TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

THE MOLECULAR MECHANISMS OF RAPID THERMAL ADAPTATION IN EUROPEAN GRAYLING

Tiina Sävilammi

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AII OSA – TOM. 406 | BIOLOGICA – GEOGRAPHICA – GEOLOGICA | TURKU 2024

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

THE MOLECULAR MECHANISMS OF RAPID THERMAL ADAPTATION IN EUROPEAN GRAYLING

Tiina Sävilammi

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AII OSA – TOM. 406 | BIOLOGICA –GEOGRAPHICA – GEOLOGICA | TURKU 2024

University of Turku

Faculty of Science
Department of Biology
Biology
Doctoral Programme in Biology, Geography and Geology

Supervised by

Professor, Craig Primmer
University of Helsinki
Helsinki, Finland

Assistant Professor, Spiros Papakostas
International Hellenic University
Thessaloniki, Greece

Reviewed by

Associate Professor, Adam Miller
School of Life and Environmental Sciences
Deakin University
Warrnambool, Australia

Docent, Lumi Viljakainen
University of Oulu
Oulu, Finland

Opponent

Associate Professor, Morten Limborg
Globe Institute
University of Copenhagen
Copenhagen, Denmark

ISBN 978-951-29-9593-6 (PRINT)
ISBN 978-951-29-9594-3 (PDF)
ISSN 0082-6979 (Print)
ISSN 2343-3183 (Online)
Painosalama, Turku, Finland 2024

“Some day, it will make sense.”

A quote written in a saddle pad of the best pony ever.

UNIVERSITY OF TURKU
Faculty of Science
Department of Biology
Biology
TIINA SÄVILAMMI: Molecular mechanisms of rapid thermal adaptation in European grayling
Doctoral Dissertation, 136 pp.
Doctoral Programme in Biology, Geography and Geology
February 2024

ABSTRACT

Adaptation is a key evolutionary process which may be important to maintain population viability during rapid environmental changes. The European grayling (*Thymallus thymallus*) is a freshwater salmonid that effectively colonises new environments despite low genetic diversity. Several new grayling populations were recently established from a single ancestral gene pool in a Norwegian water system. Phenotypic responses to colder or warmer temperature (thermal origin) during spawning and early development have been reported but the molecular mechanisms underlying these changes are often poorly understood.

The aim of this thesis was to investigate the molecular mechanisms of thermal adaptation in this Norwegian grayling system using multi-omics approaches. A prior common-garden experiment, where embryos from several populations from warmer and colder thermal origins were reared in higher and lower tempearatures, was utilised. (I) RNA-sequencing study of the common-garden populations revealed that the gene expression response was mostly plastic (to rearing temperature) but an adaptive signal (to thermal origin) existed in specific gene modules. (II) A hybrid synteny-guided chromosome-level genome assembly confirmed that graylings have a unique chromosomal setup among salmonids, with the karyotypes having evolved through pericentric inversions and one chromosomal fission rather than multiple fusions, which are typical to other salmonids. It also revealed a similarly distinctive transposable element content in comparison to Atlantic salmon (*Salmo salar*). (III) Whole-genome cytosine sequencing of the common-garden populations uncovered common methylation patterns among populations from similar thermal origins, and almost absent plasticity. (IV) A whole-genome population genomics study revealed candidate SNP loci for thermal adaptation. Biological processes associated with thermal origin overlapped between the molecular levels.

The results describe how rapid thermal adaptation may be manifested by altering standing genetic variation and epigenetic variation. They highlight the value of applying “omic” methodologies when studying adaptive biodiversity e.g. for fisheries management and conservation applications.

KEYWORDS: gene expression, epigenetic variation, genetic variation, methylation, adaptation, ecological evolution, salmonids, grayling, rapid adaptation

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Biologian laitos
Biologia
Tiina Sävilammi: Harjuksen nopean lämpöadaptaation molekyylimekanismit
Väitöskirja, 136 s.
Biologian, maantieteen ja geologian tohtoriohjelma
Helmikuu 2024

TIIVISTELMÄ

Adaptaatio on evoluution avainprosessi, joka voi olla tärkeä populaation elinkelpoisuuden säilymiseksi nopeiden ympäristönmuutosten aikana. Harjus (*Thymallus thymallus*) on makean veden lohikala joka kolonisoi tehokkaasti uusia ympäristöjä matalasta geneettisestä diversiteetistä huolimatta. Useita uusia harjuspopulaatioita syntyi nopeasti yhdestä geenipoolista norjalaisessa vesistöjärjestelmässä. Fenotyyppisiä vasteita kylmempään tai lämpimämpään lämpötilaan (alkuperäislämpötila) kudun ja aikaisen kehityksen aikana on raportoitu, mutta niiden taustalla olevat molekyylimekanismit ovat huonosti ymmärrettyjä.

Tämän väitöskirjan aiheena oli tutkia lämpöadaptaation molekyylimekanismeja tässä norjalaisessa systeemissä käyttämällä montaa eri “omics”-lähestymistapaa. Tutkimuksessa käytettiin hyväksi aiempaa yhteispuutarhakoetta, jossa kasvatettiin useista suhteellisen lämpimään ja kylmään lämpötilaan adaptoituneista populaatioista peräisin olevia alkioita lämpimämmissä ja matalammissa lämpötiloissa. (I) Yhteispuutarhapopulaatioiden RNA-sekvensointi paljasti, että geenieksпressio oli pääosin plastista (vastasi kasvatuslämpötilaan) mutta adaptiivinen signaali (vaste alkuperäislämpötilaan) löytyi tietyistä geenimoduleista. (II) Harjuksen genomisekvenssi selvitettiin kromosomitasolle synteniaa hyödyntävän hybridimenetelmän avulla. Se vahvisti, että harjuksen uniikki kromosomisto on kehittynyt kromosomien perisentristen inversioiden ja yhden fission, eikä lohikaloille tyypillisten kromosomifuusioiden, seurauksena. Myös atlantinlohesta (*Salmo salar*) poikkeava liikkuvien elementtien kokoonpano havaittiin. (III) Yhteispuutarhakoepopulaatioiden genomien sytosiinisekvensointi paljasti yhteneviä epigeneettisiä markkeriyhdistelmiä samankaltaisista alkuperäislämpötiloista peräisin olevissa populaatioissa, ja lähes puuttuvan plastisuuden. (IV) Koko genomin populaatiogenomiikkakokeessa havaittiin lämpötila-adaptaation kandidaattilokuksia. Alkuperäislämpötilaan liittyvät biologiset prosessit olivat osittain samoja eri molekyylitasojen välillä.

Tulokset selventävät miten nopea lämpöadaptaatio voi seurata olemassaolevan geneettisen muuntelun ja epigeneettisen muuntelun kautta, ja korostavat “omics”-lähestymistapojen arvoa adaptiivisen biodiversiteetin tutkimisessa mm. kalakantojen ja muiden lajien suojelemiseksi.

ASIASANAT: geenieksпressio, epigeneettinen muuntelu, geneettinen muuntelu, metylaatio, adaptaatio, ekologinen evoluutio, lohikalat, harjus, nopea evoluutio

Table of Contents

List of Original Publications

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

I Mäkinen, H. M., Sävilammi, T., Papakostas, S., Leder, E., Vøllestad, L. A. and Primmer, C. R. Modularity Facilitates Flexible Tuning of Plastic and Evolutionary Gene Expression Responses during Early Divergence. *Genome Biology and Evolution*, 2017; 10: 77-93.

II Sävilammi, T., Primmer, C. R., Varadharajan, S., Guyomard, R., Guiguen, Y., Sandve, S. R., Vøllestad, L. A., Papakostas, S. and Lien, S. The Chromosome-Level Genome Assembly of European Grayling Reveals Aspects of a Unique Genome Evolution Process Within Salmonids. *G3 Genes|Genomes|Genetics*, 2019; 9: 1283–1294.

III Sävilammi, T., Papakostas, S., Leder, E. H., Vøllestad, L. A., Debes, P. V. and Primmer, C. R. Cytosine methylation patterns suggest a role of methylation in plastic and adaptive responses to temperature in European grayling (*Thymallus thymallus*) populations. *Epigenetics*, 2021; 16: 271-288.

IV Sävilammi, T., Papakostas, S., Leder, E. H., Vøllestad, L. A. and Primmer, C. R. Signals of polygenic selection in a salmonid meta-population undergoing contemporary thermal adaptation. *Manuscript*.

1 Introduction

To survive in changing environments, populations need to adapt to new environmental conditions. A gene flow barrier between populations may eventually lead to an emergence of new species over a long period of time. Evolution is also well documented on ecological timescales [1] and can occur, for example, when environmental factors alter the fitness of individuals in a population, leading to the selection of a subset of individuals and, eventually, phenotypic changes at the population level over generations. Although environmental changes are often slow, allowing for gradual evolutionary changes, they can also be abrupt and the corresponding magnitude of phenotypic change equally high. Under favourable conditions, some traits, such as antibiotic resistance in bacteria, may evolve rapidly. In bacteria, antibiotic resistance may persist in response to recurrent use of antibiotics, and the presence of bacteria with high adaptation potential due to intrinsic properties such as horizontal gene transfer or other mechanisms that increase plasticity [2]. However, the limits of the pace of evolution in natural populations under different conditions, such as high gene flow or low genetic variation, is often unknown. Phenotypic responses have been reported in response to environmental changes in complex traits over surprisingly short time scales (or generations of individuals). For example, wild guppies (*Poecilia reticulata*) from high predation habitats evolved towards phenotypes resembling those observed in low-predation habitats when transplanted to naturally low predation environment in just eight generations [3]. Convergent phenotypic changes in ecological and morphological traits evolved in just 2-10 generations in wild European whitefish (*Coregonus lavaretus*) populations in response to transplantation to new habitats with more diverse diet [4]. Although some examples of rapid phenotypic evolution are already well understood, the molecular basis of rapid adaptation under different circumstances is less well understood.

Besides evolution, the other key mechanism that may lead to adaptive phenotypes is phenotypic plasticity [5]. Plasticity is a genetically pre-programmed phenotypic response of an organism. One genotype may produce multiple phenotypic outcomes depending on the environmental conditions. Plasticity is often an important response to improve fitness both during the developmental period of an

individual, and through its lifetime in the continuously changing natural environment. Plasticity may continue to be a key mechanism for a population to cope with environmental change during the initial generations adaptating to a new/changed environment [6]. However, the limits of plasticity are thought to be genetically determined, thus restricting its fitness benefits. Evolution of plasticity may be common in circumstances such as the initial stages of colonisation of new environments [6]. In the long term, ecological evolution may also lead to assimilation of plasticity; replacement of a plastic response with a genetically fixed response, which has evolved through ecological evolution [5]. While plasticity may kick-start the adaptation process by allowing individuals to survive in an otherwise unfavourable environment, it can also hinder selection by altering the distribution of phenotypes in a population, thus hiding non-beneficial genetic variation from natural selection [5].

1.1 Adaptation through genetic mechanisms

Some phenotypic traits are monogenic, which means that a phenotype is entirely determined by a single-locus genotype. For example, color variants of many animals such as the horse (*Equus caballus*) are often monogenic, summarized in [8]. Other traits may be controlled by 'major-effect loci'. An example of a major-effect locus is a single nucleotide polymorphism (SNP) in a genomic region encoding gene *vgll3* in Atlantic salmon (*Salmo salar*). The locus controls almost 40 % of the variation in maturation age [7]. Under certain conditions, adaptive alleles of large-effect loci may become highly abundant or even fixed in a population. According to population genetic theory, so-called selective sweeps may alter also the surrounding neutral mutations as a beneficial mutation increases to fixation [8]. However, the fixation of even such large-effect loci may be hindered e.g. by balancing selection when multiple alleles are being selected simultaneously [8]. In the case of the *vgll3* locus of the Atlantic salmon, the selection of alleles is sex-dependent and variation persists [7].

Many adaptive traits are polygenic, meaning that they are controlled by many genomic loci [8]. In addition to the major effect locus explaining a large part of the variation in the maturation timing of Atlantic salmon, another study found that a total of 116 SNPs, which associated to various genes and were dispersed over 22 of the 26 Atlantic salmon chromosomes, contributed to maturation timing. Together with the major-effect *vgll3* locus, they explained up to 61 % of the phenotypic variation [9]. Adaptive alleles of polygenic traits may have sweep-like footprints surrounding the large-effect loci [8]. However, when adaptation is polygenic, equally high fitness may result from multiple different combinations of the causal loci [8]. The extreme interpretation of polygenic inheritance is the omnigenic theory, which states that

almost all genes contribute to each polygenic trait through either direct effects to the phenotype or indirect effects through genes that they interact with [8], [10]. Indeed, genes or gene products can be arranged to tightly interconnected 'small world' networks, where any two genes are connectable through only a few intermediate interactions, resulting in a very high number of causal genes which may have very small individual effect sizes. Thus, interpreting gene sets through co-regulated gene modules may be a useful approach when studying polygenic adaptation [11].

Adaptive evolution can also take place through larger-scale rearrangements of chromosomes, which can ultimately shape chromosome identities and karyotypes [12]. Sometimes, rearrangements as large as whole-genome duplications may lead to accelerated evolution, which may be followed by multiple speciation events [13]. This has been observed in radiation events of many species groups, such as the rise of >20,000 species of teleost fish and >350,000 species of flowering plants [13]. A genome duplication is a single mutation event that alters the whole genome [14]. Although what happens at the initial stages of genome duplication and rediploidisation remain enigmatic because successful genome duplication events are often ancient, one theory suggests that the genome enters to a tetraploid state where the duplicated chromosome pairs are randomly selected for segregation during meiosis [12]. The state is imbalanced because the chromosomes do not have stabile 1:1 pairs during meiosis [12]. This imbalance may be unfavourable and selected against, which is a process that is called rediploidisation [12]. During rediploidisation, natural selection may favor mutations that re-identify pairs from the duplicated chromosome copies during meiosis by increasing dissimilarity between them. Thus, the polyploid state starts to revert back towards diploidy [12], [14]–[16].

Salmonids are a textbook example of such a genome tetraploidisation event. A whole-genome duplication of the ancestral salmonid took place 80-100 million years ago and doubled the haploid (1N) number of unique chromosomes from 25 to 50 small chromosomes [17], [18]. The duplicated ancestral chromosomal karyotypes likely persisted for 40-50 million years (**Figure 1**) [18]. After that, the three salmonid subfamilies diverged almost simultaneously [18]. Lineage-specific genomic rediploidisation resulted in the rise of species barriers by genetic incompatibility, and distinct karyotype sets evolved in the lineages [19], followed by adaptive radiation of numerous new species [13], [20]. The lineage diversification co-occurred with a period of environmental cooling which may have also contributed [18]. The number of new salmonid species was particularly high in the lineages where anadromy evolved (Coregoninae and Salmoninae) [18]. The karyotypes of most salmonids evolved through sequential Robertsonian fusions of chromosomes (**Figure 2**) [21]. An extreme example of this is the Atlantic salmon genome, where the haploid chromosome count has reduced from 50 to 29 [21]. Whereas chromosomal fusions are the trend in the two more abundant salmonid subfamilies (Salmoninae and

Coregoninae), karyotype evolution has been distinct in the small subfamily of graylings (Thymallinae). Early karyotypic studies (summarised in [21]) suggested, that the grayling genomes evolved through pericentric inversions (**Figure 2**) and chromosomal fissions [21].

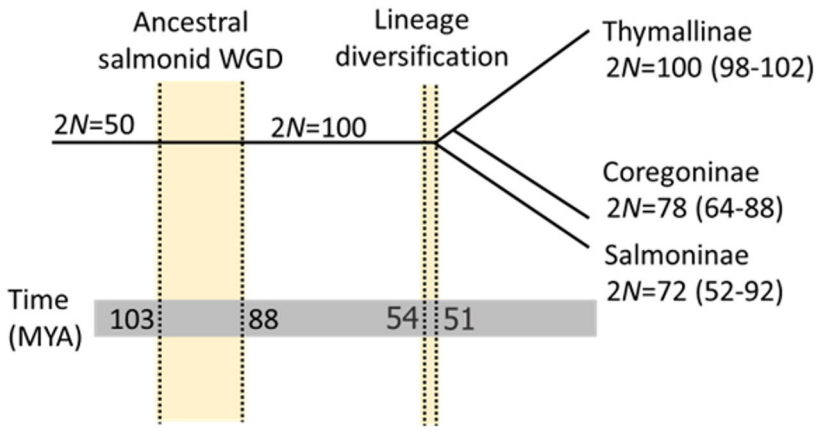


Figure 1. The salmonid-specific whole-genome duplication event (WGD) and the following nearly simultaneous lineage diversification (note that the figure is schematic and not scaled by time). The 95% confidence intervals of those events suggested by [18] are depicted with yellow bars. The grey horizontal bar represents time millions of years ago (MYA). The 2*N* chromosome number of the ancestral and derived lineages are shown in the relevant positions in the lineage. In the derived lineages, mean and range of observations from [21] are shown.

A possible explanation for the selective advantage of the grayling-type karyotypes under some circumstances is explained by the hypothesis formulated by Qumsiyeh [22], who suggested that small chromosomes with high recombination rates may be favoured in species in variable habitats where increased genetic variability and decreased proportion of fixed alleles is beneficial. Freshwater environments are typically more fragmented and unpredictable than large and constant sea environments [21]. Supporting the hypothesis, freshwater salmonids have been reported with relatively smaller chromosome sizes and higher numbers of chromosomes than the anadromous salmonids (**Figure 1**). In the sea, lower amount of genetic variation and larger probability of allele fixation may have been advantageous [21]. Although the evolution of anadromous life history in the other salmonids has been suggested as a key novelty behind the great species diversification of the salmonids, staying in the freshwater habitat may have led to a contrasting selective advantage favouring conserved, grayling-type karyotypes.

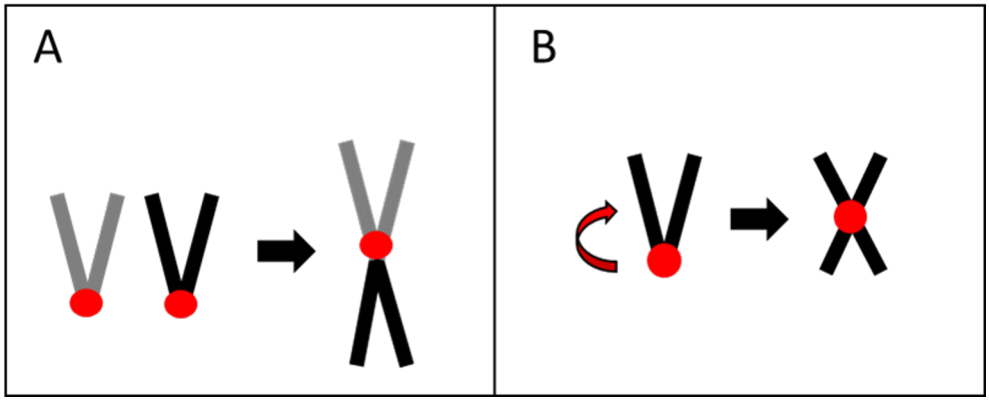


Figure 2 Chromosomal rearrangement types, which have been reported in salmonids. In a Robertsonian fusion (A), the centromeres (depicted with red) of two telocentric chromosomes (depicted with grey and black colour) fuse on the centromeres and the number of chromosomes decreases due to the formation of a large metacentric chromosome, while the total number of chromosome arms remains the same. In a pericentric inversion, the centromere-containing portion (red) of a single telocentric chromosome (black) is inverted (red arrow), and the centromere location is shifted from the chromosome end to the middle (B). In the resulting small metacentric chromosome, the chromosome identity is maintained while the number of chromosome arms increases. The chromosomes are depicted at meiosis, where they have replicated and are still joined at the centromere region.

1.2 Adaptation through non-genetic mechanisms

Besides plasticity and genetic mechanisms, also non-genetic inheritance mechanisms may contribute to the adaptation process. So-called epigenetic markers are non-sequence modifications in the DNA that have the potential to be inherited. They usually function by altering gene expression without being encoded in the nucleotide sequence. Epigenetic markers include histone modifications, small RNAs, and methylation of nucleotides in the DNA [23]. Through the regulation of gene expression, they contribute to important processes, which are listed in [24]. The processes include chromosomal silencing for dosage compensation, and genomic imprinting of genes from one parent. During differentiation, epigenetic pattern formation determines cell phenotypes, and modifications continue to accumulate during the whole lifespan of an organism in response to environmental changes. Epigenetic mechanisms also serve as a memory of the past gene-regulatory events. They may provide a more dynamic response to a changing environment than DNA evolution, sometimes affecting across several generations [24], [25]. The embryonic period may be a hotspot for methylation differentiation as the methylation levels are being reprogrammed. In fish, the methylation patterns are first erased in the germ line cells and then regenerated during early embryonic development [26], [27].

Cytosine methylation in the CpG loci (a cytosine followed by a phosphate bond and a guanine; a C-G sequence in the DNA) is one of the most important and most broadly studied epigenetic mechanism for many reasons such as the well-established high-throughput sequencing methodology, and the widely reported responsiveness to environmental factors via various biological processes [28], [29]. CpG cytosine methylation in the upstream regulatory region (particularly when a CpG island is present) may silence the expression of a gene [30]. In fish, altered CpG methylation patterns have been reported in response to a wide range of environmental factors and life history traits. Both nucleotide evolution and altered CpG methylation were correlated with morphology and the diversity of the diet in newly established natural populations of European whitefish (*Coregonus lavaretus*) [4]. CpG methylation changes have also been associated with smoltification in the rainbow trout (*onchorynchus mykiss*) [31]. In the mangrove killifish (*Kryptolebias marmoratus*), unstimulating rearing environment induced methylation changes which persisted over generations [32]. Even subtle (2-4°C) temperature changes in the rearing environment have been reported to induce CpG methylation -related changes particularly at the early life stages in fish. Altered methylation patterns were reported for larval, but not juvenile, European sea bass (*Dicentrarchus labrax*) [33], and differential expression levels of the methylation machinery were reported for embryonic Atlantic salmon in response to altered environmental temperature [34].

1.3 European grayling (*Thymallus thymallus*)

European grayling (*Thymallus thymallus*) (referred to as 'grayling' from here on) is a relatively long-lived species. The species is spring-spawning, inhabits various types of river and lake environments and preferentially migrates to stream outlets to spawn [35]. In many European populations, spawning occurs early in the spring, typically during March-May. In freshwater fish species of temperate areas, the length of the first growth season is critically important as large body size and high energy reserves predict overwinter survival [36]. During spawning, eggs are deposited a few centimetres below the gravel surface, where they develop until hatching, after which the yolk-sac soon disappears. Already at the swim-up stage, larvae leave the hatching site [37]. Early growth depends on the environmental temperature sum over the developmental and larval period [38]–[40]. The larvae move downstream to the lake very early in development, during the first summer. Thus, they experience potentially different thermal environments in the river only during early life-history stages. The juvenile fish then stay in the lake until maturation [37]. Maturation timing varies between habitats but is approximately six years [37]. Mature individuals are iteroparous and have been reported with high homing propensity [37].

An early genetic study using microsatellites showed that after the latest glaciation period, grayling re-populated Northern Europe from south to north from two distinct Pleistocene refugia, leading to eastern lineage (inhabiting Estonia, Finland and north-west Russia) and western lineage (inhabiting Norway and Sweden) [41]. The study suggested substantial genetic differentiation despite short geographic distances between populations, and low genetic diversity within populations. Thus, the initial hypothesis was that grayling may have limited migration potential and adaptation capacity.

Despite the initial expectations of challenges in adaptation due to low genetic diversity, grayling populations of Europe have been less negatively affected by human activity than their American counterparts, Arctic grayling (*Thymallus arcticus*) due to lower selectivity of the spawning sites [35]. Regardless, the conservation status of both major grayling species are stated as 'least concern' [42], [43]. However, although grayling population status is listed as 'good' in the northern European populations, populations in central and southern Europe are listed as being in 'poor' or 'bad' condition [42].

1.4 The study system

Koskinen *et al.* [44] and many others have reported that grayling has a remarkable capacity to colonise and adapt to novel environments. The focal study system of my thesis work is a fascinating example of such adaptation potential whereby multiple grayling populations in a Norwegian water system have apparently experienced rapid and ongoing thermal adaptation. The populations have adapted to significantly lower water temperatures during spawning and the first growth season than more southern European populations. The spawning time is delayed until June and decreased ability to develop in temperatures above 12°C has been reported, in contrast to the more southern populations [45]. In the system, several new populations have been established in a relatively short time span of about 30 generations from a single ancestral source [37]. Microsatellite markers have revealed that the populations typically have low effective population sizes and share a history of population bottlenecks [40], [44]. Many of the subpopulations also experience strong gene flow due to sharing a common habitat outside the breeding season [40]. In this kind of situation, random drift may dominate over directional selection. However, strong evidence supports natural selection on several traits, suggesting that selection has contributed to shaping the populations of the system [44]. The study system is located within and around Lake Lesjaskogsvatnet in central Norway and new populations have been established due to both human-induced and natural colonization events [37]. Both sympatric subpopulations, and geographically distinct populations now inhabit several lakes

and streams of the system. Although the populations are geographically closely located, a large difference has been reported in the water temperature sums between the locations in June—August, when spawning and early development of the larvae take place [46]. This is due to various topographical reasons, the size of the tributary, and the snow melt timing of small glaciers [45]. Thus, the (sub)populations may be classified as relatively warmer or colder-origin based on their home river temperature.

The colonisation history of the grayling populations in the system is unusually well covered by historical records. Shortly, the ancestral population inhabits River Gudbrandsdalslågen, which is in close proximity to Lesjaskogsvatnet. The ancestral population is of warmer thermal origin. Some individuals from the ancestral population colonised the south basin of Lake Lesjaskogsvatnet in 1880's following a construction of a temporary canal, which was later closed again [37]. Grayling dispersed through Lesjaskogsvatnet and established several spawning stocks in both the south and north basins of the lake, which have numerous river outlets where both warmer- and colder-thermal origin subpopulations originated. Subpopulation structure has been observed despite strong gene flow between the sympatric subpopulations using microsatellites [40], [47].

Later, some individuals were relocated from Lesjaskogsvatnet to a small nearby Lake Hårrtjønn (1910; warmer-origin). From Hårrtjønn, grayling dispersed naturally through a stream to Lake Aursjøen (1920's; colder-origin river outlet sampled). Finally, graylings colonised Lake Osbumagasinet (1954, warmer-origin) via a temporary tunnel from Lake Aursjøen. Detailed information about the dispersal events, years, generations, and thermal origins are given in **Table 1** and in [37].

The spawning time of the colder-origin populations is delayed up to four weeks in comparison to warmer-origin populations [48]. Several differences between the (sub)populations have been reported to exceed the expected neutral evolution rate, suggesting adaptive evolution [44], [46], [49]. Interestingly, some of the differences are related to the thermal origin. Egg size is larger in the warmer-origin than in the colder-origin populations [48]. Embryonic survival is the highest in a rearing temperature which resembles the natal conditions [44]. Other traits related to the thermal origin include muscle mass development, yolk conversion rate and growth rate at different temperatures [38], [44], [45]. Assuming a rough estimate of six years for the generation time [37], the first dispersal event in the study system took place only ca. 30 generations prior to sampling of individuals in the 2010's for the common garden experiment where samples for study I and III originated.

1.5 The aims of this thesis

Although phenotypic differentiation has been described between the (sub)populations of the study system, and microsatellite studies have suggested some level of genetic differentiation, molecular evidence of thermal adaptation has been limited due to the low number of markers which were used in those microsatellite-based studies (see previous section). The aim of this thesis was to utilise grayling tissue samples which had been collected during several years of intense field work and experiments, and to apply multi-omics approaches to increase understanding of the molecular mechanisms behind the thermal adaptation of grayling at the whole-genome level. We aimed to study gene expression, epigenetic and nucleotide variation to gain broader understanding of the processes. As low levels of genetic variation had been reported among the populations, we sought to answer the question of whether the observed phenotypic differentiation was mostly under plastic (suggesting limited further adaptation capacity), epigenetic, or genetic, regulation (suggesting increased further adaptation capacity).

By studying the changes at the various molecular levels, we sought to improve knowledge of the relevant molecular patterns in rapid adaptation processes in natural populations. By doing this, we aimed to increase the ability to estimate the pace at which species can overcome new environmental challenges, and the mechanisms that allow them to achieve this. The questions are timely due to anthropogenic environmental changes, such as global climate warming, which has been predicted to raise the global temperature significantly in just a few decades and cause increased unpredictability of the weather and other environmental conditions, while the populations simultaneously encounter increasing stress by the fragmentation of habitats due to human action [50].

In study I, gene expression levels were studied using a transcribed RNA-sequencing (RNAseq) approach. Both adaptive and plastic components shaping the expression levels were studied. In study II, prior studies on grayling karyotype evolution and genome sequence were complemented by assembling the first chromosomal level version of the grayling genome sequence. Further, the aim was to study the genomic rearrangements that have led to the distinctive genome evolution process in the grayling subfamily among salmonids. Importantly, the grayling genome sequence was a prerequisite for studies III and IV, in which molecular variation in the genomic DNA was studied. In study III, the patterns of epigenomic variation and plastic and evolutionary components explaining the variation were examined. In study IV, adaptive nucleotide evolution patterns between warmer- and colder thermal origin subpopulations were studied in more detail than in the previous microsatellite studies. We further compared the different molecular levels of variation to gain the first insights into whether the same or different biological processes were targeted by the different molecular levels of variation.

2 Materials and Methods

The sequencing efforts for the studies were based on samples, which had been collected between 2006 and 2018 by several collaborators (**Table 1**). Included were multiple relatively closely related (sub)populations from the Norwegian grayling system, originating from a single ancestral population (study I, III and IV), and more distantly related populations from Norway, Finland and France (study II and IV). The samples included individuals from both western and eastern northern European lineages, and a southern lineage, all of which have been separated from each other for a long time based on early microsatellite studies, e.g. [41].

2.1 The common-garden experiment

To assess the relative magnitudes of the plastic component (response to rearing temperature regardless of the source population) and the adaptive component (differences in the plastic response between warmer- and colder-origin populations) of thermal adaptation, embryonic samples from a previous common garden experiment in were utilised where multiple grayling sub-populations from colder and warmer thermal origins were raised in reciprocal temperatures. Population-by-rearing temperature effect on larval growth was measured subsequently to hatching of the embryos. The experimental protocol for the common garden experiment is described in detail in an earlier publication [38]. The common garden experiment was repeated in 2013 for sample collection for our sequencing efforts. In the 2013 experiment, mature individuals from two colder and three warmer thermal origin populations were caught from the river outlets (**Table 1**) during the spawning season. Eggs and sperm were extracted under anaesthesia and transported on ice to the University of Oslo experimental facility. A pool of sperm from several (four to six) males was used to fertilise a pool of eggs from several (four to five) females of each population. The offspring from each population were reared in relatively colder (aiming to 6°C), medium (aiming to 8°C) and warmer (aiming to 10°C) temperatures. The water temperature was intended to remain constant from fertilisation until sampling. However, due to practical reasons beyond our control, the temperatures of the colder and medium treatment were more similar than planned. The realised temperatures were 6.7°C, 7.0°C and 10.2°C for low, medium,

and high temperature treatments, resulting in reciprocal thermal treatments that resembled the natal and non-natal condition, as opposed to three distinct temperatures. The embryos were sampled at multiple time points, which were selected either based on visually evident stages: eyes-stage, on average 113 degree-days (dd) with standard deviation of 7.2 dd over all rearing temperatures; and hatching stage, on average 181 dd with standard deviation of 17.2 dd over all rearing temperatures; and two intermediate time points which were selected by predicted degree-days (approximately 140 dd and 162 dd). Embryonic samples were stored in -80°C until sample preparation for the molecular studies (study I and III).

2.2 (Sub)populations and individuals used for the studies

The various grayling populations and sampling times used for the project are summarised in **Table 1**. For the gene expression and methylation studies (studies I and III), embryos from the common garden experiment were used, including individuals from both warmer and colder rearing temperature treatments and warmer and colder thermal origins. Two and three warmer-origin populations were used for study I and III, respectively, and two colder-origin populations for both study I and III. The biological replicates for study I included groups of four—five embryos from the 140 dd sampling point both for the realised 7.0°C and 10.2°C rearing temperatures. For study III, the group sizes were two—three individuals from the hatching-stage sampling point, and from both 7.2°C and 10.2°C rearing temperatures (note that the developmental stage is erroneously reported in the publication of study III). Although unarguably only single-time-point glimpses of the developing individuals, the observed embryonic gene expression and methylation patterns were taken from stages during the developmental period when individuals are currently exposed to contrasting temperatures, and when many differences have been observed between the (sub)populations in the study system (e.g. [38]–[40], [44]–[46], [49]). Using larvae also minimised stochastic processes that might alter the molecular patterns during life span. We also considered prior research suggesting that the teleost early-life history methylation patterns are re-established and achieve equivalency to the parental genome already at an early (blastula) stage [26], [51], [52].

For the reference genome assembly (study II), one adult male grayling from River Glomma (Evenstad, Norway; 61.42 N 11.09 E) was used. The sampling location is approximately 200 kilometres south-east from Lesjaskogsvatnet. The DNA of that individual had previously been used in a scaffold-level genome assembly [53]. The genome assembly (study II) was guided by a linkage map which was constructed from a library of restriction site associated DNA (RAD) tags. The

library had been previously obtained from a single family of grayling from the Rhine River (Obenheim, France). The family consisted of the female and male parent, 44 male offspring, and 69 female offspring. The DNA had been fragmented using SbfI restriction enzyme, following sequencing with the RAD methodology [54].

For the population genomics study (study IV), samples from three warmer- and three colder-origin subpopulations from Lesjaskogsvatnet were used. The samples were different individuals than the samples obtained from the common-garden experiment and used in study I and III. Rather, they had been collected from the nature during various years (**Table 1**). Two reference populations were also utilised including the ancestral population from Lågen, which is part of the River Gudbrandsdalslågen system, and a distantly related Finnish reference population of hatchery-maintained individuals from Puruvesi population. 23-24 individuals from each population were sequenced.

Table 1. The study populations.

POPULATION	DIVERGENCE TIME	ROLE	THERMAL ORIGIN	USED IN STUDIES (SAMPLING YEAR)
River Glomma, Norway	NA	for genome assembly	NA	II (2012)
River Gudbrandsdalslågen, Norway	-	ancestral	warmer	
- Otta	-	ancestral	warmer	I, III (2013)
- Lågen	-	ancestral	warmer	IV (2013)
Lake Lesjaskogsvatnet, Norway	1880's	sympatric	both	
- South basin	1880's	sympatric	both	
- Sandbekken	1880's	sympatric	warmer	IV (2007, 2008)
- Valåe	1880's	sympatric	colder	I, III (2013), IV (2007, 2008)
- Hyrion	1880's	sympatric	colder	IV (2007, 2008)
- North basin	1880's	sympatric	both	
- Brandlåe	1880's	sympatric	colder	IV (2006)
- Steinbekken	1880's	sympatric	warmer	III (2013), IV (2007, 2008)
- Store Skotte	1880's	sympatric	warmer	IV (2007, 2008)
- Hårrtjønn	1910's	parapatric	warmer	I, III (2013)
- Kvita	1920's	parapatric	colder	I, III (2013)
Lake Puruvesi, Finland	NA	outgroup	NA	II (2018)
River Rhine, France	2010's	for genome assembly	NA	II (NA)

2.3 RNA and DNA extraction and sequencing

A phase separation method with TRI reagent was used to extract total mRNA separately from the other molecules (total DNA and proteins) for study I. Salt extraction protocol [55] was used for the DNA extractions (studies III and IV), followed by RNAse treatment to exclude mRNA from the DNA samples. Full embryos were used in most extractions, except for the sequencing for genome assembly (study II) and some samples for study IV, where DNA of juvenile or adult individuals were sequenced. To avoid lane effects, sample libraries were split to multiple lanes or sample order was randomised before sequencing. Whole-genome sequencing was used for all data sets except for the RAD sequences for study II, where DNA fragments were sequenced thought the genome.

For RNA sequencing (study I), we used 100 base paired-end reads and the Illumina HiSeq 2000 platform. The sequencing depth of each transcript depended on the abundance of the corresponding mRNA in that sample to allow for the transcript abundance quantification. For the genome assembly (study II), DNA from one adult individual was sequenced using PacBio RS2 platform to obtain long reads at 19x depth. Reads with >10.000 base length and 5x coverage were then used in scaffold assembly where they were combined with 150 base paired-end reads, which had been previously sequenced from the DNA of the same grayling individual. Those short reads had been sequenced using Illumina HiSeq 2000 platform for an earlier scaffold-level genome assembly [53]. The RAD markers (study II) from the grayling family had been previously sequenced using 100 base single-end sequencing with Illumina HiSeq 2500 platform. For the cytosine methylation study (study III), 75 base paired-end reads were sequenced with Illumina HiSeq 3000 platform, targeting to 12x coverage per sample. Prior sequencing, MethylCode Bisulphite Conversion Kit was used to convert unmethylated cytosines in the DNA fragments to uracils, which were consequently sequenced as thymines, making the methylation levels in the cytosine loci quantifiable. For the population genomics study (study IV), the DNA was sequenced using BGIseq platform and 150 base paired-end reads at 5x depth per individual.

The reads were sequenced, quality-controlled and adapters were trimmed at sequencing centres. Subsequently, the quality of the reads was further inspected using FastQC, and the quality was controlled by trimming the reads using ConDeTri [56] or fastp [57]. Raw reads were stored in the Sequence Read Archive and are publicly available through NCBI servers.

2.4 Sequence assembly

Next, a summary is provided for the assembly steps of the various sequence libraries. The main sequence assembly efforts were the *de novo* reference transcriptome assembly and the following reference-based mapping of the reads from multiple individuals (study I); and the chromosome-level genome sequence assembly (study II), which was the reference for later DNA-resequencing efforts of multiple individuals (studies III and IV).

To evaluate the gene expression levels before a reference genome was available (study I), a *de-novo* transcriptome assembly was required. Pooled sequences from the ancestral population were used for *in-silico*-normalisation to reduce the computational cost by excluding kmers over 50x coverage, and the transcriptome was assembled using Trinity 2.0.4 [58]. To avoid extracting incompletely spliced transcripts and assembly error artefacts, only transcripts expressed at least at 10x were used for further analysis. Potential protein-coding regions of at least 100 bases long were predicted using TransDecoder [59]. To exclude transcript sequences with only minor differences with the major splice variants, transcripts that resulted into identical protein coding sequence were merged with CD-HIT [60].

To quantify the expression levels of each sampled individual, the RNA-seq reads (study I) were re-assembled against the *de novo* transcriptome using Bowtie2 [61] with settings that avoid mappings to long indels to prevent incorrect mappings to wrong splice variants [59]. SNP genotypes were called using BCFtools [62].

For the chromosome-level genome sequence assembly (study II), multiple assembly steps were required (**Figure 3**). First, the long reads were merged to consensus reads using Canu assembler, which aims to correct sequencing errors by combining the base information from overlapping reads [63]. The consensus reads were further merged to scaffolds in a hybrid assembly using PBJelly2 [64]. PBJelly2 first searches for adjacent consensus reads by aligning the short-read pairs to the long consensus reads. It then fills the gaps between the adjacent long consensus reads using base information of the short-read sequences.

To further combine the scaffolds, a linkage map for grayling was constructed. To obtain the RAD markers, the RAD marker sequences were mapped against the genomic scaffolds using Bowtie2 and the genotypes were called [61]. The RAD marker order was then predicted using Lep-MAP2 software, which estimates the probability of recombination between any two markers by comparing the parental SNP genotypes to those observed in the offspring [65]. The software then uses the probabilities to infer linkage groups of markers (chromosomes) and finally orders the markers within each linkage group [65]. Two of the original grayling linkage groups shared markers between chromosome pairs, a situation interpreted as residual

tetrasomy which cannot be automatically handled by the software. Thus, linkage information was not utilised for these chromosome pairs.

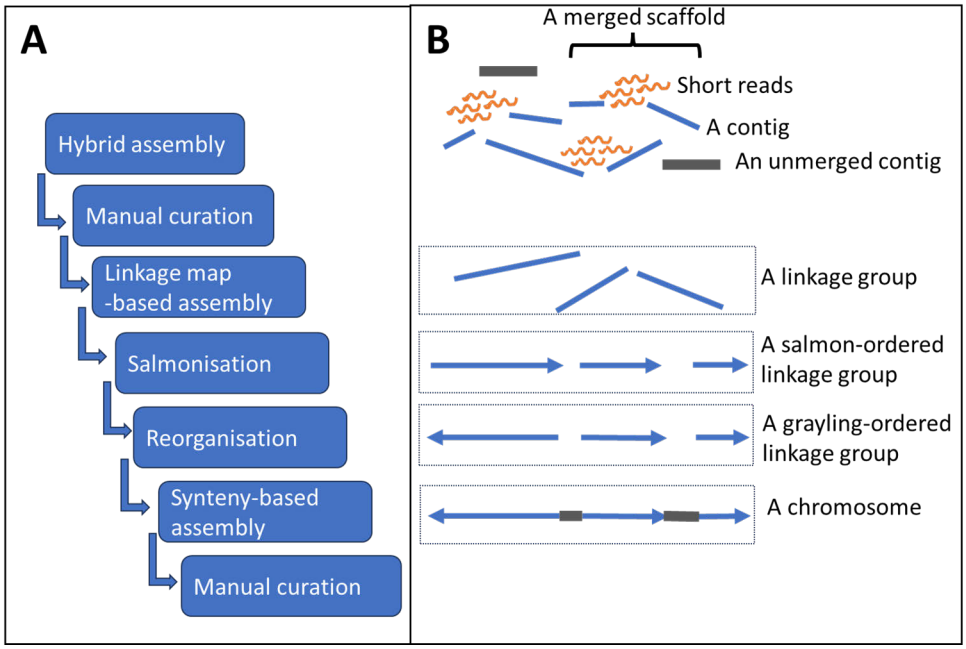


Figure 3. The process of the hybrid genome assembly, described as the key steps of the assembly process (A). The process is also described as a schematic presentation of the input Pacbio-based contigs (blue), which were merged with the help of short reads (orange) and ordered with the unassembled contigs (dark grey) during multiple subsequent steps up to chromosome-level (B).

To complete the scaffold assembly, synteny between grayling and Atlantic salmon was used to add more information about the scaffold order within chromosomes heuristically. For this, the grayling genome scaffolds were ordered by aligning the scaffolds against the Atlantic salmon genome sequence using MuMMER [66]. Then, the scaffolds which contained RAD markers were assigned to grayling chromosomes based on the associated linkage group. The initial ordering of the RAD-marker containing scaffolds within chromosomes was drafted based on their ordering in the Atlantic salmon genome (a process called salmonisation). Map distances between markers were calculated for the salmonised order. The grayling scaffolds were reorganised by breaking the salmonised chromosomes between pairs of adjacent scaffolds with unusually large map length increases between the associated markers, suggesting deviations from the salmonised order. All possible combinations of the order and orientation of the scaffolds within chromosomes were tested and the combination with the shortest total map length was selected. The final

ordering of the scaffolds in the grayling chromosomes was completed using manual curation, especially for the two chromosome pairs with residual tetrasomy, and three chromosome pairs with high similarity between the pairs (2A-2B, 11A-11B, 23A-23B). Finally, the scaffolds were merged into full-length chromosomes using repeats of the N (unknown) nucleotide between the scaffolds.

To quantify cytosine methylation levels (study III), the bisulphite sequencing output was assembled with Bismark bisulphite mapper v. 0.16.1 against the chromosome-level genome sequence [67]. Then, CpG methylation levels were extracted. As the information content of a CpG locus can be read symmetrically from both strands by the molecular machinery (a CG sequence is also a CG sequence when it is reverse complemented), methylation information from both strands was combined. Only a subset of the CpG loci with at least 16 samples sequenced at 8–30 combined read coverage were included for further analysis. Potential SNP variation was identified from the bisulphite assembly using BS-SNPer [68], and extracted using the SAMtools SNP calling pipeline. Potential C/G polymorphisms were excluded both during the Bismark assembly and based on the predicted SNP variation.

The genomic reads (study IV) were mapped to the chromosome-level genome sequence using BWA mem v. 0.7.17 [69]. SNPs were extracted using SAMtools v. 1.9 and BCFtools v. 1.10 [62]. Rare allele SNP loci with MAF < 0.05 among all populations and within Lesjaskogsvatnet were excluded.

The mRNA sequences from study I were utilised also in studies III-IV by updating the *de novo* transcript assembly with a reference-based assembly of the reads against the new chromosome-level genome assembly with TopHat assembler v. 2.1.1 and counting the expression levels using HTSEQ-count v. 0.9.0 [70], [71].

2.5 Annotation

The predicted *de novo* transcriptome (study I) was annotated using reciprocal best BLASTP hits of the predicted grayling protein coding sequences in at least one of the reference proteomes including multiple reference species. The reference proteomes for zebrafish (*Danio rerio*), stickleback (*Gasterosteus aculeatus*) and cod (*Gadus morhua*) were extracted from the Ensembl [72] database, and the reference proteome for Atlantic salmon from GenBank [73].

To name the grayling chromosomes meaningfully, that is, consistently with homeology with the teleost ancestor, the grayling chromosome pairs were compared to the Northern pike (*Esox lucius*) chromosomes [74], [75] using MUMmer [76]. The reference gene annotations for the assembled genome sequence (study II) were created by combining *de novo* gene predictions, made using various gene finder

software, and a library of expressed grayling transcripts, which had been generated during prior genome sequencing efforts [77]. The gene "sexually dimorphic on the Y-chromosome" (*sdY)* has been reported as male-specific in other salmonids [78]. The gene was located using rainbow trout (*Oncorhynchus mykiss*) gene sequence as a query against the grayling genome sequence in a BLASTN search. The transposable element annotation of the genome sequence (study II) included a more complex approach, which combined both *de novo* element predictions based on the assembled chromosome sequence content, and elements matching those previously found in Atlantic salmon [12].

The CpG methylation loci (study III) annotations were based on overlaps with the grayling reference genes from study II. The grayling genes were matched to the orthologous genes of a model species zebrafish based on the best matches in the Ensembl database [79] using BLASTP comparisons. Similarly, functional annotations for the CpG loci (study III) were assigned for each locus based on overlapping functional region category:

- upstream untranslated regions (5'UTR),
- promoters,
- protein-coding regions,
- introns and
- downstream untranslated regions (3'UTR).

The loci located outside functional gene regions were annotated as intergenic.

In study IV, gene and functional region annotations were made for the SNPs using the grayling gene annotations from study II and SNPEff software [63]. This approach resulted in even more detailed functional annotations, also including annotations such as

- synonymous and missense mutations of the protein-coding region
- upstream regulatory regions and
- downstream regulatory regions.

2.6 Gene expression (study I)

Transcript expression levels were corrected for length bias and the expression levels were extracted based on reads mapping to each transcript using eXpress software [80]. A residual-based error correction protocol was used to remove unwanted variation, implemented in the function RUVr in the R package RUVSeq, and assuming that the covariates of interest were not correlated with the unwanted variation [81]. Then, interactions of genes in a co-expression network were predicted

using correlations between the expression levels of each gene pair. Co-regulated genes were then clustered into 'modules' of co-regulated genes using a hierarchical clustering method implemented in the WGCNA package [82]. The module assignments of each gene were assessed based on support of at least 70 % of the generated module assignments in a hundred bootstrap replicates. Enriched gene ontology categories and PFAM protein domains were sought from the gene lists of the robust modules using STRING database [83].

The global and per-module transcript expression profiles for each grayling individual were summarised using a principal component analysis. The eigenvalue along the most significant principal component was used as the individual profile value. Then, the significance of the potential explanatory variables associated with the expression profiles were assessed using ANOVA. The potential explanatory variables included population as a proxy for neutral variation; rearing temperature as a proxy for plasticity; and population-by-temperature interaction as a proxy for adaptive evolution.

2.6.1 Comparisons between selection-affected and neutral divergence

Genetic differentiation associated with thermal adaptation was approximated using the Q_{ST} statistic. In the absence of gene-wise heritability (h^2) estimates, the Q_{ST} estimates of each gene were based on the observed phenotypic divergence in the gene expression levels between populations (P_{ST}), and ratio c/h^2, where c and h^2 represent the additive phenotypic variance (σ^2) between (σ^2_B) and within (σ^2_W) populations, respectively, as defined in [84]. For the c/h^2 ratios, a parameter space with three scenarios was searched to obtain a range of Q_{ST} values complementary to each scenario: 1) c/h^2=0.5 as a scenario where the phenotypic variation is primarily explained by environmental or non-additive factors; 2) c/h^2=1 as a scenario where both non-additive and additive factors have a similarly sized roles; and 3) c/h^2=1.5 as a scenario where the phenotypic variation is primarily determined by additive genetic variation [84]. To estimate σ^2_B and σ^2_W, a fixed effect of thermal treatment and a random effect of population were added, respectively. These explanatory variables were used to fit a linear mixed model for each gene. Finally, the c/h^2 ratio, σ^2_B and σ^2_W were plugged to equation

$$P_{ST} = \left(\frac{c}{h^2}\sigma^2_B\right)/\left(\frac{c}{h^2}\sigma^2_B + 2\sigma^2_W\right) \quad (1)$$

to obtain the Q_{ST} estimate intervals for each gene.

A genome scan was used for the SNPs extracted from the re-mapped transcripts to detect candidate loci which are potentially under natural selection. The Bayesian candidate approach compares generalised logistic regression models with a

population (neutral) component shared between all loci, to models which contain the population component and a SNP-specific (selection) component [85].

2.7 Genome assembly (study II)

The presence of residual tetrasomy was assessed using the observations of shared linkage maps. The presence of highly similar regions (assumed recent residual tetrasomy), which has been reported in other salmonids, was assessed with manual inspection of the grayling-to-salmon alignments.

Transposable element abundancies in the grayling and the Atlantic salmon genomes were compared with a linear model *log2(element abundancy in salmon+1) ~ log2(element abundancy in grayling+1*) which predicted the element response assuming that the abundances of each element were equal between the two species. Highly diverged transposable elements with residuals exceeding ±1.96 standard deviations from zero were interpreted as differentially abundant.

2.8 CpG methylation (study III)

The methylation levels of each CpG locus and sampled individual were estimated based on methylated and unmethylated read counts. Then, the CpG loci were assigned as hypomethylated, intermediately methylated or hypermethylated based on the average methylation percentage of <20 %, 20—80 % and >80 % across all individuals. The potential explanatory variables for the CpG methylation patterns were assessed using multiple methods at different genomic resolution, from global sample-wise overall methylation profiles to locus-specific CpG methylation rates. The methods are summarised next.

Genome-wide methylation profiles of each sampled individual were acquired using two methods. First, principal component analysis was used to summarise the three molecular data sets without any prior information of the explanatory variables. The two most explanatory principal components were used to compare the average pairwise Euclidean distances between individuals from each population pair and the differences were verified using MANOVA and Tukey's *post-hoc* test. Second, non-parametric distance-based redundancy analysis and permutation tests were used to explain pairwise distances between individuals with prior information of multiple explanatory variables. The variables tested were the divergence order, thermal origin (evolutionary effect), rearing temperature (plastic effect) and sex of each embryo. To compare the molecular profiles at different levels of molecular variation, we also generated profiles at the SNP level (utilising the SNPs extracted from the bisulphite-sequencing reads (study III) and at the gene expression level, utilising the re-mapped RNA-sequencing reads (study I).

Genome-wide methylation level differences between population-by-rearing temperature groups of the sampled individuals were evaluated using *t*-tests.

To test for the presence of overrepresented consistent methylation patterns as a signal for convergent adaptive response, the abundances of consistently changed loci were compared between the two pairs of populations with warmer-to-colder colonisation (Otta—Valåe and Hårrtjønn—Kvita) to the abundances of inconsistently changed loci between the population pairs. Only the subset of loci where methylation level changed substantially (at least 50 %) were considered. It was expected that consistent changes occur either as a parallel adaptive response to altered environmental temperature or drift, whereas inconsistent changes would be related to drift. The overrepresentation of consistent in comparison to inconsistent changes was verified using the Chi-squared test. Similarly, to test for the presence of a plastic methylation response, the abundances of the loci with large consistent methylation level changes to rearing temperature were compared to the abundances of inconsistently changed loci.

Site-specific CpG methylation level responses were further studied with an epigenome scan approach. The scan aimed to detect both developmental plasticity (response to rearing temperature) and evolved differences in the developmental plasticity in response to selection (population-by-rearing temperature response) explaining the methylation variation among individuals. A mixed logistic regression model was repeated for each CpG locus to explain the numbers of methylated and unmethylated reads with fixed effects of developmental temperature and sex, a random intercept of population, and random slopes for the population-by-developmental temperatures. The significance of each explanatory variable was tested using likelihood ratio tests.

Multiple methods were also used to assess the functional relevance of the plastic or adaptive CpG loci. First, to understand which functional regions were most affected by the plastic and adaptive phenotypic response, the abundances of the observed plastic (from the epigenome scan) and selected (consistently changed) CpG methylation changes in different functional regions of genes were compared to the expected random distribution based on the overall abundancies of the CpG loci in each functional region. The abundancies of plastic and potentially selected CpG loci in the promoters (500 bases upstream from the 5'UTR), 5'UTR regions, protein coding sequences, and 3'UTR regions of genes were evaluated using Chi-squared tests.

To define the functions of gene sets associated with characteristic upstream regulatory regions, and genes associated with plastic or evolutionary methylation response, gene ontology enrichment tests were performed for the corresponding gene lists using standard hypergeometric tests. The lists contained genes

- with constantly hypo- or hypermethylated and CpG-abundant or -poor upstream regulatory region (four lists),
- associated with plastic CpG methylation response to rearing temperature in the epigenome scan (one list) and,
- associated with adaptive CpG methylation response between populations (population-by-rearing temperature response in the epigenome scan; one list).

For the constantly hypo- and hypermethylated gene lists, a constant methylation status was required in at least five associated CpG loci in the upstream regulatory sequence, respectively. Intermediately methylated loci were excluded from this analysis. The hypo- and hypermethylated gene lists were further divided by the CpG-abundancy of the upstream regulatory regions using the median CpG abundancy of each list as a cut-off value. A combination of all four lists was used as a background.

For the gene lists with plastic and adaptive methylation response, all genes with multiple CpG loci detected with plastic or population-by-rearing temperature effect in the epigenome scan were used as the foreground lists. All genes with multiple CpG loci assessed in the epigenome scan were used as the background.

We further studied the enriched term specificity of the gene lists. The terms enriched in the gene lists were compared. The levels of specificity of the terms were assessed: first, the level of specificity of each term was calculated as the proportion of parental terms of the total associated (both parental and offspring) terms [86]. Then, the differences in the term specificity between the three term lists were evaluated using ANOVA and Tukey's *post-hoc* test (unpublished data).

2.9 Population genomics (study IV)

Shared ancestry between the Finnish outgroup, the ancestral population, and the subpopulations of Lesjaskogsvatnet was first assessed using ADMIXTURE analysis [87]. The presence of neutral divergence between subpopulations was further estimated by detecting isolation-by-distance from the presence of a correlation between pairwise genetic diversity estimates [$F_{ST}/(1-F_{ST})$] and geographic distances between subpopulations.

Three genome scan methods were used to detect differentiated genomic regions between the grayling populations with contrasting thermal origins. First, PCAdapt was used to detect highly diverged loci in the absence of explanatory variables. We used Cattell's rule to determine the number of principal components used. Second, a latent factor mixed model (LFMM) was used to detect diverged loci between the thermal origins while simultaneously controlling for the neutral divergence between individuals, modelled as latent factors. Third, OutFLANK was used to infer the F_{ST}

distribution for the loci. For the first two genome scan methods, the *P*-value thresholds for the candidate loci were defined according to the suggestive genome-wide significance limit of $P<5*10^{-5}$, suggested in [88]. The OutFLANK method estimates the empirical *Q*-value threshold from the F_{ST} distribution.

The abundance of consistently changed loci among the subpopulations was evaluated from the allele frequency changes of variable loci in all possible warmer-to-colder subpopulation pair comparisons. The overrepresentation of the observed consistently changed loci in comparison to random occurrences was estimated using a permutation test. The null distribution was obtained by randomising the SNP order of the compared population pairs for a hundred permutations.

For gene ontology term enrichments associated with thermal adaptation candidate loci, lists of genes associating with candidate SNPs in PCAdapt, LFMM and OutFLANK analysis were compared to all genes associated to genotyped SNPs using classic Kolmogorov-Smironov tests implemented in TopGO package [89]. To determine whether the potentially adaptive variation was novel it was assessed whether the top-10 candidate loci of each candidate locus test 1) for each chromosome 2) for the overall top-10 lists were already present in the ancestral population.

The average candidate locus significances for the functional regions and the more often neutral intergenic regions were measured for the functional regions with at least 5,000 overlapping SNPs.

2.10 Comparisons between biological processes associated with plasticity, adaptive methylation and adaptive nucleotide variation

Biological processes associated with CpG methylation variation (study III) were compared to the processes associated with nucleotide variation (study IV). Similarly, a list of both processes and genes associated with plastic gene expression variation were compared to the processes and gene lists associated with the candidate nucleotide loci (study IV). The process and gene lists were determined by combining the results of all three genome scans. In the absence of comparable biological process list from the prior gene expression study (study I), we re-mapped the RNA-seq reads from study I to the grayling reference genome (study II) and extracted the enriched biological processes using Kolmogorov-Smironov tests implemented in TopGO package. Finally, we visualised the LFMM candidate -related biological processes using SimRel semantic similarity and Revigo v. 1.8.1 [90].

3 Results

3.1 Sequence assembly

The *de novo* transcriptome assembly resulted in 142,653 final predicted transcripts in 109,102 predicted genes. 61,190 of the transcripts contained unique protein-coding sequences. 2,458 SNPs were extracted from the re-mapped mRNA sequence data (study I).

The prior contig-level genome assembly [53] consisted of 24,369 contigs. After consensus contig generation and hybrid assembly, the number of scaffolds decreased to 18,265. 6,076 RAD markers associated with 54 % of the genome scaffolds. After further assembly utilising the linkage maps, salmonisation and manual curation, the final chromosome-level assembly consisted of 51 chromosomes. 18 pericentric (containing the centromere) and 24 paracentric (not containing the centromere) inversion events were predicted in 34 of the grayling chromosomes. No fusions of ancestral chromosomes were detected. Instead, one fission event was observed, increasing the ancestral 1N chromosome number by one from 50 to 51 (study II). The final estimated genome size of grayling was 1.5 gigabase pairs.

The bisulphite sequencing assembly covered 9,663,307 loci with variable methylation levels observed among the individuals after filtering. Moreover, 290,705 loci were unmethylated in all individuals. 3,465,289 of the CpG loci remained in the filtered data set in all individuals. 78,012 SNPs with no missing observations were extracted (study III). The population genomic sequence assembly (study IV) resulted in 4,454,829 and 3,433,357 SNPs variable after filtering among all populations and populations within Lesjaskogsvatnet, respectively.

The transcriptome reassembly from the mRNA sequences of the gene expression data set (study I) resulted in 22,526 reference-based transcripts.

3.2 Annotation

Of the *de novo* transcripts (study I), 19,461 were annotated against at least one reference proteome. After removing the transcripts with zero expression level in

some of the samples for conveniency, the final expression data set consisted of 16,622 annotated protein-coding transcripts.

Using the northern pike as a proxy for the non-duplicated teleost ancestor of salmonids, we were able to match almost all ancestral (pike) chromosomes to grayling chromosomes in 1:2 ratio. Subsequently, I was able to name the grayling chromosomes according to the naming convention of the pike (chromosomes 1-25) by adding extension 'A' and 'B' for the grayling chromosome duplicates, except for the ancestral chromosome 13 where one chromosomal fission event resulted in the grayling chromosomes 13A, 13B and 13C. After annotating the genes in the grayling genome sequence (study II), the male-specific sdY gene was in the chromosome 11A, which was determined as the grayling sex chromosome. 47.4% of the grayling genome sequence consisted of transposable elements, which was comparable to the coverage observed in Atlantic salmon (52.3%) (study II).

Of the almost ten million CpG methylation loci in study III, 1,559,048 (15.6 %) were associated with a functional gene region. The corresponding amount of the 3.4 million SNP loci (study IV) was 1,095,935 (32.0 %).

3.3 Gene expression (study I)

36.1 % of the transcripts were assigned to six robust co-expressed gene modules during module construction. The first principal component explained 25.9 % of the variation. At the genome-wide gene expression profile level, a proportion of the variation of the gene expression profiles was explained by a population-by-temperature interaction ($F_{3,27}$ = 11.3, P<0.001), suggesting evolution of reaction norms. However, the plastic main effect ($F_{1,27}$ = 208.1, P<0.001) was the most relevant in the data while the population effect was minor ($F_{3,27}$ = 26.8, P<0.001). In the per-module expression profiles, plasticity was detected in two modules and evolution of reaction norms in two modules (**Table 2**). The latter modules also had relatively higher average Q_{ST} estimates than the overall mean (**Table 2**). Four of the modules contained enriched gene ontology terms (**Table 2**).

Mean Q_{ST} estimates for the 0.5, 1 and 1.5 c/h^2 ratios were 0.024 (ranging between 0–0.570), 0.044 (ranging between 0–0.726), and 0.062 (ranging between 0–0.799), respectively. Most transcript-wise Q_{ST} estimates were over zero, indicating phenotypic divergence between populations. However, only two transcripts had significantly greater Q_{ST}-values than the mean F_{ST} of 0.128 (ranging between 0 and 0.531), indicating divergence higher than expected by neutral level of evolution. Similarly, no SNPs were detected using the Bayesian candidate locus approach. Thus, we determined that the evidence for gene expression evolution was inconclusive.

Table 2. Six robust modules of co-expressed gene sets in the grayling populations, having evolved under rapid thermal adatation pressure for c.a. 30 grayling generations.

MODULE	NUMBER OF GENES	NUMBER OF GO ENRICHMENTS, EXAMPLES OF HIGHLIGHTED TERMS	FACTORS EPLAINING MODULE -BASED SIGNUTURES (P<0.001)	MEAN Q_{ST} (COMPARISON TO THE OVERALL MEAN)
black	223	15; muscle fibre development	rearing temperature	0.037 (smaller)
blue	1,500	55; methylation	population-by-rearing temperature	0.123 (larger)
brown	1,133	-	-	0.020 (smaller)
green	740	33; response to stress, response to stimulus, nervous system development	rearing temperature	0.041 (smaller)
red	302	-	-	0.016 (smaller)
turquoise	2101	7; embryonic organ development	population-by-rearing temperature	0.137 (larger)

3.4 Genome assembly (study II)

Residual tetrasomy (chromosomal regions where quartets of chromosomes are still recombining) has been reported in multiple salmonid species [75]. Some of those regions were also present in the grayling, particularly in the grayling chromosome pairs 9A-9B and 25A-25B. The other grayling chromosome pairs with residual tetrasomy reported in the Atlantic salmon (2A-2B, 11A-11B, 23A-23B) had close between-pair homeolog sequence similarity.

The individual transposable element abundancies were generally very similar between the two species, except for 14 Atlantic salmon -specific and three grayling-specific elements, covering 83.6 and 0.2 megabase pairs of the respective genome.

3.5 CpG methylation (study III)

The embryonic grayling CpG methylome was generally highly methylated with 76.8 % mean methylation level over all loci. The majority of 72.1 % of the loci were hypermethylated, 19.7 % intermediately methylated and only 8.2 % hypomethylated. The methylation levels were typically high, except in the upstream regulatory regions of genes, which included both abundant hyper- and hypomethylated loci.

At the methylation and nucleotide level, the principal component -based molecular profiles clustered individuals of populations together. Moreover, populations from colder thermal origin clustered at the methylation level, and sympatric populations from Lesjaskogsvatnet at the nucleotide level. Contrastingly,

thermal origin did not explain the divergence at the gene expression level (**Table 3**). The distance-based redundancy analysis identified divergence order and thermal origin as explanators at the nucleotide and methylation level and rearing temperature at the gene expression level (**Table 4**). Lower methylation levels (with $P < 0.0001$) were observed in four of the five studied populations in the colder rearing environment.

Table 3 Variance explained by the two most expressive principal components for the three levels of molecular variation, used as molecular profiles of the individuals, and mean distances (Euclidean distances from PC1 and PC2 eigenvalues) between the populations. The distances above zero are presented only for comparisons with Tukey's adjusted $P<0.05$. The distances are based on the complete observations for the four or six individuals from the study populations of the methylation study III (nucleotide and methylation markers) and the re-mapped gene expression data from study I (**Table 1**).

MOLECULAR LEVEL	PC1	PC2	PAIRWISE DISTANCE WITHIN WARMER ORIGIN	PAIRWISE DISTANCE WITHIN COLDER ORIGIN	PAIRWISE DISTANCE BETWEEN ORIGINS
nucleotide	7.2 %	6.2 %	78, 118, 48	55	79, 112,0*, 64, 53,92
methylation	5.1 %	4.7 %	148, 164, 150	0	134, 134, 71*, 87, 81, 64
gene expression	51.3 %	14.2 %	0	0	0, 0, 0, 0

Table 4 Variance explained by axis 1 and axis 2 from a distance-based redundancy analysis, and the significance levels of the divergence order, warmer or colder thermal origin and rearing temperature. The distances are based on the complete observations for the two or three individuals from the warmer and colder rearing temperature in each study population of the methylation study III (nucleotide and methylation level) and the gene expression study I (**Table 1**).

MOLECULAR LEVEL	VARIANCE EXPLAINED		SIGNIFICANCE		
	axis 1	axis 2	divergence order	thermal origin	rearing temperature
nucleotide	6.4 %	5.2 %	***	**	
methylation	4.6 %	4.2 %	***	**	
gene expression	34.0 %	2.7 %		.	***

'***' for $P < 0.001$, '**' for $P < 0.01$, and '.' for $P < 0.1$.

Among the loci with major methylation level changes between the population pairs with colder-to-warmer colonisation history, the allele frequencies of the majority of

715 loci were consistently changed ($\chi^2_1 = 82.3$, $P < 0.0001$) between population pairs while the frequencies of only 408 loci changed inconsistently. However, no such enrichment was observed among the consistently plastic loci.

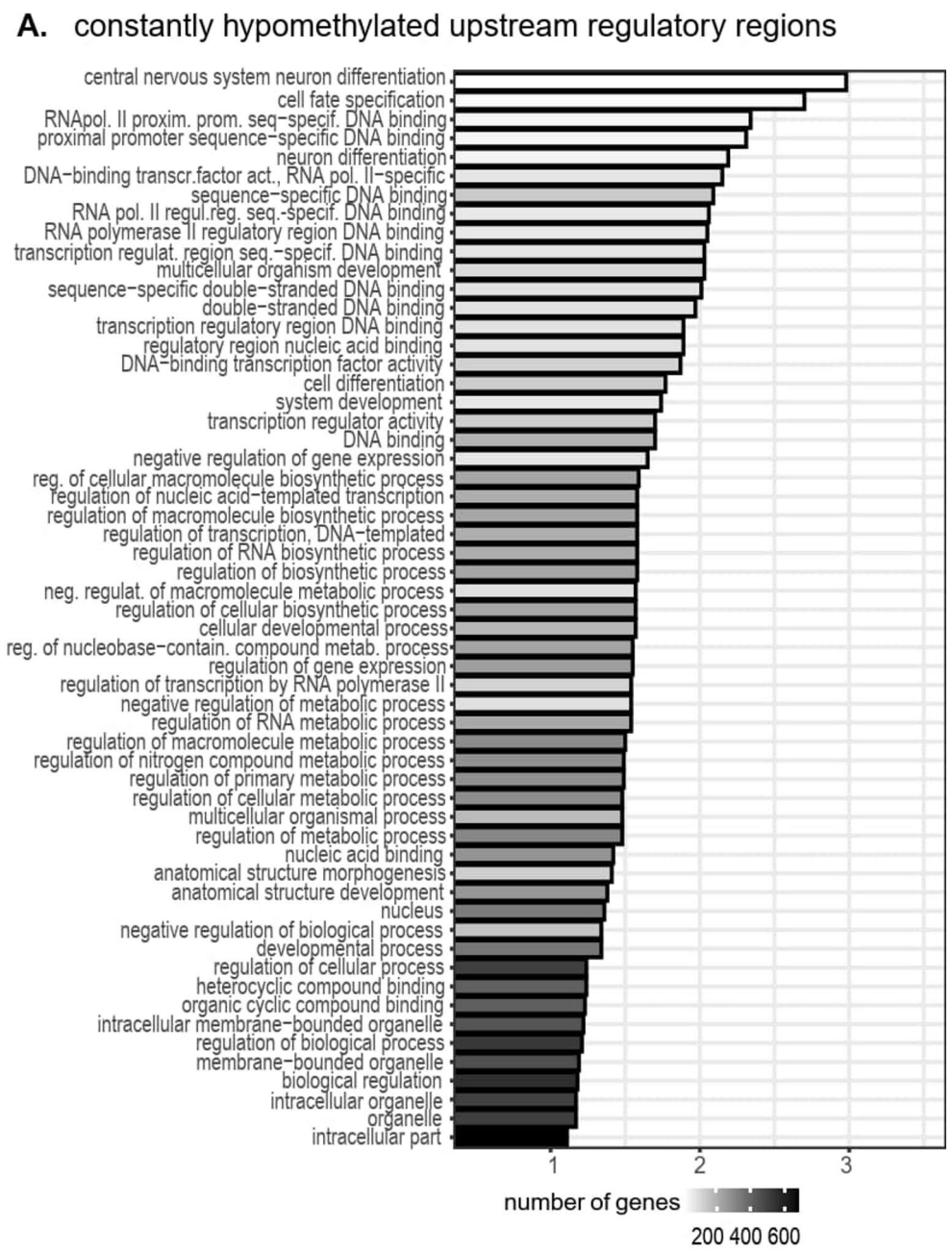
The epigenome scan highlighted 21,566, 25,980 and 72 loci that were best described including random population term, population-by-temperature interaction term, or both, respectively. The remaining 882,756 loci were best described without random effects. Plastic effects explained a portion of the methylation level variation of 1,806 loci in 943 genes for the rearing temperature, and 2,271 in 1,277 genes for the sex.

The consistently changed CpG methylation loci were abundant in the coding sequences and the 3'UTR region but depleted from the upstream regulatory region. On the contrary, the plastic loci were often located in the upstream regulatory regions.

The lists of genes with CpG-abundant and systematically hypo- or hypermethylated upstream regions consisted of 1,709 and 1,477 genes, respectively. The corresponding lists of the CpG-poor regions included 2,094 and 2,324 genes. Gene ontology term enrichments were scarce among genes with CpG-poor upstream regulatory regions (only eight and 15 terms found in the gene lists with hypo- and hypermethylated upstream regions, respectively). In contrast, 69 and 58 terms were enriched in the corresponding CpG-abundant genes (**Figure 4**). Among those, the hypomethylated loci associated with terms such as transcription regulation and DNA binding in organelles within the cell. The terms related to basic developmental processes such as the development of nervous system and anterior-posterior-axis pattern, and to metabolic processes. In contrast, hypermethylation related to more specific terms such as cell signalling and cell-cell adhesion functions in the receptors and the cell surface membrane, and muscle fibre binding. The associated processes included muscle fibre assembly and activation in response to calcium-signalling.

43 terms were enriched among the genes with differentially evolved methylation response (population-by-rearing temperature interaction; **Figure 4**). Those terms included terms related to various binding functions of molecules such as GTPAses, ions and muscle subunits, binding to signal sequence, transporter enzyme functions and motor activity, and localisation in the myosin complex. The processes also included membrane depolarisation during action potential and developmental processes. There were no term enrichments for the plastic gene set either for rearing temperature or for sex.

Term specificity was studied between the three term lists which were reasonably long (the lists of terms related to CpG-rich and hypo- or hypermethylated promoters, and the term list related to genes with population-by-rearing temperature interaction). The enriched terms were more specific in the hypermethylated promoters (adj. $P<0.001$) and in terms associated with population-by-rearing temperature interaction (adj. $P=0.004$) in comparison to the hypomethylated promoters.


A. constantly hypomethylated upstream regulatory regions
central nervous system neuron differentiation
cell fate specification
RNApol. II proxim. prom. seq−specif. DNA binding
proximal promoter sequence−specific DNA binding
neuron differentiation
DNA−binding transcr.factor act., RNA pol. II−specific
sequence−specific DNA binding
RNA pol. II regul.reg. seq.−specif. DNA binding
RNA polymerase II regulatory region DNA binding
transcription regulat. region seq.−specif. DNA binding
multicellular organism development
sequence−specific double−stranded DNA binding
double−stranded DNA binding
transcription regulatory region DNA binding
regulatory region nucleic acid binding
DNA−binding transcription factor activity
cell differentiation
system development
transcription regulator activity
DNA binding
negative regulation of gene expression
reg. of cellular macromolecule biosynthetic process
regulation of nucleic acid−templated transcription
regulation of macromolecule biosynthetic process
regulation of transcription, DNA−templated
regulation of RNA biosynthetic process
regulation of biosynthetic process
neg. regulat. of macromolecule metabolic process
regulation of cellular biosynthetic process
cellular developmental process
reg. of nucleobase−contain. compound metab. process
regulation of gene expression
regulation of transcription by RNA polymerase II
negative regulation of metabolic process
regulation of RNA metabolic process
regulation of macromolecule metabolic process
regulation of nitrogen compound metabolic process
regulation of primary metabolic process
regulation of cellular metabolic process
multicellular organismal process
regulation of metabolic process
nucleic acid binding
anatomical structure morphogenesis
anatomical structure development
nucleus
negative regulation of biological process
developmental process
regulation of cellular process
heterocyclic compound binding
organic cyclic compound binding
intracellular membrane−bounded organelle
regulation of biological process
membrane−bounded organelle
biological regulation
intracellular organelle
organelle
intracellular part
1
2
3
number of genes
200 400 600

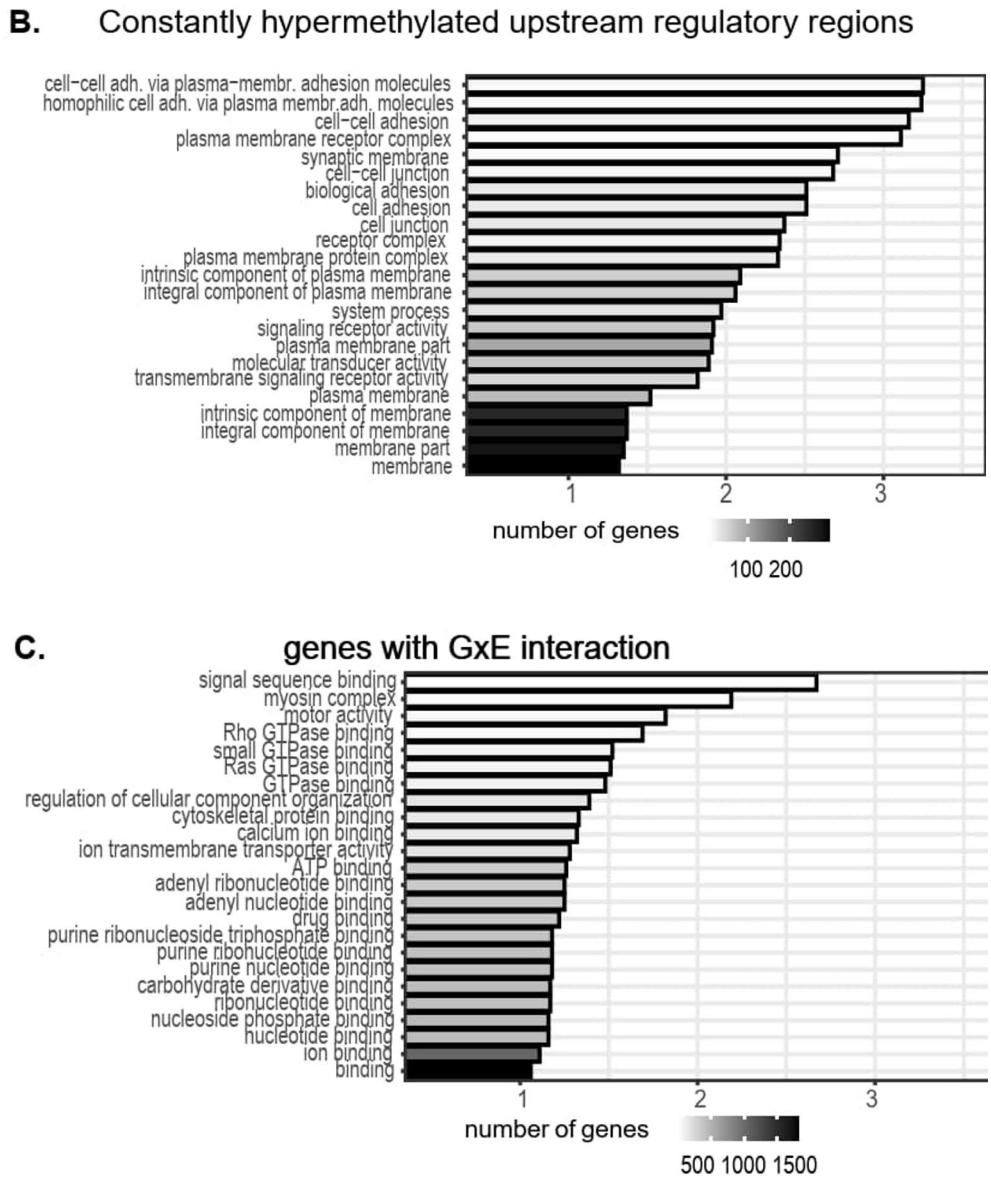


Figure 4. Representative (terms selected using Revigo [90]) gene ontology enrichments (with *FDR*<0.01; along the x-axes) of the gene lists with hypomethylated promoters (embryonic gene expression on; A), constantly hypermethylated promoters (embryonic gene expression off; B), and genes with population-by-rearing temperature -interaction in the metylation levels (C).

3.6 Population genomics (study IV)

The ancestry analysis identified three clusters, including the Finnish reference population, the ancestral population, and the subpopulations within

Lesjaskogsvatnet. Isolation by distance was found among the Lesjaskogsvatnet subpopulations (r=0.508, P=0.03).

Within Lesjaskogsvatnet, the first six principal components explained 54.7 % of the SNP data and explained the variation according to the Cattell's rule. The PCAdapt, LFMM and OutFLANK genome scan approaches identified 17,220, 325 and 26,456 candidate loci and 150, 135 and 106 enriched biological processes, respectively.

The observed frequency of consistently changed loci between the warmer-to-colder subpopulation comparisons was 20.8 %, which was greater than the 10.6 % of loci expected by chance (P = 0.001). A minority of 18 % of the variable loci in Lesjaskogsvatnet were potentially novel (variation not observed among the ancestral population samples). Almost all top candidate loci (50 of 51, 46 of 51 and 47 of 51 of the per-chromosome top candidate loci among the PCAdapt, LFMM, and OutFLANK candidate loci, respectively; and 26 of 26 of the overall top candidate loci) were already present in the ancestral population, suggesting selection primarily from standing genetic variation rather than from *de novo* mutations.

Compared to the intergenic regions, 5'UTR regions and synonymous coding sequence variants had notably higher average significance levels. Interestingly, the most elevated functional regions by significance were splice sites and 3'UTR regions.

3.7 Comparisons between biological processes associated with plasticity, adaptive methylation and adaptive nucleotide variation

82 of the 145 biological processes related to plastic gene expression (study I sequences remapped in study IV) and 20 of the 32 processes related to adaptive methylation (study III) were also present among the 210 candidate SNP -related processes (study IV). Similarly, 1,707 of the 5,823 plastic genes were also among the 8,213 candidate SNP -associated genes.

4 Discussion

Besides the many phenotypic changes which have been previously reported in the study populations in response to temperature, protein expression levels have been associated with both plastic and evolutionary response in the system [91]. As expected, an extensive plastic response of rearing temperature altered the gene expression levels, i.e, the grayling embryos acclimated to the developmental environment (study I). However, contrary to the prior expectation based on the observed protein expression differentiation between populations from contrasting thermal origins [39], the genome-wide gene expression profiles were mostly canalised. Thus, the individual gene expression profiles were mainly explained by rearing temperature and only two genes exceeded the neutral divergence rate in the Q_{ST}/F_{ST} comparisons (study I). While the strong effect of plasticity on gene expression levels is clear, the evolutionary signal may have been underestimated as the selection on SNPs harvested from transcriptome sequencing data may not have been close to neutral. I discovered this later (study IV) when observing that many of the functionally classified SNPs within the transcribed regions had elevated signal of adaptive divergence in comparison to the non-transcribed regions. The bias caused by non-neutral SNPs may also have contributed to the apparent lack of signal for natural selection of the expression levels using the Bayesian approach and the Q_{ST}/F_{ST}-comparisons of study I. Post-translational modifications of proteins are also potential factors that may have an important but so far unassessed role in the process.

Although adaptive signals were non-conclusive when assessed using single genes (the Q_{ST} estimates overlapped with the F_{ST} distributions), some of the co-expressed gene modules were more informative: we found two gene modules with both population-by-rearing temperature interactions in the module-specific gene expression profiles, and high phenotypic divergence (Q_{ST}) in comparison to other modules. The clustering of genes into modules also helped to overcome the problem of multiple testing, which I often found very problematic in the whole-genome data sets. Moreover, the genes in these modules also shared functions, suggested by module-specific enriched terms such as methylation and embryonic development. Modularity may have facilitated gene expression evolution of key functions of thermal adaptation by encapsulating the expression response within the two relevant

modules while maintaining homeostasis of the other functions in the rest of the modules.

For my thesis, finalising the genome assembly was an important step that enabled further research on the adaptation process in the system, as the assembly was used as the reference sequence for the later sequencing efforts. My research provided sequence-level evidence that the grayling lineage has undergone unique karyotype evolution, which is distinct from other salmonids. The almost perfect one-to-two matches between pike and grayling chromosomes, and one fission event, verified the prior karyotype findings which suggested that the grayling genome has been mostly conserved at the chromosome identity level since the whole-genome duplication of salmonids. By detecting chromosomal rearrangements, I verified that the necessary karyotype differentiation between duplicated chromosome pairs after genome tetraploidisation has progressed through intra-chromosomal inversions rather than (otherwise) salmonid-type between-chromosome fusions. In addition to pericentric inversions, which were previously anticipated based on visual inspection of karyotypes, the data suggested that also paracentric inversions, which are usually invisible to visual karyotype determination, may be abundant. Besides verifying the prior hypotheses of the karyotype evolution in the grayling subfamily, and facilitating further genomic research of the species, the assembly of this distinctive genome serves as a valuable addition to the collection of chromosome-level salmonid genomes, which have been published for 16 species in the NCBI Genome database to date (last accessed 11th December 2023). These salmonid genomes may be utilised for comparative genomic research of the iconic fish family of salmonids to study key evolutionary questions such as what are the forces driving genome evolution and the impacts of genome duplication.

Although large-scale differences are evident between the genomes of grayling and other salmonids, I also observed many similarities at smaller scales. First, like in Atlantic salmon [12], also the grayling linkage map suggested the presence of ongoing and recent residual tetrasomy in two and three pairs of grayling chromosomes, respectively. This suggests that the residual tetrasomy is not specific to the salmon-type genome evolution processes among the family. Second, a male-specific gene sdY was on chromosome 11A. The region was present in the known male individual which I used for the genome assembly. In the subsequent population-level studies (III and IV), the observed sex status of the adult individuals from Puruvesi matched the predictions based on sdY coverage (unpublished data).

While transposable elements have often been suggested as a contributing factor in the rediploidisation process [12]**,** I did not find many grayling-specific elements which could have caused differential karyotype evolution as the occurrences of the grayling-specific elements were very low. Contrastingly, a significant amount of 80

megabase pairs the Atlantic salmon genome is covered by DNA-transposons such as certain Tc1-Mariner types, with a major suspected role in the salmon-type rediploidisation [12]. Although I could not determine whether there is any true causality in the presence or absence of transposable elements, the lack of such elements in grayling is a more plausible hypothetical mechanism rather than the rare grayling-specific elements catalysing the grayling-type karyotype evolution. Although speculative, the observed differences in transposable element content between grayling and Atlantic salmon are a potential level of variation which may have facilitated the karyotype evolution of small grayling-type chromosomes which is distinctive in comparison to the large recombinant karyotypes of anadromous salmonids, as predicted by Qumsiyeh [22].

I observed a different response to thermal origin in the methylation than in the gene expression profiles. Contrary to my prior expectation, plasticity was nearly absent at the methylation level suggesting, that the methylome was not highly sensitive to the rearing temperature, at least at the natural-resembling temperature range and within generation (excluding the germ-line methylation changes). Instead, I found that the methylation profiles were diverged between (sub)populations. Interestingly, I found common patterns among (sub)populations from similar thermal origins, suggesting that epigenetic responses to altered temperature may have played a key role in the adaptation process in grayling, allowing them to expand and maintain their range and contemporary climatic niche breadth. Although some portion of the methylation divergence may reflect genetic drift, the observed clustering of the methylation profiles of the colder-adapted, but not warmer-adapted populations may highlight the increased importance of methylation variation under novel environmental conditions. A similar but not identical pattern was present in the nucleotide profiles of study III samples with the difference being that sympatric populations from Lesjaskogsvatnet, rather than colder thermal origin, were the most similar. This suggested that the methylation response consists at least partly of pure epialleles, meaning those not induced by the underlying nucleotide variation. To what extent the variation is unlinked to the genotypes, however, remains a topic of further study. A methylation response may be faster (divergence may appear within generation) [34] than a nucleotide response due to many reasons, such as the time lag (in generations) during which new favourable nucleotide allele combinations are brought together, the background selection of maladaptive nucleotide loci due to linked adaptive loci, and the scarcity of novel beneficial nucleotide variation [8].

Although the global methylation profile analysis did not uncover plasticity, some plastic differences were observed at the whole-methylome level. The observed global hypomethylation in the colder rearing temperature may be a stochastic process, such as methylome erosion under stress, or an adaptive response such as

compensation for the overall reduction in the pace of gene expression in cold. Although determining the underlying reason was beyond the scope of this study, the observation is consistent with prior research [92], [93] and may be of importance in the study system.

The observations of population-by-rearing temperature interactions in the site-specific differential methylation analysis were abundant, and consistently changed methylation loci between population pairs with warmer-to-colder adaptation history were enriched. These results supported my hypothesis of convergent epigenetic adaptation to temperature among populations with similar thermal environment. However, there was a tendency for consistent changes to occur within protein-coding and downstream sequences, rather than in the upstream regulatory sequences. This was rather surprising as my prior expectation was for changes to occur in upstream regulatory regions, as they are commonly reported as hotspots for local adaptation at the nucleotide level (e.g. [94]), and as gene expression is under epigenetic regulation [24]. However, also the importance of coding sequence and intron methylation have been acknowledged in the regulation of mRNA splicing of genes [24], which is a potential candidate mechanism for further study of epigenetic regulation in the system. Although cytosine methylation studies are often concentrated on the well-understood functional regions, some studies have investigated the role of other regions such as 3'UTR. 3'UTR methylation has been linked to gene and protein expression changes, but the functional relevance is still unclear [95]–[97].

The gene list analyses (study III) summarised the functions which may be under epigenetic control in grayling juveniles. Term enrichments were much more abundant in the lists of genes with CpG-rich than CpG-depleted promoters. CpG-rich promoters may associate with genes that may require long-term control, such as developmental and tissue-specific genes [30]. Such control may restrict the expression within certain cell types or time points. In the CpG-rich promoters, hypomethylation associated with relatively general terms such as gene expression and biosynthetic processes. The general terms may have broad relevance across different tissues. In contrast, hypermethylation associated with more specific terms related to cell-cell-signalling in the cellular membranes and, interestingly, muscle cells. The latter set of terms may be related to spatially and temporally restricted developmental functions [98].

The set of gene ontology terms related to genes with a population-by-rearing temperature interaction may provide a molecular explanation for some of the phenotypic differences that prior research has identified between grayling populations from different thermal origins. Also, those terms were relatively specific. The most enriched terms were related to binding of various biomolecules. Some of the functions may be easily linked with thermal adaptation. For example,

terms related to muscle tissue are in line with the observed differences in the density and composition of muscle mass between individuals from contrasting thermal origins both in grayling and in other salmonids [34], [45]. Membrane depolarisation during action potential was the most enriched term, although only suggestively significant after multiple test correction. The small number of genes in the category may explain the borderline significance. Despite this uncertainty, the function is particularly interesting as action potential is important both for muscle cells in general and for more specific functions such the heartbeat, which is a key teleost regulatory mechanism to maintain homeostasis when the body temperature changes [99]. Among the top candidate genes, we also found an L-type calcium channel subunit *cacna1d* (LOC106583449; score = 1,712; *e*-value < 0.0001), which is required to activate cell types such as heart tissue [100].

As the core units of muscle tissue and regulators of muscle contraction were found both among the individual genes with site-specific adaptive response and in enriched terms, the regulation of muscle tissue is a potential key phenotypic feature altered by epigenetic regulation in the grayling embryos. The absence of terms related to collagen, which is the most abundant protein structure in vertebrates, suggests that this observation is not biased by the abundance of myofibril tissue in the body.

The lack of gene ontology enrichments confirmed the observed low importance of plasticity at the methylation level. Also, the consistently plastic loci between warmer-to-colder rearing temperatures in the populations were not more abundant than what can be expected due chance. Thus, the combined evidence does not support a major role of CpG methylation as a regulator for plasticity in the juvenile teleosts.

Generally, the results of the fine-resolution population-genomic study agreed with the prior microsatellite studies. Low overall level of genetic differentiation was observed among the sympatric subpopulations of Lesjaskogsvatnet (pairwise F_{ST} ranging from 0.000 to 0.005). Similar levels of differentiation were previously estimated for the grayling subpopulations of Lesjaskogsvatnet using microsatellite data (pairwise F_{ST} estimates ranging from -0.006 to 0.036) [40]. Likewise, the overall pairwise F_{ST} was 0.002–0.030 among recently established whitefish populations [4] and 0.017-0.032 in guppies [3]. However, the observed isolation by distance among the populations of Lesjaskogsvatnet confirmed that the populations have indeed differentiated in response to spatial isolation despite high gene flow.

The genome scan revealed some major *P*-value peaks with significant differentiation between warmer and colder thermal origin and suggested that besides cytosine methylation, also nucleotide variation may be an important mechanism in the thermal adaptation process. For example, sortin nexin 2 (*snx2*) was located close

to a top candidate SNP in both LFMM and OutFLANK analyses. An important follow-up step to better understand the role of such loci in the thermal adaptation process would be to determine their effect sizes with respect to thermal adaptation -related traits. Many small peaks with suggestive genome-wide significance were also present, which could be a signal of minor effect loci, insufficient statistical power, or simply drift. Twenty five candidate SNP -associated genes (1 % of all candidate SNP -associated genes identified with at least one genome scan approach) were detected in all of the three genome scan approaches. Although available information on their individual functions was scarce, gene ontology analysis shed more light on the functional significance of these candidate SNPs. Seventy two terms (increasing the overlap to 34 % of all enriched terms) were found in all three genome scans. The terms were related to cell signalling, organisation, transmembrane transport, and embryonic development. Interestingly, terms related to skeletal system and striated muscle development were also present. Although low level of genetic variation has been suggested in the grayling system, adaptive variation may be localised in functionally important regions of the genome. The adaptive signal of the candidate SNPs was much stronger in some functional regions of genes, particularly in the splice sites and downstream sequences, but also upstream regions and coding sequences. This was interesting as the downstream sequences of genes are rarely identified as hotspots of functional importance. However, the observation was consistent with the increased abundance of potentially adaptive methylation loci within and downstream of the coding sequence. Adaptive variation may often target functions with alternative splicing, post-transcriptional and translational regulatory roles, [101], [102] rather than changing the structure or the expression level of a gene.

When comparing the lists of biological processes related to plastic gene expression and adaptive methylation variation to the processes related to adaptive nucleotide variation, a significant overlap was found suggesting that the gene expression level differentiation related to those processes may be simultaneously under genetic and epigenetic control. Future detailed comparisons of the different molecular levels of variation would benefit from sampling the same individuals for all molecular markers simultaneously. The increased similarity of annotations between molecular levels from the gene to the functional level complements the observed increase in plastic signal at the level of gene modules rather than single genes, and highlights the importance of unified systems-level approaches in the research efforts aiming to understand local adaptation.

5 Summary and Conclusions

The grayling populations of Lesjaskogsvatnet arose recently from a common ancestral gene pool and are now distributed across a strong temperature gradient pointing to the possibility of rapid thermal adaptation. Prior research has revealed consistent phenotypic changes between populations with contrasting thermal origins. The phenotypic changes have appeared even in subpopulations which are sympatric outside the breeding and early development season, resulting in high levels of gene flow. Samples had been collected from the study system across several years of work by multiple collaborators. As a result, I was able to study multiple large molecular data sets consisting of hundreds of individuals using multi-omics approaches.

The gene expression response was most distinctively affected by the embryonic rearing temperature and the attempts to detect adaptive signal at the level of single gene expression were inconclusive. This was surprising as clear protein-level divergence in the grayling system in response to thermal origin has been reported. However, an adaptive gene expression response was revealed in two co-regulated gene modules, suggesting that small changes in the expression levels of many genes may contribute to a phenotypic effect.

The chromosome-level genome assembly confirmed that the grayling genome has evolved through chromosomal rearrangements such as pericentric inversions, and one chromosome fission, which has also increased the number of chromosomes and chromosomal arms. Natural selection may have affected grayling karyotype evolution through mechanisms such as the lack of the transposable element types which have been associated with chromosomal fusions in the Atlantic salmon. The genome evolution processes typical to grayling may have been favoured for the increased mutation rate.

The genome-wide methylation data revealed that, unlike gene expression, the methylation levels were almost non-plastic i.e. not affected by rearing temperature. Instead, common methylation patterns among populations from similar thermal environments were detected using multiple different approaches, including principal component profiles, the overrepresentation of convergently changed methylation loci between thermal origins, and many individual loci affected by population-by-rearing temperature interaction.

Population genomics data confirmed early microsatellite studies by showing that although the overall differentiation (F_{ST}) was very low within Lesjaskogsvatnet, a signal of isolation by distance was present in neutral genetic divergence between subpopulations. Multiple fine-resolution genome scan approaches repeatedly identified many candidate regions around the genome, suggesting that also nucleotide evolution may have a role in the thermal adaptation process. I found evidence that selection may primarily act on standing genetic variation. Besides the large-effect loci which were detected, most of the peaks had only suggestive genome-wide significance. Thus, an overall substantial amount of variation is indeed present in the genome. Both methylation and nucleotide variation were enriched in some functional gene regions, such as coding and 3'UTR sequences, suggesting that those may be hotspots for local adaptation.

The findings of my thesis shed light on how changes in complex gene networks may produce targeted adaptive responses. By combining the evidence I conclude, that phenotypic responses at the juvenile stage may be produced by a subset of gene modules. I found candidate adaptive loci suggesting, that heritable molecular patterns may have shaped the standing genetic variation of the ancestral gene pool. The observed variation mitigates the previous uncertainty of whether sufficient variation is present for adaptive evolution to have a role in the adaptation process of grayling by revealing abundant nucleotide variation in a subset of genomic locations. The observed epigenetic profiles were similar yet not identical to the genetic profiles, suggesting that interplay between both epigenetic and genetic differentiation may be important for biological processes contributing to rapid local adaptation. Slight changes in a subset of key loci seem to contribute to a profound change at the functional level. The findings highlight the potential benefit of the application of genomic methodologies both when studying populations for fisheries management and also in the conservation efforts of other species, e.g. when determining the conservation entities. Also, the need to carefully assess the limitations of small-scale genetic markers is evident based on the results. Besides nucleotide variation, epigenetic variation may be an important level of biodiversity that contributes to the evolutionary potential and resilience of the populations, however the mechanisms of epigenetic regulation need more research.

Acknowledgements

The main financial supporter for this project was the Academy of Finland (projects 287342 and 302873), while support for various stages of the project was also provided by other instances such as The Norwegian Research Council (project 177728) and Turku University Foundation. The computational resources were provided by the CSC - IT Center for Science.

I would also like to acknowledge the many people who have supported me throughout the completion of this PhD thesis over many years. While the journey was difficult, I am proud of the work that I have accomplished and grateful for the lessons learned along the way.

Firstly, I would like to thank my supervisors, Professor Craig Primmer and Assistant Professor Spiros Papakostas. I appreciate your willingness to provide guidance and feedback throughout the thesis process. I am grateful for the support. Also the contributions of all other collaborators have been essential to my success and I thank all co-authors for the fruitful discussions. Thank you Sari Järvi and Professor Jon Brommer for giving me the last push to get this thesis finally out of my system. I would also like to acknowledge the staff and faculty at both the University of Turku and the University of Jyväskylä, who have provided me with a rigorous academic environment. I am grateful for the resources and support that were available to me.

I would also like to express my appreciation to my family and friends for sharing the full landscape of life, which has balanced the academic challenges. I thank my family for the encouragement and for believing in me, and my dear sons Jussi and Aaro, who will always be my largest achievements. Thank you Jani for everything. I am very grateful for having all old and new friends. I tried to make a list but could not bear forgetting any of you from it, so you just have to trust on my word that every one of you is important. Finally, I thank the people from the stables for sharing the most wonderful adventures with me. I am sure the best are yet to come.

11.10.2023

Tiina Sävilammi

List of References

[1] O. Savolainen, M. Lascoux, and J. Merilä, 'Ecological genomics of local adaptation', *Nat Rev Genet*, vol. 14, no. 11, pp. 807–820, 2013, doi: 10.1038/nrg3522.

[2] E. Christaki, M. Marcou, and A. Tofarides, 'Antimicrobial resistance in bacteria: mechanisms, evolution, and persistence', *J Mol Evol*, vol. 88, pp. 26–40, 2020, doi: 10.1007/s00239-019-09914-3.

[3] M. J. van der Zee *et al.*, 'Rapid genomic convergent evolution in experimental populations of Trinidadian guppies (*Poecilia reticulata*)', *Evolution Letters*, vol. 6, no. 2. pp. 149–161, 2022. doi: 10.1002/evl3.272.

[4] M. Crotti, E. Yohannes, I. J. Winfield, A. A. Lyle, C. E. Adams, and K. R. Elmer, 'Rapid adaptation through genomic and epigenomic responses following translocations in an endangered salmonid', *Evol Appl*, vol. 14, no. 10, pp. 2470–2489, Oct. 2021, doi: 10.1111/eva.13267.

[5] R. J. Fox, J. M. Donelson, C. Schunter, T. Ravasi, and J. D. Gaitán-Espitia, 'Beyond buying time: The role of plasticity in phenotypic adaptation to rapid environmental change', *Philos. Trans. R. Soc. Lond., B, Biol. Sci*, vol. 374, no. 1768, p. 20180174, 2019, doi: 10.1098/rstb.2018.0174.

[6] R. Lande, 'Evolution of phenotypic plasticity in colonizing species', *Mol Ecol*, vol. 24, no. 9, pp. 2038–2045, 2015, doi: 10.1111/mec.13037.

[7] N. J. Barson *et al.*, 'Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon', *Nature*, vol. 528, no. 7582, pp. 405–408, Dec. 2015, doi: 10.1038/nature16062.

[8] N. Barghi, J. Hermisson, and C. Schlötterer, 'Polygenic adaptation: a unifying framework to understand positive selection', *Nat Rev Genet*, vol. 21, no. 12, pp. 769–781, 2020, doi: 10.1038/s41576-020-0250-z.

[9] M. Sinclair-Waters *et al.*, 'Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon', *Genet. Sel. Evol*, vol. 52, no. 1, pp. 1–11, 2020, doi: 10.1186/s12711-020-0529-8.

[10] E. A. Boyle, Y. I. Li, and J. K. Pritchard, 'An Expanded View of Complex Traits: From Polygenic to Omnigenic', *Cell*, vol. 169, no. 7, pp. 1177–1186, 2017, doi: 10.1016/j.cell.2017.05.038.

[11] I. Höllinger, P. S. Pennings, and J. Hermisson, 'Polygenic adaptation : From sweeps to subtle frequency shifts', *PLoS Genet*, vol. 15, no. 3, p. e1008035, 2019, doi: 10.1371/journal.pgen.1008035.

[12] S. Lien *et al.*, 'The Atlantic salmon genome provides insights into rediploidization', *Nature*, vol. 533, no. 6020, pp. 200–205, 2016, doi: 10.1038/nature17164.

[13] Y. Van De Peer, S. Maere, and A. Meyer, 'The evolutionary significance of ancient genome duplications', *Nat Rev Genet*, vol. 10, no. 10, pp. 725–732, 2009, doi: 10.1038/nrg2600.

[14] M. Sémon and K. H. Wolfe, 'Consequences of genome duplication', *Curr Opin Genet Dev*, vol. 17, no. 6, pp. 505–512, 2007, doi: 10.1016/j.gde.2007.09.007.

[15] S. Ohno, *Evolution by Gene Duplication*. Springer, Berlin, Heidelberg, 1970. doi: 10.1007/978-3-642-86659-3.

[16] A. L. Hufton and G. Panopoulou, 'Polyploidy and genome restructuring: a variety of outcomes', *Curr Opin Genet Dev*, vol. 19, no. 6, pp. 600–606, 2009, doi: 10.1016/j.gde.2009.10.005.

[17] F. W. Allendorf and G. H. Thorgaard, 'Tetraploidy and the Evolution of Salmonid Fishes', in *Evolutionary Genetics of Fishes*, B. J. Turner, Ed., Boston, MA: Springer US, 1984, pp. 1–53. doi: 10.1007/978-1-4684-4652-4_1.

[18] D. J. Macqueen and I. A. Johnston, 'A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification', *Proc. Royal Soc. B*, vol. 281, p. 20132881, 2014, doi: 10.1098/rspb.2013.2881.

[19] L. H. Rieseberg, 'Chromosomal rearrangements and speciation', *Trends Ecol Evol*, vol. 16, no. 7, pp. 351–358, 2001, doi: 10.1016/s0169-5347(01)02187-5.

[20] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer, 'The gain and loss of genes during 600 million years of vertebrate evolution.', *Genome Biol*, vol. 7, no. 5, p. R43, 2006, doi: 10.1186/gb-2006-7-5-r43.

[21] R. Phillips and P. Ráb, 'Chromosome evolution in the Salmonidae (Pisces): an update.', *Biol Rev Camb Philos Soc*, vol. 76, no. 1, pp. 1–25, 2001, doi: 10.1017/s1464793100005613.

[22] M. B. Qumsiyeh, 'Evolution of Number and Morphology of Mammalian Chromosomes', *J.Hered*, vol. 85, no. 6, pp. 455–465, 1994, doi: 10.1093/oxfordjournals.jhered.a111501.

[23] K. J. F. Verhoeven, B. M. VonHoldt, and V. L. Sork, 'Epigenetics in ecology and evolution: What we know and what we need to know', *Mol Ecol*, vol. 25, no. 8, pp. 1631–1638, 2016, doi: 10.1111/mec.13617.

[24] E. J. Duncan, P. D. Gluckman, and P. K. Dearden, 'Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype?', *J Exp Zool B Mol Dev Evol*, vol. 322, no. 4, pp. 208–220, 2014, doi: 10.1002/jez.b.22571.

[25] S. Ecker, V. Pancaldi, A. Valencia, S. Beck, and D. S. Paul, 'Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity', *BioEssays*, vol. 40, no. 2, pp. 1–11, 2018, doi: 10.1002/bies.201700148.

[26] M. E. Potok, D. A. Nix, T. J. Parnell, and B. R. Cairns, 'Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern', *Cell*, vol. 153, no. 4, pp. 759–772, 2013, doi: 10.1016/j.cell.2013.04.030.

[27] I. S. Andersen, A. H. Reiner, H. Aanes, P. Aleström, and P. Collas, 'Developmental features of DNA methylation during activation of the embryonic zebrafish genome', *Genome Biol*, vol. 13, no. 7, p. R65, 2012, doi: 10.1186/gb-2012-13-7-r65.

[28] M. Turpin and G. Salbert, '5-methylcytosine turnover: Mechanisms and therapeutic implications in cancer', *Front Mol Biosci*, vol. 9, p. 976862, 2022, doi: 10.3389/fmolb.2022.976862.

[29] S. Kumar, V. Chinnusamy, and T. Mohapatra, 'Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond', *Front Genet*, vol. 9, p. 640, 2018, doi: 10.3389/fgene.2018.00640.

[30] A. M. Deaton and A. Bird, 'CpG islands and the regulation of transcription', *Genes Dev*, vol. 25, no. 10, pp. 1010–1022, 2011, doi: 10.1101/gad.2037511.

[31] M. R. Baerwald *et al.*, 'Migration-related phenotypic divergence is associated with epigenetic modifications in rainbow trout', *Mol Ecol*, vol. 25, no. 8, pp. 1785–1800, 2016, doi: 10.1111/mec.13231.

[32] W. M. Berbel-Filho, N. Berry, D. Rodríguez-Barreto, S. Rodrigues Teixeira, C. Garcia de Leaniz, and S. Consuegra, 'Environmental enrichment induces intergenerational behavioural and epigenetic effects on fish', *Mol Ecol*, vol. 29, no. 12, pp. 2288–2299, 2020, doi: 10.1111/mec.15481.

[33] D. Anastasiadi, N. Díaz, and F. Piferrer, 'Small ocean temperature increases elicit stage-dependent changes in DNA methylation and gene expression in a fish, the European sea bass', *Sci Rep*, vol. 7, no. 1, p. 12401, Dec. 2017, doi: 10.1038/s41598-017-10861-6.

[34] E. Burgerhout, M. Mommens, H. Johnsen, A. Aunsmo, N. Santi, and O. Andersen, 'Genetic background and embryonic temperature affect DNA methylation and expression of myogenin and muscle development in Atlantic salmon (*Salmo salar*)', *PLoS One*, vol. 12, no. 6, pp. 1–15, 2017, doi: 10.1371/journal.pone.0179918.

[35] T. G. Northcote, 'Comparative biology and management of Arctic and European grayling (Salmonidae, *Thymallus*)', *Rev Fish Biol Fish*, vol. 5, pp. 141–194, 1995, doi: 10.1007/BF00179755.
[36] T. P. Hurst, 'Causes and consequences of winter mortality in fishes', *J Fish Biol*, vol. 71, no. 2, pp. 315–345, 2007, doi: 10.1111/j.1095-8649.2007.01596.x.
[37] T. O. Haugen and L. A. Vøllestad, 'A century of life-history evolution in grayling', *Genetica*, vol. 112–113, pp. 475–491, 2001, doi: 10.1023/A:1013315116795.
[38] G. Thomassen, N. J. Barson, T. O. Haugen, and L. A. Vøllestad, 'Contemporary divergence in early life history in grayling (*Thymallus thymallus*)', *BMC Evol Biol*, vol. 11, no. 1, p. 360, 2011, doi: 10.1186/1471-2148-11-360.
[39] S. Papakostas *et al.*, 'Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions', *Nat Commun*, vol. 5, p. 4071, 2014, doi: 10.1038/ncomms5071.
[40] C. Junge *et al.*, 'Strong gene flow and lack of stable population structure in the face of rapid adaptation to local temperature in a spring-spawning salmonid, the European grayling (*Thymallus thymallus*)', *Heredity (Edinb)*, vol. 106, no. 3, pp. 460–471, 2011, doi: 10.1038/hdy.2010.160.
[41] M. T. Koskinen, J. Nilsson, A. J. Veselov, A. G. Potutkin, E. Ranta, and C. R. Primmer, 'Microsatellite data resolve phylogeographic patterns in European grayling, *Thymallus thymallus*, Salmonidae', *Heredity (Edinb)*, vol. 88, pp. 391–401, 2002, doi: 10.1038/sj/hdy/6800072.
[42] J. Freyhof, ' The IUCN Red List of Threatened Species 2011', 2011.
[43] J. Freyhof and M. Kottelat, 'The IUCN Red List of Threatened Species 2008'.
[44] M. T. Koskinen, T. O. Haugen, and C. R. Primmer, 'Contemporary fisherian life-history evolution in small salmonid populations.', *Nature*, vol. 419, no. 6909, pp. 826–30, 2002, doi: 10.1038/nature01029.
[45] K. D. Kavanagh, T. O. Haugen, F. Gregersen, J. Jernvall, and L. A. Vøllestad, 'Contemporary temperature-driven divergence in a Nordic freshwater fish under conditions commonly thought to hinder adaptation.', *BMC Evol Biol*, vol. 10, no. 1, p. 350, 2010, doi: 10.1186/1471-2148-10-350.
[46] T. O. Haugen, 'Early survival and growth in populations of grayling with recent common ancestors - Field experiments', *J Fish Biol*, vol. 56, no. 5, pp. 1173–1191, 2000, doi: 10.1006/jfbi.2000.1238.
[47] N. J. Barson, T. O. Haugen, L. A. Vøllestad, and C. R. Primmer, 'Contemporary isolation-by-distance, but not isolation-by-time, among demes of European grayling (*Thymallus thymallus, linnaeus*) with recent common ancestors', *Evolution*, vol. 63, no. 2, pp. 549–556, 2009, doi: 10.1111/j.1558-5646.2008.00554.x.
[48] F. Gregersen, T. O. Haugen, and L. A. Vøllestad, 'Contemporary egg size divergence among sympatric grayling demes with common ancestors', *Ecol Freshw Fish*, vol. 17, no. 1, pp. 110–118, 2008, doi: 10.1111/j.1600-0633.2007.00264.x.
[49] T. O. Haugen and L. A. Vøllestad, 'Population differences in early life-history traits in grayling', *J Evol Biol*, vol. 13, pp. 897–905, 2000.
[50] IPCC Working group II, *Climate Change 2014 - Impacts, Adaptation, and Vulnerability, Part B: Regional Aspects*. Geneva: Cambridge University Press, 2014. doi: 10.1007/s13398-014-0173-7.2.
[51] L. Jiang *et al.*, 'Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos', *Cell*, vol. 153, no. 4, pp. 773–784, 2013, doi: 10.1016/j.cell.2013.04.041.
[52] C. Best *et al.*, 'Epigenetics in teleost fish: From molecular mechanisms to physiological phenotypes', *Comp Biochem Physiol B Biochem Mol Biol*, vol. 224, pp. 210–244, 2018, doi: 10.1016/j.cbpb.2018.01.006.
[53] S. Varadharajan *et al.*, 'The grayling genome reveals selection on gene expression regulation after whole-genome duplication', *Genome Biol Evol*, vol. 10, no. 10, pp. 2785–2800, 2018, doi: 10.1093/gbe/evy201.

[54] A. Amores, J. Catchen, A. Ferrara, Q. Fontenot, and J. H. Postlethwait, 'Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication', *Genetics*, vol. 188, no. 4. pp. 799–808, 2011. doi: 10.1534/genetics.111.127324.

[55] S. M. Aljanabi, I. Martinez, S. Rural, W. C. P. Norte, and C. E. P. Brasilia, 'Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques', *Nucleic Acids Res*, vol. 25, no. 22, pp. 4692–4693, 1997, doi: 10.1093/nar/25.22.4692.

[56] L. Smeds and A. Künstner, 'ConDeTri - A Content Dependent Read Trimmer for Illumina', *PLoS One*, vol. 6, no. 10, p. e26314, 2011, doi: 10.1371/ journal.pone.0026314.

[57] S. Chen, Y. Zhou, Y. Chen, and J. Gu, 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, 2018, doi: 10.1093/bioinformatics/bty560.

[58] B. J. Haas *et al.*, '*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nat Protoc*, vol. 8, pp. 1–43, 2013, doi: 10.1038/nprot.2013.084.

[59] B. J. Haas, 'TransDecoder code repository'. Accessed: Dec. 30, 2017. [Online]. Available: http://transdecoder.github.io

[60] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, 'CD-HIT: Accelerated for clustering the next-generation sequencing data', *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[61] B. Langmead and S. L. Salzberg, 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, vol. 9, no. 4, pp. 357–359, 2012, doi: 10.1038/nmeth.1923.

[62] P. Danecek *et al.*, 'Twelve years of SAMtools and BCFtools', *Gigascience*, vol. 10, no. 2, p. giab008, 2021, doi: 10.1093/gigascience/giab008.

[63] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, 'Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation', *Genome Res*, vol. 27, no. 5, pp. 722–736, 2017, doi: 10.1101/gr.215087.116.

[64] A. C. English *et al.*, 'Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology', *PLoS One*, vol. 7, no. 11, pp. 1–12, 2012, doi: 10.1371/journal.pone.0047768.

[65] P. Rastas, F. C. F. Calboli, B. Guo, T. Shikano, and J. Merilä, 'Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example.', *Genome Biol Evol*, vol. 8, no. 1, pp. 78–93, 2016, doi: 10.1093/gbe/evv250.

[66] S. Kurtz *et al.*, 'Versatile and open software for comparing large genomes.', *Genome Biol*, vol. 5, no. 2, p. R12, 2004, doi: 10.1186/gb-2004-5-2-r12.

[67] F. Krueger and S. R. Andrews, 'Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics*, vol. 27, no. 11, pp. 1571–1572, 2011, doi: 10.1093/bioinformatics/btr167.

[68] S. Gao *et al.*, 'BS-SNPer: SNP calling in bisulfite-seq data', *Bioinformatics*, vol. 31, no. 24, pp. 4006–4008, 2015, doi: 10.1093/bioinformatics/btv507.

[69] H. Li and R. Durbin, 'Fast and accurate long-read alignment with Burrows–Wheeler transform', *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010, doi: 10.1093/bioinformatics/btp698.

[70] S. Anders, P. T. Pyl, and W. Huber, 'HTSeq-A Python framework to work with high-throughput sequencing data', *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015, doi: 10.1093/bioinformatics/btu638.

[71] C. Trapnell *et al.*, 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocoll*, vol. 7, pp. 562–578, 2012, doi: 10.1038/nprot.2012.016.

[72] F. Cunningham *et al.*, 'Ensembl 2015', *Nucleic Acids Res*, vol. 43, no. D1, pp. D662–D669, Jan. 2015, doi: 10.1093/nar/gku1010.

[73] D. A. Benson *et al.*, 'GenBank', *Nucleic Acids Res*, vol. 46, no. Database issue, p. D41, 2018.

[74] E. B. Rondeau *et al.*, 'The genome and linkage map of the northern pike (*Esox lucius*): Conserved synteny revealed between the salmonid sister group and the neoteleostei', *PLoS One*, vol. 9, no. 7, 2014, doi: 10.1371/journal.pone.0102089.

[75] B. J. G. Sutherland *et al.*, 'Salmonid Chromosome Evolution as Revealed by a Novel Method for Comparing RADseq Linkage Maps', *Genome Biol Evol*, vol. 8, no. 12, pp. 3600–3617, 2016, doi: 10.1093/gbe/evw262.

[76] A. L. Delcher, S. L. Salzberg, and A. M. Phillippy, 'Using MUMmer to Identify Similar Regions in Large Sequence Sets', *Curr Protoc Bioinformatics*, vol. 1, pp. 10–3, 2003, doi: 10.1002/0471250953.bi1003s00.

[77] S. Varadharajan *et al.*, 'The grayling genome reveals selection on gene expression regulation after whole-genome duplication', *Genome Biol Evol*, vol. 10, no. 10, pp. 2785–2800, 2018, doi: 10.1093/gbe/evy201.

[78] A. Yano *et al.*, 'The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids', *Evol Appl*, vol. 6, no. 3, pp. 486–496, Apr. 2013, doi: 10.1111/eva.12032.

[79] A. Frankish *et al.*, 'Ensembl 2018', *Nucleic Acids Res*, vol. 46, no. D1, pp. D754–D761, 2017, doi: 10.1093/nar/gkx1098.

[80] A. Roberts and L. Pachter, 'Streaming fragment assignment for real-time analysis of sequencing experiments', *Nat Methods*, vol. 10, no. 1, pp. 71–73, 2012, doi: 10.1038/nmeth.2251.

[81] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, 'Normalization of RNA-seq data using factor analysis of control genes or samples', *Nat Biotechnol*, vol. 32, no. 9, pp. 896–902, 2014, doi: 10.1038/nbt.2931.

[82] P. Langfelder and S. Horvath, 'WGCNA: An R package for weighted correlation network analysis', *BMC Bioinformatics*, vol. 9, p. 559, 2008, doi: 10.1186/1471-2105-9-559.

[83] D. Szklarczyk *et al.*, 'The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible', *Nucleic Acids Res*, vol. 45, no. D1, pp. D362–D368, 2017, doi: 10.1093/nar/gkw937.

[84] J. E. Brommer, 'Whither P_{st}? The approximation of Q_{st} by P_{st} in evolutionary and conservation biology', *J Evol Biol*, vol. 24, no. 6, pp. 1160–1168, 2011, doi: 10.1111/j.1420-9101.2011.02268.x.

[85] M. Foll and O. Gaggiotti, 'A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective', *Genetics*, vol. 180, no. 2, pp. 977–993, 2008, doi: 10.1534/genetics.108.092221.

[86] B. Louie, S. Bergen, R. Higdon, and E. Kolker, 'Quantifying protein function specificity in the gene ontology', *Stand Genomic Sci*, vol. 2, no. 2, pp. 238–244, 2010, doi: 10.4056/sigs.561626.

[87] D. H. Alexander and K. Lange, 'Enhancements to the ADMIXTURE algorithm for individual ancestry estimation', *BMC Bioinformatics*, vol. 12, 2011, doi: 10.1186/1471-2105-12-246.

[88] R. K. Hammond *et al.*, 'Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci', *Elife*, vol. 10, p. e62206, Jan. 2021, doi: 10.7554/eLife.62206.

[89] A. Alexa and J. Rahnenführer, 'Gene set enrichment analysis with topGO', *Bioconductor Improv*, vol. 27, pp. 1–26, 2009.

[90] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, 'REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms', *PLoS One*, vol. 6, no. 7, p. e21800, 2011, doi: 10.1371/journal.pone.0021800.

[91] H. Mäkinen, S. Papakostas, L. A. Vøllestad, E. H. Leder, and C. R. Primmer, 'Plastic and Evolutionary Gene Expression Responses Are Correlated in European Grayling (*Thymallus thymallus*) Subpopulations Adapted to Different Thermal Environments', *J. Hered.*, vol. 107, no. 1, pp. 82–89, 2015, doi: 10.1093/jhered/esv069.

[92] K. H. Skjærven, K. Hamre, S. Penglase, R. N. Finn, and P. A. Olsvik, 'Thermal stress alters expression of genes involved in one carbon and DNA methylation pathways in Atlantic cod

embryos', *Comp Biochem Physiol A Mol Integr Physiol*, vol. 173, pp. 17–27, 2014, doi: 10.1016/j.cbpa.2014.03.003.

[93] D. C. H. Metzger and P. M. Schulte, 'Persistent and plastic effects of temperature on dna methylation across the genome of threespine stickleback (*Gasterosteus aculeatus*)', *Proc. Royal Soc. B*, vol. 284, no. 1864, p. 20171667, 2017, doi: 10.1098/rspb.2017.1667.

[94] D. L. Stern and V. Orgogozo, 'The loci of evolution: How predictable is genetic evolution?', *Evolution*, vol. 62, no. 9, pp. 2155–2177, 2008, doi: 10.1111/j.1558-5646.2008.00450.x.

[95] G. Maussion *et al.*, 'Functional DNA methylation in a transcript specific 3′ UTR region of TrkB associates with suicide', *Epigenetics*, vol. 9, no. 8, pp. 1061–1070, 2014, doi: DOI: 10.4161/epi.29068.

[96] L.-J. Zhang *et al.*, 'Expression and epigenetic dynamics of transcription regulator Lhx8 during mouse oogenesis', *Gene*, vol. 506, no. 1, pp. 1–9, 2012, doi: 10.1016/j.gene.2012.06.093.

[97] M. H. McGuire *et al.*, 'Pan-cancer genomic analysis links 3'UTR DNA methylation with increased gene expression in T cells', *EBioMedicine*, vol. 43, pp. 127–137, 2019, doi: 10.1016/j.ebiom.2019.04.045.

[98] N. Elango and S. V. Yi, 'DNA methylation and structural and functional bimodality of vertebrate promoters', *Mol Biol Evol*, vol. 25, no. 8, pp. 1602–1608, 2008, doi: 10.1093/molbev/msn110.

[99] I. A. Johnston, 'Environment and plasticity of myogenesis in teleost fish', *J. Exp. Biol*, vol. 209, no. 12, pp. 2249–2264, 2006, doi: 10.1242/jeb.02153.

[100] M. Vornanen, H. A. Shiels, and A. P. Farrell, 'Plasticity of excitation-contraction coupling in fish cardiac myocytes', *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*, vol. 132, no. 4, pp. 827–846, 2002, doi: 10.1016/S1095-6433(02)00051-X.

[101] D. Griesemer *et al.*, 'Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution', *Cell*, vol. 184, no. 20, pp. 5247-5260.e19, 2021, doi: 10.1016/j.cell.2021.08.025.

[102] J. J. Turunen, E. H. Niemelä, B. Verma, and M. J. Frilander, 'The significant other: Splicing by the minor spliceosome', *Wiley Interdiscip Rev RNA*, vol. 4, no. 1. pp. 61–76, Jan. 2013. doi: 10.1002/wrna.1141.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2024

ISBN 978-951-29-9593-6 (PRINT)
ISBN 978-951-29-9594-3 (PDF)
ISSN 0082-6979 (Print)
ISSN 2343-3183 (Online)