed here fulfill this requirement in that subsequent research can use them to investigate if any polymorphic marker or variant in the genome is associated with a trait of interest. This association of phenotypic data with observable characteristics is a way to identify causal genes.

Tools for Genome Editing Technologies

Genome editing (GE) refers to a set of technologies that give the ability to perform specific changes at targeted genomic sites of interest [202]. Notable genome editing technologies comprise transcription activator-like effector nucleases (TALEN) and CRISPR-associated protein-9 nuclease (Cas9). The potential of GE to improve aquaculture breeding and production has been successfully demonstrated in diverse aquaculture species of *Salmonidae* [203], *Cyprinidae* [204, 205] and *Siluridae* [206] fish families. In all these examples, species-specific genomics information and tools have been harnessed to use these technologies efficiently. Therefore, this work paves the way to future applications in pikeperch and other percids species where genome-based technologies in aquaculture still fall short.

Tools for Sex and Complex Traits Discrimination

In aquaculture, it is often crucial to know the sex of individuals because some species show differentiated or sex-specific straits, e.g., growth rate. To that end, practical tools are needed to determine sex and characterize sex-specific growth. High-density STRs and SNPs polymorphic markers developed here offer a considerable opportunity for investigating sex-linked markers, which facilitates future selection and broodstock management strategies.

Tools for the Development of SNPs Chips

The comprehensive SNPs panel developed in this work could serve as a predictive tool for developing SNPs chips using significantly predictive SNPs between groups of individuals. For instance, SNPs displaying significant heterozygous genotypic differences between disease resistant and robust pikeperch and more susceptible individuals of a given cohort can now be identified. Short oligonucleotides encompassing a subset of significant SNPs, e.g., 1000 SNPs per Mb, could be immobilized on an SNPs array as allele-specific oligonucleotides. These high-density SNPs array are most frequently used in aquaculture for genome-wide association studies, selection of complex traits, and in studies aiming to detect loss of heterozygosity in broodstocks.

22.2. Perspectives and Future Directions

This Ph.D. thesis is a substantial contribution to the pikeperch genome project. However, the elucidation of the pikeperch genome, with the underlying resources developed here, is not the end of the project. Rather, this opens promising perspectives for in-depth functional genomics studies, which are crucial for biological inference and a better understanding of pikeperch biology.

The newly developed pikeperch reference genome provides essential genomic resources to investigate genomic and transcriptional hallmarks for the phenotypic characteristics of this species. Additionally, it paves the way for a comprehensive analysis of genome resequencing data from larger cohorts of individuals to enlighten our understanding of the genomic basis of natural selection and how pikeperches adapt to changing environments. These insights will ultimately aid stock management programs and bring pikeperch genomic research to the next level. Moreover, the genomic tools and resources developed in this work will be essential for addressing the genomic effects of smart selection related to breeding programs and thus, enhance the efficiency of selection for improved growth, disease resistance, and other economically important traits for its aquaculture.

A future direction shall be the functional analysis of key genes that have advantageous implications in pikeperch farming. Ultimately, a customized selection scheme that builds upon utilizing the genome information from both wild and captive populations should be developed. For a viable and sustainable future in the aquaculture, these approaches will be inherent components towards the construction of a successful breeding program for pikeperch. Finally, future domestication research projects using modern genomic tools should also benefit from this work.

However, for the developed genomics resources to effectively be helpful, they need to be combined with phenotypic data and other identifiable traits correlated with genomic markers and functional features. This work has neither addressed these kinds of functional associations nor validated the function of the predicted genes and genomic markers.

22.3. General Conclusion

This doctoral thesis aimed to develop and establish essential species-specific genomic resources and genetic tools for the valued farmed fish pikeperch, an emerging aquaculture species for the European aquaculture sector. To unveil comprehensive and accurate genomic information for this commercial species, I applied state-of-the-art next-generation sequencing technologies and integrated different high-throughput data and bioinformatics methods to build and annotate reference-quality genome and transcriptome data. These data are the first quantitative and qualitative foundation for future in-depth and targeted large-scale genomics research in pikeperch. These initial resources were leveraged in subsequent analyses to characterize the genome-wide atlas of gene expression and co-expression networks and functional modules across multiple individuals and tissues. Moreover, genomics and transcriptomics information was used to establish the landscape of genetic variation, including SNPs, STRs, and SVs in broodstock families, including several hundred individuals. Finally, comparative genome analyses provided critical insights into the recent positive selection and evolutionary life history of pikeperches within the *Percidae* fish family.

The reported genomic resources will enhance genomics, genetics, and breeding research in pikeperch. More specifically, the data generated and insights gained through this thesis will serve as essential resources to a broader community of genomicists to investigate the causal links between genome and relevant aquaculture traits. It additionally lays the fundamental framework for understanding the genomic architecture of these traits, thereby significantly enhancing the species-specific breeding programs, rearing, and production practices. Ultimately, these species-specific resources provide crucial insights susceptible to improving pikeperch production efficiency and sustainability in the aquaculture sector.

The findings, output data, and resources have been validated, structured, and disseminated for easy use by the scientific community in peer-reviewed publications and public genomic data repositories, respectively, following the FAIR principles.

22.4. Data and Resources Availability

Genomics Resources and Data

The raw sequencing data generated in this project is available at the NCBI Sequence Read Archive (SRA) under Accession Number PRJNA626522. The annotated assembly of *Sander lucioperca* is available at the NCBI GenBank under the Accession Number GCA_008315115.2. Raw genomic reads are available at SRA BioSample-accession SAMN12618724.

Transcriptomic Resources and Data

The raw RNA-Seq reads are openly available at NCBI SRA (BioProject PRJNA752979). The transcriptome shotgun assembly has been deposited at DDBJ/EMBL/GenBank under the accession GJIW00000000. The version described in this paper is the first version, GJIW01000000. Codes used for data analysis as well as generated figures, tables, and extended methods are available on github (`https://github.com/bbalog87/Pikeperch_transcriptomics`, accessed on 28 September 2021).

Pipeline Codes and other Resources

Codes and pipelines used in the analyses as well as extensive methods and data are available on the following Github repositories:

ngs_containers set of singularity and docker containers for NGS analysis pipelines.
– `https://github.com/bbalog87/ngs_containers`

Pikeperch_transcriptomics pipelines and scripts fo multitissues transcriptome assembly of pikeperch, tissue expression atlas, co-expression and positive selection analyses.
– `https://github.com/bbalog87/Pikeperch_transcriptomics`

SLUC_assembly Scripts and resources for pikeperch (*S. lucioperca*) genome project
– `https://github.com/bbalog87/SLUC_assembly`

23. Bibliography

[1] "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb. 2001.

[2] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time DNA Sequencing from Single Polymerase Molecules," *Science*, vol. 323, pp. 133–138, Jan. 2009.

[3] E. Pennisi, "Search for Pore-fection," *Science*, vol. 336, pp. 534–537, May 2012.

[4] S. E. Levy and R. M. Myers, "Advancements in Next-Generation Sequencing," *Annual Review of Genomics and Human Genetics*, vol. 17, pp. 95–115, Aug. 2016.

[5] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner, "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes," *Science*, vol. 297, pp. 1301–1310, Aug. 2002.

[6] B. Star, A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrøm, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, O. F. Wetten, A. Lanzén, R. Winer, J. Knight, J. H. Vogel, B. Aken, O. Andersen, K. Lagesen, A. Tooming-Klunderud, R. B. Edvardsen, K. G. Tina, M. Espelund, C. Nepal, C. Previti, B. O. Karlsen, T. Moum, M. Skage, P. R. Berg, T. Gjøen, H. Kuhl, J. Thorsen, K. Malde, R. Reinhardt, L. Du, S. D. Johansen, S. Searle, S. Lien, F. Nilsen, I. Jonassen, S. W.

Omholt, N. C. Stenseth, and K. S. Jakobsen, “The genome sequence of Atlantic cod reveals a unique immune system,” *Nature*, vol. 477, pp. 207–210, Aug. 2011.

[7] H. Abdelrahman, M. ElHady, A. Alcivar-Warren, S. Allen, R. Al-Tobasei, L. Bao, B. Beck, H. Blackburn, B. Bosworth, J. Buchanan, J. Chappell, W. Daniels, S. Dong, R. Dunham, E. Durland, A. Elaswad, M. Gomez-Chiarri, K. Gosh, X. Guo, P. Hackett, T. Hanson, D. Hedgecock, T. Howard, L. Holland, M. Jackson, Y. Jin, K. Khalil, T. Kocher, T. Leeds, N. Li, L. Lindsey, S. Liu, Z. Liu, K. Martin, R. Novriadi, R. Odin, Y. Palti, E. Peatman, D. Proestou, G. Qin, B. Reading, C. Rexroad, S. Roberts, M. Salem, A. Severin, H. Shi, C. Shoemaker, S. Stiles, S. Tan, K. F. Tang, W. Thongda, T. Tiersch, J. Tomasso, W. T. Prabowo, R. Vallejo, H. van der Steen, K. Vo, G. Waldbieser, H. Wang, X. Wang, J. Xiang, Y. Yang, R. Yant, Z. Yuan, Q. Zeng, and T. Zhou, “Erratum to: Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research,” *BMC Genomics*, vol. 18, p. 235, Mar. 2017.

[8] P. Xu, X. Zhang, X. Wang, J. Li, G. Liu, Y. Kuang, J. Xu, X. Zheng, L. Ren, G. Wang, Y. Zhang, L. Huo, Z. Zhao, D. Cao, C. Lu, C. Li, Y. Zhou, Z. Liu, Z. Fan, G. Shan, X. Li, S. Wu, L. Song, G. Hou, Y. Jiang, Z. Jeney, D. Yu, L. Wang, C. Shao, L. Song, J. Sun, P. Ji, J. Wang, Q. Li, L. Xu, F. Sun, J. Feng, C. Wang, S. Wang, B. Wang, Y. Li, Y. Zhu, W. Xue, L. Zhao, J. Wang, Y. Gu, W. Lv, K. Wu, J. Xiao, J. Wu, Z. Zhang, J. Yu, and X. Sun, “Genome sequence and genetic diversity of the common carp, cyprinus carpio,” *Nature Genetics*, vol. 46, pp. 1212–1219, Sep. 2014.

[9] M. Y. Laghari, P. Lashari, X. Zhang, P. Xu, N. T. Narejo, B. Xin, Y. Zhang, and X. Sun, “QTL mapping for economically important traits of common carp (Cyprinus carpio L.),” *J Appl Genet*, vol. 56, pp. 65–75, Feb. 2015.

[10] G. Lu and M. Luo, “Genomes of major fishes in world fisheries and aquaculture: Status, application and perspective,” *Aquaculture and Fisheries*, vol. 5, no. 4, pp. 163–173, 2020.

[11] M. Tine, H. Kuhl, P. A. Gagnaire, B. Louro, E. Desmarais, R. S. Martins, J. Hecht, F. Knaust, K. Belkhir, S. Klages, R. Dieterich, K. Stueber, F. Piferrer, B. Guinand, N. Bierne, F. A. Volckaert, L. Bargelloni, D. M. Power, F. Bonhomme, A. V. Canario, and R. Reinhardt, “European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation,” *Nat Commun*, vol. 5, p. 5770, Dec. 2014.

[12] A. Figueras, D. Robledo, A. Corvelo, M. Hermida, P. Pereiro, J. A. Rubiolo, J. Gómez-Garrido, L. Carreté, X. Bello, M. Gut, I. G. Gut, M. Marcet-Houben, G. Forn-Cuní, B. Galán, J. L. García, J. L. Abal-Fabeiro, B. G. Pardo, X. Taboada, C. Fernández, A. Vlasova, A. Hermoso-Pulido, R. Guigó, J. A. Álvarez Dios,

A. Gómez-Tato, A. Viñas, X. Maside, T. Gabaldón, B. Novoa, C. Bouza, T. Alioto, and P. Martínez, "Whole genome sequencing of turbot (Scophthalmus maximus; Pleuronectiformes): a fish adapted to demersal life," *DNA Res*, vol. 23, pp. 181–192, Jun. 2016.

[13] J. M. YÃ¡Ã±ez, R. D. Houston, and S. Newman, "Genetics and genomics of disease resistance in salmonid species," *Frontiers in Genetics*, vol. 5, Nov. 2014.

[14] J. P. Lhorente, J. A. Gallardo, B. Villanueva, M. J. Carabaño, and R. Neira, "Disease resistance in Atlantic salmon (Salmo salar): coinfection of the intracellular bacterial pathogen Piscirickettsia salmonis and the sea louse Caligus rogercresseyi," *PLoS One*, vol. 9, no. 4, p. e95397, 2014.

[15] G. D. Wiens, Y. Palti, and T. D. Leeds, "Three generations of selective breeding improved rainbow trout (oncorhynchus mykiss) disease resistance against natural challenge with flavobacterium psychrophilum during early life-stage rearing," *Aquaculture*, vol. 497, pp. 414–421, Dec. 2018.

[16] C. P. Moraleda, D. Robledo, A. P. Gutiérrez, J. Del-Pozo, J. M. Yáñez, and R. D. Houston, "Investigating mechanisms underlying genetic resistance to Salmon Rickettsial Syndrome in Atlantic salmon using RNA sequencing," *BMC Genomics*, vol. 22, p. 156, Mar. 2021.

[17] Y. Wang, Y. Lu, Y. Zhang, Z. Ning, Y. Li, Q. Zhao, H. Lu, R. Huang, X. Xia, Q. Feng, X. Liang, K. Liu, L. Zhang, T. Lu, T. Huang, D. Fan, Q. Weng, C. Zhu, Y. Lu, W. Li, Z. Wen, C. Zhou, Q. Tian, X. Kang, M. Shi, W. Zhang, S. Jang, F. Du, S. He, L. Liao, Y. Li, B. Gui, H. He, Z. Ning, C. Yang, L. He, L. Luo, R. Yang, Q. Luo, X. Liu, S. Li, W. Huang, L. Xiao, H. Lin, B. Han, and Z. Zhu, "The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation," *Nat Genet*, vol. 47, pp. 625–631, Jun. 2015.

[18] N. I. of Health (NIH), "Fact-sheets." https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data, 2021. [Online; downloaded 19-July-2021].

[19] N. Wang, X. Xu, and P. Kestemont, "Effect of temperature and feeding frequency on growth performances, feed efficiency and body composition of pikeperch juveniles (sander lucioperca)," *Aquaculture*, vol. 289, pp. 70–73, Apr. 2009.

[20] N. Alexi, D. V. Byrne, E. Nanou, and K. Grigorakis, "J Sci Food AgricInvestigation of sensory profiles and hedonic drivers of emerging aquaculture fish species," *J Sci Food Agric*, vol. 98, pp. 1179–1187, Feb. 2018.

[21] P. Kestemont, K. Dabrowski, and R. C. Summerfelt, *Biology and Culture of Percid Fishes: Principles and Practices.* Springer, 1st ed., Oct. 2015.

[22] J. Dalsgaard, I. Lund, R. Thorarinsdottir, A. Drengstig, K. Arvonen, and P. B. Pedersen, "Farming different species in RAS in nordic countries: Current status and future perspectives," *Aquacultural Engineering*, vol. 53, pp. 2–13, 2013.

[23] A. M. Samarin, M. Blecha, D. Bytyutskyy, and T. Policar, "Post-Ovulatory Oocyte Ageing in Pikeperch (sander lucioperca l.) and its Effect on Egg Viability Rates and the Occurrence of Larval Malformations and Ploidy Anomalies," *Turkish Journal of Fisheries and Aquatic Sciences*, vol. 15, pp. 429–435, 2015.

[24] D. Pyanov, A. Delmukhametov, and E. Khrustalev, "Pike-perch Farming in Recirculating Aquaculture Systems (RAS) in the Kaliningrad Region," in *9th Baltic Conference on Food Science and Technology "Food for Consumer Well-Being" (FoodBalt 2014)*, FoodBlat 2014, pp. 315–317, 2014.

[25] R. Rufchaei, S. Nedaei, S. H. Hoseinifar, S. Hassanpour, M. Golshan, and M. S. Bourani, "Improved growth performance, serum and mucosal immunity, haematology and antioxidant capacity in pikeperch (sander lucioperca) using dietary water hyacinth (eichhornia crassipes) leaf powder," *Aquaculture Research*, vol. 52, pp. 2194–2204, Dec. 2020.

[26] M. Bercsényi, B. Urbányi, M. Bódis, and T. Müller, "Comparison of growth in pike-perch (sander lucioperca) and hybrids of pike-perch (s. lucioperca) × volga pike-perch (s. volgensis)," *Israeli Journal of Aquaculture - Bamidgeh*, Jan. 2011.

[27] FAO, "Fishery and aquaculture statistics. global production by production source 1950–2017," FishstatJ, 2019.

[28] P. Kestemont, X. Xueliang, N. Hamza, J. Maboudou, and I. I. Toko, "Effect of weaning age and diet on pikeperch larviculture," *Aquaculture*, vol. 264, pp. 197–204, Apr. 2007.

[29] S. Baekelandt, B. Redivo, S. N. Mandiki, T. Bournonville, A. Houndji, B. Bernard, N. E. Kertaoui, M. Schmitz, P. Fontaine, J.-N. Gardeur, Y. Ledoré, and P. Kestemont, "Multifactorial analyses revealed optimal aquaculture modalities improving husbandry fitness without clear effect on stress and immune status of pikeperch sander lucioperca," *General and Comparative Endocrinology*, vol. 258, pp. 194–204, Mar. 2018.

[30] M. Szkudlarek and Z. Zakeś, "Effect of stocking density on survival and growth performance of pikeperch, sander lucioperca (l.), larvae under controlled conditions," *Aquaculture International*, vol. 15, pp. 67–81, Jan. 2007.

[31] T. Policar, M. Blecha, J. Křištan, J. Mráz, J. Velíšek, A. Stará, V. Stejskal, O. Malinovskyi, P. Svačina, and A. M. Samarin, “Comparison of production efficiency and quality of differently cultured pikeperch (sander lucioperca l.) juveniles as a valuable product for ongrowing culture,” *Aquaculture International*, vol. 24, pp. 1607–1626, Aug. 2016.

[32] R. D. Houston, T. P. Bean, D. J. Macqueen, M. K. Gundappa, Y. H. Jin, T. L. Jenkins, S. L. C. Selly, S. A. M. Martin, J. R. Stevens, E. M. Santos, A. Davie, and D. Robledo, “Nat Rev GenetHarnessing genomics to fast-track genetic improvement in aquaculture,” *Nat Rev Genet*, vol. 21, pp. 389–409, Jul. 2020.

[33] D. Żarski, A. Le Cam, J. Nynca, C. Klopp, S. Ciesielski, B. Sarosiek, J. Montfort, J. Król, P. Fontaine, A. Ciereszko, and J. Bobe, “Mol Reprod DevDomestication modulates the expression of genes involved in neurogenesis in high-quality eggs of Sander lucioperca,” *Mol Reprod Dev*, vol. 87, pp. 934–951, Sep. 2020.

[34] X. Han, Q. Ling, C. Li, G. Wang, Z. Xu, and G. Lu, “Characterization of pikeperch (sander lucioperca) transcriptome and development of SSR markers,” *Biochemical Systematics and Ecology*, vol. 66, pp. 188–195, Jun. 2016.

[35] J. Guo, C. Li, T. Teng, F. Shen, Y. Chen, Y. Wang, C. Pan, and Q. Ling, “Construction of the first high-density genetic linkage map of pikeperch (sander lucioperca) using specific length amplified fragment (SLAF) sequencing and QTL analysis of growth-related traits,” *Aquaculture*, vol. 497, pp. 299–305, Dec. 2018.

[36] M. Kottelat and J. Freyhof, *Handbook of European Freshwater Fishess.* Kottelat, 2007.

[37] E. Eschbach, A. W. Nolte, K. Kohlmann, P. Kersten, J. Kail, and R. Arlinghaus, “Population differentiation of zander (Sander lucioperca) across native and newly colonized ranges suggests increasing admixture in the course of an invasion,” *Evol Appl*, vol. 7, pp. 555–568, May 2014.

[38] B. B. Collette and P. Banarescu, “Systematics and Zoogeography of the Fishes of the Family Percidae ,” *Journal of the Fisheries Research Board of Canada*, vol. 34, no. 10, pp. 1450–1463, 1997.

[39] M. Frisk, P. V. Skov, and J. F. Steffensen, “Thermal optimum for pikeperch (sander lucioperca) and the use of ventilation frequency as a predictor of metabolic rate,” *Aquaculture*, vol. 324-325, pp. 151–157, Jan. 2012.

[40] A. E. Vinogradov, “Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship,” *Cytometry*, vol. 31, pp. 100–109, feb 1998.

[41] T. Goldammer and M. B. Klinkhardt, “Karyologische Studien an verschiedenen Süßwasserfischen aus brackigen Küstenwässern der südwestlichen Ostsee. V. Der Zander (Stizostedion lucioperca (Linnaeus, 1758),” *Zool.*, vol. 3/4, no. 228, pp. 129–139, 1992.

[42] N. S. Nagpure, A. K. Pathak, R. Pati, I. Rashid, J. Sharma, S. P. Singh, M. Singh, U. K. Sarkar, B. Kushwaha, R. Kumar, and S. Murali, “Fish Karyome version 2.1: a chromosome database of fishes and other aquatic organisms,” *Database (Oxford)*, vol. 2016, 2016.

[43] J. Kitano and C. L. Peichel, “Turnover of sex chromosomes and speciation in fishes,” *Environ. Biol. Fishes*, vol. 94, no. 3, pp. 549–558, 2012.

[44] D. Tsaparis, D. Kyriakis, K. Ekonomaki, S. Darivianakis, P. Fontaine, and C. S. Tsigenopoulos, “A first step for sustainable breeding programmes in pikeperch (sander lucioperca) through the evaluation of the genetic variation in domesticated broodstocks and natural populations,” in *12. International Symposium on Genetics in Aquaculture*, vol. Volume 472 (Special Issue: SI) of *ISGA XII*, (Santiago de Compostela, Spain), ELSEVIER SCIENCE BV, Jun. 2015.

[45] D. K. Sipos, G. Kovács, E. Buza, K. Csenki-Bakos, Á. Ősz, U. Ljubobratović, R. Cserveni-Szücs, M. Bercsényi, I. Lehoczky, B. Urbányi, and B. Kovács, “Comparative genetic analysis of natural and farmed populations of pike-perch (sander lucioperca),” *Aquaculture International*, vol. 27, pp. 991–1007, Apr. 2019.

[46] N. Poulet, P. Balaresque, T. Aho, and M. Björklund, “Genetic structure and dynamics of a small introduced population: the pikeperch, Sander lucioperca, in the Rhône delta,” *Genetica*, vol. 135, pp. 77–86, Jan. 2009.

[47] M. Salminen, M. L. Koljonen, M. Säisä, and J. Ruuhijärvi, “Genetic effects of supportive stockings on native pikeperch populations in boreal lakes–cases, three different outcomes,” *Hereditas*, vol. 149, pp. 1–15, Feb. 2012.

[48] L. S. Pereira, A. A. Agostinho, and K. O. Winemiller, “Revisiting cannibalism in fishes,” *Reviews in Fish Biology and Fisheries*, vol. 27, pp. 499–513, Sep. 2017.

[49] G. Fan, Y. Song, L. Yang, X. Huang, S. Zhang, M. Zhang, X. Yang, Y. Chang, H. Zhang, Y. Li, S. Liu, L. Yu, J. Chu, I. Seim, C. Feng, T. J. Near, R. A. Wing, W. Wang, K. Wang, J. Wang, X. Xu, H. Yang, X. Liu, N. Chen, and S. He, “Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K),” *GigaScience*, vol. 9, Aug. 2020. giaa080.

[50] F. Pfeiffer, C. Gröber, M. Blank, K. Händler, M. Beyer, J. L. Schultze, and G. Mayer, “Systematic evaluation of error rates and causes in short samples in next-generation sequencing,” *Sci Rep*, vol. 8, p. 10950, Jul. 2018.

[51] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, “Quake: quality-aware detection and correction of sequencing errors,” *Genome Biol*, vol. 11, no. 11, p. R116, 2010.

[52] J. J. Kasianowicz, E. Brandin, D. Branton, D. W. Deamer, and D. W. Deamer, “Characterization of individual polynucleotide molecules using a membrane channel,” *Proc Natl Acad Sci U S A*, vol. 93, pp. 13770–13773, Nov. 1996.

[53] F. J. Rang, W. P. Kloosterman, and J. de Ridder, “From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy,” *Genome Biol*, vol. 19, p. 90, Jul. 2018.

[54] A. M. Wenger, P. Peluso, W. J. Rowell, P. C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C. S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, and M. W. Hunkapiller, “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome,” *Nat Biotechnol*, vol. 37, pp. 1155–1162, Oct. 2019.

[55] T. C. Conway and A. J. Bromage, “Succinct data structures for assembling large genomes,” *Bioinformatics*, vol. 27, pp. 479–486, Jan. 2011.

[56] P. E. C. Compeau, P. A. Pevzner, and G. Tesler, “How to apply de bruijn graphs to genome assembly,” *Nature Biotechnology*, vol. 29, pp. 987–991, Nov. 2011.

[57] J. il Sohn and J.-W. Nam, “The present and future ofde novowhole-genome assembly,” *Briefings in Bioinformatics*, p. bbw096, Oct. 2016.

[58] K.-J. Räihä and E. Ukkonen, “The shortest common supersequence problem over binary alphabet is NP-complete,” *Theoretical Computer Science*, vol. 16, no. 2, pp. 187–198, 1981.

[59] P. A. Pevzner, H. Tang, and M. S. Waterman, “An Eulerian path approach to DNA fragment assembly,” *Proc Natl Acad Sci U S A*, vol. 98, pp. 9748–9753, Aug. 2001.

[60] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, “Assembly of long, error-prone reads using repeat graphs,” *Nat Biotechnol*, vol. 37, pp. 540–546, May 2019.

[61] H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, and H. Li, “Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm,” *Nat Methods*, vol. 18, pp. 170–175, Feb. 2021.

[62] S. Nurk, B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon, R. Grothe, K. H. Miga, E. E. Eichler, A. M. Phillippy, and S. Koren, "HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads," *Genome Res*, vol. 30, pp. 1291–1305, Sep. 2020.

[63] A. M. Wenger, P. Peluso, W. J. Rowell, P. C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C. S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, and M. W. Hunkapiller, "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome," *Nat Biotechnol*, vol. 37, pp. 1155–1162, Oct. 2019.

[64] K. Shafin, T. Pesout, R. Lorig-Roach, M. Haukness, H. E. Olsen, C. Bosworth, J. Armstrong, K. Tigyi, N. Maurer, S. Koren, F. J. Sedlazeck, T. Marschall, S. Mayes, V. Costa, J. M. Zook, K. J. Liu, D. Kilburn, M. Sorensen, K. M. Munson, M. R. Vollger, J. Monlong, E. Garrison, E. E. Eichler, S. Salama, D. Haussler, R. E. Green, M. Akeson, A. Phillippy, K. H. Miga, P. Carnevali, M. Jain, and B. Paten, "Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes," *Nat Biotechnol*, vol. 38, pp. 1044–1053, Sep. 2020.

[65] M. Pop, "Genome assembly reborn: recent computational challenges," *Brief Bioinform*, vol. 10, pp. 354–366, Jul. 2009.

[66] A. V. Zimin, D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marçais, J. A. Yorke, J. Dvořák, and S. L. Salzberg, "Hybrid assembly of the large and highly repetitive genome of aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm," *Genome Research*, vol. 27, pp. 787–792, Jan. 2017.

[67] J. O. Korbel and C. Lee, "Genome assembly and haplotyping with Hi-C," *Nat Biotechnol*, vol. 31, pp. 1099–1101, Dec. 2013.

[68] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions," *Nat Biotechnol*, vol. 31, pp. 1119–1125, dec 2013.

[69] T. Kawakami, L. Smeds, N. Backström, A. Husby, A. Qvarnström, C. F. Mugal, P. Olason, and H. Ellegren, "A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution," *Molecular Ecology*, vol. 23, pp. 4035–4058, Jun. 2014.

[70] "A reference standard for genome biology," *Nature Biotechnology*, vol. 36, pp. 1121–1121, Dec. 2018.

[71] A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, C. Lee, B. J. Ko, M. Chaisson, G. L. Gedman, L. J. Cantin, F. Thibaud-Nissen, L. Haggerty, I. Bista, M. Smith, B. Haase, J. Mountcastle, S. Winkler, S. Paez, J. Howard, S. C. Vernes, T. M. Lama, F. Grutzner, W. C. Warren, C. N. Balakrishnan, D. Burt, J. M. George, M. T. Biegler, D. Iorns, A. Digby, D. Eason, B. Robertson, T. Edwards, M. Wilkinson, G. Turner, A. Meyer, A. F. Kautt, P. Franchini, H. W. Detrich, H. Svardal, M. Wagner, G. J. P. Naylor, M. Pippel, M. Malinsky, M. Mooney, M. Simbirsky, B. T. Hannigan, T. Pesout, M. Houck, A. Misuraca, S. B. Kingan, R. Hall, Z. Kronenberg, I. Sović, C. Dunn, Z. Ning, A. Hastie, J. Lee, S. Selvaraj, R. E. Green, N. H. Putnam, I. Gut, J. Ghurye, E. Garrison, Y. Sims, J. Collins, S. Pelan, J. Torrance, A. Tracey, J. Wood, R. E. Dagnew, D. Guan, S. E. London, D. F. Clayton, C. V. Mello, S. R. Friedrich, P. V. Lovell, E. Osipova, F. O. Al-Ajli, S. Secomandi, H. Kim, C. Theofanopoulou, M. Hiller, Y. Zhou, R. S. Harris, K. D. Makova, P. Medvedev, J. Hoffman, P. Masterson, K. Clark, F. Martin, K. Howe, P. Flicek, B. P. Walenz, W. Kwak, H. Clawson, M. Diekhans, L. Nassar, B. Paten, R. H. S. Kraus, A. J. Crawford, M. T. P. Gilbert, G. Zhang, B. Venkatesh, R. W. Murphy, K. P. Koepfli, B. Shapiro, W. E. Johnson, F. Di Palma, T. Marques-Bonet, E. C. Teeling, T. Warnow, J. M. Graves, O. A. Ryder, D. Haussler, S. J. O'Brien, J. Korlach, H. A. Lewin, K. Howe, E. W. Myers, R. Durbin, A. M. Phillippy, and E. D. Jarvis, "Towards complete and error-free genome assemblies of all vertebrate species," *Nature*, vol. 592, pp. 737–746, Apr. 2021.

[72] M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov, "BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes," *Molecular Biology and Evolution*, Jul. 2021.

[73] Z. J. Liu, ed., *Bioinformatics in Aquaculture.* John Wiley & Sons, Ltd, Apr. 2017.

[74] V. Dominguez Del Angel, E. Hjerde, L. Sterck, S. Capella-Gutierrez, C. Notredame, O. Vinnere Pettersson, J. Amselem, L. Bouri, S. Bocs, C. Klopp, J. F. Gibrat, A. Vlasova, B. L. Leskosek, L. Soler, M. Binzer-Panchal, and H. Lantz, "Ten steps to get started in Genome Assembly and Annotation," *F1000Res*, vol. 7, 2018.

[75] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo,

R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, p. 160018, Mar. 2016.

[76] R. Leinonen, H. Sugawara, and M. S. and, "The sequence read archive," *Nucleic Acids Research*, vol. 39, pp. D19–D21, Nov. 2010.

[77] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, pp. D61–D65, Jan. 2007.

[78] B. Gjerde, H. Gjøen, and B. Villanueva, "Optimum designs for fish breeding programmes with constrained inbreeding mass selection for a normally distributed trait," *Livestock Production Science*, vol. 47, pp. 59–72, Dec. 1996.

[79] J. T. D.-Y. Ma), M. Rye, Y.-X. Wang, K.-S. Yang, H. B. Bentsen, and T. Gjedrem, "Genetic improvement of tilapias in china: Genetic parameters and selection responses in growth of nile tilapia (oreochromis niloticus) after six generations of multi-trait selection for growth and fillet yield," *Aquaculture*, vol. 322-323, pp. 51–64, Dec. 2011.

[80] J. Thodesen and T. Gjedrem, *Breeding programs on Atlantic salmon in Norway: lessons learned.* No. 38744 in Working Papers, The WorldFish Center, 2006.

[81] P. V. Khang, T. H. Phuong, N. K. Dat, W. Knibb, and N. H. Nguyen, "Genetic Evaluation, Experiences, and Challenges," *Front Genet*, vol. 9, p. 191, 2018.

[82] K. Janssen, H. Saatkamp, and H. Komen, "Cost-benefit analysis of aquaculture breeding programs," *Genet Sel Evol*, vol. 50, p. 2, Jan. 2018.

[83] K. R. Zenger, M. S. Khatkar, D. B. Jones, N. Khalilisamani, D. R. Jerry, and H. W. Raadsma, "Genomic selection in aquaculture: Application, limitations and opportunities with special reference to marine shrimp and pearl oysters," *Frontiers in Genetics*, vol. 9, Jan. 2019.

[84] "Anaconda software distribution," 2020.

[85] mamba org, "The Fast Cross-Platform Package Manager." https://github.com/mamba-org/mamba, 2021. [Online; accessed 19-July-2021].

[86] G. M. Kurtzer, V. Sochat, and M. W. Bauer, “Singularity: Scientific containers for mobility of compute,” *PLOS ONE*, vol. 12, p. e0177459, May 2017.

[87] D. Merkel, “Docker: lightweight linux containers for consistent development and deployment,” *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

[88] S. Andrews, “Fastqc: A quality control tool for high throughput sequencing data,” *Online*, Apr. 2010.

[89] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “fastp: an ultra-fast all-in-one FASTQ preprocessor,” *Bioinformatics*, vol. 34, pp. i884–i890, Sep. 2018.

[90] G. Marcais and C. Kingsford, “A fast, lock-free approach for efficient parallel counting of occurrences of k-mers,” *Bioinformatics*, vol. 27, pp. 764–770, Mar. 2011.

[91] R. Li, W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. A. Ryder, F. C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T. W. Lam, S. M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang, “The sequence and de novo assembly of the giant panda genome,” *Nature*, vol. 463, pp. 311–317, Jan. 2010.

[92] M. Hozza, T. Vinař, and B. Brejová, “How Big is That Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra,” in *String Processing and Information Retrieval (SPIRE)* (C. S. Iliopoulos, S. J. Puglisi, and E. Yilmaz, eds.), vol. 9309 of *Lecture Notes in Computer Science*, (London, UK), pp. 199–209, Springer, Sep. 2015.

[93] G. W. Vurture, F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, and M. C. Schatz, “GenomeScope: fast reference-free genome profiling from short reads,” *Bioinformatics*, vol. 33, pp. 2202–2204, Jul. 2017.

[94] G. H. Shin, Y. Shin, M. Jung, J. M. Hong, S. Lee, S. Subramaniyam, E. S. Noh, E. H. Shin, E. H. Park, J. Y. Park, Y. O. Kim, K. M. Choi, B. H. Nam, and C. I.

Park, "First Draft Genome for Red Sea Bream of Family Sparidae," *Front Genet*, vol. 9, p. 643, 2018.

[95] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, and T. Itoh, "Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads," *Genome Res.*, vol. 24, pp. 1384–1395, Aug. 2014.

[96] B. Liu, Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, and W. Fan, "Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects," *arxiv.org*, Aug. 2013.

[97] Y. Lin, J. Yuan, M. Kolmogorov, M. W. Shen, M. Chaisson, and P. A. Pevzner, "Assembly of long error-prone reads using de Bruijn graphs," *PNAS*, vol. 113, p. 643, Dec. 2016.

[98] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754–1760, Jul. 2009.

[99] J. Catchen, A. Amores, and S. Bassham, "Chromonomer: A tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny," *G3 Genes|Genomes|Genetics*, vol. 10, pp. 4115–4128, Nov. 2020.

[100] A. Rhie, B. P. Walenz, S. Koren, and A. M. Phillippy, "Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies," *Genome Biol*, vol. 21, p. 245, Sep. 2020.

[101] J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit, "RepeatModeler2 for automated genomic discovery of transposable element families," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 9451–9457, Apr. 2020.

[102] J. M. Crescente, D. Zavallo, M. Helguera, and L. S. Vanzetti, "Mite tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–10, 2018.

[103] W. Bao, K. K. Kojima, and O. Kohany, "Repbase update, a database of repetitive elements in eukaryotic genomes," *Mobile Dna*, vol. 6, no. 1, pp. 1–6, 2015.

[104] I. Numanagic, A. S. Gökkaya, L. Zhang, B. Berger, C. Alkan, and F. Hach, "Fast characterization of segmental duplications in genome assemblies," *Bioinformatics*, vol. 34, pp. i706–i714, Sep. 2018.

[105] R. Feron, M. Zahm, C. Cabau, C. Klopp, C. Roques, O. Bouchez, C. Ech?, S. Valière, C. Donnadieu, P. Haffray, A. Bestin, R. Morvezen, H. Acloque, P. T. Euclide, M. Wen, E. Jouano, M. Schartl, J. H. Postlethwait, C. Schraidt, M. R. Christie, W. A. Larson, A. Herpin, and Y. Guiguen, "Characterization of a Y-specific duplication/insertion of the anti-Mullerian hormone type II receptor gene based on a chromosome-scale genome assembly of yellow perch, Perca flavescens," *Mol Ecol Resour*, vol. 20, pp. 531–543, Mar. 2020.

[106] R. L. Moran, J. M. Catchen, and R. C. Fuller, "Genomic Resources for Darters (Percidae: Etheostominae) Provide Insight into Postzygotic Barriers Implicated in Speciation," *Mol Biol Evol*, vol. 37, pp. 711–729, Mar. 2020.

[107] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, Sep. 1997.

[108] G. S. Slater and E. Birney, "Automated generation of heuristics for biological sequence comparison," *BMC Bioinformatics*, vol. 6, p. 31, Feb. 2005.

[109] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "AUGUSTUS: ab initio prediction of alternative transcripts," *Nucleic Acids Res.*, vol. 34, pp. W435–439, Jul. 2006.

[110] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol.*, vol. 268, pp. 78–94, Apr. 1997.

[111] W. H. Majoros, M. Pertea, and S. L. Salzberg, "Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders," *Bioinformatics*, vol. 20, no. 16, pp. 2878–2879, 2004.

[112] I. Korf, "Gene finding in novel genomes," *BMC bioinformatics*, vol. 5, no. 1, pp. 1–9, 2004.

[113] V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, and M. Borodovsky, "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training," *Genome research*, vol. 18, no. 12, pp. 1979–1990, 2008.

[114] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nat. Methods*, vol. 12, pp. 357–360, Apr. 2015.

[115] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nat Protoc*, vol. 7, pp. 562–578, Mar. 2012.

[116] B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell, and J. R. Wortman, “Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments,” *Genome Biol.*, vol. 9, p. R7, Jan. 2008.

[117] T. M. Lowe and P. P. Chan, “tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes,” *Nucleic Acids Res.*, vol. 44, pp. W54–57, Jul. 2016.

[118] K. Lagesen, P. Hallin, E. A. Roedland, H. H. Staerfeldt, T. Rognes, and D. W. Ussery, “RNAmmer: consistent and rapid annotation of ribosomal RNA genes,” *Nucleic Acids Res.*, vol. 35, no. 9, pp. 3100–3108, 2007.

[119] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, “miRBase: microRNA sequences, targets and gene nomenclature,” *Nucleic Acids Res.*, vol. 34, pp. D140–144, Jan. 2006.

[120] M. R. Friedlander, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, “miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades,” *Nucleic Acids Res.*, vol. 40, pp. 37–52, Jan. 2012.

[121] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, *et al.*, “Interproscan 5: genome-scale protein function classification,” *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, 2014.

[122] S. Lien, B. F. Koop, S. R. Sandve, J. R. Miller, M. P. Kent, T. Nome, T. R. Hvidsten, J. S. Leong, D. R. Minkley, A. Zimin, F. Grammes, H. Grove, A. Gjuvsland, B. Walenz, R. A. Hermansen, K. von Schalburg, E. B. Rondeau, A. Di Genova, J. K. Samy, J. Olav Vik, M. D. Vigeland, L. Caler, U. Grimholt, S. Jentoft, D. I. Våge, P. de Jong, T. Moen, M. Baranski, Y. Palti, D. R. Smith, J. A. Yorke, A. J. Nederbragt, A. Tooming-Klunderud, K. S. Jakobsen, X. Jiang, D. Fan, Y. Hu, D. A. Liberles, R. Vidal, P. Iturra, S. J. Jones, I. Jonassen, A. Maass, S. W. Omholt, and W. S. Davidson, “The Atlantic salmon genome provides insights into rediploidization,” *Nature*, vol. 533, pp. 200–205, May 2016.

[123] W. Ding, X. Zhang, X. Zhao, W. Jing, Z. Cao, J. Li, Y. Huang, X. You, M. Wang, Q. Shi, and X. Bing, “A chromosome-level genome assembly of the mandarin fish (siniperca chuatsi),” *Frontiers in Genetics*, vol. 12, Jun. 2021.

[124] L. Guo, H. Yao, B. Shepherd, O. J. Sepulveda-Villet, D.-C. Zhang, and H.-P. Wang, “Development of a genomic resource and identification of nucleotide diversity of yellow perch by RAD sequencing,” *Frontiers in Genetics*, vol. 10, Oct. 2019.

[125] J. Ghurye, A. Rhie, B. P. Walenz, A. Schmitt, S. Selvaraj, M. Pop, A. M. Phillippy, and S. Koren, "Integrating hi-c links with assembly graphs for chromosome-scale assembly," *PLOS Computational Biology*, vol. 15, p. e1007273, Aug. 2019.

[126] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, and A. M. Phillippy, "The complete sequence of a human genome," May 2021.

[127] K. J. Martin and A. B. Pardee, "Identifying expressed genes," *Proceedings of the National Academy of Sciences*, vol. 97, pp. 3789–3791, Apr. 2000.

[128] S. Chandhini and V. J. R. Kumar, "Transcriptomics in aquaculture: current status and applications," *Reviews in Aquaculture*, vol. 11, pp. 1379–1397, Oct. 2018.

[129] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nat Biotechnol*, vol. 29, pp. 644–652, May 2011.

[130] E. Bushmanova, A. D., L. A., and A. D. Prjibelski, "rrnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data," *GigaScience*, vol. 8, pp. 1–13, Aug. 2019.

[131] S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea, "Transcriptome assembly from long-read RNA-seq alignments with StringTie2," *Genome Biol*, vol. 20, p. 278, Dec. 2019.

[132] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," *Nat Protoc*, vol. 11, pp. 1650–1667, Sep. 2016.

[133] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, pp. 3150–3152, Dec. 2012.

[134] K. Nakasugi, R. Crowhurst, J. Bally, and P. Waterhouse, "Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana," *PLoS One*, vol. 9, no. 3, p. e91776, 2014.

[135] D. Gilbert, "Gene-omes built from mRNA seq not genome DNA," *F1090Research 5:1695*, 2013.

[136] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, and P. Bork, "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses," *Nucleic Acids Research*, vol. 47, pp. D309–D314, Nov. 2018.

[137] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, pp. 3094–3100, May 2018.

[138] G. Pertea and M. Pertea, "GFF utilities: GffRead and GffCompare," *F1000Research*, vol. 9, p. 304, Apr. 2020.

[139] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Jan. 2013.

[140] R. Vera Alvarez, L. S. Pongor, L. Mariño-Ramírez, and D. Landsman, "TPMCalculator: one-step software to quantify mRNA abundance of genomic features," *Bioinformatics*, vol. 35, pp. 1960–1962, 06 2019.

[141] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian, "BMC BioinformaticsTiGER: a database for tissue-specific gene expression and regulation," *BMC Bioinformatics*, vol. 9, p. 271, Jun. 2008.

[142] M. D. Chikina, C. Huttenhower, C. T. Murphy, and O. G. Troyanskaya, "PLoS Comput BiolGlobal prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans," *PLoS Comput Biol*, vol. 5, p. e1000417, jun 2009.

[143] A. B. Bentz, E. K. Dossey, and K. A. Rosvall, “Gen Comp EndocrinolTissue-specific gene regulation corresponds with seasonal plasticity in female testosterone,” *Gen Comp Endocrinol*, vol. 270, pp. 26–34, Jan. 2019.

[144] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli, “Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification,” *Bioinformatics*, vol. 21, pp. 650–659, Mar. 2005.

[145] J. E. Mank, L. Hultin-Rosenberg, M. Zwahlen, and H. Ellegren, “Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression,” *Am Nat*, vol. 171, pp. 35–43, Jan. 2008.

[146] N. Kryuchkova-Mostacci and M. Robinson-Rechavi, “A benchmark of gene expression tissue-specificity metrics,” *Brief Bioinform*, vol. 18, pp. 205–214, Mar. 2017.

[147] B. Y. Liao and J. Zhang, “Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution,” *Mol Biol Evol*, vol. 23, pp. 1119–1128, Jun. 2006.

[148] D. J. McCarthy, Y. Chen, and G. K. Smyth, “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation,” *Nucleic Acids Res*, vol. 40, pp. 4288–4297, May 2012.

[149] A. Jain and G. Tuteja, “TissueEnrich: Tissue-specific gene enrichment analysis,” *Bioinformatics*, vol. 35, pp. 1966–1967, Jun. 2019.

[150] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pont?n, “Proteomics. Tissue-based map of the human proteome,” *Science*, vol. 347, p. 1260419, Jan. 2015.

[151] P. S. T. Russo, G. R. Ferreira, L. E. Cardozo, M. C. Bürger, R. Arias-Carrasco, S. R. Maruyama, T. D. C. Hirata, D. S. Lima, F. M. Passos, K. F. Fukutani, M. Lever, J. S. Silva, V. Maracaja-Coutinho, and H. I. Nakaya, “CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses,” *BMC Bioinformatics*, vol. 19, p. 56, Feb. 2018.

[152] X. Huang, X. G. Chen, and P. A. Armbruster, “Comparative performance of transcriptome assembly methods for non-model organisms,” *BMC Genomics*, vol. 17, p. 523, Jul. 2016.

[153] H. Hu, Y. R. Miao, L. H. Jia, Q. Y. Yu, Q. Zhang, and A. Y. Guo, "AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors," *Nucleic Acids Res*, vol. 47, pp. D33–D38, Jan. 2019.

[154] N. Cerveau and D. J. Jackson, "BMC BioinformaticsCombining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms," *BMC Bioinformatics*, vol. 17, p. 525, Dec. 2016.

[155] M. Sadat-Hosseini, M. R. Bakhtiarizadeh, N. Boroomand, M. Tohidfar, and K. Vahdati, "PLoS OneCombining independent de novo assemblies to optimize leaf transcriptome of Persian walnut," *PLoS One*, vol. 15, no. 4, p. e0232005, 2020.

[156] A. L. Ferraz, A. Ojeda, M. López-Béjar, L. T. Fernandes, A. Castelló, J. M. Folch, and M. Pérez-Enciso, "BMC GenomicsTranscriptome architecture across tissues in the pig," *BMC Genomics*, vol. 9, p. 173, Apr. 2008.

[157] A. R. Mohamed, H. King, B. Evans, A. Reverter, and J. W. Kijas, "Front GenetMulti-Tissue Transcriptome Profiling of North American Derived Atlantic Salmon," *Front Genet*, vol. 9, p. 369, 2018.

[158] X. Liao, L. Cheng, P. Xu, G. Lu, M. Wachholtz, X. Sun, and S. Chen, "Transcriptome analysis of crucian carp (carassius auratus), an important aquaculture and hypoxia-tolerant species," *PLoS ONE*, vol. 8, p. e62308, Apr. 2013.

[159] Y. Yu, J. C. Fuscoe, C. Zhao, C. Guo, M. Jia, T. Qing, D. I. Bannon, L. Lancashire, W. Bao, T. Du, H. Luo, Z. Su, W. D. Jones, C. L. Moland, W. S. Branham, F. Qian, B. Ning, Y. Li, H. Hong, L. Guo, N. Mei, T. Shi, K. Y. Wang, R. D. Wolfinger, Y. Nikolsky, S. J. Walker, P. Duerksen-Hughes, C. E. Mason, W. Tong, J. Thierry-Mieg, D. Thierry-Mieg, L. Shi, and C. Wang, "Nat CommunA rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages," *Nat Commun*, vol. 5, p. 3230, 2014.

[160] B. Li, T. Qing, J. Zhu, Z. Wen, Y. Yu, R. Fukumura, Y. Zheng, Y. Gondo, and L. Shi, "Sci RepA Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq," *Sci Rep*, vol. 7, p. 4200, Jun. 2017.

[161] J. M. Saju, M. S. Hossain, W. C. Liew, A. Pradhan, N. M. Thevasagayam, L. S. E. Tan, A. Anand, P. E. Olsson, and L. Orbán, "Cell RepHeat Shock Factor 5 Is Essential for Spermatogenesis in Zebrafish," *Cell Rep*, vol. 25, pp. 3252–3261, Dec. 2018.

[162] Y. Hu, B. Wang, and H. Du, "A review onsoxgenes in fish," *Reviews in Aquaculture*, Mar. 2021.

[163] K. K. To and L. E. Huang, "J Biol ChemSuppression of hypoxia-inducible factor 1alpha (HIF-1alpha) transcriptional activity by the HIF prolyl hydroxylase EGLN1," *J Biol Chem*, vol. 280, pp. 38102–38107, Nov. 2005.

[164] N. Pescador, Y. Cuevas, S. Naranjo, M. Alcaide, D. Villar, M. O. Landázuri, and L. Del Peso, "Biochem JIdentification of a functional hypoxia-responsive element that regulates the expression of the egl nine homologue 3 (egln3/phd3) gene," *Biochem J*, vol. 390, pp. 189–197, Aug. 2005.

[165] C. K. Mukhopadhyay, B. Mazumder, P. L. Fox, and P. L. Fox, "J Biol ChemRole of hypoxia-inducible factor-1 in transcriptional activation of ceruloplasmin by iron deficiency," *J Biol Chem*, vol. 275, pp. 21048–21054, jul 2000.

[166] M. Stange, R. D. H. Barrett, and A. P. Hendry, "The importance of genomic variation for biodiversity, ecosystems and people," *Nature Reviews Genetics*, vol. 22, pp. 89–105, Oct. 2020.

[167] L. de Los Ríos-Pérez, J. A. Nguinkal, M. Verleih, A. Rebl, R. M. Brunner, J. Klosa, N. Schäfer, M. Stüeken, T. Goldammer, and D. Wittenburg, "An ultra-high density SNP-based linkage map for enhancing the pikeperch (Sander lucioperca) genome assembly to chromosome-scale," *Sci Rep*, vol. 10, p. 22335, dec 2020.

[168] G. A. Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline," *Current Protocols in Bioinformatics*, vol. 43, Oct. 2013.

[169] T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, and Y. Erlich, "Genome-wide profiling of heritable and de novo STR variations," *Nature Methods*, vol. 14, pp. 590–592, Apr. 2017.

[170] S. S. Ho, A. E. Urban, and R. E. Mills, "Structural variation in the sequencing era," *Nature Reviews Genetics*, vol. 21, no. 3, pp. 171–189, 2020.

[171] A. C. Bertolotti, R. M. Layer, M. K. Gundappa, M. D. Gallagher, E. Pehlivanoglu, T. Nome, D. Robledo, M. P. Kent, L. L. Røsæg, M. M. Holen, T. D. Mulugeta, T. J. Ashton, K. Hindar, H. Sægrov, B. Florø-Larsen, J. Erkinaro, C. R. Primmer, L. Bernatchez, S. A. M. Martin, I. A. Johnston, S. R. Sandve, S. Lien, and D. J. Macqueen, "The structural variation landscape in 492 Atlantic salmon genomes," *Nat Commun*, vol. 11, p. 5176, Oct. 2020.

[172] S. Liu, G. Gao, R. M. Layer, G. H. Thorgaard, G. D. Wiens, T. D. Leeds, K. E. Martin, and Y. Palti, "Identification of High-Confidence Structural Variants in Domesticated Rainbow Trout Using Whole-Genome Sequencing," *Front Genet*, vol. 12, p. 639355, 2021.

[173] D. E. Larson, H. J. Abel, C. Chiang, A. Badve, I. Das, J. M. Eldred, R. M. Layer, and I. M. Hall, "svtools: population-scale analysis of structural variation," *Bioinformatics*, vol. 35, no. 22, pp. 4782–4787, 2019.

[174] C. Chiang, R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan, and I. M. Hall, "Speedseq: ultra-fast personal genome analysis and interpretation," *Nature methods*, vol. 12, no. 10, pp. 966–968, 2015.

[175] B. S. Pedersen and A. R. Quinlan, "Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls," *Gigascience*, vol. 8, no. 4, p. giz040, 2019.

[176] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," *The American Journal of Human Genetics*, vol. 81, no. 5, pp. 1084–1097, 2007.

[177] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, "A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3," *Fly*, vol. 6, no. 2, pp. 80–92, 2012.

[178] S. N. Maduna, A. Vivian-Smith, Ó. D. B. Jónsdóttir, A. K. D. Imsland, C. F. C. Klütsch, T. Nyman, H. G. Eiken, and S. B. Hagen, "Genome- and transcriptome-derived microsatellite loci in lumpfish cyclopterus lumpus: molecular tools for aquaculture, conservation and fisheries management," vol. 10, Jan. 2020.

[179] M. A. D. Dias, R. T. F. de Freitas, S. E. Arranz, G. V. Villanova, and A. W. S. Hilsdorf, "Evaluation of the genetic diversity of microsatellite markers among four strains ofOreochromis niloticus," vol. 47, pp. 345–353, Mar. 2016.

[180] J. F. Taylor, "Implementation and accuracy of genomic selection," *Aquaculture*, vol. 420, pp. S8–S14, 2014.

[181] H. Liu, Y. Jiang, S. Wang, P. Ninwichian, B. Somridhivej, P. Xu, J. Abernathy, H. Kucuktas, and Z. Liu, "Comparative analysis of catfish BAC end sequences with the zebrafish genome," *BMC Genomics*, vol. 10, p. 592, Dec. 2009.

[182] D. M. Emms and S. Kelly, "OrthoFinder: phylogenetic orthology inference for comparative genomics," vol. 20, Nov. 2019.

[183] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[184] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular biology and evolution*, vol. 26, no. 7, pp. 1641–1650, 2009.

[185] T. Britton, C. L. Anderson, D. Jacquet, S. Lundqvist, and K. Bremer, "Estimating divergence times in large phylogenetic trees," vol. 56, pp. 741–752, Oct. 2007.

[186] A. Sahm, M. Bens, M. Platzer, and K. Szafranski, "PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes," *Nucleic Acids Res*, vol. 45, p. e100, Jun. 2017.

[187] P. Xie, S. K. Yi, H. Yao, W. Chi, Y. Guo, X. F. Ma, and H. P. Wang, "Comparative transcriptome analysis reveals potential evolutionary differences in adaptation of temperature and body shape among four Percidae species," *PLoS One*, vol. 14, no. 5, p. e0215933, 2019.

[188] E. Axelsson, L. Hultin-Rosenberg, M. Brandström, M. Zwahlén, D. F. Clayton, and H. Ellegren, "Mol EcolNatural selection in avian protein-coding genes expressed in brain," *Mol Ecol*, vol. 17, pp. 3008–3017, Jun. 2008.

[189] R. Ekblom, L. French, J. Slate, and T. Burke, "Genome Biol EvolEvolutionary analysis and expression profiling of zebra finch immune genes," *Genome Biol Evol*, vol. 2, pp. 781–790, 2010.

[190] A. E. Haponski and C. A. Stepien, "Phylogenetic and biogeographical relationships of theSanderpikeperches (percidae: Perciformes): patterns across north america and eurasia," vol. 110, pp. 156–179, Jun. 2013.

[191] C. A. Stepien, J. Behrmann-Godel, and L. Bernatchez, "Evolutionary relationships, population genetics, and ecological and genomic adaptations of perch (perca)," *Biology of perch*, pp. 7–46, 2015.

[192] T. Gjedrem, N. Robinson, and M. Rye, "The importance of selective breeding in aquaculture to meet future demands for animal protein: A review," *Aquaculture*, vol. 350-353, pp. 117–129, Jun. 2012.

[193] S. M. Glasauer and S. C. Neuhauss, “Whole-genome duplication in teleost fishes and its evolutionary consequences,” *Mol Genet Genomics*, vol. 289, pp. 1045–1060, Dec. 2014.

[194] C. Palaiokostas, R. D. Houston, *et al.*, “Genome-wide approaches to understanding and improving complex traits in aquaculture species,” *Perspect. Agric. Vet. Sci. Nutr. Nat. Resour*, vol. 12, pp. 1–10, 2017.

[195] L. Orbán, X. Shen, N. Phua, and L. Varga, “Toward Genome-Based Selection in Asian Seabass: What Can We Learn From Other Food Fishes and Farm Animals?,” *Front Genet*, vol. 12, p. 506754, 2021.

[196] C. Garcia de Leaniz, I. A. Fleming, S. Einum, E. Verspoor, W. C. Jordan, S. Consuegra, N. Aubin-Horth, D. Lajus, B. H. Letcher, A. F. Youngson, J. H. Webb, L. A. Vøllestad, B. Villanueva, A. Ferguson, and T. P. Quinn, “A critical review of adaptive genetic variation in Atlantic salmon: implications for conservation,” *Biol Rev Camb Philos Soc*, vol. 82, pp. 173–211, May 2007.

[197] N. F. Thompson, E. C. Anderson, A. J. Clemento, M. A. Campbell, D. E. Pearse, J. W. Hearsey, A. P. Kinziger, and J. C. Garza, “A complex phenotype in salmon controlled by a simple change in migratory timing,” *Science*, vol. 370, pp. 609–613, 10 2020.

[198] I. COSCIA and S. MARIANI, “Phylogeography and population structure of european sea bass in the north-east atlantic,” *Biological Journal of the Linnean Society*, vol. 104, pp. 364–377, Aug. 2011.

[199] M. Salem, B. Paneru, R. Al-Tobasei, F. Abdouni, G. H. Thorgaard, C. E. Rexroad, and J. Yao, “Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout,” *PloS one*, vol. 10, no. 3, p. e0121778, 2015.

[200] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard, “Prediction of total genetic value using genome-wide dense marker maps,” *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[201] C. Lu, Y. Kuang, X. Zheng, C. Li, X. Sun, *et al.*, “Advances of molecular marker-assisted breeding for aquatic species.,” *Journal of Fisheries of China*, vol. 43, no. 1, pp. 36–53, 2019.

[202] A. A. Nemudryi, K. R. Valetdinova, S. P. Medvedev, and S. M. Zakian, “TALEN and CRISPR/Cas Genome Editing Systems: Tools of Discovery,” *Acta Naturae*, vol. 6, pp. 19–40, Jul. 2014.

[203] R. B. Edvardsen, S. Leininger, L. Kleppe, K. O. Skaftnesmo, and A. Wargelius, "Targeted mutagenesis in atlantic salmon (salmo salar l.) using the crispr/cas9 system induces complete knockout individuals in the f0 generation," *PloS one*, vol. 9, no. 9, p. e108622, 2014.

[204] M. Li, H. Yang, J. Zhao, L. Fang, H. Shi, M. Li, Y. Sun, X. Zhang, D. Jiang, L. Zhou, *et al.*, "Efficient and heritable gene targeting in tilapia by crispr/cas9," *Genetics*, vol. 197, no. 2, pp. 591–599, 2014.

[205] Z. Zhong, P. Niu, M. Wang, G. Huang, S. Xu, Y. Sun, X. Xu, Y. Hou, X. Sun, Y. Yan, *et al.*, "Targeted disruption of sp7 and myostatin with crispr-cas9 results in severe bone defects and more muscular cells in common carp," *Scientific Reports*, vol. 6, no. 1, pp. 1–14, 2016.

[206] M. Li, R. Feng, H. Ma, R. Dong, Z. Liu, W. Jiang, W. Tao, and D. Wang, "Retinoic acid triggers meiosis initiation via stra8-dependent pathway in southern catfish, silurus meridionalis," *General and comparative endocrinology*, vol. 232, pp. 191–198, 2016.

Pikeperch Genes — Functional Annotation

COG Categories Distribution

Information Storage and Processing:	
Transcription (K):	1950
RNA processing and modification (A):	643
Translation, ribosomal structure and biogenesis (J):	477
Replication, recombination and repair (L):	409
Chromatin structure and dynamics (B):	263
Total:	**3742 / 14.62%**
Cellular Processses and Signaling:	
Signal transduction mechanisms (T):	5331
Posttranslational modification, protein turnover, chaperones (O):	1591
Intracellular trafficking, secretion, and vesicular transport (U):	819
Cytoskeleton (Z):	672
Extracellular structures (W):	310
Cell cycle control, cell division, chromosome partitioning (D):	233
Defense mechanisms (V):	190
Cell wall/membrane/envelope biogenesis (M):	102
Cell motility (N):	19
Nuclear structure (Y):	10
Total:	**9277 / 36.26%**
Metabolism:	
Inorganic ion transport and metabolism (P):	573
Carbohydrate transport and metabolism (G):	554
Lipid transport and metabolism (I):	503
Amino acid transport and metabolism (E):	427
Energy production and conversion (C):	417
Secondary metabolites biosynthesis, transport and catabolism (Q):	295
Nucleotide transport and metabolism (F):	222
Coenzyme transport and metabolism (H):	104
Total:	**3095 / 12.1%**
Poorly Characterized:	
Function unknown (S):	7926
General function prediction only (R):	0
Total:	**7926 / 30.98%**

List of Publications

Peer-reviewed publications and contributions

- **Nguinkal, J.A.**; Brunner, R.M.; Verleih, M.; Rebl, A.; de Los Ríos-Pérez, L.; Schäfer, N.;Hadlich, F.; Stüeken, M.; Wittenburg, D.; Goldammer, T. **The First Highly Contiguous Genome Assembly of Pikeperch (Sander lucioperca), an Emerging Aquaculture Species in Europe**. Genes (Basel) **2019**,10

  This article reports the first released version of the pikeperch genome data, including genome assembly, complete annotation and genome sequencing data. Besides the development of the assembly and annotation workflows, my contribution also include the bioinformatics and statistical analysis of the experimental data. In addition, I have validated, structured and submitted the data in public repositories. Writing the entire manuscript is moreover part of my contribution. This paper was published in ***Genes. JIF: 4.096***

- De Los Ríos-Pérez, L.; **Nguinkal, J.A.**; Verleih, M.; Rebl, A.; Brunner, R.M.; Klosa, J.; Schäfer, N.; Stüeken, M.; Goldammer, T.; Wittenburg, D. **An ultra-high density SNP-based linkage map for enhancing the pikeperch (*Sander lucioperca*) genome assembly to chromosome-scale**. Sci Rep **2020**, 10, 22335.

  My contribution to this work consisted in the bioinformatic analysis including updating the first draft genome to chromosome-level, preforming structural and functional annotation, as well data assessment, visualization and dissemination in public repositories. In addition, I substantially contributed in drafting the manuscript whose results were published in ***Scientific Reports. JIF: 4.379***

- Nadine Schäfer, Yagmur Kaya, Henrike Rebl, Marcus Stüeken, Alexander Rebl, **Julien A. Nguinkal**, George P. Franz, Ronald M. Brunner, Tom Goldammer, Bianka Grunow & Marieke Verleih. **Insights into early ontogenesis: characterization of stress and development key genes of pikeperch (Sander lucioperca) *in vivo* and *in vitro*.** Fish Physiology and Biochemistry, volume 47, pages 515–532 **(2021)**

  This work describes the results of the study to (i) characterize and evaluate key genes for development as well as stress response in the early ontogenesis of pikeperch, (ii) to initiate the first approach to generating a cell model from pikeperch derived out of whole embryos, and (iii) to analyze the suitability of an in vitro model for studying developmental processes in pikeperch. I contributed in the data acquisition and analysis. The findings and data were published in ***Fish Physiol. and Biochem. JIF: 2.794***

- **Nguinkal, J.A.**; Verleih, M.; de los Ríos-Pérez, L.; Brunner, R.M.; Sahm, A.; Bej, S.; Rebl, A.; Goldammer, T. **Comprehensive Characterization of Multitissue Expression Landscape, Co-Expression Networks and Positive Selection in Pikeperch.** Cells **2021**, 10, 2289.

  This publication reports multissue-transcriptome data of pikeperch along with comprehensive characterization of gene co-expression network, tissue-specificity of gene expression and positive selection signatures in Percidae. I contributed in designing the experiments and data acquisition. I performed the bioinformtics analyses including data processing, visualization, and validation. I additionally wrote the manuscript which was published in ***Cells. JIF: 6.600***

Scientific Posters and Oral Presentations

- 07/2018: **Towards the draft genome sequence and annotation of *Sander lucioperca* (pikeperch)** – 3th International Congress on the Biology of Fish, July 15th-19th, 2018, University of Calgary, Alberta, Canada, P80.
  (**Poster presentation**)
- 10/2019: **Deciphering and analysis of the biodiversity of the pikeperch genome** – Aquaculture Europe 2019, Aquaculture, Vol. 2019, Berlin, Germany.
  (**Oral presentation**)
- 11/2019: **De novo Genome Assembly – A solved Problem?** – Day of doctoral students, Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany. (**Oral presentation**)
- 07/2021: **Pikeperch genome data - basis for smart farming in aquaculture** – ISAG 2021, P230 (virtual) (**Poster presentation**)

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe. Ich versichere, dass ich nur die angegebenen Quellen und Hilfsmittel verwendet habe und die benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Rostock, 20.12.2021		Julien A. Nguinkal
Ort, Datum	Unterschrift	