The genetic basis for adaptation in natural populations

SANGEET LAMICHHANEY
Abstract

Many previous studies in evolutionary genetics have been based on few model organisms that can be reared at ease in the laboratory. In contrast, genetic studies of non-model, natural populations are desirable as they provide a wider range of adaptive phenotypes throughout evolutionary timescales and allow a more realistic understanding of how natural selection drives adaptive evolution. This thesis represents an example of how modern genomic tools can be effectively used to study adaptation in natural populations.

Atlantic herring is one of the world’s most numerous fish having multiple populations with phenotypic differences adapted to strikingly different environments. Our study demonstrated insignificant level of genetic drift in herring that resulted in minute genetic differences in the majority of the genome among these populations. In contrast, a small percentage of the loci showed striking genetic differentiation that were potentially under natural selection. We identified loci associated with adaptation to the Baltic Sea and with seasonal reproduction (spring- and autumn-spawning) and demonstrated that ecological adaptation in Atlantic herring is highly polygenic but controlled by a finite number of loci.

The study of Darwin’s finches constitutes a breakthrough in characterizing their evolution. We identified two loci, ALX1 and HMGA2, which most likely are the two most prominent loci that contributed to beak diversification and thereby to expanded food utilization. These loci have played a key role in adaptive evolution of Darwin’s finches. Our study also demonstrated that interspecies gene flow played a significant role in the radiation of Darwin’s finches and some species have a mixed ancestry.

This thesis also explored the genetic basis for the remarkable phenotypic differences between three male morphs in the ruff. Identification of two different versions of a 4.5 MB inversion in Satellites and Faeders that occurred about 4 million years ago revealed clues about the genetic foundation of male mating strategies in ruff. We highlighted two genes in the inverted region; HSD17B2 that affects metabolism of testosterone and MC1R that has a key role in regulating pigmentation, as the major loci associated with this adaptation.

Keywords: Adaptive evolution, Atlantic herring, ecological adaptation, seasonal reproduction, TSHR, Darwin’s finches, natural selection, beak, ALX1, HMGA2, ruff, lek, inversion, HSD17B2, MC1R
The measure of intelligence is the ability to change

- Albert Einstein
Cover images: (1) Herring in the Baltic Sea © Riku Lumiaro/Finnish Environment Institute (2) Medium ground finch © P. R. Grant, originally published in Grant & Grant (2014) and (3) Independent male ruff dominating Satellite male at lek © Torsten Green-Petersen
List of papers

This thesis is based on the following five papers


IV. **Sangeet Lamichhaney**, Fan Han, Jonas Berglund, Chao Wang, Markus Sällman Almén, Matthew T. Webster, B. Rosemary Grant, Peter R. Grant and Leif Andersson, A beak size locus in Darwin’s finches facilitated character displacement during a drought. *Science, in press.*


* These authors contributed equally

Reprints were made with permission from the respective publishers.
List of papers (not included in the thesis)


III. Mats E. Pettersson, Marcin Kierczak, Markus Sällman Almén, Sangeet Lamichhaney and Leif Andersson, A model-free approach for detecting genomic regions of deep divergence using the distribution of haplotype distances, Manuscript submitted.


* These authors contributed equally
Table of contents

1. Introduction .................................................................................................................. 13
   1.1. Adaptation ............................................................................................................. 13
   1.2. Genetics of adaptation ......................................................................................... 14
   1.3. The ‘Genomics’ of adaptation .............................................................................. 15
   1.4. Methods to identify loci underlying adaptation .................................................... 15
       1.4.1. Forward genetics approaches ...................................................................... 15
       1.4.2. Reverse genetics approaches ...................................................................... 18
       1.4.3. Candidate genes approaches ...................................................................... 19
   1.5. Background to the thesis projects ....................................................................... 20
       1.5.1. Atlantic herring ............................................................................................ 20
       1.5.2. Darwin’s finches ......................................................................................... 22
       1.5.3. Ruff .............................................................................................................. 25

2. Methods ......................................................................................................................... 28
   2.1. Sampling and collection of ecological data ............................................................. 28
   2.2. Whole-genome sequencing ................................................................................... 29
   2.3. Variant discovery and downstream analysis ......................................................... 29
   2.4. Genome-wide screen for genetic differentiation ................................................... 30
   2.5. Underlying genetic basis for adaptive divergence ................................................. 30
   2.6. Further characterization of candidate loci ............................................................. 31

3. Results and Discussion ............................................................................................... 32
   3.1. Herring .................................................................................................................. 32
       3.1.1. Paper I .......................................................................................................... 32
       3.1.2. Paper II ........................................................................................................ 34
   3.2. Darwin’s finches .................................................................................................... 39
       3.2.1 Paper III ......................................................................................................... 39
       3.2.2. Paper IV ...................................................................................................... 42
   3.3. Ruff ....................................................................................................................... 44
       3.3.1. Paper V ......................................................................................................... 44

4. Conclusion and future perspectives ........................................................................... 48

5. Acknowledgements ..................................................................................................... 53

6. References .................................................................................................................... 55
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALX1</td>
<td>Aristaless-like homeobox 1</td>
</tr>
<tr>
<td>BGI</td>
<td>Beijing Genomics Institute</td>
</tr>
<tr>
<td>CALM</td>
<td>Calmodulin</td>
</tr>
<tr>
<td>CENPN</td>
<td>Centromere protein N</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered regularly-interspaced short palindromic repeats</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>EMSA</td>
<td>Electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>ESR2a</td>
<td>Estrogen receptor 2a</td>
</tr>
<tr>
<td>FGF10</td>
<td>Fibroblast growth factor 10</td>
</tr>
<tr>
<td>FOXC1</td>
<td>Forkhead box C1</td>
</tr>
<tr>
<td>FST</td>
<td>Fixation index</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>GSC</td>
<td>Goosecoid homeobox</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome wide association studies</td>
</tr>
<tr>
<td>HCE</td>
<td>High choriolytic enzyme</td>
</tr>
<tr>
<td>HMGA2</td>
<td>High mobility group AT-hook 2</td>
</tr>
<tr>
<td>HSD17B2</td>
<td>Hydroxysteroid (17-Beta) dehydrogenase 2</td>
</tr>
<tr>
<td>IBD</td>
<td>Identity by descent</td>
</tr>
<tr>
<td>InDels</td>
<td>Insertions and deletions</td>
</tr>
<tr>
<td>KB</td>
<td>Kilobase</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>MB</td>
<td>Megabase</td>
</tr>
<tr>
<td>MC1R</td>
<td>Melanocortin 1 receptor</td>
</tr>
<tr>
<td>MRCA</td>
<td>Most recent common ancestor</td>
</tr>
<tr>
<td>mDNA</td>
<td>Mitochondrial deoxyribonucleic acid</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PRLR</td>
<td>Prolactin receptor</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
<tr>
<td>QTL</td>
<td>Quantitative trait loci</td>
</tr>
<tr>
<td>RDH14</td>
<td>Retinol dehydrogenase 14</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SciLifeLab</td>
<td>Science for Life Laboratory</td>
</tr>
<tr>
<td>SDR42E1</td>
<td>Short chain dehydrogenase/reductase family 42E, member 1</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SOX11</td>
<td>SRY (Sex Determining Region Y)-Box 11</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>TSHR</td>
<td>Thyroid stimulating hormone receptor</td>
</tr>
<tr>
<td>UTR</td>
<td>Un-translated region</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Adaptation

Living organisms inhabit diverse ecosystems on earth and these natural environments present a multitude of biotic and abiotic challenges. (1). Adaptation is an evolutionary process whereby a species becomes more able to live in such challenging environments (2). Every species that has successfully adapted to a certain habitat possesses some unique adaptive traits that may be structural, behavioral and physiological or can involve life-history parameters. Structural adaptations are morphological features of an organism (for instance, large beaks in certain Darwin’s finch species increase their ability to crack larger seeds). Behavioral adaptations are a set of actions that an individual carries out to increase its chance of survival and reproduction (for instance, unique lekking\(^1\) behavior in different morphs of male ruff allows them an increased access to females for mating). Physiological adaptation comprises of the response of a species to a particular stimulus (for instance, local adaptation of Atlantic herring populations to salinity gradients in the Baltic sea).

Adaptive evolution is driven by natural selection and is one of the important process that explains the diversity of life we find in nature (3). The origin of this diversity on earth and the mechanisms that drive species to change through time have fascinated scientists and general public since ancient times (4). One of the key challenges in evolutionary biology to date is to understand how natural selection drives evolution (5). A species can adapt to either abiotic environment (e.g. climate change), to another species (e.g. a competitor) or a combination of both abiotic and biotic forces which makes mechanistic dissection of adaptive evolution empirically difficult (6).

---

\(^1\) An aggregation of males with competitive displays to entice the visiting females for mating
1.2. Genetics of adaptation

Many biologists have contributed to our current understanding of adaptive evolution and no one is more famous than Charles Darwin, whose theory of evolution provided compelling arguments about the power of natural selection to produce remarkable phenotypic adaptation (7). But, Darwin himself did not know about the actual mode of heredity, i.e. “genes”, and later on, only after the Mendel’s laws of inheritance were accepted, genetics was incorporated into evolutionary theory (8). Genetic variation in a population provides flexibility to adapt to changing environment and is crucial for survival of the population over time. Genetic diversity is regarded as the fundamental resource upon which adaptation and speciation will depend (9).

Clues underlying specific evolutionary processes responsible for adaptation are known to be hidden in the “genes” that control organismal phenotypes (7). Therefore, knowledge about the genetic basis of adaptive traits has become an important requirement to understand the majority of hypotheses regarding the genetics of adaptation (10). Historical developments in the field of adaptation genetics have gone through three major transitions. The first era saw theoretical developments in population and quantitative genetics including the codification of population genetics theory by Fisher, Wright and Haldane (11–13). Dobzhansky (2) led the application of population genetic principles to natural populations. This era set a theoretical framework for the utilization of genetic data that was soon to follow.

The second era was characterized by a gradual increment in the availability of genetic data. Developments in protein electrophoresis methods opened up the door for assessing the levels of genetic variation in wider variety of organisms rather then being limited to certain model species (14). There was an increased interests in natural populations, asking questions about the levels of genetic variation for ecologically important traits. Later on, the development of Sanger sequencing methods became one of the landmarks of this era.

The third era began with the publication of first metazoan genome sequence (15) in 1998. In this last decade, we have experienced an enormous progress in high-throughput sequencing technologies. Their availability has allowed sequencing genomes at population scales and generating toolkits that has facilitated integration of ecological and genomic data (16). Applications of these methodological advances have now started to answer some of the previously elusive long-standing questions on adaptation and evolutionary biology (17).
1.3. The ‘Genomics’ of adaptation

Progress in genetic studies of adaptive evolution until recently had been constrained by the lack of resolution and the absence of a genomic perspective (16). Ironically, the majority of our knowledge on such genetic mechanisms was only based on a few model organisms that could be reared at relative ease in the laboratory, away from their natural environment (1). Genomic characterization of natural populations was always desirable as they provide a wider range of phenotypes across evolutionary timescales that allows a more realistic understanding on the origins of genetic variation and how natural selection drives adaptive evolution. Next Generation Sequencing (NGS) methods have not only made the existing genetic techniques cheaper and faster, but have provided ample opportunities for genomic studies on any ecologically interesting species (17). In addition, widespread availability of these technologies has also enhanced the collaborative efforts of diverse research communities working in a particular organismal system with multi-dimensional expertise (18). Cost-effective opportunities for individual research groups to create genomic resources for their favorite species are changing the research landscape for many evolutionary biologists and ecologists.

1.4. Methods to identify loci underlying adaptation

An important objective in the majority of the current studies in the genetics of adaptation is to identify the loci associated with adaptive phenotypes (19). The challenging aspect of narrowing down genomic regions of interest is usually done using three major approaches, namely forward genetics, reverse genetics and candidate gene studies; either alone or often combined (20, 21).

1.4.1. Forward genetics approaches

Forward genetics is a method to identify genomic regions of interest in cases where there is prior knowledge regarding the associated phenotype. These approaches perform genetic characterization of the adaptive trait in cases where the phenotypic response to selection has already been quantified. The methods include QTL\textsuperscript{2} mapping using pedigrees or association mapping (GWAS), both of which screen large number of molecular markers across the genomes of individuals that are segregating for adaptive trait of interest (19).

\textsuperscript{2} Quantitative trait locus, a genomic region that is linked to, or contains, the genetic factors that control a particular phenotype
QTL mapping (Figure 1) attempts to create recombinant populations and performs linkage analysis to map genomic loci that are associated with the trait of interest (23). These approaches have been extensively used in humans, livestock, as well as in other model organisms in the past (24). Unfortunately, applicability of such studies has been limited due to the requirement of large families to obtain enough recombinant offspring (19). In addition, mapping traits that are shaped by many minimal effect loci (i.e. complex traits) have had much less success (23) with QTL mapping approaches.
**Figure 2** Genome wide association studies (GWAS) compares common single nucleotide polymorphisms between two groups with different phenotypes. The results of GWAS are most often presented as a manhattan plots showing significance of trait-association across the genome. Adapted from (25) © 2014 Schierding, Cutfield and O’Sullivan.

GWAS (Figure 2) is the examination of genetic variants across the genome of different individuals to identify if any variant is associated with the phenotype of interest by exploiting linkage disequilibrium as measured across several hundred thousand markers (26). GWAS is used for mapping genes within a single population as opposed to a controlled intercross done in QTL mapping studies. In the last few years, such studies have led to many scientific discoveries about genes and biological pathways associated with certain diseases in human as well as new biological insights in other species (23). However, bridging the gap between loci identified using GWAS and the actual identification of causal variants is still challenging. In addition, in many cases, these studies have only been able to identify a small fraction of the genetic variation associated with the traits of interest (27).
1.4.2. Reverse genetics approaches

Reverse genetics is a method to identify genes in cases where there is no prior knowledge regarding the associated phenotype. These methods do not require quantification of phenotypic traits and are based on genome-wide screening of markers to detect footprints of selection (28). Other traditional approaches of reverse genetics not mentioned in this thesis include mutagenesis screens, or direct transgenic modification of candidate genes.

When a new mutation arises in a population that increases the fitness of carriers, natural selection will favor such individuals and thus allele frequency of the beneficial mutation increases over time (Figure 3A). Due to a process termed genetic hitchhiking, neutral alleles that are nearby to this beneficial mutation also tend to change their allele frequency and may gradually lead to fixation (29). This phenomenon will result in a reduction of genetic diversity resulting in the signature of a selective sweep. Genome-wide scans of nucleotide diversity\(^3\) (\(\pi\)), Tajima’s \(D\)\(^4\) and patterns of linkage disequilibrium\(^5\) are commonly used measures for detecting such signatures of selection within a single population (Figure 3B).

Fixation index \((F_{ST})\)\(^6\) is the most widely used method to identify loci showing genetic differentiation between populations. The availability of high density SNP genotyping chips and whole-genome sequencing have now allowed the development of new algorithms that account for haplotype structures and not only individual SNP frequency alone, for quantifying genetic differentiation among populations (30, 31).

---

3 The average number of nucleotide differences per site for a group of DNA sequences sampled  
4 Difference between two estimates of genetic diversity; mean number of pairwise differences and number of segregating sites  
5 Non-random association of alleles at different loci  
6 Ratio of average number of differences between pairs of chromosomes sampled within a population with average number sampled between different populations
Figure 3 Selective sweep screen (A) Under selection, a beneficial mutation will rise in frequency and in addition, nearby linked alleles also “hitchhike” along with it to higher frequency creating a region of selective sweep © 2008 Nature Education (B) Effect of selective sweep on genetic variation, Adapted from (32) © Annual Reviews, 2005.

One major challenge of such studies is to differentiate signatures of natural selection from signals generated due to genetic drift\(^7\), population bottlenecks\(^8\) and other demographic factors (33). There has been a lot of effort in recent times to develop statistical procedures to better distinguish signals of natural selection from genetic drift (34). Still there is much debate on the proportion of genomic loci under selection and drift (32). Studies on model organisms where genetic drift plays a non-significant role in comparison to other mechanisms that shape up the adaptive evolution could address many issues of this debate.

1.4.3. Candidate gene approaches

In contrast to forward and reverse genetics methods that attempt to scan the entire genome, candidate gene approaches focus on pre-specified genes. In the majority of cases, these genes have already been associated with similar phenotype in other species (35). In addition, such genes are also selected based on a *priori* knowledge regarding their function and possible biological impact on the particular phenotype.

These methods are relatively cheap and quick to perform and they have been extensively used in human (36) and domestic animals (37). Currently, as knowledge about genes and their interactions are increasingly becoming available, these studies are moving away from simplistic *single gene to trait*

---

\(^7\) change in the relative frequency of allele in a population due to random sampling of organisms

\(^8\) sharp reduction in population size (and thereby variation in gene pool) due to environmental events (such as floods, landslides, diseases, droughts) or human activities
approach to analysis of multiple genes that are potentially involved in biological pathways associated with a certain trait of interest (38).

However, for the majority of adaptive traits of interest, particularly in natural populations, only a handful of candidate genes are available to date. In addition, these approaches tend to get biased towards well-studied candidate genes. Due to the higher chances of false positives and negatives, results of these studies are also difficult to be replicated in subsequent follow-up studies (36). In today’s post-genomic era, rather than using candidate gene methods as a first step in genetic association studies, these have been effectively used to characterize regions already narrowed down by forward or reverse genetics approaches.

1.5. Background to the thesis projects

After the arrival of next-generation sequencing technologies, there have been revolutionary changes in population genetics research. The possibility of data collection from genomes of many individual at affordable costs have become a reality, even for smaller research groups (39). The generation of unprecedented amounts of genomic data has changed the scenario from exploring only a limited “snapshot” of the genome (for instance, mtDNA, microsatellites or few nuclear markers) to development and usage of genome-wide high-density markers (40). This has allowed inference of the evolutionary history and identification of genetic loci associated with a particular trait in any species at very high resolution. In addition, cost effective approaches of sequencing in “pools” of multiple individual DNA originating from a certain population have been successfully applied to infer demographic history and screen genomes for signatures of selection (41–43). This PhD project was motivated by the progress achieved in areas of genome analysis in the last decade. It aims to harness the power of high-throughput sequencing technologies to characterize genetic diversity in natural populations and understand the genetic basis of adaptive phenotypes.

1.5.1. Atlantic herring

Atlantic herring is a schooling, pelagic fish species that inhabits both sides of the North Atlantic Ocean and the Baltic Sea (Figure 4). It is one of the most abundant marine species with the number of individuals in a herring school being up to a billion (44).
The fecundity in herring females is about 30,000–70,000 eggs per female (45) and herring spawn in huge schools as females release eggs and males release clouds of milt simultaneously in the water. Atlantic herring migrate over great distances in open Sea and Ocean to the feeding grounds during the majority of the year and return back to their specified locations during spawning seasons from early spring to late fall.

Atlantic herring has been critical for the economic development of northern Europe in the past (46) and possess an important presence in scandinavian cuisine. Until imported food became available, salted herring was a crucial source of protein and energy, in particular during winter periods. In addition to human consumption, it is also widely used in fish feed and currently ranks among one of the largest fishery industries in the world (47).

As the habitat of Atlantic herring ranges throughout Baltic Sea and Atlantic Ocean, it shows remarkable adaptation to environmental variables as, for instance, salinity, temperature, light, feed resources and predators are strikingly different between the Baltic Sea and Atlantic Ocean. This has allowed the formation of distinct populations that are locally adapted to a particular environment (48). Several herring stocks have been recognized in the past (Figure 5) based on their morphological features, life history parameters, spawning locations and spawning time (49).
Despite clear phenotypic differences, no or only minute genetic differences between these stocks had been documented. Earlier studies based on limited genetic markers (allozymes (14, 50), microsatellites and SNPs (51–53)) have shown limited evidence of genetic differentiations between ecologically divergent stocks or populations with different spawning seasons (autumn-spawners and spring-spawners). Our aim was to generate high quality genome resources for this species and understand the genetic basis of population differentiations due to local adaptation.

1.5.2. Darwin’s finches

Darwin’s finches, studied on the Galápagos archipelago, have historic importance in the field of evolutionary biology as they provided some of the fundamental insights into processes of natural selection and adaptive radiation⁹ (54). This archipelago is geologically young and rose from the ocean less than five million years ago (55) and has experienced significant changes in its geography in the past due to regular shifts in sea level during periods of glaciation (56). These changes have resulted in a striking diversity of flora,

---

⁹ A process in which organisms diversify rapidly into multitude of new forms, particularly when changes in environment make new resources available, create new challenges and open environmental niches
fauna and overall landscape in the archipelago (56). These islands range from dry zones with prickly cacti, littoral zones with bushes and grasses to dense forests in humid zones with high elevation and rainfall (57). The ability of Darwin’s finches to adapt, survive and evolve in such diversified habitats have made them a classic textbook example for a variety of concepts in evolutionary biology (58).

Apart from a single Darwin’s finch species on Cocos Island, there were 14 recognized species in the Galápagos archipelago (Figure 6). Phylogenetic reconstructions from mtDNA suggested that Darwin’s finches were monophyletic and derived from a mainland common ancestor that colonized the archipelago about 2-3 million years ago (58). Evolution in Darwin’s finches is characterized by rapid adaptation to an unstable and challenging environment leading to ecological diversification and speciation. This has resulted in striking diversity in their phenotypes (for instance, beak types, body size, plumage, feeding behavior and song types) (Extended Data table 1, paper III). Beaks are one of the most diversified features in these birds and are well adapted to the type of food they eat; ranging from fine needle-like beaks in warbler finches that are perfect for picking up insects; long, sharp and pointed beaks in cactus finches for probing into cactus or deep, broad and blunt beaks in large ground finches suited for cracking large nuts and seeds (Figure 6).
Darwin’s finches have always been the subject of intense research; however, we knew little about how processes such as natural selection and gene flow have shaped the patterns of polymorphisms and divergence across the genomes of these birds to produce such astonishing phenotypic variation. Previous studies on the evolutionary history of Darwin’s finches were either based on mtDNA (58) or few nuclear microsatellites (59) and complete genomes of all 14 recognized species were as yet unexplored. In this study, we have characterized genetic diversity within and between each of these species in a genome-wide scale. We have explored their genomes to identify loci that most likely contributed to morphological diversification and understand their roles in adaptive evolution of these iconic birds.
1.5.3. Ruff

Ruff is a medium-sized wading bird that breeds in marshes and wet meadows across the Palearctic zone. It is a migratory bird that spends winter in the tropics (mainly in Africa) and breeds in the wetlands of northern Europe and Siberia during summer (Figure 7).

![Figure 7 Habitat of ruff across the year © Jimfbleak/Wikimedia Commons/CC-BY-SA-3.0/GFDL](image)

Ruffs show a marked sexual dimorphism, with males being larger than females (Supplementary Figure 4, paper V). In addition, males develop distinctive breeding plumage (ornamental feathers around their neck and head tufts) during the breeding season and show one of the most remarkable mating behaviors (60, 61). There are three strikingly different male morphs (Independents, Satellites and Faeders) that differ in behavior, plumage color and body size (Figure 8). Independents (80-95% of males) show spectacular diversity in color of ruff/head tufts and vigorously defend territories at leks\(^\text{10}\). Satellites (5-20% of males) usually have white ruff/head tufts, do not defend territories and display submissive behavior at leks. Faeder is a rare (< 1% of males) third morph, mimicking females by its smaller size and female-like plumage. Independents attract females by their elaborate performance display, high degree of aggression and strong colored plumage. Satellites allow Independents to dominate them, in exchange for getting closer to females visiting the territories occupied by Independents. Faeders being a female-mimic, get uninterrupted access to mating territories due to this disguise and

\(^{10}\) An aggregation of males with competitive displays to entice the visiting females for mating
attempt to mate with females accordingly (62). Adaptations in each of these male morphs allow them to take advantage of their unique courtship behavior to reproduce and survive.

Previous studies in ruff captive populations have suggested that the genetic polymorphism associated with development of either Satellite or Independent was an autosomal, dominant single locus with mendelian mode of inheritance (64, 65). Later on, following discovery of the third morph Faeder (66), the possibility of a third dominant allele at the same locus was supposed to account for the development of female-mimicking Faeder males (67). In addition, a microsatellite marker near MC1R was found to be linked with ornamental plumage in these birds (63) but a recent study did not find any coding changes in MC1R that were associated with plumage color variation (68).
It was an enigma, how could a single genetic locus explain such complex differences in these birds. In this study, we aimed to characterize the genetic architecture underlying phenotypic differences in these three male morphs and get insights into scenarios of adaptive changes that possibly could have led to the evolution of this spectacular mating system.
2. Methods

2.1. Sampling and collection of ecological data

Herring samples for paper I were collected between 1978-1980. The samples were originally used for studying genetic differentiation among herring populations based on isozyme loci using starch gel electrophoresis (14, 50). Interestingly, we could reuse them for whole-genome sequencing after 30 years. Hence these samples have literally witnessed transitions in genomic technologies in last few decades. For paper II, we used additional herring samples collected recently in 2012-2013 by our collaborators. Muscle tissues of each fish were used for DNA isolations. For transcriptomics studies, RNA extractions from multiple tissues (for instance, skeletal muscle, brain, eye, spleen, intestine and kidney) were used. For paper III and IV, Darwin’s finch samples were collected on the Galápagos archipelago and Cocos Island by our collaborators. Similarly, closely related tanagers were collected on mainland Barbados to be used as out-group. The birds were captured in mist nets and released after collecting blood samples on FTA papers. DNA was isolated from pieces of these FTA papers. For paper V, our collaborators provided ruff population samples and DNA from a fresh blood sample required for genome assembly was collected from the Helsinki Zoo. The DNA/RNA extractions were done using current standard protocols (for details, see methods for each paper).

A successful study on adaptation genomics is only possible by integration of the data from long-term life history and other ecological traits, history of biogeography and field experimental studies. Our collaborations with ecologists and evolutionary biologists helped us not only to collect samples for genetic studies but, most importantly, allowed us to access their rich ecological knowledgebase, establish research hypothesis and experimental designs. In addition, they were also crucial for interpreting the results and deciding on future directions.
2.2. Whole-genome sequencing

All studies within this thesis involved whole-genome sequencing of multiple populations of a particular species. We used two different approaches for genome sequencing of samples (individually or pooled). Individual sequencing is a standard approach where DNA of every individual is sequenced separately. For pooled sequencing, DNA of multiple individuals in each population was pooled together in equimolar concentrations and the resulting pooled DNA is sequenced. Sequence coverage\textsuperscript{11} is an important aspect that determines degree of confidence for discovery of genetic variants in such studies. For individual sequencing, we targeted the sequence coverage to \(\sim 10X\), and for pooled sequencing to \(\sim 30X\). Choice of individual or pooled sequencing depends on the research question. Pooled sequencing is a cost-effective method to scan genomes of multiple individuals across many populations to estimate allele frequency differences and identify most informative SNPs whereas individual sequencing provides additional advantages of performing haplotype based demographic inferences and other intra- and inter-population genetics analysis. We used Illumina next generation sequencers for whole-genome sequencing, which is a commonly used platform for these studies.

2.3. Variant discovery and downstream analysis

In the current era of wide-scale applications of NGS technologies in almost every biological discipline, methods and pipelines for analyzing these data have become quite robust and streamlined. After receiving the raw data (as short sequence reads of \(\sim 125\) bp in FASTQ format) from Illumina sequencing platforms, we first started the quality control (QC) steps by removing low quality sequences (generally at the start and end of the reads), trimming Illumina sequence adaptors, checking for possible contaminations and other errors during library preparation and sequencing protocols. These QC-passed sequences were then aligned against a reference genome assembly and sequence alignments for each individual/pool were compared against each other to identify genetic variants. We used stringent in-house filtering pipelines to select high-confidence calls from raw genetic variants. Single nucleotide polymorphisms (SNPs) were the most commonly used variants for genetic characterizations of the populations. Apart from SNPs, we also screened for other genetic variants such as small insertions and deletions (Indels) and large structural variants (inversion, deletion, duplication and translocation).

\textsuperscript{11} average number of reads sequenced for each base in the genome
2.4. Genome-wide screen for genetic differentiation

SNPs were used to generate the phylogenetic trees, which demonstrated vital information regarding genetic diversity within and between populations. These trees also provided genome-wide perspectives on the evolutionary history of populations. Phylogenetic relationships also guided us towards the choice of methods to be used for screening genetic differentiation among these populations. In addition, statistics generated from such phylogenetic inferences were also used to estimate time since divergence between different populations and from the most recent common ancestor (MRCA).

In our studies, we grouped populations according to a particular ecological adaptation they shared, for instance, Atlantic herring populations adapted to certain salinity conditions or spawning seasons (paper I and II); Darwin’s finches with certain beak shapes (paper III) or beak sizes (paper IV) or male morphs in ruff with different reproductive strategies (paper V). We then compared the allele frequency at each SNP across genomes of these groups to identify regions with increased genetic divergence between them. Such regions with high genetic divergence were potential candidate loci associated with certain phenotypic traits that allowed populations to adapt to a particular environment.

2.5. Underlying genetic basis for adaptive divergence

Once genomic regions with increased divergence between populations were identified, an important step ahead was to identify candidate genes/mutations within the region that might contribute to phenotypic differences. In majority of the cases, it was challenging because these genomic regions were often large and contains multiple genes. Hence, such genomic regions needed careful, fine scale examination using different methods of bioinformatics and functional annotations of the genetic variants before coming to any robust conclusions.

Genetic variation at candidate loci under selection for particular adaptive trait of interest must be shared between populations possessing a similar trait. To identify this shared genetic variation, we looked into the haplotypes within these regions in each population. Scanning haplotypes from multiple populations allowed us to locate the core IBD$^{12}$ segment associated with a phenotype. Having narrowed down the candidate region, we further annotated variants within this IBD segment to get overview of gene content and

$^{12}$ A DNA segment in two or more individuals that is identical by descent due to same ancestral origin
genomic distribution of these changes (e.g. intergenic, upstream/downstream, intronic, synonymous or non-synonymous) to understand what particular genomic category has been enriched in the candidate regions. Such analysis helped us to understand relative importance of genetic variations in regulatory and coding sequences, which has been one of the long-standing questions in evolutionary genetics.

We also analyzed gene ontology (GO\textsuperscript{13}) terms for genes within these candidate regions to understand if any of the GO terms are over (or under) represented. These analyses provided information regarding biological pathways associated with candidate regions and their contributions in the development of an adaptive phenotype.

We also examined sequence conservation of these genetic variants across published vertebrate sequence databases. The genetic variants showing high divergence between populations that were also highly conserved across the vertebrates were likely to possess functional significance underlying ecological adaptation. In addition, comparison against out-group populations allowed us to identify derived (or ancestral) state of the candidate locus in each population. Such analysis helped to understand evolution of candidate regions associated with the trait of interest.

2.6. Further characterization of candidate loci

Once genomic regions that were highly differentiated between populations were identified using bioinformatics approaches, we further selected diagnostic SNPs from each of these regions. These SNPs were genotyped in a set of individuals from each population using either a single SNP genotyping assay or custom-made SNP chip, depending upon the number of SNPs being used. Genotyping individual samples allowed us to confirm associations detected by whole-genome sequencing and to study the differentiated region in more detail. Genotyping data was also used to examine individual haplotype types in populations that were only sequenced in pools before, providing a better characterization of the differentiated locus. In the presence of quantitative phenotypic measurements, these data were also used to perform regression analysis to study their association with phenotypic diversity (\textit{paper III and IV}) and estimate the amount of phenotypic variance explained by the candidate locus.

\textsuperscript{13} Standard description of gene products in terms of their associated cellular component, biological processes and molecular functions (104)
3. Results and Discussion

3.1. Herring

3.1.1. Paper I

Previous genetic studies in herring were limited to allozymes, mtDNA, microsatellites or few numbers of SNPs and high resolution genomic resources were lacking. A major hurdle undertaking such studies in any non-model species like herring is to generate a high quality genome assembly. Generating long insert size libraries for large eukaryotic genomes are still costly and prone to error. Additionally, creating linkage maps and anchoring the correct order of sequences along the chromosomes is technically challenging. In this study, we developed a cost effective strategy that does not require a high-quality genome assembly for such genome-wide studies in non-model species (Figure 9).

![Exome assembly pipeline](image-url)

*Figure 9 Exome assembly pipeline, adapted from (69)*

The pipeline of this method first required a transcriptome, which was generated by cDNA sequencing of skeletal muscle from a Baltic herring. Furthermore, it utilized whole-genome sequences for alignment against transcripts
and added pieces of flanking sequences on both sides of exons to extend the transcriptome into an exome$^{14}$ assembly. An exome assembly provided better alignability of genomic reads at the exon/intron boundaries in comparison to a transcriptome assembly and also covered additional features of the genome by adding flanking intronic/intergenic sequences. Although this exome assembly was only ~ 6-8 % of whole genome (Table S1, paper I), most importantly it captured part of the coding sequences and flanking non-coding regions in the genome. This exome assembly was further used as a reference to characterize genetic variation in Atlantic herring populations.

Figure 10 Example of SNP showing no significant differentiation (A) and highly significant differentiation (B) among populations

$^{14}$ exons obtained from assembled mRNA transcripts with flanking sequences derived from genomic reads
We carried out whole-genome sequencing in eight pools of 50 fish each, sampled from geographic locations across Baltic Sea and Atlantic Ocean (Figure 1A, paper I). Mapping genomic reads against the exome assembly identified ~440,000 SNPs among these populations. Consistent with the results of low genetic differentiation among herring populations from previous studies, the great majority of these SNPs did not show significant allele frequency differences (Figure 10A). A small set of SNPs (2-3%) showed highly significant allele frequency differences (Figure 10B) and a phylogenetic tree based on these SNPs showed clear separation of populations (Figure 2B, paper I). We validated these results from pooled sequencing by individual genotyping of the same fish samples using a selected set of “neutral” and “differentiated” SNPs (Figure 2G, S2 and S3, paper I). These results provided for the first time, evidence of genetically differentiated populations in Atlantic herring.

A highly debated question in population genetics is whether the observed genetic differentiation is due to drift or natural selection. We did a simulation study to compare distribution of $F_{ST}$ in our data with the one expected for selectively neutral loci. The results indicated that there was a significant excess of loci showing extreme $F_{ST}$ values in our genome-wide screen than was expected under a genetic drift model (Figure D, paper I). The results confirmed that genetic drift has played a subordinate role in herring populations, consistent with their large population size and natural selection is possibly the major factor underlying genetic differentiation in herring.

This study was an important step in genetic characterization of population structure in Atlantic herring. Identifying loci under natural selection from the background of genetic differentiation due to drift has always been a challenge and Atlantic herring could be used as an excellent model system to address such issues.

3.1.2. Paper II

*Paper I* provided us a preliminary screenshot regarding SNPs showing significant allele frequency differences. In order to characterize the total number of independent loci in the genome that showed strong genetic differentiations among herring populations, to understand the patterns of genetic differentiation and annotate such genomic regions, a genome assembly was required. In this study, we generated a high quality herring genome assembly in collaboration with research groups from BGI, China and SciLifeLab, Uppsala.
Atlantic herring are exposed to diverse ecological conditions (for instance, salinity differences across Baltic Sea and Atlantic Ocean) and show a considerable variation in spawning time. In the context of genetic drift playing insignificant role in this species, herring provided us with a unique opportunity to explore the genetic basis of ecological adaptation in natural populations. In addition to the eight populations sequenced for paper I, we carried out whole-genome sequencing of 11 more populations across Baltic Sea and Atlantic Ocean and also included a population of Pacific herring as an outgroup (Figure 1A, paper II). The reference herring genome assembly was used to analyze genetic differentiation among these 20 populations.

![Figure 11 Neighbor-joining tree of Atlantic herring populations using ~6.5 million SNPs across the whole genome](image)

Previous studies have described divergence of Atlantic and Pacific herring as allopatric speciation due to geologic evolution of the Arctic-North Atlantic basin (70). Using the data from mtDNA, we estimated the date of split between Pacific and Atlantic herring to be ~2.2 million years ago. A phylogenetic tree provided a good indication on the genetic similarity of Atlantic herring populations and their divergence from Pacific herring (Figure 1C, paper II).

Although phylogeny of Atlantic herring was almost star-shaped, there were indications of genetic differentiation among populations (Figure 11). For instance, autumn-spawners from Baltic Sea clustered separately from spring-
spawning populations, which was possibly due to certain genomic regions that were distinct between autumn- and spring-spawning herring. In order to identify such regions, we analyzed the allele frequencies of each SNP across the genomes and identified ~100 independent loci that showed significant allele frequency differences between autumn- and spring-spawners (Figure 4A, paper II).

![Figure 12 Genetic differentiations across the TSHR locus between autumn- and spring-spawners](image)

The strongest signal overlapped TSHR that has an established role in photoperiodic regulation of reproduction in birds and mammals (71) (Figure 12). One of the major differences between populations that either spawn in spring or autumn was their response to varying day lengths and genetic variation in TSHR locus possibly had an important contribution towards this adaptation. Other strong signals included CALM, SOX11 and ESR2a (Figure 4C, paper II) that also had important functions in reproductive biology. The strong genetic differentiation at these loci indicated that they were major loci associated with preference of spawning season for a particular population.

Using similar methods, we identified ~500 independent loci showing strong genetic differences between populations adapted to Atlantic Ocean and Baltic Sea (Figure 3A, paper II). One of the important differences between these two habitats is the presence of a steep salinity gradient. It changes from 35‰ in the marine water in Atlantic Ocean to ~20-32‰ in Kattegat/Skagerrak zone and down to ~3-12‰ across the brackish Baltic Sea. Hence, certain genetic differences between these populations must be associated with salinity tolerance. Interestingly, one of the strong differentiated region overlapped PRLR (Figure 13) which had previously been associated with osmoregulation in fish (72). Another example of strongly differentiated region was around hatching enzyme HCE, whose association with salinity had been reported previously (73). HCE might be an important locus for adaptation, as hatching eggs in brackish condition for a marine fish must be a challenging stage of development.
The great majority of loci associated with either difference in choice of spawning season or adaptation to different habitats (Atlantic and Baltic) showed strong differentiation signals but no noticeable differences in the surrounding regions (Figure 12, 13). This result was consistent with low genetic differentiations in most regions of the genome and natural selection associated with ecological adaptation causing strong genetic differentiation at particular loci among herring populations.

There had been a lot of discussion in the past regarding relative importance of coding and regulatory mutations associated with a change in particular phenotype (74). We analyzed genomic distribution of loci showing strong differentiations and identified strong enrichment of non-synonymous as well as changes in UTR’s and 5kb upstream and downstream of the gene (Figure 6, paper II). The result indicated that both coding and regulatory changes contributed to adaptation in herring. The question regarding which of them contributed more could perhaps only be answered once the effect size of each of these loci is quantified.

The majority of loci that showed strong genetic differentiations occurred in larger blocks (Figure 12). We considered two possible explanations for occurrence of such larger blocks. One, it may be a result of recent selective sweep leading to genetic hitchhiking of neutral variants linked with causal
mutations. Second, such regions may include multiple causal mutations associated with the adaptive trait and to avoid the favoured haplotype to break up, there is selection for suppressed recombination and thus leading to large haplotype blocks. To explore more into these hypotheses, we compared nucleotide diversity in highly differentiated regions against random regions of the genome. The nucleotide diversity was significantly higher in larger blocks of highly differentiated regions compared to random regions both within and between divergent populations (Figure 5, paper II). These results indicated that highly differentiated regions occurring in larger blocks may possess multiple causal mutations and possibly were maintained during an evolutionary process rather than being a result of genetic hitchhiking due to recent selective sweep.

An important aspect of this study was the estimation of allele frequency in each population using a pooled-sequencing approach. Since the relative contribution of each individual in a pool might not be equal, an obvious question that might arise was regarding the reliability of these estimates. Therefore, results obtained from whole genome sequencing were verified by genotyping ~30 individual fish from each population using a custom-made 70K SNP chip. The same individual samples used for pooled–sequencing were used for genotyping. There was an excellent correlation in the allele frequency estimates from whole genome pooled sequencing and individual genotyping (Figure 3-S1, paper II), which indicated the reliability of allele frequency estimates obtained from pooled sequencing.

This paper constitutes a major advance regarding the genetic characterization of Atlantic herring. The availability of high-quality genome assembly and the annotation database of genes and population variants will be important resources for further studies in this species. Most importantly, it has provided a comprehensive view on the genetic architecture underlying ecological adaptations in herring and these results could be of seminal influence for designing adaptation studies in any natural population. This study has demonstrated that variation in spawning time in herring populations is not just a phenotypic plasticity but genetic factors must contribute. It has highlighted herring as a unique model organism where a large proportion of the genome shows minute genetic differentiation between populations. In contrast, natural selection is strong enough to cause genetic differences at loci underlying adaptation.
3.2. Darwin’s finches

3.2.1 Paper III

In this study, we have done extensive genomic characterization of the entire Darwin’s finch radiation (Figure 6) by whole-genome sequencing 120 individual birds that included all currently recognized species and two of their close mainland relatives. For certain species, we had sequenced samples from multiple islands (Extended data table 2, paper III) to understand how populations adapt and evolve under different environmental conditions. In contrast to herring, a draft genome assembly from one species of Darwin’s finches (medium ground finch) had been already generated (75) and we used this genome as a reference for our study.

Traditional taxonomy of Darwin’s finches was based on their morphology (Figure 14A) and previous phylogenetic studies have been limited to mtDNA and few microsatellites. Using our dataset of ~45 million sites that were variable within or between populations, we generated a high-resolution phylogenetic tree that was in general consistent with current taxonomy but had several interesting new features (Figure 14B).
The populations of sharp-beaked ground finches (*Geospiza difficilis*) from six different islands did not form a monophyletic group but rather separated into three distinct clusters. This result was consistent with an earlier version of taxonomy where these three groups were classified as separate species (77, 78). An ABBA-BABA test\(^\text{15}\) designed to study the evidence of introgression confirmed that *G. difficilis* populations from Wolf and Darwin is-

\(^{15}\) Study of the pattern of shared derived variants using a four-taxon test (105)
lands were species of mixed ancestry. The results indicated that majority of their genome is possibly shared with *G. magnirostris* or a similar close relative and portion of the genome affecting phenotypic characters was derived from ancestral *G. difficilis* population (Figure 2b, paper III). Interestingly, beaks in *G. difficilis* populations from Wolf are not as sharp-pointed as other populations but demonstrate slightly curved-shaped mandible which is a characteristic feature in blunt-beaked birds such as *G. magnirostris* (Figure 15 A). This typical beak morphology possibly supports the hypothesis of its mixed ancestry.

![Image A](image1.png)  
**Figure 15** Some species of Darwin’s finches, images reproduced from "How and Why Species Multiply: The Radiation of Darwin’s Finches by Peter R. Grant & B. Rosemary Grant, Copyright © 2008 Princeton University Press. Reprinted by permission.

Similarly, populations of the large cactus Finch (*G. conirostris*) from two islands, Genovesa and Española did not cluster together in the phylogenetic tree (Figure 14B). This result was consistent with differences in their beak shape, as the population from Genovesa has pointed beaks like *G. scandens* whereas Española population has blunt beaks similar to *G. magnirostris* (Figure 15B). Due to these discrepancies with traditional phylogeny, we proposed a revised taxonomy for sharp-beaked ground finches and large cactus finches (Supplementary text, paper III).

The most striking morphological diversity among Darwin’s finches are their beaks (Figure 6), which have been a popular subject of interest since the famous visit to Galápagos by Charles Darwin almost 200 years ago. Genome-wide comparisons among closely related species with blunt and pointed beaks (*G. magnirostris*/*G. conirostris-Española* vs. *G. conirostris-*
Genovesa/G. difficilis-Wolf, Figure 15) identified 15 candidate regions with striking allele frequency differences between these birds (Figure 3a, paper III). There were six genes (CALM, GSC, RDH14, ALX1, FGF10 and FOXC1) within these 15 regions that had been previously associated with craniofacial and/or beak development in birds and mammals (79–85).

The strongest signal was a 240-kb region overlapping ALX1 (Figure 3a,b paper III) where blunt beaked G. magnirostris and G. conirostris were homozygous for one haplotype (Figure 3d, paper III). Even though six populations of G. difficilis were distinct for major part of their genome (Figure 14B), they all shared the “pointed” beak haplotype at this locus, as per their species name suggests (sharp-beaked ground finch). A phylogenetic tree based on this 240 kb region indicated deep divergence between “blunt” and “pointed” beak haplotypes that occurred soon after the split between warbler and non-warbler finches (Figure 3c, paper III).

The ALX1 locus was segregating in medium ground finches (G. fortis) (Figure 3c, d paper III). Interestingly, field observations had shown that there was considerable beak shape diversity among medium ground finches (86). This provided an opportunity to study segregation at the ALX1 locus within a species. We genotyped an additional (n=62) medium ground finch for a diagnostic SNP and observed significant association with beak shape (Figure 3e, paper III). The two homozygotes had significantly different beak shapes measurements whereas heterozygotes were more intermediate, suggesting an additive effect of this locus on beak shape.

This study is a classic example on the usage of recent advancements in population-scale genomics coupled with detailed knowledge on ecology and evolutionary history collected over decades from extensive fieldwork to generate new insights into some age-old questions in evolutionary biology.

3.2.2. Paper IV
Multiple species may coexist in geographically overlapping habitat. Whenever there is competition between them for limiting resources, differences among species are accentuated, allowing divergence of their resource exploiting traits. Such ecological character displacement (87, 88) has been regarded as an important component of speciation. It has been difficult to obtain evidence of these events in nature (89, 90) and in many cases it has only been inferred. But, in a landmark study by Peter and Rosemary Grant (91), they actually saw it happening. Due to severe drought on the Daphne island in Galápagos during 2004-05, food availability declined, newly arrived large ground finches with large beaks dominated the feeding ground and with near removal of large seeds, larger-beaked birds among medium
ground finch did not have enough food to survive. Smaller-beaked birds survived better as they could eat small seeds that were relatively unrewarding to large beaked birds. This trait was passed onto future generations, as there was a strong shift towards smaller beak size in the next two generations.

Beak size was identified as a major factor affecting survival of medium ground finches during this episode of drought (91). Survivors and non-survivors did not differ in beak shape; hence as expected ALX1 locus that regulates variation in beak shape (paper III) was not associated with survival. In this study, we took advantage of a unique feature in Darwin’s finches; presence of triplet species only differing in size-related traits; large, medium and small ground and tree finches (Figure 1A, paper IV) to understand the genetic basis associated with beak size diversity and its role in character displacement during this particular episode of drought.

We sequenced ten additional birds from each of these large, medium and small ground and tree finch species to ~10X coverage and conducted a genome-wide FST screen that revealed seven independent genomic regions showing significant genetic differences among these birds (Fig 2A, Table S2, Fig S1, paper IV). The strongest signal was a ~525 kb region overlapping HMGA2, which was an excellent candidate locus for differences in growth patterns, as this gene is associated with human height, craniofacial distances (92), growth retardation in pygmy mouse (93) and body size in dogs (94). The ~525 kb HMGA2 locus revealed two major haplotypes associated with large and small birds that split even before the divergence of warbler and non-warbler finches about a million years ago (Figure 1C, paper IV). Most species appeared to be fixed for one haplotype but as expected, this locus was segregating in medium ground and tree finches.

As body and beak size are correlated traits, an important question was whether this HMGA2 locus was primarily associated with body size or beak size, or both. Similar as in paper III, we genotyped a diagnostic SNP from the HMGA2 locus in 133 additional medium ground finches and the results demonstrated a highly significant association with beak size rather than body size (Figure 2E, paper IV). Similar to the ALX1 locus, HMGA2 also had an additive effect with heterozygotes showing intermediate beak size compared with the two homozygotes.

Further, in order to investigate the possible role of HMGA2 locus in character displacement episode during the drought in 2004-05, we genotyped 71 medium ground finches from Daphne that had experienced the drought. Thirty-seven of these birds had survived the drought and 34 had died. Indi-
viduals with the SS genotype (associated with small beaks) survived best, LL individuals (associated with large beaks) survived worst and heterozygous had an intermediate survival rate (*Table 2F, paper IV*). These results indicated that *HMGA2* was a major locus controlling beak size that had facilitated rapid diversification of an adaptive trait during a character displacement episode.

*Paper III and IV* together have provided evidence of two loci with major effects on beak morphology (shape and size) that were the most important traits involved in major evolutionary changes during the adaptive radiation of Darwin’s finches.

### 3.3. Ruff

#### 3.3.1. Paper V

Ruff shows a remarkable lekking behavior where three different male morphs have evolved a unique courtship behavior to attract and gain access to females that are ready to mate. In order to understand the genetic basis of phenotypic differences among male morphs, we first generated a genome assembly from an independent male and further whole-genome sequenced an additional 15 *Independents*, nine *Satellites* and a single *Faeder* male. Comparing genomes of *Independents* and *Satellites* identified a strongly differentiated 4.5 MB region (*Figure 16, upper panel*). Apart from this region, there was no significant differentiation between these male morphs in the rest of genome (*Figure 1c, paper V*) indicating that this might be the locus controlling the observed phenotypic differences between the morphs.
Presence of a large region showing genetic differentiation may be explained by structural changes, in particular inversions. These structural changes reduce frequency of recombination in a region and allow accumulation of multiple changes leading to a large region of genetic differentiation. We screened for the presence of such structural changes throughout the genome using the sequence data and identified a 4.5 MB inversion in Satellites, perfectly overlapping with the differentiated region identified in the genome-wide screen (Figure 16, middle panel). A diagnostic PCR-based test confirmed the break points of this inversion and showed that all Satellites and Faeders were heterozygous for this inversion (Figure 16, lower panel). The inversion disrupts the CENPN, a gene that encodes the centromere protein N and its inactivation has severe deleterious effect in other species (96, 97). Therefore, we hypothesized that this inversion was recessive lethal and maintained by balancing selection. A recent study on pedigrees from a captive ruff population has now confirmed that this inversion was in fact recessive lethal (98).
As the inverted region was large with ~90 genes, it was challenging to identify candidate genes that might be associated with phenotypic differences. *Independents* have higher level of circulating testosterone in comparison to *Satellites* and *Faeders* (98) which possibly is associated with their aggressive behavior in leks. Hence, genes within this inversion that were associated with steroid metabolism were obvious candidate genes to look into.

Additional screening of structural changes within this inversion identified three heterozygous deletions in *Satellites* and *Faeders* that clustered in vicinity of HSD17B2 and SDR42E1 genes and deleted evolutionary conserved sequences (*Figure 2d, 2e and Supplementary Figure 6, paper V*). These two genes have important roles in the metabolism of sex hormones; for instance, HSD17B2 catalyzes conversion of testosterone to less active keto-forms. We postulated that these deletions constitute regulatory mutations affecting expression of these genes. For instance, the deletions around HSD17B2 in *Satellites* and *Faeders* possibly leads to overexpression of this gene, which enhances the degradation of testosterone leading to submissive behavior in leks. We also identified additional structural changes unique to *Faeders* (*Supplementary Figure 7, paper V*), which may contribute to its female-mimicking phenotype.

Apart from differences in behavior among the male morphs, they also possess striking diversity in plumage color (*Figure 8B*). One of the candidate genes associated with pigmentation, *MC1R*, was located within the inverted region. Coding changes in *MC1R* have been associated with plumage variation in several species of birds (*Figure 3, paper V*). We found that the *Satellite MC1R* sequence consisted of four missense mutations at positions that were identical among birds and mammals (*Figure 3, paper V*). We proposed that *MC1R* allele on *Satellite* chromosome and possible altered metabolism of sex hormones was associated with white color of ornamental feathers in *Satellites*. It was surprising that a previous study in ruff (68) did not identify any coding changes in *MC1R* that were associated with plumage color, probably due to errors in phenotyping or sequencing.

A closer examination of the inverted region revealed disruptions in the strong genetic differentiation between *Satellite* and *Independent* chromosomes at two regions (*Figure 2a, paper V*). Remarkably, a comparison of *Satellite* and *Faeder* chromosomes showed a mirror image with these two regions showing high genetic differentiation. We postulated that the *Satellite* chromosome arose by one or two rare recombination events between an *Independent* and a *Faeder*-like chromosome leading to such pattern of genetic differentiations (*Figure 2f, paper V*). Estimation of evolutionary age indicated that first inversion event occurred ~3.8 million years ago whereas second recombination event must have happened ~520,000 years ago.
This study has demonstrated how structural changes like an inversion of a chromosomal section can have profound effect on the evolutionary trajectory of an organism. The inversion constituted the starting point of an evolutionary process leading to suppression of recombination and allowing for accumulation of subsequent adaptive changes that ultimately led to the formation of three different male morphs in ruff.
4. Conclusion and future perspectives

The genetic basis of adaptive traits has been an elusive long-standing question in the field of evolutionary biology. But, progress in such studies were always hampered by the absence of a genomic perspectives and lack of enough genetic markers. Enormous progress in high-throughput sequencing technologies in the last decade has allowed sequencing genomes at a population scale and generated toolkits that have facilitated the integration of ecological and genomic data for any species. This PhD work has successfully utilized these methodological advancements in the areas of whole-genome sequencing and has provided some unique insights into the adaptive traits that have amazed evolutionary biologists for decades.

First, this thesis has characterized the genetic architecture for ecological adaptations in Atlantic herring in unprecedented detail that may have important implications for sustainable herring fishery management. Our results have established herring as a model organism to study natural selection where the noise due to genetic drift is of minor importance. Second, this thesis has revisited the iconic Darwin’s finches from Galápagos islands and revealed clues about the genetic foundation associated with phenotypic diversity in beak shape and size that played a key role in shaping up the evolution of Darwin’s finches. Finally, the study on ruff has highlighted the importance of large genomic structural changes in shaping up an adaptive trait, which we believe will be a textbook example for the evolution of alternative mating strategies in animals.

This thesis opens up a plethora of possibilities for future work. Herring populations mix when they migrate for feeding outside their spawning seasons. Fisheries, in general catch fish from such a mix and methods to distinguish a particular stock in these mixed catches are not well established. This might lead to overfishing of certain stocks, which has always been a challenging issue for herring fisheries. The comprehensive list of genetic markers identified in our study can be used to develop genotyping panels that can be used to characterize and classify respective herring stocks outside their spawning seasons. Such genotyping panels can be very useful tools in stock assessment programs of Atlantic herring.
Our findings on the genomic regions that show strong differentiation between autumn- and spring-spawning herring have indicated that spawning preference is not just influenced by the environmental conditions (for instance, availability of planktons, certain water temperature) but genetic factors have an important role to determine when does a particular population spawn. Diagnostic genetic markers from such regions can be used to develop genetic tools to distinguish autumn and spring spawning herring that would aid in proper management of stocks with different spawning preference. Candidate genes identified within these regions include some of those that have key roles in photoperiodic regulation of reproduction in birds and mammals. Genetic basis of photoperiodic regulation is not so well studied in fish and our results can provide opportunities for detailed studies in these aspects. Photoperiodic manipulations have been used to control body growth, egg production and fertilization success in aquaculture for commercial purposes. The candidate genes and associated genetic markers identified in our study can be of interest for marker-assisted selective breeding programs for these commercial aquaculture species.

This study has addressed only two of the major questions in Atlantic herring (a) adaptation of a marine species to low salinity and (b) genetic factors associated with timing of reproduction. There is a long list of research questions that can be explored using the data generated in this study. For instance, populations in Baltic Sea and Atlantic Ocean are also adapted to different feed resources, pathogen content, predators and temperature/light conditions. Whole-genome datasets generated in this study can be utilized to study the underlying genetic basis of such adaptations.

Autumn-spawning herring populations from southern-Baltic Sea used in our study that were sampled in 1970’s have been reported to have begun to disappear in recent times (99). An interesting aspect would be to study whether the whole population simply disappeared due to overfishing or if there was a change in spawning preference in this population. Additional sequencing of newly inhabited spring-spawning herring in these areas and their comparisons to the autumn-spawning samples collected in 1970’s might throw light into this important biological question.

The ongoing study in Atlantic herring could include additional geographic regions that would broaden present knowledge regarding the evolutionary history and population structure of this species. For instance, there are autumn- and spring-spawning herring on the west side of the Atlantic Ocean as well. An interesting follow-up study would be to find out whether the results of genetic regions associated with spawning could be replicated in the West-Atlantic populations or if a different set of genetic variants control these phenotypes in those populations. These comparisons would allow us to study
parallel evolution of genetic factors associated with a common ecological adaptation in geographically distant populations.

Similarly, there are different populations of herring in the White Sea and the taxonomic status of these populations is not very clear. Occurrence of remote populations of Pacific herring in the White Sea is known, whereas Atlantic herring also penetrates the White Sea from the west (100). A study of mitochondrial DNA in these populations had indicated evidence of introgression from Atlantic herring to Pacific herring in northern European waters (101). Whole-genome studies of these populations will certainly add a new dimension regarding ecological adaptation of two similar species (Pacific and Atlantic herring) in similar environmental conditions and identify if there are genomic signatures of introgression in these populations.

Our work on Darwin’s finches has revealed some of the underlying genetic variations explaining the diversity in beak shape and size which, probably is the most widely used illustration of morphological adaptation for exploitation of available food resource. In addition to beak diversity, Darwin’s finches also possess diversity in types of song they sing. Finch mating dynamics is known to be under the influence of song and previous studies have associated song evolution to speciation and adaptive radiation in Darwin’s finches (102, 103). Extensive on-field recordings of song for many species have been done and it would be interesting to utilize the sequence data we have generated to explore the genetics behind the vocal signals in species with contrasting song types.

Our results have indicated that certain populations of the same Darwin’s finch species that reside in different islands do not share similar genetic architecture. This probably is associated with the adaptation to variable environmental conditions on each island. For certain species, we have not yet explored the populations from multiple islands (for instance, 12 different established populations of medium ground finches have been described in the entire Galápagos archipelago). It would be useful to sequence additional populations of these species to investigate whether they behave as a monophyletic clade or they possess different genetic architectures, as was the case for G. difficilis and G. conirostris populations in our study.

Our results have also provided evidence of widespread interspecific gene flow throughout the radiation. Field studies have shown evidence of interspecies hybridization between G. fortis and G. scandens on Daphne Island in Galápagos archipelago that has resulted in large changes in average beak shape over the period of 30 years. Whole genome sequencing of these birds from pre- and post-introgression period would allow us to trace the genomic changes in these species as a result of introgression and understand if the
previously identified locus (i.e. ALX1) is the only genetic factor involved in this phenotypic shift or if there are additional loci involved. In addition, similar introgressive hybridization has led to the formation of an entirely new reproductively isolated lineage in Daphne Island. Whole genome sequencing of these samples will provide further insights on the importance of introgression among Darwin’s finches.

Our study in ruff has provided evidence of how major structural changes in the genome can lead to evolution of adaptive phenotypes. Results have indicated that an inversion itself and further accumulation of genetic changes within this inversion possibly altered the metabolism of sex hormones and pigmentation patterns, which were key for evolution of three male morphs. However, the molecular mechanisms behind these hypotheses remain yet to be unveiled. As a subsequent follow up, tissue samples can be collected from these birds during the breeding season and transcriptomics studies can be carried out. This will allow us to understand which of the genes within the inversion show differential expression among these male morphs during the breeding season. For instance, we have hypothesized that candidate genes such as HSD17B2 is over-expressed in Satellites leading to increased conversion of testosterone to less active keto-forms. These gene expression studies could test for such hypothesis generated from our current study. In depth analysis of differentially expressed genes will also provide clues behind the metabolic pathways associated with these phenotypic changes. In addition, transcriptome data can also be used to improve the current version of ruff genome annotation.

Even though genetic basis of alternative mating strategies was our major biological question in ruff, there are additional aspects that are yet unexplored. There is, for instance, a striking amount of phenotypic diversity in plumage color within Independents. We have access to collections of tissue samples and photographs from Independents with wide variety of plumage. Comparing genome sequences of the birds with different plumage types will be very useful in understanding the genetic mechanisms behind the plumage diversity in Independent males.

All three studies in this thesis have identified major loci associated with adaptive traits of interest. From a molecular genetics perspective, one important task moving forward is to identify causal variants and understand the cellular and molecular mechanisms that are involved in generating the phenotypic effects observed. The majority of markers in our study showing strong genetic differentiations are located either upstream or downstream of the gene, probably indicating that they may play an important role in gene regulation. Electrophoretic mobility shift assays (EMSA) is a technique that can be used to screen allele-specific gel shift differences in order to study
altered DNA-protein interactions and thus identify potential transcription factor binding sites. These in-vitro studies will allow us to diagnose if such variants are candidate causal mutations that contributed to the phenotypic differences. Similarly, functional assays can be developed using wild type and mutated constructs to characterize missense mutations in the candidate genes showing strong genetic differentiations. Genome editing tools such as CRISPR/Cas9 have generated considerable excitement in recent times for precise and reliable targeted changes in the genome. We can make use of such methods to manipulate candidate causal mutations in an appropriate animal model to mimic the trait of interest and study phenotype-genotype correlations in a controlled environment.

In today’s post-genomic era, high-throughput methods that translate vast wealth of data generated by genomics or transcriptomics projects into profiles at each stage of the flow of biological information from DNA → RNA → proteins → protein interactions are in their infancy. The technological advancements of these methods and their wise usage in the future will allow us to characterize the entire multi-dimensional space of biomolecules that will provide a more complete picture of how an adaptive phenotype has evolved. As gene products function together in biosynthetic networks, these results can complement the work of scientists studying related networks in any species, even well beyond the field of adaptation genomics.
5. Acknowledgements

Leif, there are no enough words to explain how grateful I am to get an opportunity to work with you. You have been an inspirational mentor to me. I can now proudly say that I have worked with you, one of the best geneticists in the world. The way you identify the research questions and design brilliant ideas to solve them, have inspired and motivated me, literally everyday. Apart from being an outstanding scientist, you are also a great human being. Although you have a large number of personnel working with you on multiple projects, your tireless efforts to organize everything and taking care of everyone is one of those unique qualities, which I guess is a rarity. The research excellence and amazing environment in your group made my PhD never a “work” for me, but a sort of fun adventurous trip to unearth some hidden mysteries of the nature.

I had a fantastic opportunity to receive supervision from Calle and Alvaro, who are among the best in our business. Before I arrived in Sweden, I was just an ordinary veterinarian with limited knowledge in bioinformatics and research skills. I have learnt most of the workflow skills required for my PhD from both of you. Thank you for all your in-depth suggestions and critical insights that has been the backbone of my PhD.

The cherishing aspect of this PhD has been the excellent collaborations with experts from all over the world. To get an opportunity to work alongside Peter and Rosemary Grant on the iconic Darwin’s finches has been one of those “dream come true” moments to me and I would like to extend my sincere gratitude to them for sharing their expertise and their vast wealth of knowledge. I also take this opportunity to thank Nils, Linda, Arild, Michele, Fredrik and Jacob for all the fruitful collaborations.

I am very grateful to Matt for all the insightful discussions and suggestions. You have been my primary resource for any technical questions regarding population genetics. Thank you for allowing me to knock your office door anytime I needed. Very special thanks to Nima, Jonas, Fan, Doreen, Mats, Susanne Kerje, Markus, Manfred, Görel and Marta for your valuable contribution to my thesis. Thank you Ulla, Jessica and Ulrika for all your help on wet-lab stuffs whenever I needed them. I have enjoyed the company
of brilliant researchers within our department and I am thankful to Patric, Andreas, Erik, Marcin, Ronnie and Miguel for your expert suggestions whenever I needed them. I would also like to thank Björn, Jacques, Marc and the whole SciLifeLab BILS, WABI and sequencing platforms for being extremely helpful collaborators. I am also grateful to all the personnel at IMBIM administration for taking care of all the practical aspects of my PhD program.

I would like to take this opportunity to thank Göran Andersson, my first mentor in Sweden. It was your suggestion that I should contact Leif for my master’s thesis project and the rest is history. Thank you for the valuable advices and insights every time I got an opportunity to meet you. Special thanks for sharing your rare books about Darwin’s finches. I would also like to thank Kerstin for being an inspiring personality and a source of motivation.

I have been blessed with friends and colleagues who have made my experience in the lab very special. Nima and Shady, we three have shared a long partnership, way back from our master’s. A heartfelt thanks for all your support throughout, which has played an important role in my growth as a researcher. I have had some quality times with Jonas, Freyja, Iris, Matteo, Fan, Chao, Fabiana, Doreen, Dora, Anna, Ann, Angela, Sam, Sharadha, Martin, Axel, Rakan, Ginger, Jessika, Chungang, Khurram and Shumaila both in and outside the lab and it has been a lot of fun. I am fortunate to know you all. A big thanks to my “SLU gang”, Lin, Merina, Agnese, Chrissy, Daniela, Bingjie, Sandrine, Nancy, Andre and Viviane for sharing all the good times together.

Living far way from your home country makes you alone and miss your family. But love, warmth and support I received from Malin and her whole family here in Sweden made sure that I didn’t miss my home too often. Thank you not only for providing a physical space in your house where I could spend all these years in Sweden, but also accepting me as a part of your extended family. I would also like to thank Maite, Regine and Matthias for your friendship, support and care.

Most importantly, none of this would have been possible without the love and blessing of my family. I would like to thank my wife Durga for standing besides me in all rough and happy times and for her continued love, support and encouragement. Special thanks for all your patience that is needed to take care of a crazy husband like me.
6. References


86. Grant, PR Grant BR (2014) 40 Years of Evolution: Darwin’s Finches on Daphne Major Island (*Princeton University Press*).


A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)