UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1301

Exploring genetic diversity in natural and domestic populations through next generation sequencing

NIMA RAFATI

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2017

ISSN 1651-6206
ISBN 978-91-554-9821-4
urn:nbn:se:uu:diva-315032

Dissertation presented at Uppsala University to be publicly examined in B42, BMC, Husarg. 3, Uppsala, Thursday, 30 March 2017 at 13:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Craig Primmer (University of Turku).

Abstract

Rafati, N. 2017. Exploring genetic diversity in natural and domestic populations through next generation sequencing. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 1301. 62 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9821-4.

Studying genetic diversity in natural and domestic populations is of major importance in evolutionary biology. The recent advent of next generation sequencing (NGS) technologies has dramatically changed the scope of these studies, enabling researchers to study genetic diversity in a whole-genome context. This thesis details examples of studies using NGS data to: (i) characterize evolutionary forces shaping the genome of the Atlantic herring, (ii) detect the genetic basis of speciation and domestication in the rabbit, and, (iii) identify mutations associated with skeletal atavism in Shetland ponies.

The Atlantic herring (*Clupea harengus*) is the most abundant teleost species inhabiting the North Atlantic. Herring has seasonal reproduction and is adapted to a wide range of salinity (3-35‰) throughout the Baltic Sea and Atlantic Ocean. By using NGS data and whole-genome screening of 20 populations, we revealed the underlying genetic architecture for both adaptive features. Our results demonstrated that differentiated genomic regions have evolved by natural selection and genetic drift has played a subordinate role.

The European rabbit (*Oryctolagus cuniculus*) is native to the Iberian Peninsula, where two rabbit subspecies with partial reproductive isolation have evolved. We performed whole genome sequencing to characterize regions of reduced introgression. Our results suggest key role of gene regulation in triggering genetic incompatibilities in the early stages of reproductive isolation. Moreover, we studied gene expression in testis and found misregulation of many genes in backcross progenies that often show impaired male fertility. We also scanned whole genome of wild and domestic populations and identified differentiated regions that were enriched for non-coding conserved elements. Our results indicated that selection has acted on standing genetic variation, particularly targeting genes expressed in the central nervous system. This finding is consistent with the tame behavior present in domestic rabbits, which allows them to survive and reproduce under the stressful non-natural rearing conditions provided by humans.

In Shetland ponies, abnormally developed ulnae and fibulae characterize a skeletal deformity known as skeletal atavism. To explore the genetic basis of this disease, we scanned the genome using whole genome resequencing data. We identified two partially overlapping large deletions in the pseudoautosomal region (PAR) of the sex chromosomes that remove the entire coding sequence of the *SHOX* gene and part of *CRLF2* gene. Based on this finding, we developed a diagnostic test that can be used as a tool to eradicate this inherited disease in horses.

Keywords: Ecological adaptation, seasonal reproduction, Atlantic herring, domestication, speciation, rabbit, skeletal atavism, Shetland ponies, NGS, SMRT sequencing, genome, transcriptome, assembly, structural variation, genetic diversity, HCE, TSHR, SHOX, CRLF2

Nima Rafati, Department of Medical Biochemistry and Microbiology, Box 582, Uppsala University, SE-75123 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-554-9821-4
urn:nbn:se:uu:diva-315032 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-315032)

“The knowledge of anything is not acquired or complete unless it is known by its causes.”

Avicenna

To my family

Cover images: Herring by Zem Linki, rabbit by José Blanco-Aguiar, and horse by Lisa Andersson. Designed by Yasaman Azodifar.

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

I. Lamichhaney S*, Martinez Barrio A*, **Rafati N***, Sundström G*, Rubin C-J, Gilbert E.R, Berglund J, Wetterbom A, Laikre L, Webster M.T, Grabherr M, Ryman N, and Andersson L, (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences* U.S.A. 109:19345–19350.

II. Martinez Barrio A*, Lamichhaney S*, Fan G*, **Rafati N***, Pettersson M, Zhang H, Dainat J, Ekman D, Höpner M, Jern P, Martin M, Nystedt B, Liu X, Chen W, Liang X, Shi C, Fu Y, Ma K, Zhan X, Feng C, Gustafson U, Rubin C-J, Sällman Alme M, Blass M, Folkvord A, Laikre L, Ryman N, Ming-Yuen Lee S, Xu X, and Andersson L, (2016) The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *Elife* 5:1–32.

III. **Rafati N***, Blanco-Aguiar J.A, Rubin C-J, Sayyab S, Sabatino S.J, Afonso S, Feng C, Celio Alves P, Villafuerte R, Ferrand N, Andersson L, and Carneiro M, The early stages of species formation revealed by a genomic map of clinal variation across the European rabbit hybrid zone. *Manuscript.*

IV. Carneiro M*, Rubin C-J*, Di Palma F, Albert F.W, Alföldi J, Martinez Barrio A, Pielberg G, **Rafati N**, Sayyab S, Turner-Maier J, Younis S, Afonso S, Aken B, Alves J.M, Barrell D, Bolet G, Boucher S, Burbano H.A, Campos R, Chang J.L, Duranthon V, Fontanesi L, Garreau H, Heiman D, Johnson J, Mage R.G, Peng Z, Queney G, Rogel-Gaillard C, Ruffier M, Searle S, Villafuerte R, Xiong A, Young S, Forsberg-Nilsson K, Good J.M, Lander E. S, Ferrand N, Lindblad-Toh K, and Andersson L, (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345: 1074–1079.

V. **Rafati N***, Andersson L.S*, Mikko S, Feng C*, Raudsepp T*, Pettersson J, Janecka J, Wattle O, Ameur A, Thyreen G, Eberth J, Huddleston J, Malig M, Bailey E, Eichler E.E, Dalin G, Chowdary B, Andersson L, Lindgren G, and Rubin C-J, (2016) Large deletions at the SHOX locus in the pseudoautosomal region are associated with skeletal atavism in Shetland ponies. *G3 Genes|Genomes|Genetics* 6: 2213–2223.

* These authors contributed equally

Related works by the Author

(Not included in this thesis)

I. Saenko S.V, Lamichhaney S, Martinez Barrio A, **Rafati N**, Andersson L, and Milinkovitch M, (2015) Amelanism in the corn snake is associated with the insertion of an LTR-retrotransposon in the OCA2 gene. *Scientific Reports* 5: 17118.

II. Seroussi E, Cinnamon Y, Yosefi S, Genin O, Smith J.G, **Rafati N**, Bornelöv S, Andersson L, and Friedman-Einat M, (2015) Identification of the Long-Sought Leptin in Chicken and Duck: Expression Pattern of the Highly GC-Rich Avian leptin Fits an Autocrine/Paracrine Rather Than Endocrine Function. *Endocrinology* 157: 737–751.

III. Feng C*, Pettersson M*, Lamichhaney S*, Rubin C-J, **Rafati N**, Caisni M, Folkvord A, and Andersson L, Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *Manuscript submitted.*

IV. Shumaila S, **Rafati N**, Carneiro N, Andersson G, Andersson L, Rubin C-J, A computational method for detection of structural variants using Deviant Reads and read pair Orientation: DevRO. *bioRxiv*.

V. Zamani N*, Torabi Moghadam B*, **Rafati N***, Lamichhaney S*, Sundström G, Lantz H, Martinez Barrio A, Komorowski J, Clavijo B.J, Jern P, and Grabherr MG, A local web interface for protein and transcript aligners: Smörgås. *Manuscript submitted*.

* These authors contributed equally

Contents

Abbreviations

aCGH	array comparative genomic hybridization
AFLP	Amplified fragment length polymorphism
AR	Androgen receptor
BAC	Bacterial artificial chromosome
CALM	Calmodulin
CHRM3	Cholinergic receptor muscarinic 3
CNV	Copy number variation
CRISPR	Clustered regularly interspaced short palindromic repeats
CRLF2	Cytokine receptor-like factor 2
ddPCR	Droplet digital PCR
DNA	Deoxyribonucleic acid
DNA-seq	DNA sequencing
EIF4G1	Eukaryotic translation initiation factor 4 gamma 1
EMSA	Electrophoretic mobility shift assay
eQTL	Expression quantitative trait loci
eRNA	Enhancer RNA
EST	Expressed sequence tag
ESTR2a	Estrogen receptor 2a
FMN2	Formin 2
Fst	Fixation index
GO	Gene ontology
GRIK2	Glutamate receptor, ionotropic, kainate 2
GWAS	Genome wide association study
HCE	High choriolytic enzyme
INDEL	Insertion-Deletion
ISS	Idiopathic short stature
Kb	Kilo base
KDM6B	Lysine-specific demethylase 6B
KIT	Tyrosine kinase
KLF4	Kruppel like factor 4
lncRNA	Long non-coding RNA
LWD	Léri-Weill dyschondrosteosis
Mb	Mega base
miRNA	Micro RNA
mtDNA	Mitochondrial DNA

mya	Million year ago
NGP	Next generation phenotyping
NGS	Next generation sequencing
NR6A1	Nuclear receptor subfamily 6, group A, member 1
OAT	Ornithine aminotransferase
PABPC1L2A/B	Poly(A) binding protein cytoplasmic 1 like 2A and B
PAR	Pseudoautosomal region
PAX2	Paired box 2
PCR	Polymerase chain reaction
piRNA	Piwi-interacting RNA
QC	Quality control
QTL	Quantitative trait loci
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RRGS	Reduced representation genome sequence
SA	Skeletal atavism
SHOX	Short stature homeobox
SLC12A3	Solute carrier family 12 (sodium/chloride transporter), member 3
SMRT	Single molecule real time
SNP	Single nucleotide polymorphism
snRNA	Small nucleotide RNA
SOX11	SRY(Sex determining region Y)-box11
SOX2	SRY(Sex determining region)-box 2
SV	Structural variation
TF	Transcription factor
TSHR	Thyroid stimulating hormone receptor
TSS	Transcription start site
TTC21B	Tetratricopeptide repeat domain 21B
UTR	Untranslated rigon

Introduction

Fluctuating biotic and abiotic interactions result in the continual necessity for species to adapt to environmental change. Darwin described these variation as an essential component for evolution [1]. Natural selection acting on genetic variation yields adaptive changes that are transmitted to the next generation [2, 3]. Inherited adaptive features may be in the form of physiological, behavioral, or morphological traits. An example of an adaptive physiological characteristic is tolerance of salinity in a marine species. An example of an adaptive morphological characteristic is color patterning, such as melanism in the peppered moth. Meanwhile, an example of an adaptive behavioral trait is the collective shoaling behavior exhibited by many fish. Evolution may also occur via the artificial selection of desired traits. Examples of this are height in horses, and tameness in domestic animals such as the dog and domestic rabbit.

When Darwin published his influential work on evolution, he was unaware of the mechanism behind the inheritance of characteristics. Rediscovery of Mendel's law of inheritance, in the early 20th century, revealed the mode of inheritance for phenotypes such as hair color controlled by a single gene (known as monogenic inheritance). This breakthrough introduced "genetics" as a tool to study the link between hereditary information and phenotypes (*e.g.* [4]). However, most biological traits show a polygenic inheritance that is controlled by multiple genes, often in combination with environmental factors. The study of such traits had to await the development of biometric methods developed by the founders of population genetics namely Wright, Haldane, and Fisher [5].

Advances in molecular techniques and the discovery of new markers fundamentally transformed biological research. Molecular markers enabled researchers to screen for genetic variation and differentiation within and between populations and species. In the last decade, the emergence of next generation sequencing (NGS) technologies has provided a new avenue to study genetic variation at an unprecedented resolution. These technologies produce large amounts of data that allow the exploration of genomes, epigenomes, transcriptomes, and proteomes. Such studies are no longer limited to model organisms, and whole genome sequencing of diverse living systems has become feasible at a reasonable cost. Indeed, the application of "pan-

omics[a]" data has yielded valuable insights into population histories, the genetic basis of speciation, adaptation, and diseases across a very large range of taxa [6].

Studying genetic diversity

Understanding the evolutionary forces shaping genetic diversity has implications in agriculture and animal breeding strategies [7], human and animal health [8], and the conservation of endangered species [9]. Natural selection drives adaptive evolution, in which individuals well-adapted to their environmental conditions are more likely to survive and reproduce than less well-adapted individuals. Early studies on adaptive evolution were based on protein markers (allozymes). These studies uncovered protein polymorphisms in natural populations and humans (reviewed in [10]). Later, DNA-markers were used to explore genetic variation, including restriction fragment length polymorphisms (RFLPs) and amplified fragment length polymorphisms (AFLPs). Microsatellites are another class of polymorphic marker that took full advantage of PCR-methods. Microsatellites have been widely used for studying population structure, paternity testing, and constructing genetic maps in many species [11, 12]. Finally, single nucleotide polymorphisms (SNPs) are today, by far the most commonly utilized and convenient markers in genetic studies (*e.g.* [13-15]).

Until recently, most studies on genetic diversity were restricted to a limited number of loci due to the cost of genotyping. Yet, decreases in the costs and development of genotyping techniques made it possible to screen thousands of markers using SNP chips, which is the most widely used method for genome wide association studies (GWAS). The development of NGS technologies allowed to sequence whole genomes and expand genetic analysis beyond a small subset of genes or genomic regions.

Whole genome screening has been focused mostly on humans and model organisms, for which reference genome assemblies were available. Whole genome assembly for non-model organisms was not trivial due to the costs and limitations associated with sequencing and computation. As an alternative to generate a reference assembly, researchers have often used available reference genomes from closely related species. However, this approach may be error prone because of mapping biases and chromosomal rearrangements. In the absence of a reference genome, a reduced portion of the genome can be used as a reference. There are three major methods for constructing a partial genome assembly (Table 1).

[a] Using collection of NGS data from genome, epigenome, transcriptome, and proteome to study biological mechanisms in multidimensional space.

Table 1. *Major methods for constructing a partial genome assembly.*

Method	Limitations
Sequencing transcriptome or expressed sequence tags (EST)	-Assembly is fragmented
Exon capture	-Prior knowledge about gene models is needed, but a closely related species gene model can be used
Reduced representation genome sequencing (RRGS)	-It may preclude many potentially informative markers and may fail to reveal genetic differentiation at high resolution

The key characteristic of most NGS technology is the generation of large amounts of sequence, typically in fragments each in the range of 100-200 base pairs long. Processing and assembling such data demand considerable computational resources and present many challenges. One of the main challenges is controlling sequencing errors. This can be overcome by increasing sequencing depth. In complex regions, such as large repeats and tandem duplications, there is still limited application of these technologies. Recently, Single Molecule Real Time (SMRT) sequencing technology has opened a new realm in characterizing complex genetic structures by producing longer reads (average length greater than 10 kb and up to 60 kb). One of the latest sequencing technologies is optical mapping, which is a technique for mapping the order of restriction enzyme sites over several millions of bases in length. This method has wide applications to: (i) validate genome assemblies, (ii) complete draft genome assemblies, and, (iii) to detect large structural changes.

NGS application for studying genetic diversity

By reducing costs of large assays and improving quantity and quality of sequencing output, the application of NGS has become the gold standard in evolutionary biology studies. Genetic variation can be explored by whole genome sequencing (WGS) in individuals and populations. A cost-effective approach to studying genetic diversity within and between populations is "pooled sequencing" (papers I-V); DNA from several individuals is pooled in equimolar quantities and sequenced at fairly high depth to infer allele frequencies and identify genetic differentiation between groups. With this approach, one can screen and quantify different forms of variation (see next section), characterize their effect on coding sequences, evaluate their association with adaptive traits [16], disease status [17], and identify footprints of selective forces [13, 18].

In addition to the genome (DNA-seq), NGS have been utilized to study the transcriptome. RNA sequencing (RNA-seq) provides a large amount of

information that offers to enrich our understanding of gene expression and regulation. This technique has overcome limitations associated with microarrays, such as low sensitivity and specificity, probe cross-hybridization, and the ability to detect novel genes. RNA-seq is used to catalog long non-coding RNAs (lncRNA), micro-RNA (miRNA), and small RNA (for instance snRNA and piRNA) that are involved in protein translation and chromatin modulation [19]. Moreover, RNA-seq provides information about transcriptional start sites (TSS) as well as enhancer RNAs (eRNAs), which play an important role in transcriptional regulation [20].

RNA-seq data can be used to build a transcriptome map by two methodologies, *de novo* transcriptome assembly in the absence of a reference genome (paper I), and genome-guided transcriptome assembly when a reference genome assembly is available (paper II-V). Other applications of RNA-seq are novel transcript/isoform discovery as well as detection of gene fusions in cancer [21]. Quantifying gene expression has been a distinct application of RNA-seq in molecular biology (paper III). Using these data we can explore differences in expression associated with disease [22], adaptation [23, 24], and domestication [25]. Recently, RNA-seq has been used to identify polymorphisms, perform allelic imbalance analyses [26], and detect expression quantitative trait loci (eQTL) with a broader dynamic range, for instance in speciation studies [27, 28].

Classes of genetic variation

Two common classes of DNA polymorphisms detected by WGS are SNPs and small insertions and deletions (INDELs). SNPs have been extensively used in linkage and association studies of diseases [29] as well as genomic selection for economic traits in animals and plants [30, 31]. SNP data have also been widely used to study population structure and adaptive traits in natural and domestic populations [14, 32, 33].

Structural variants (SVs) constitute unbalanced forms[a] of variation such as copy number variation (CNV), insertion, duplication, deletion, and balanced forms[b] including translocation and inversion (Figure 1) [34]. CNVs constitute a substantial fraction of SVs and some have functional significance. For instance, a CNV at the *KIT* gene is associated with white spotting in pigs [18]. In addition to CNVs, inversions may contribute to the evolution of adaptive traits in natural populations [35]. As an example, inversions associated with local adaptation have been reported in stickleback [36]. In humans, SVs are also common and account for ~1% of genome variation [37].

[a] In unbalanced structural variation, DNA segments are lost or gained.
[b] In balanced structural variation, the location or orientation of a DNA segment is changed without losing or gaining new DNA sequence.

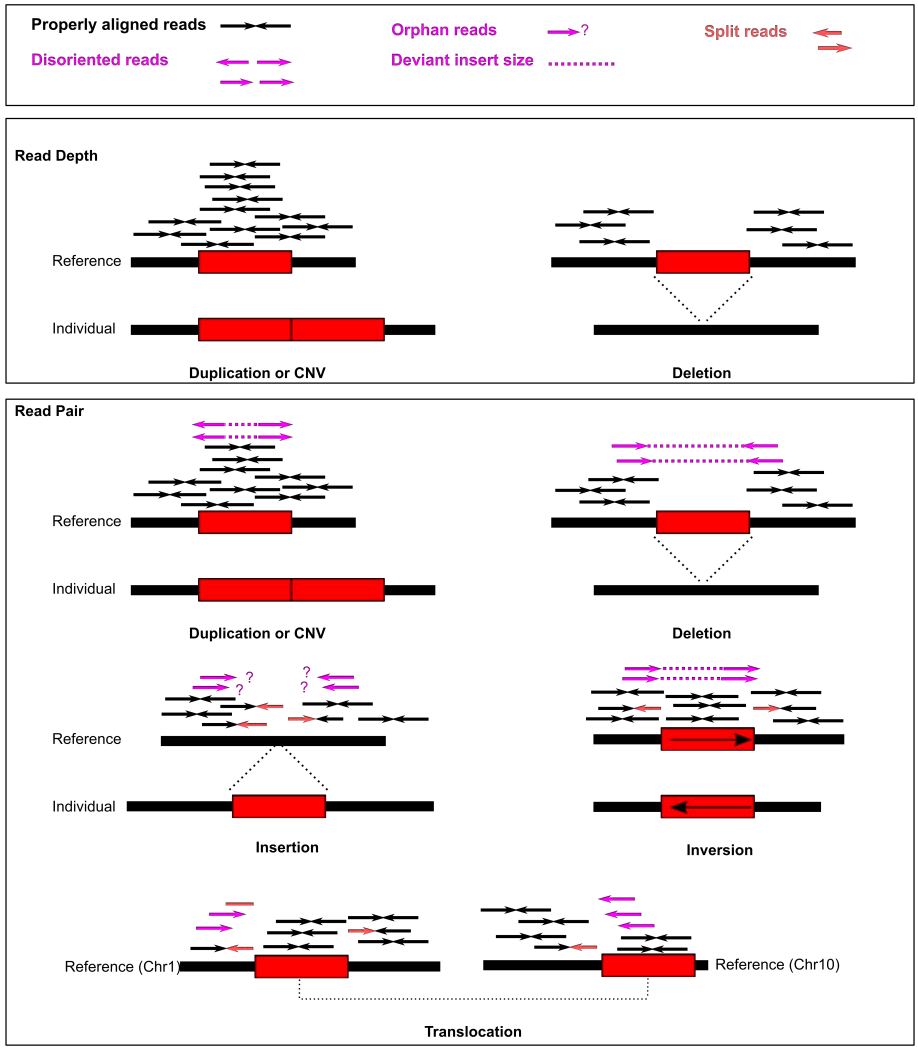

Figure 1. Different classes of SVs and methods commonly used for their detection. Read Depth: detecting structural variation using read depth. Read Pair: detecting structural variation using read-pair information including read orientation and insert size[a].

Identification of loci under selection

Mutations introduce variation into the genome and, depending on their influence on an individual's fitness and the potential effects of genetic drift, their frequency in the population will rise or fall. Furthermore, the fate of any new mutation is also influenced by selection at linked loci. Allele frequencies at

[a] Insert size is referred to distance between read pairs.

neutral sites tend to change if there is a nearby beneficial mutation, under a phenomenon known as genetic hitchhiking [38]. Depending on the strength of selection, genetic diversity is reduced as the selected variant increases in frequency and finally becomes fixed in the entire population forming a "hard sweep" (Figure 2A and B). In a "soft sweep" as opposed to hard sweep, selection may act on standing genetic variation or on multiple new mutations. Consequently, selection will only cause moderate reduction in genetic diversity at linked neutral loci (Figure 2C-F).

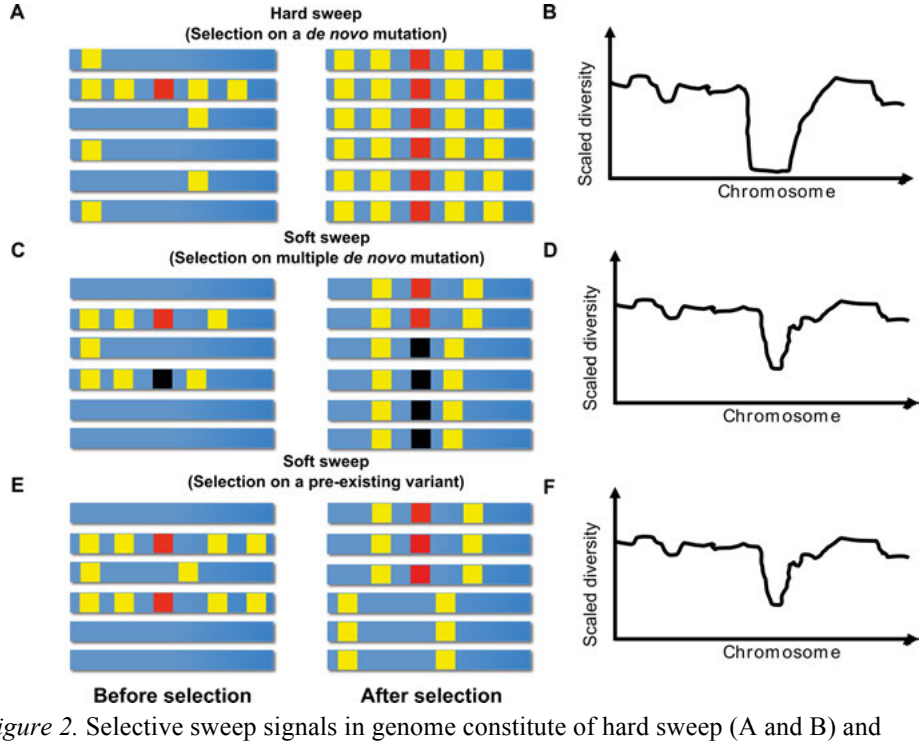

Figure 2. Selective sweep signals in genome constitute of hard sweep (A and B) and soft sweep (C-F).

Advanced sequencing technologies put at our disposal the ability to screen the entire genome of many species and individuals and explore the evolutionary forces producing and maintaining genetic diversity among populations. Depending on the sequencing data and the availability of reference genome or transcriptome, different approaches are employed to make use of this information. Prior information such as morphometric data (*e.g.* pigmentation pattern) or ecological parameters (*e.g.* salinity gradient or habitat) complements this analysis. Two methods commonly exploited to study association between genetic factors and phenotypes are quantitative trait loci (QTL) mapping and GWAS.

QTL mapping attempts to detect and locate signals associated with traits of interest using genotyping data of individuals in pedigrees or in experimental crosses. The success of QTL studies lies in having a sufficient number of individuals for the trait under study and accurate phenotyping data. GWAS attempts to use linkage disequilibrium between genetic markers and causal mutations contributing to phenotypic variation by contrasting one group as control and another group with the phenotype of interest (*e.g.* with disease status). Both methods have been successfully used in many studies to identify loci underpinning adaptive traits without requiring whole genome data (*e.g.* [33, 39]). However, both approaches share some limitations. In QTL mapping, the genome is scanned at low resolution and unanchored sequences are usually ignored [40]. In addition, dissecting the genetic basis of complex traits where many loci with small effect are involved remains challenging [40, 41].

WGS can complement these methods by generating unprecedented amounts of data to assess genetic diversity at high resolution. To detect differentiated regions, we can simply quantify intra- and inter-population genetic diversity by generating different statistics (Table 2).

Table 2. *Commonly used statistics to estimate levels and patterns of inter- and intra-population genetic diversity.*

Intra-population statistics		
	Tajima's D	The difference between the average number of pairwise differences and total number of segregating sites
	Linkage disequilibrium (LD)	Nonrandom occurrence of alleles at different loci
	Nucleotide diversity (π)	The average frequency of nucleotide differences between two random sequences from one population
	Haplotype statistics (EHH and iHH)	Series of statistics that detect the decay of LD.
Inter-population statistics		
	Wright fixation index (*Fst*)	Most commonly used statistic displaying allele frequency differences between populations.
	d_{xy}	Scales divergence between two populations by capturing the mean number of variable sites in a randomly selected sequence

In an ecological context, where measuring the phenotype of individuals is limited or difficult, other approaches based on ecological parameters are employed to study evolutionary forces underlying genetic divergence between populations. To demonstrate adaptation across gradients of abiotic and biotic environmental factors, we can study clinal patterns of variation (paper II). Bayesian generalized linear mixed modeling (Bayenv) [42, 43] is a

method to detect clinal variation in allele frequencies while accounting for population structure by using a variance-covariance matrix of allele frequency. In addition, clinal variation analysis can be used to identify differentiated regions along a latitude or geographic range (paper III).

The interplay between stochastic (random genetic drift) and deterministic (natural selection) evolutionary forces shapes genetic differentiation and a major goal in evolutionary biology is to reveal the relative importance of these factors. Studying living systems, in which genetic drift has a subordinate role, provides an opportunity to reveal the action of natural selection.

Methods

Sequencing and collection of phenotypes/ecological data

The success of genetic studies lies in the collection of samples with accurate phenotypic/ecological information combined with high quality genetic data. In most of the studies in this thesis, the availability of such data facilitated the analysis and interpretation of results.

Herring population structure has been the focus of several studies over the last few decades. During the 1980s [44, 45], a number of Baltic and Atlantic populations were explored to characterize population structure and the genetic relationship between Atlantic and Baltic herring. In paper I and II, using NGS data, we studied genetic differentiation on a genome-wide scale among the population samples collected in the 1980s together with additional samples collected during 2012 and 2013. In addition to sample collection, accurate information about ecological parameters allowed us to study the genetic basis of ecological adaptation in this species.

In paper IV, samples were collected from domestic and wild rabbits in the Iberian Peninsula and Southern France. The same material was used for the speciation study (paper III) where we also generated transcriptome data from the two subspecies and their hybrids.

Genetic studies in domestic animals are facilitated because of phenotypic diversity and availability of pedigree data. In paper V, we performed a whole genome scan comparing horses with a skeletal disorder against a pool of unaffected controls. Furthermore, we genotyped individuals with pedigree information and evaluated the relative importance of identified loci.

Quality control

NGS is amenable to studies of a wide range of biological problems and standardized methods have been established to analyze these data. Since such data are prone to sequencing errors, identifying quality issues that may impact downstream analyses and interpretation is crucial. In quality control (QC) steps, reads with low Phred-scaled[a] quality scores were removed and reads were trimmed for sequencing adapters.

[a] Phred score shows the quality of each sequenced base and sequencing error rate.

DNA-seq

In all papers presented in this thesis, we generated WGS data from both individuals and pools. For pooled sequencing data, we pooled DNA from a number of individuals in equimolar concentration and sequenced them at high depth (~10-30X). The individual sequencing data were generated at ~7-12X/sample.

From pooled data, we could identify variants (SNPs and SVs) and infer allele frequencies in populations. Individual sequencing data, in addition to variant discovery, provided haplotype information that was used for demographic inferences. WGS data for herring (paper I and II) and horse (paper V) were generated with Illumina instruments at SciLifeLab, (Uppsala, Sweden). Rabbit WGS data (paper III and IV) were primarily generated with Genome Analyzer II (Illumina) machines at the Broad Institute (Boston, U.S.A). For paper V, we conducted long-read sequencing of horse BAC-clones by SMRT sequencing technology at SciLifeLab (Uppsala) and at the Eichler lab (Seattle, U.S.A). After QC, genomic aligners (*e.g.* BWA [46]) were used to align filtered data to a reference genome (paper II-V) and exome assembly (paper I).

RNA-seq data

In papers I-IV, we generated RNA-seq data for annotation, transcriptome assembly, and differential expression analysis. All RNA-seq data were generated with Illumina instruments at SciLifeLab (Uppsala, Sweden). For paper I, due to the absence of a genome assembly, we generated a *de novo* muscle transcriptome assembly that was later extended to an exome assembly (see paper I for more details).

For the herring genome annotation, we used RNA-seq data for liver, skeletal muscle, and kidney (paper II). In the rabbit genome project (paper IV), we annotated the genome using two independent pipelines, Ensembl, and a custom pipeline using RNA-seq data from ten tissues and human orthologs.

Since some of the RNA-seq reads span exon-intron boundaries, we need to introduce gaps in reads to have accurate alignment over intronic regions. Genomic-aligners penalize large gaps and do not perform well in aligning the RNA-seq reads. Hence, we applied specific tools capable of performing gapped alignment. Depending on the availability of annotation data, we used different pipelines to improve or to annotate the genome/transcriptome. After aligning reads, we quantified expression of annotated genes by generating fragment counts[a] to perform downstream statistical analysis where we performed pairwise comparisons between different groups (paper III).

[a] Fragment count is a commonly used measure to infer expression levels. This value is a count of a properly aligned pairs.

Variant calling

Aligned read data were used to discover SNPs and INDELs by comparison to a reference genome assembly. Several tools are available to call variants that are based on different statistical methods (FreeBayes [47], VarScan [48], and Genome analysis toolkit [49]). The genome analysis toolkit (GATK) is one of the most commonly used tools and was the approach employed in the majority of our studies.

Unlike small variants, characterizing and genotyping SVs remains challenging because these events are enriched near repetitive elements and usually have complex structures. Earlier surveys based on array comparative genomic hybridization (aCGH) and cloned-based approaches were limited to a small number of samples [50]. These methods are dependent upon a reference genome and can precisely detect deletions over insertions. They have been successfully used to detect balanced SVs (inversions and translocations) [51]. NGS data helped to overcome these shortcomings and enabled us to screen and genotype more complex SVs with finer resolution at breakpoints by implementing novel algorithms. There are different methodologies for SV detection that are mainly based on "Read Depth" information for detecting unbalanced changes (CNVs, duplications, and deletions), or "Read Pair" information, including read orientation and insert size, for detecting balanced and unbalanced SVs (Figure 1). The read pair approach has limitations in quantifying CNVs. We used a Read Depth method in paper II, III, and IV, and tools based on Read Pair methods (delly [52], breakdancer [53], DevRO [54]) in paper II and III.

Downstream analysis of candidate loci

We further characterized identified genes (from differential gene expression analysis or associated genes in genome-wide scans) for their function by statistical assessment of gene ontology (GO) annotation. GO is a collection of gene product properties classified within cellular components, molecular functions, and biological processes [55]. In these analyses, we assess gene lists for enrichment (over or under-representation) of GO terms (papers II-IV). In addition to GO terms, we also evaluated the association of candidate genes with phenotypic data from human and mouse (papers III and IV). Moreover, we assessed the relative importance of coding and regulatory changes in relation to genetic differentiation between populations (papers I-IV).

WGS data can be prone to different biases due to sequencing and mapping errors or technical errors during library preparation that can affect conclusions and interpretation of results [56]. To validate our findings from pooled sequencing data, we used different genotyping/sequencing methods. In paper I and II we developed custom-made SNP arrays (5k and 70k SNPs,

respectively) and genotyped the same individuals sequenced in pools. These data could also provide useful information about haplotypes within differentiated regions. In paper IV, we generated target sequence capture from a number of individuals and validated identified sweep regions with high correspondence. For identified CNVs in papers II and V, we developed CNV assays and confirmed the association between genotype and phenotype/ecological parameter.

Background to papers

Earlier studies, by using protein and DNA markers, provided insights concerning genetic variation and population structure, but typically only considered very small proportions of the genome [2, 10]. Unlike former methods, NGS technologies allowed us to query whole genomes rather than a small subset of regions or genes. Most early genome studies focused on humans and model organisms, due to the huge investment required to develop genome resources. Studying non-model systems became more feasible as sequencing costs decreased. In this PhD thesis, we utilized cost-effective approaches for sequencing individuals and pools of individuals to explore the genetic architecture and evolutionary forces that shape the genomes of individuals in natural populations, as well as to detect the footprints of speciation and domestication on a genome-wide scale.

Ecological adaptation in Atlantic herring

The Atlantic herring (*Clupea harengus*) is a pelagic fish and the fifth largest global fishery as judged by the amount of catch [57]. Herring is an obligate schooler and voyages oceanwide in large schools (Figure 3A). From at least the medieval period, the herring fishery has been one of the most important and valuable natural resources in Northern Europe also called as "silver of the sea" [58, 59]. Herring appears to display natal homing behavior; it travels long distances for feeding and returns to its natal site for spawning (58). Herring spawns in huge schools releasing thousands of eggs and sperms in water; each female can release more than 70,000 eggs during a spawning season [60].

Given the economic importance of herring, its population structure has been studied for years. Populations were historically classified using morphological differences, life history parameters, spawning localities, and spawning time (Figure 3B) [61]. Their adaptation to heterogeneous environments (*e.g.* range of salinity) and their variation in seasonal reproduction are unique features that have amazed researchers for decades. These characteristics have been the subject of previous research, yet, so far they have provided little evidence about genetic differentiation between herring populations. In this project, we employed NGS technology to generate a high

quality reference genome and to identify genome regions under selection for local adaptation as well as regions associated with the timing of spawning.

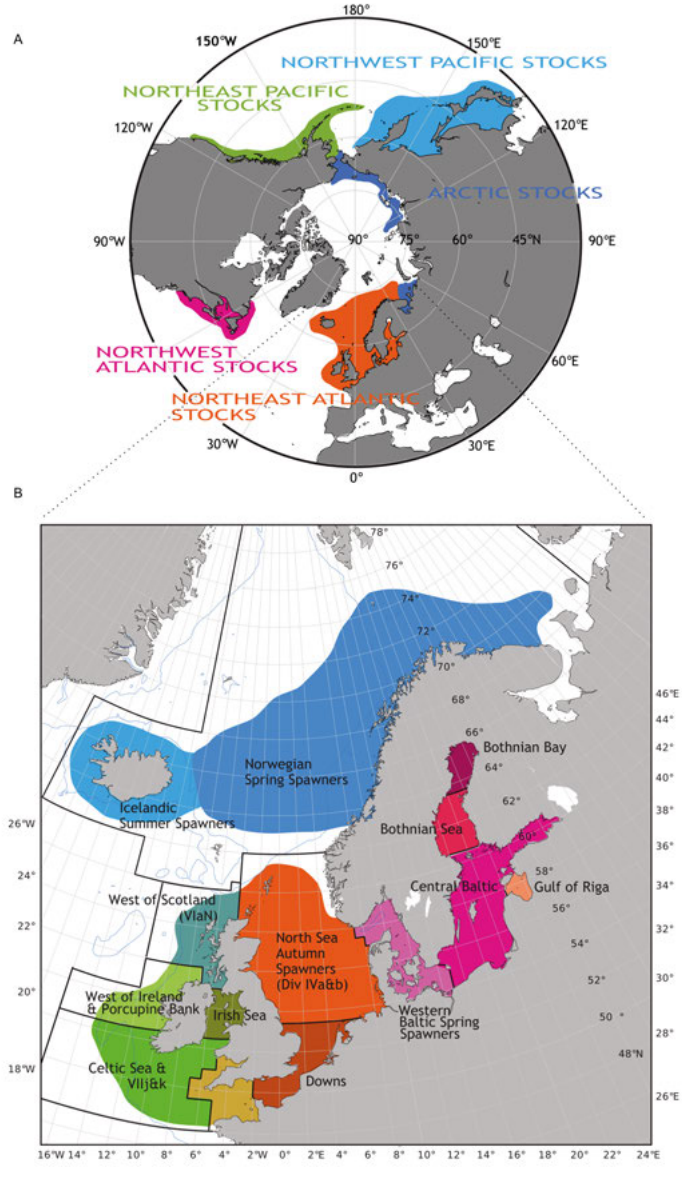

Figure 3. Herring population distribution around the globe. A) Distribution of herring stocks in the Northern hemisphere. B) Herring stocks classified based on locality and/or spawning season © C Zimmermann, www.clupea.net.

Rabbit speciation and domestication

The European rabbit (*O. cuniculus*) is one of the species within the order Lagomorpha. Rabbits are the main prey of various predators (*e.g.* foxes, European lynx, and a large number of birds) [62]. The rabbit was historically restricted to the Iberian Peninsula, where two subspecies evolved in the mid-Pleistocene [63, 64]. However, they were distributed around the globe during human voyages, and later by a specific interest in keeping them either for food or as pets. Breeding of rabbits resulted in domestication and now many distinct breeds are raised for meat, wool, fur production, and as pets [65]. The rabbit has also been widely used as a laboratory animal for biomedical purposes (*e.g.* for antibody production) [66]. Compared to other mammals, the rabbit has a high reproduction rate. Rabbits have been extremely successful in adapting to different ecological settings. To obtain insights into the genetic basis of reproductive isolation between the two rabbit subspecies of the Iberian Peninsula and to detect footprints of domestication, we performed genome-wide screens using WGS data.

Speciation is a gradual evolutionary process often resulting in substantial genetic differentiation between diverging entities. Identifying the most important factors during the initial stages of speciation is an important topic in evolutionary biology. There are a number of studies based on highly diverged species that do not hybridize under natural conditions. In these systems, many genetic differences may have accumulated after speciation [67-69]. Thus, studying distinct taxa that naturally hybridize because speciation has not yet been completed provides insights into the genetic architecture of reproductive isolation during the initial stages of speciation. There are two subspecies of rabbit in the Iberian Peninsula (*O. cuniculus algirus* and *O. cuniculus cuniculus*) and male hybrids display subfertility (paper IV, Figure 1b). To study the genetic bases of reproductive isolation between these two subspecies, we conducted WGS of different populations across the hybrid zone and explored the pattern of gene expression among purebreds and hybrids with subfertility.

Agriculture and animal domestication have been critical for the development of human societies and civilization. Domestication is a process by which plants or animals are adapted to the conditions that humans provide via selective breeding [70, 71]. This selection, termed "methodological selection" by Darwin, has left pronounced differences between domestic animals and their wild progenitors [70, 72-74]. Standing genetic variation together with artificial selection in the large populations of domestic animals across the globe have led to a rich reservoir of phenotypic variation[a] in domesticated species.

[a] These phenotypic changes are behavioral, morphological, or physiological.

The genetic factors underpinning the initial stages of animal domestication are poorly understood. Domestication is expected to have targeted animal behavior, allowing domestic animals to survive and reproduce in the stressful conditions that human provide. In the rabbit, differences between domestic and wild populations represent an excellent opportunity to decipher the genetic bases underlying domestication (Figure 4).

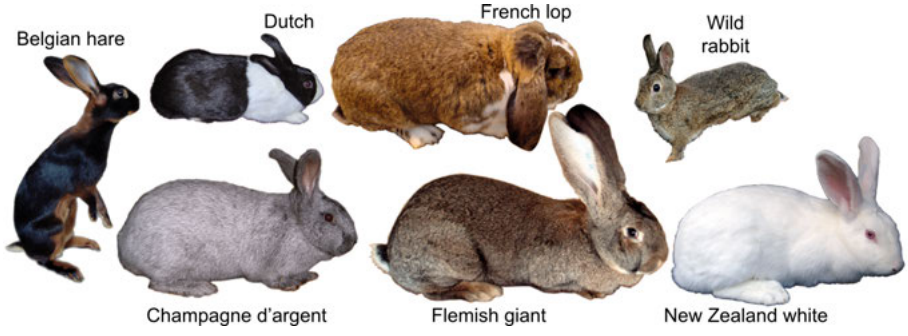

Figure 4. Phenotypic variation in domestic and wild rabbits, adapted from [75] (Reprinted with permission from AAAS).

Unlike other domesticated species, rabbits were domesticated relatively recently and the geographic origin of domestication is well known. As recently as ~1400 years ago, monasteries in southern France started to keep rabbits in captivity which eventually resulted in the domestication of rabbits (paper IV, Figure 1B) [76].

Today's wild populations are believed to resemble the wild ancestors of the domestic rabbit [77]. This well-defined origin is an advantage to provide insights into primary genetic changes during domestication. In this study, we performed WGS of pooled samples from domestic breeds and wild populations distributed across the Iberian Peninsula and Southern France.

Skeletal Atavism in Shetland ponies

Horses, like other domesticated animals, have drastically contributed to shape human societies and civilizations. Humans have bred and trained different horse breeds to utilize their capabilities in war, transportation, pleasure, and work. For instance, because of small size and relatively large strength, Shetland ponies were primarily bred to be used in mines and agriculture. Different horse traits have been the target of selection for a long time leading to the present marked genetic and phenotypic differences between breeds.

In the 1950s a skeletal deformity was described in Shetland ponies (reviewed in [5]). Fused radius and ulna and fused tibia and fibula were observed in affected foals (Figure 5). This skeletal formation resembled one present approximately 15 million years ago (mya), in the ancestors of modern equids (reviewed in [78]). The reappearance of phenotypic characteristics previously seen at an earlier evolutionary time is referred to as an atavism. Therefore, this defect in Shetland ponies is known as skeletal atavism (SA). Other examples of atavisms are supplementary nipples in humans [79] or hind limbs in whales [80].

Since the 1960s, there have been multiple reports of the occurrence of SA in the Netherlands, UK, and Sweden. Previously published data were consistent with an autosomal recessive mode of inheritance but the causative mutations of the disease were unknown [81, 82]. Understanding the genetic basis of skeletal atavism can be used in breeding programs to avoid foals being born with the disorder.

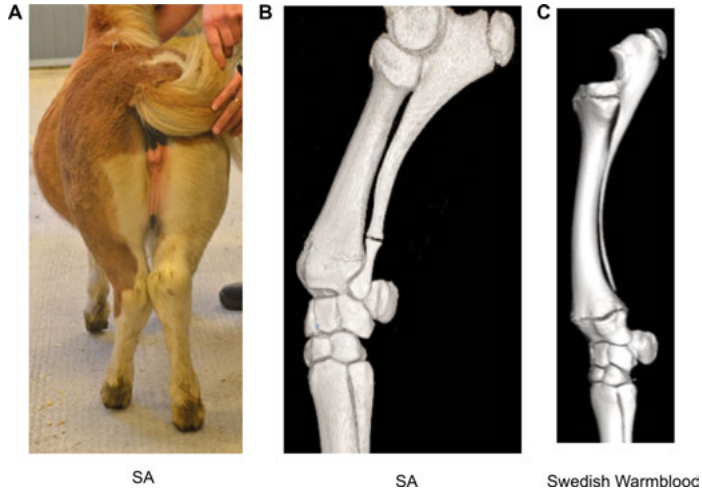

Figure 5. 16-week-old Shetland pony with SA. A) Caudal view when standing, B) Computed tomography scans of the left front limb of the SA with complete radius and ulna, C) Computed tomography scan of a healthy foal with normal development of radius and ulna. Adapted from [17] (Reprinted with permission from Genetics Society of America).

Results and Discussion

Ecological adaptation in the Atlantic herring

After the last glaciation approximately 10,000 years ago [83], Atlantic herring populations expanded and adapted to the brackish water of the Baltic Sea. Due to considerable fluctuation in annual catches [57], much interest has focused on studying the status of herring population stocks. Early molecular studies based on allozymes, microsatellites, SNPs, and mitochondrial DNA (mtDNA) found conspicuously low genetic differentiation between herring stocks in the Atlantic Ocean and in the Baltic Sea [44, 45, 84-87]. We expanded this analysis to a genome-wide scale by using NGS technologies, as are described in following two papers.

Paper I

Genome-wide studies require a high-quality draft genome assembly, which is a limitation for non-model species. The main challenges involved are constructing large insert size libraries to span repetitive elements and ordering the assembled sequences along the chromosomes. In the absence of a genome assembly for the Atlantic herring and to reduce the complexity of assembly, we developed a new pipeline to construct a partial genome assembly by combining RNA-seq and WGS data ("exome assembly") (Figure 6).

In this method, we first assembled a transcriptome from the RNA-seq data generated from skeletal muscle of a Baltic herring. Then we aligned genomic reads from one of the pools (Gulf of Bothnia; paper I, Figure 1A) to the muscle transcriptome. From this alignment, we extracted aligned pairs and orphan reads[a] to perform mini assemblies per contig. The final output of this pipeline was transcripts with extended exons consisting of coding and non-coding (intronic/promoter) regions. Using this approach, we could capture a larger proportion of the genome (~6%) and thus achieve a better alignment at exon/intron boundaries with accompanying improvements in variant calling.

[a] The term "orphan reads" refers to reads where only one of the pairs is aligned. Using this pipeline, we extracted orphan reads with their unmapped reads, together with properly mapped pairs to perform a mini assembly.

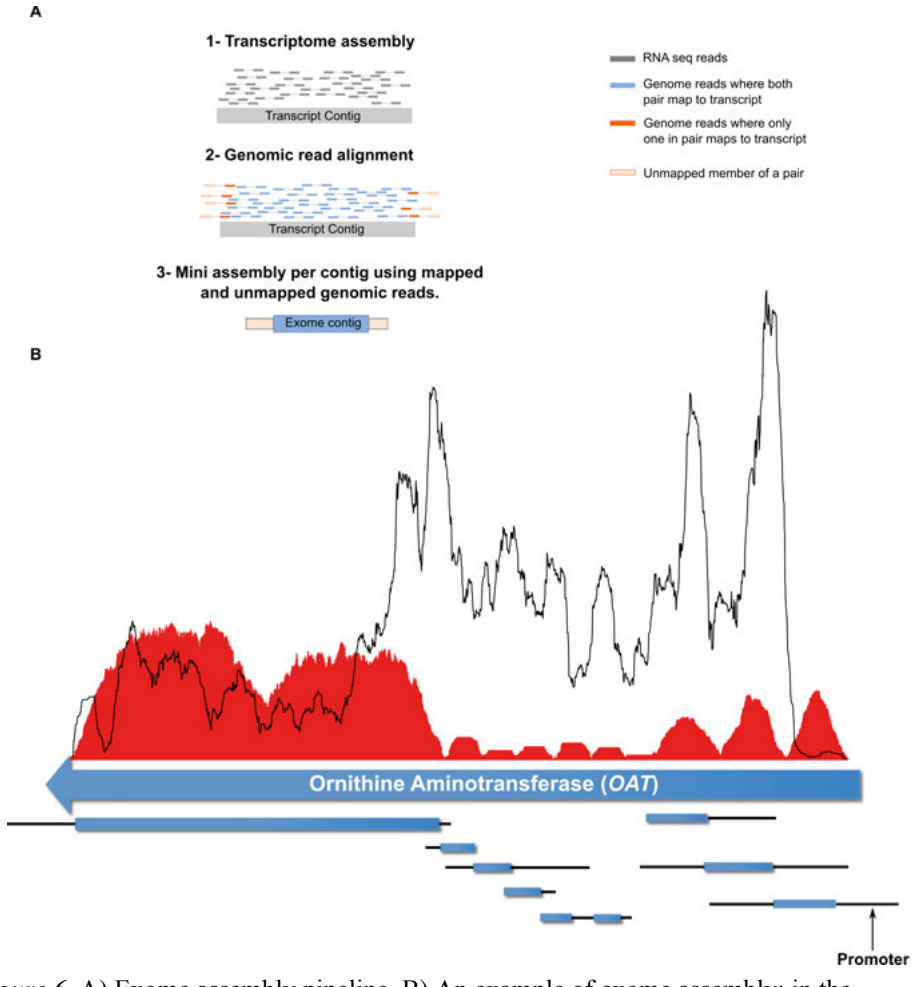

Figure 6. A) Exome assembly pipeline. B) An example of exome assembly; in the *OAT* gene we could extend coding region (blue boxes) into introns and promoter regions (black lines). The promoter was predicted by promoter software (v2.0) [88]; DNA depth (red) and RNA depth (black). Adapted from [16].

We sequenced eight pools each consisting of 50 fish from different localities in the Baltic Sea and the Atlantic Ocean (paper I, Figure 1A). These samples were collected between 1970s and 1980s (paper I, Table 1). We mapped these genomic reads to the exome assembly and called SNPs, resulting in ~400,000 polymorphic sites. Estimated heterozygosity was very similar among populations, consistent with low genetic differentiation between sampled regions (paper I, Table 1). Most SNPs showed no significant difference between populations while a small subset of SNPs (n=3,847) showed marked allele frequency differences (Figure 7 shows an example of significant SNP). A phylogenetic tree based on significant markers clearly shows separation between populations (paper I, Figure 2B).

We validated allele frequency estimation by individual genotyping of the same fish included in pooled sequencing using a subset of differentiated (outlier) and non-differentiated (neutral) SNPs (paper I, Figures 2G, S2, and S3).

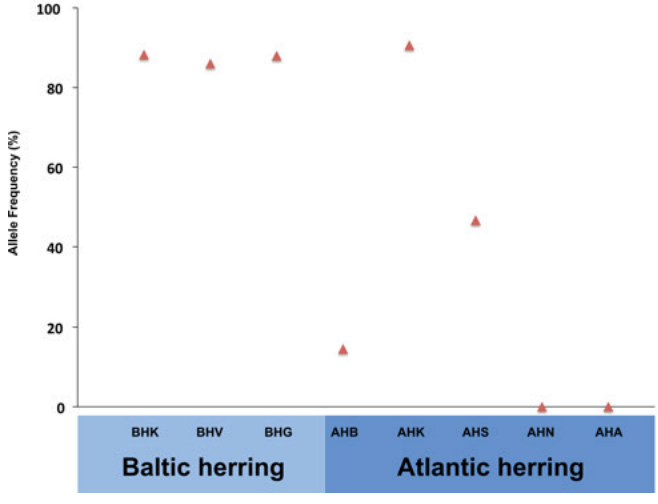

Figure 7. An example of a SNP with significant allele frequency difference among populations.

To explore possible colocalization of the identified signals in the absence of a genome assembly, we mapped the exome sequences containing significant markers on the stickleback (*Gasterosteus aculeatus*) genome. Herring sequences were spread across the stickleback genome but we identified few regions where significant signals were clustered. For instance, 98 SNPs in 14 genes formed a cluster on a chromosome corresponding to stickleback chrXIV (paper I, Figure 3A)

Genetic differentiation may be caused by genetic drift or selection. To distinguish the relative importance of these two forces, we performed a simulation study to compare expected and observed F_{st} distributions for selectively neutral SNPs. This resulted in a highly significant excess of observed loci with extreme F_{st} compared to the expected distribution under a drift model for selectively neutral alleles. This analysis confirmed that natural selection has been the main force causing genetic divergence between herring populations while genetic drift has played a subordinate role (paper I, Figure 2D-F).

In this paper, we presented an approach to studying population structure in a non-model organism. This methodology can be applied to other species lacking a reference genome assembly. However, recently developed sequencing methods can generate long read sequence data by which a draft genome assembly can be generated at lower costs. Furthermore, in this study we highlighted a number of markers associated with adaptation to environ-

mental niche and spawning time. To further characterize the identified signals and better understand the associated genes and loci, a more comprehensive transcriptome study from several tissues and a high-quality reference genome assembly was required.

Paper II

In paper I, we constructed a muscle exome assembly for the Atlantic herring and identified sites with marked allele frequency differences between populations distributed across the Baltic Sea and Atlantic Ocean. The organization of identified regions on the chromosomal level was unclear due to the highly fragmented nature of the exome assembly. To clarify the genetic architecture of differentiated regions and annotation of associated regions, we generated a high-quality draft genome assembly and annotated this genome using transcriptome data from three tissues (muscle, liver, and kidney). In addition to previously described populations in paper I, we sequenced 11 additional populations across the Baltic Sea, Atlantic Ocean, and a Pacific population as an outgroup (paper II, Figure 1A and Table 2). Furthermore, we sequenced 16 individuals (eight Baltic and eight Atlantic) at ~10X coverage (paper II, Table 2). This collection of populations from different geographical locations provided a great opportunity to study local adaptation to diverse ecological conditions (such as salinity) as well as seasonal spawning.

We assembled the genome by using short read sequencing of paired-end and mate-pair libraries (paper II, Supplementary File1A). The current assembly size is ~808 Mb which is close to the estimated genome size [89, 90] with scaffold N50[a] equal to 1.84 Mb and 23,336 protein-coding genes. For quality evaluation of the genome assembly, we performed core gene analysis and comparative analysis against other fish species gene sets. These analyses suggested that the current assembly has a high degree of completeness and quality compared to other fish genome assemblies (paper II, Table 1).

We next aligned the population data on the genome assembly and called SNPs. After stringent filtering we found 8.8 million high quality SNPs among all samples and 6.04 million high quality SNPs excluding Pacific herring. Phylogenetic analysis revealed an ancient split between Pacific and Atlantic herring. We estimated the approximate time of the split to be ~2.2 mya, based on mitochondrial cytochrome B data (paper II, Figure 1C). In contrast, Atlantic and Baltic samples form a star-like phylogenetic tree consistent with minute genetic differentiation between populations [16, 45]. Despite the low divergence, populations were clustered largely according to their geographic locations (paper II, Figure 1C). Three populations fell between the Baltic and Atlantic/south Baltic samples, two of these were au-

[a] N50 is a statistic for assembly qulity. It is referred to the shortest sequence at 50% of the assembly size.

tumn spawners (BÄH and BF; paper II, Table 2) indicative of genetic differentiation between populations separated by seasonal reproduction. The third population (KT; paper II, Table 2) is located between the Kattegat and the Baltic Sea where populations migrate for feeding and it may represent a mixed population because it was sampled outside of spawning time.

To explore loci associated with the recent niche expansion into brackish waters after the last glaciation, we compared allele frequency differences between the Baltic (all pools from Baltic) and Atlantic (Atlantic Ocean, Skagerrak, and Kattegat) populations using a χ^2 test. For this screen, we excluded autumn-spawners, as a confounding factor. We identified 46,045 SNPs with significant allele frequency differences ($P < 1{\times}10^{-10}$; paper II, Figure 3A). These sites were clustered in 472 independent loci distributed across the genome. One of the largest associated regions is located on scaffold218 (~119 kb), and included only a small number of genes. These sites partially overlapped with previously identified significant markers in our exome study that had much lower resolution due to the fragmented nature of the assembly (Figure 8).

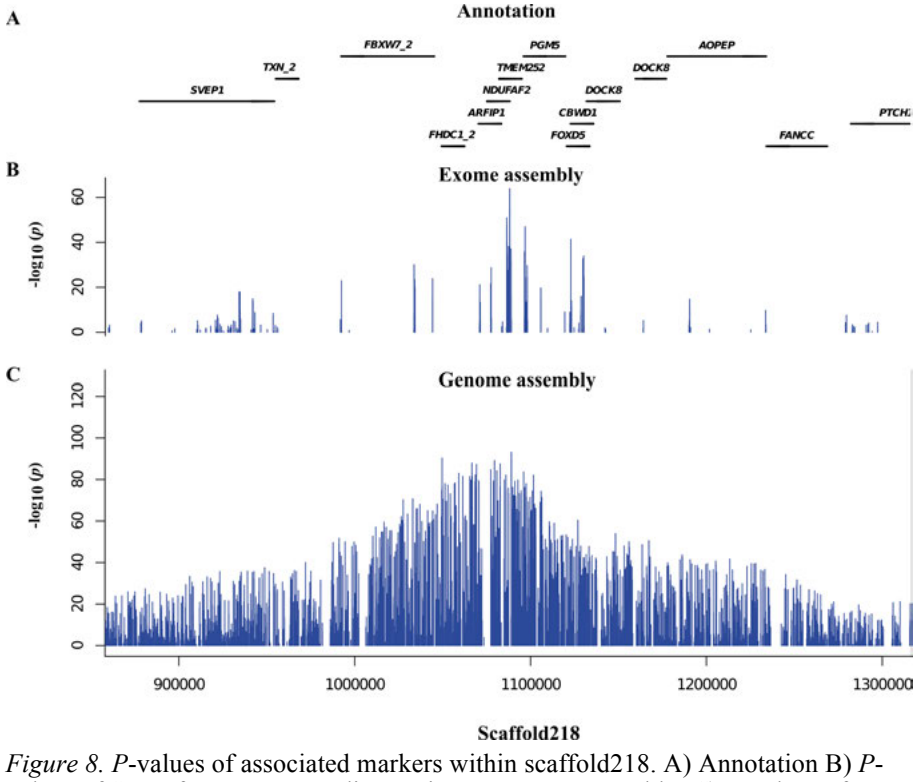

Figure 8. *P*-values of associated markers within scaffold218. A) Annotation B) *P*-values of SNPs from corresponding region on exome assembly. C) *P*-values of SNPs from whole genome screen.

Herring populations persist under the influence of a variety of environmental factors such as temperature, food resources (zooplankton), and predators, but the most substantial variation is probably in salinity. We studied the correlation between salinity gradient and allele frequency differences among populations using Bayenv 2.0 [42]. In general, there was a good agreement between Bayenv output and χ^2 tests. In the identified regions, there were 21 genes previously associated with hypertension in human and 36 genes with significant expression differences in sticklebacks kept in freshwater and seawater (paper II, supplementary File 3A). One of the strongest loci was a region downstream of *solute carrier family 12* (*sodium/chloride transporter*) *member 3* (*SLC12A3*) (paper II, Figure 3E). This gene has a key role in osmoregulation and is associated with hypertension in humans. Moreover, this gene showed differential expression in kidney tissue between sticklebacks kept in fresh and marine water [24].

In our structural variation screen, we identified a CNV that partially overlaps the *high choriolytic enzyme* (*HCE*) gene. This massive CNV is negatively correlated with salinity gradient (paper II, Figure 3D). HCE (also denoted as fish hatching enzyme) is a protease involved in egg hatching and is known to be under adaptive selection in response to salinity [91, 92]. Adaptation of a marine fish to spawning in brackish water may have been particularly challenging and therefore adaptation to this new environment might entail novel changes during egg hatching.

Herring spawning time has long been debated to be under the influence of environmental factors and the possible contribution of genetic factors has been unknown [61, 93]. To address this question, we compared allele frequencies between three autumn and ten spring spawning populations (paper II, Table 2). This screen resulted in more than 10,000 significant sites with marked differences in allele frequency between spring and autumn spawners ($P < 1\times10^{-10}$; paper II, Figure 4A). Phylogenetic analysis based on significant markers showed an independent cluster of spring spawners compared to autumn spawners (paper II, Figure 4B). Our strongest signal was close to *Thyroid Stimulating Hormone Receptor* (*TSHR*), which has a key role in photoperiodic regulation of reproduction in birds and mammals [94-96]. Moreover, other significant signals fell close to *CALM*, *SOX11*, and *ESR2a*, which are all known to be involved in the regulation of reproduction [97, 98]. These strong associations demonstrate that genetic factors underlie seasonal reproduction in herring populations.

To validate our findings based on pooled sequencing data, we developed a custom 70k SNP chip and genotyped the same individuals used for pooled sequencing. There was an excellent correlation in allele frequencies between resequencing and genotyping data (paper II, Figure 3-S1).

In the screens for genetic differentiation between Atlantic and Baltic herring, and between spring and autumn spawners, we identified large blocks spanning multiple genes. Large blocks may result from suppression of re-

combination *e.g.* due to inversions, as previously described in sticklebacks [36]. To explore the contribution of inversions to this pattern of evolution in the differentiated regions, we sequenced four ~3.3 kb mate-pair libraries (two Atlantic and two Baltic individuals). None of the identified inversions coincided with highly differentiated regions associated with the timing of reproduction or adaptation to the Baltic Sea (paper II, Figure 3C, 4C). Thus, we concluded that inversions do not appear to have a major role in the development of large haplotype blocks in herring.

We hypothesized two models for the presence of large haplotype blocks associated with genetic differentiation:

i Hitchhiking: large blocks can evolve by genetic hitchhiking of neutral sites linked with causal variants.
ii Haplotype evolution: multiple causal variants may be maintained by natural selection. In this model, selection acts on a given combination of causal variants leading to evolution of large haplotype blocks.

A hitchhiking model predicts a low level of genetic diversity in the differentiated regions. Under a haplotype evolution model, however, the nucleotide diversity in differentiated regions, within and between populations, can be as high or even higher than neutral regions. We compared nucleotide diversity of the 30 most differentiated regions against random regions within and between one Baltic (Kalix) and one Atlantic (Bergen) population. Interestingly, in these differentiated regions, we found significantly higher diversity compared to neutral regions (paper II, Figure 5A, B). Therefore, these results imply that large haplotype blocks harbouring multiple causal mutations are evolved as a result of natural selection on a combination of causal variants.

The relative contribution of coding and non-coding mutations in altering phenotypes has been under debate for decades [99]. To evaluate the contribution of highly differentiated SNPs to ecological adaptation, we analyzed the distribution of SNPs in coding and non-coding elements across the genome. Based on this analysis, we found a significant enrichment for non-synonymous changes as well as UTRs and (5kb) upstream and downstream of coding sequences. This suggests that both regulatory and coding changes have contributed to local adaptation. However, the effect size of these changes must be addressed by more rigorous phenotyping and genotype-phenotype evaluation in individuals.

We expanded our screen from transcriptome to genome to improve our understanding of population structure and genetic diversity for two adaptive traits in a natural population. Identifying differentiated markers associated with seasonal reproduction can help the fishery industry and complement existing methods (*e.g.* otolith structure) for stock assessment. In addition, the

genetic information generated in this study (paper I and II) will be useful for future studies in this species.

Rabbit speciation and domestication

Paper III

The two rabbit subspecies split from a common ancestor ~1.8 mya, but this was followed by episodes of contact between them due to climatic fluctuations (paper III, Figure 1b) [64]. Recent studies based on both nuclear markers [100-102] and mitochondrial DNA [64] identified a limited number of regions with high differentiation between the two subspecies. To determine islands of differentiation on a genome-wide scale, we carried out allele frequency analysis using WGS data of populations distributed across the Iberian Peninsula (obtained from paper IV [75]). Our data consists of 11 sampled populations across the hybrid zone between *O. c. cuniculus* and *O. c. algirus* (Figure 9 and paper III, S3 Table). As male hybrids show subfertility, we further investigated expression patterns in testis in hybrids together with pure subspecies individuals (Table 3).

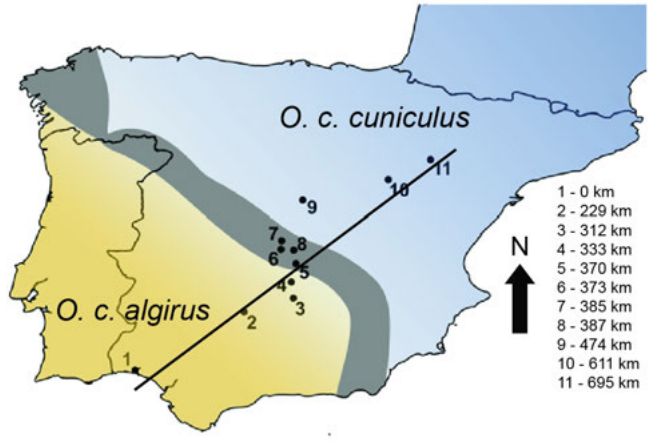

Figure 9. Sampled localities across the hybrid zone (in grey) between *O. c. cuniculus* and *O. c. algirus*. Numbers correspond to locations of populations (paper III, S3 Table).

The genome scan was restricted to ~4 million high quality SNPs. The phylogenetic tree generated from these SNPs was in general consistent with the geographical distribution of populations (paper III, Figure 1d). Allele frequency changes were variable from sharp to more steady shifts across the genome (paper III, Figure 1e, f) demonstrating the semipermeable nature of the two genomes at this early stage of speciation.

Table 3. Information of sequenced samples for expression analysis.

Genotype	Mother	Father	#individuals
ca	*c*	*a*	2
ac	*a*	*c*	3
cc-ca*	*c*	ca	5
cc	*c*	*c*	3
aa	*a*	*a*	3
		Total	**16**

a= *O. c. algirus*
c= *O. c. cuniculus*
*= Backcross

We used clinal variation analysis to measure the relative level of introgression across the hybrid zone. We identified regions with abrupt changes in allele frequencies that are likely to be involved in genetic barriers. The slope and center estimates coincided with the narrow region of contact between subspecies (~372 km from the most southerly located sample) (paper III, Figure 1g). Regions showing reduced introgression represented ~4% of the genome, and ranged in size between ~17 kb to ~5 Mb from, where the largest regions were located on chrX and three autosomal centromeres (chr3, chr4, and chr7). Both mtDNA and chrY displayed the steepest clines compared to other regions throughout the genome. We detected 253 loci with a sharp transition presenting regions of reduced introgression (RRI) across the genome (paper III, Figure 2a). We found 84 additional RRI on unassigned scaffolds (chrUn) but most of them may be disrupted because of gaps in the sequence.

Enrichment analysis of genes in RRI resulted in significant association with chromatin and gene regulation activities, as well as hybrid fertility phenotypes in knocked-out mice (paper III, Table 1). Moreover, we found three out of the five enriched transcription factor DNA sequence motifs to be binding sites for the nuclear receptor subfamily 6, group A, member 1 (*NR6A1*) and the androgen receptor (*AR*). NR6A1 has an important function in male and female fertility by regulating transcription of postmeiotic genes in testis [103] and the paracrine interaction between oocyte and somatic cells [104]. AR inactivation disrupts spermatogenesis and development of mature sperm leading to male sterility [105].

We next explored interactions between proteins located in RRI by using DAPPLE [106]. We focused on 436 genes with one-to-one orthologs with humans. The interactions between these genes were significantly higher than by chance ($P < 0.001$) (paper III, S1 Table). Most of these interactions were between proteins from different chromosomes implying that the identified pattern is not due to clusters of genes belonging to the same gene family

(paper III, Figure 2a). This suggests that hybrid incompatibility is formed by intergenic interactions and is in line with the proposed Bateson-Dobzhansky-Müller model for speciation [107-109].

We further searched for coding changes in genes within RRI that can potentially alter the function and structure of proteins. We only included sites with marked allele frequency differences between subspecies (allele frequency differences ≥ 0.70). A considerable proportion of regions did not contain any protein-coding genes, suggesting that these regions harbor regulatory elements that may control expression of proximal genes. More than half of the genes in RRI (57.9%) did not possess any coding changes, but in the remaining set of genes, we detected 404 changes (401 nonsynonymous and 3 splice mutations) (paper III, S3 File).

Chromosomal rearrangements have frequently been reported to promote reproductive isolation that lead ultimately to speciation [110, 111]. To investigate the potential role of chromosomal rearrangement in reproductive isolation of rabbit subspecies, we generated long insert mate-pair libraries (~4.5 kb) from two individuals per species. Close to 5% of detected variants (62 loci) overlapped RRI and the number of variants was not significantly more than expected by chance. This result suggests that SVs may not be a key mechanism involved in hybrid incompatibilities in rabbits. However, we identified two interesting loci overlapping RRI on chr16 and chrX (paper III, Figure 3a and b). A 150 kb duplication in *O. c. algirus* on chr16 located upstream of *formin 2* (*FMN2*) and downstream of *cholinergic receptor muscarinic 3* (*CHRM3*) (paper III, Figure 3a). FMN2 is an important protein required for polar body extrusion during female germ cell meiosis [112]. CHRM3 is a member of the G protein-coupled receptor family with a key role in bladder detrusor contractions and pupillary constriction [113]. The second locus contained an inversion of ~116 kb on chromosome X overlapping *poly(A) binding protein cytoplasmic 1 like 2A* and *B* (*PABPC1L2A* and *PABPC1L2B*). The encoded proteins together with other regulators (such as *EIF4G1*) is involved in regulation of translation during spermatogenesis [114].

F1 hybrid and backcrossed males have smaller testis and a higher incidence of sperm morphoanomalies [115]. To investigate gene expression differences in testis between hybrids and both subspecies, we used RNA-seq data from 16 individuals (Table 3). We found a large number of differentially expressed genes for which backcrosses showed the highest number compared to other F1 hybrids (paper III, Figure 4a). The large extent of expression variation among contrasts may be due to the differential cell content of heterogeneous tissue such as testis. However, testis histology analysis did not reveal any striking difference between groups (F1, backcross, and pure subspecies individuals) [115].

We next inferred mode of inheritance by comparing the level of expression between hybrids and parental individuals. Inheritance of expression can

be dominant, additive, and transgressive (overdominant and underdominant) (paper III, Figure 4b). Both groups of F1s showed higher numbers of genes with an additive (~35-36%) and a dominant (~46-51%) mode while in back-crosses most of the genes (88%) were classified as transgressive. Genes with an underdominant mode were six times more frequent than genes with an overdominant mode (binomial test, $P < 2.0\times10^{-16}$; Table 4).

Table 4. Number and proportion of differentially expressed genes following an additive, dominant, or transgressive mode of inheritance.

Pattern	**ac**	**ca**	**cc-ac**
Additive	246 (36.22%)	229 (35.01%)	116 (3.89%)
Dominant	322 (46.42%)	334 (51.07%)	227 (7.61%)
Misregulation (overdominance)	66 (9.72%)	62 (9.00%)	401 (13.00%)
Misregulation (udnerdominance)	45 (6.63%)	29 (4.00%)	2238 (75.00%)

The "large X-effect" has been repeatedly reported as a key factor in the evolution of reproductive barriers [116-118]. This theory suggests that mis-expression of genes residing on chrX can contribute to hybrid sterility and speciation. In contrast to previous studies, we did not find overrepresentation of differentially expressed genes on chrX in any comparison (Fisher's exact test $P > 0.23$, for all contrasts) (paper III, Figure 4c). Furthermore, despite an enrichment of identified RRI on chrX, we did not observe overrepresentation of differentially expressed genes overlapping these regions. This pattern is at odds with other studies showing that chrX plays a disproportionately role in hybrid male fertility (for instance [28, 119]). Our results suggest that reduced fertility among mammals have evolved via different mechanisms.

To explore the potential role of regulatory divergence at RRI, we looked for co-localization of differentially expressed and/or mis-expressed genes within these regions ($\pm$100 kb). This revealed no significant overrepresentation of genes coinciding near RRI compared to random expectation (Fisher's exact test, $P > 0.19$, for all contrasts). Hence, our results suggest that the observed pattern of differentially expressed genes largely located outside RRI may often be secondary effects due to disrupted regulatory pathways.

An important aspect of this paper was the application of clinal analysis to identify regions of the genome underlying partial reproductive isolation in the two rabbit subspecies. Our results provided a genome-wide screen across the rabbit hybrid zone and demonstrated that ~4% of the genome show a very strong shift in allele frequencies at the known hybrid zone. Diverged regions are distributed across the genome demonstrating the polygenic nature of speciation in rabbit. This pattern was previously reported in mice, where many loci seem to impede gene flow among mouse species [120]. Another important finding in this paper was the high level of gene mis-

expression in backcrosses where the majority of differentially expressed genes showed an underdominance mode of inheritance.

Paper IV

A previous genetic study provided us with a preliminary picture regarding footprints of domestication in the rabbit genome [77]. To expand this previous analysis and further characterize genetic changes during domestication, the Broad Institute generated a high-quality draft genome assembly using Sanger sequencing data. The draft genome assembly size is 2.6 Gb with a scaffold N50 size of 35.9 Mb and it was annotated using the Ensembl pipeline, as well as a custom pipeline for annotation of UTRs and non-coding RNAs.

To identify regions under selection during domestication, we conducted whole genome sequencing (at ~10x coverage) of 11 Iberian and three French wild populations as well as six domestic breeds (paper IV, Figure 1B). To deduce the ancestral state of alleles, we sequenced the snowshoe hare (*Lepus americanus*). We found more than 50 million SNPs and 5 million INDELs after stringent filtering. Our screen based on identity scores[a], using SNP data, revealed that domestic populations are closely related to wild populations from southern France where domestication is believed to have taken place (paper IV, Figure S1A). There is also a strong correlation in allele frequency at most loci between domestic and wild French populations (paper IV, Figure S1B).

Nucleotide diversity was remarkably high in wild populations ranging from 0.6 to 0.9%. Our nucleotide diversity analysis showed a reduction in heterozygosity after Iberian wild rabbits colonized southern France and suggested that a second drop in genetic diversity occurred during domestication (paper IV, Figure 1C). We screened for regions under selection using two approaches, F_{st} and pooled heterozygosity (H) in 50 kb windows (hereafter referred to as *Fst-H*), and SweepFinder [121]. Both methods detected more than 70 sweep regions (78 by *Fst-H* and 74 by SweepFinder) (paper IV, Figure 2A). We validated these results by targeted sequence capture (6 Mb) on an independent set of individuals from wild French and domestic populations. There was a high level of agreement between detected signals both in captured and pooled resequencing data (more than 70% of sweep regions were replicated).

One of the identified selective sweeps overlaps the 3'-end of the *glutamate receptor, ionotropic, kainate 2* (*GRIK2*) (paper IV, Figure 2B). The signal overlaps a sequence that is conserved in 29 mammals suggesting func-

[a] Identity score shows relatedness between populations. It is calculated by comparing the reference allele frequency of individual SNPs in each sequenced pool to the reference pool in a window.

tional importance of this segment. *GRIK2* is highly expressed in the brain and has been shown to be associated with recessive mental retardation in humans [122]. We identified two selective sweeps close to *SOX2* encoding a transcription factor that is involved in maintenance of stem cells [123] (paper IV, Figure 2C).

We next quantified the differentiation between domestic and wild rabbits (ΔAF) and explored their distribution in coding and non-coding conserved regions. We found a significant enrichment for highly differentiated SNPs (ΔAF >0.45) in UTRs, coding, and conserved elements (paper IV, Figure 2D) but the difference at very high ΔAF SNPs ($\Delta AF \geq 0.8$) suggested that changes at regulatory sites have had a more prominent role during domestication compared to changes in coding sequences.

We did not find any highly significant SNP that was expected to cause complete gene inactivation (*e.g.*, nonsense or frame-shift mutation), consistent with studies in chicken [13] and pigs [18]. Thus, we propose that there is no single (domestication) gene responsible for rabbit domestication. Among the missense mutations, we did not find any fixed difference (ΔAF=1) between wild and domestic populations. There were only 14 SNPs showing very high differentiation ($\Delta AF \geq$ 0.9) in coding regions. Considering low sequence conservation and similar physiochemical properties of altered amino acids, we assume that most of these mutations resulted from hitchhiking rather than as a direct response to selection. However, we identified two missense mutations with high conservation among more than 40 vertebrate species. The first mutation is Gln^{813} ➔ Arg^{813} in *tetratricopeptide repeat domain 21B protein* (*TTC21B*), which is involved in hedgehog signaling. The second mutation is Arg^{1627} ➔ Trp^{1627} in *lysine-specific demethylase 6B* (*KDM6B*), which is involved in *HOX* gene regulation. Both genes play an important role during development.

We performed gene ontology enrichment analysis examining genes located within 1 Mb of highly differentiated SNPs ($\Delta AF \geq$ 0.8). The most statistically significant categories were involved in brain and nervous system cell development (paper IV, Table 1). In the biological process, “cell fate commitment” showed the highest enrichment (enrichment factor = 4.9) (paper IV, Database S3). To examine genes associated with this term, we performed electrophoretic mobility shift assay (EMSA) using nuclear extracts from mouse embryonic stem cell–derived neural stem and DNA probes for highly differentiated SNPs near genes *SOX2*, *KLF4*, and *PAX2* (paper IV, Figure 3). EMSA resulted in clear gel shift differences between domestic and wild-type alleles. This demonstrated altered DNA-protein interactions, for 7 out of 17 polymorphic sites near *SOX2* and *PAX2*, that show striking differences in allele frequency between wild and domestic rabbits.

Our power to detect deletions unique to domestic rabbits was limited because the reference assembly was generated from a domestic rabbit. Similar to SNPs, we did not find a consistent pattern of differentiation between do-

mestic and wild rabbits in copy number, but a few number of loci showing copy number variation were identified (paper IV, Database S5).

In this study, we did not find many fixed difference between wild and domestic rabbits, but rather shifts in allele frequencies at many loci. The result is consistent with a polygenic background of domestication and soft sweep modes of selection that acted on standing genetic variation primarily in regulatory elements. One of the initial steps in domestication is to change behavior, to allow animals to tolerate human presence and survive in the new environment provided by humans (in other words tameness). Tameness has a complex genetic background and it is unlikely that alteration in just a few genes can lead to this dramatic change in the behavior of domestic animals. Therefore, we propose that domestication has primarily targeted behavior by means of genetic changes at many loci.

Paper V

Skeletal atavism affects both front and hind limbs as well as severe deformation in the hind hock and knee (Figure 6 and paper V, Figure 1). Movement progressively becomes worse with age and in most cases the horse is euthanized at an early stage of life. Previously published data were consistent with an autosomal recessive mode of inheritance [81]. To better understand the genetic basis of this disease, we first performed a GWAS using the EquineSNP50 BeadChip, but this analysis did not reveal any significant association between genetic markers and disease status. We therefore conducted whole genome sequencing from six affected individuals (at ~7X depth) and a pool of male individuals without a history of siring atavistic individuals (at ~56X depth).

After SNP calling, we scanned the genome for fixed differences between cases and the pool of controls. We identified 25 SNPs of which two were located on chr1 and the rest on unanchored scaffolds (chrUn). Three of the cases (2, 3, and 4) were homozygous for variant alleles while the other three cases did not show any coverage over these sites, suggesting the presence of a deletion in chrUn. Most of the SNPs were located on the same unanchored scaffold (chrUn00036:26,645,953-26,779,752) consisting of 19 contigs separated by large gaps. We screened these regions for depth and found that cases 2, 3, and 4 had just 50% of the expected read coverage and the other three cases lacked any coverage. Despite this variation in depth of coverage, the control pool showed normal depth along this region. This pattern demonstrated that cases 1, 5, and 6 are homozygous for a large deletion over this region while the other three cases are hemizygous (hereafter referred as Del-1). In addition, all cases shared another deletion partially overlapping Del-1 (hereafter referred as Del-2) (Figure 10). We observed this depth pattern scattered on chrUn removing ~160 kb of sequence in certain cases.

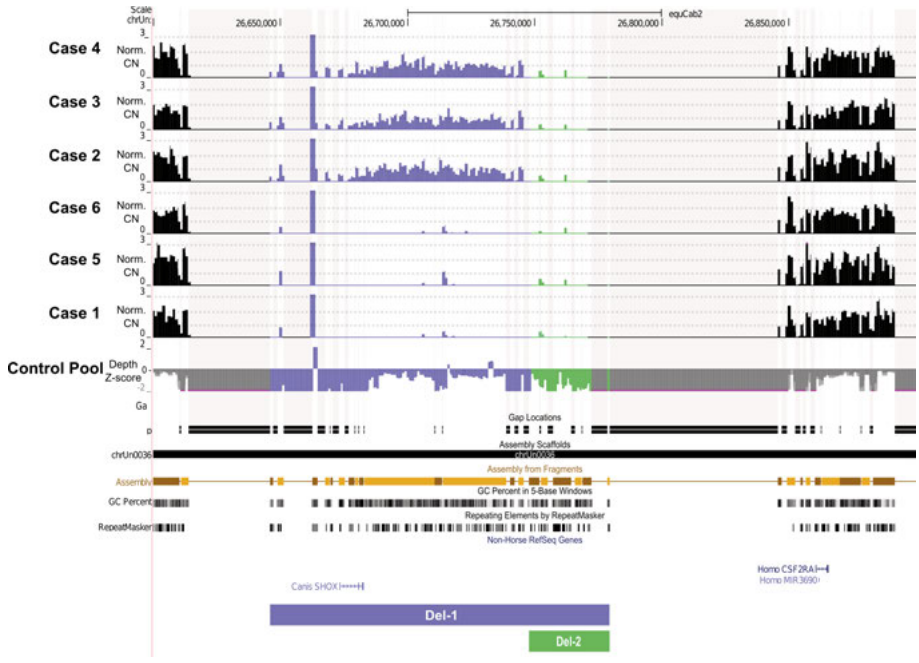

Figure 10. Identified deletions in chrUn: 26.62-26.9 Mb in UCSC genome browser. Normalized copy numbers of SA cases and Z-score values of control pool depth. Adapted from [17] (Reprinted with permission from Genetics Society of America).

This region is poorly annotated and the only coding sequence identified by similarity was the *short stature homeobox* (*SHOX*) gene from different species. *SHOX* encodes a homeobox transcription factor involved in growth and expressed during developmental stages. *SHOX* is located in the pseudoautosomal region (PAR) of the sex chromosomes in eutherian mammals [124] but it is absent from rodent genomes [125] (paper V, Figure 3).

The horse genome (EquCab2.0) is highly fragmented in the identified region and our attempt to improve the assembly using short-read data failed mostly due to the highly repetitive nature and high GC content of this region. Additionally, our attempts to bridge the contigs of this region using PCR and chromosome walking proved futile. To improve the assembly, we sequenced nine bacterial artificial chromosome (BAC) clones overlapping this segment using SMRT sequencing technology (paper V, Table 1). We tried to extend assembled contigs from five BAC clones based on overlapping similarities (paper V, Figure 4A) and this resulted in three BAC-derived consensus contigs (paper V, Figure 4B).

We generated a pseudogenome assembly by merging the new BAC-derived consensus contigs with EquCab2.0. We next aligned the short reads from all cases and the pool of controls to this pseudogenome assembly and screened for depth. We identified the Del-1 depth pattern on BAC-C1 for ~97 kb and a presumed breakpoint flanked by a stretch of TGGA repeats

(paper V, Figure 4B). The Del-2 depth pattern was observed on the two other contigs (BAC-C2 and BAC-C3) where all cases lacked depth compared to the control pool with normal depth (paper V, Figure 4B). We annotated BAC-derived consensus sequences by GENSCAN [126] as well as previously reported RNA-seq data from 24 samples [127]. Based on this annotation, we found a new coding sequence within Del-2 on BAC-C3 coding for *cytokine receptor-like factor 2* (*CRLF2*). This gene is located downstream of *SHOX* in humans (paper V, Figure 3) [124]. Based on this information we estimated the size of Del-2 to be 60-80 kb, and the uncertainty is due to evident mis-assembly in BAC-C2 (paper V, Figure 4B).

We validated our findings and estimated the allele frequencies after designing a TaqMan copy number assay and used this for genotyping by droplet digital PCR (ddPCR). We genotyped two groups of individuals, comprising of cases, obligate carriers, potential carriers[a] (18 American and 63 Swedish samples), versus a random set of Swedish Shetland ponies (94 samples). Figure 11 shows the genotyping result of the first set where individuals have been clustered according to their status. In the random set, we observed ~12% *Del-1* or *Del-2* carriers, but no *Del-1* or *Del-2* homozygotes nor compound *Del-1/Del-2* heterozygotes (paper V, Figure 5B).

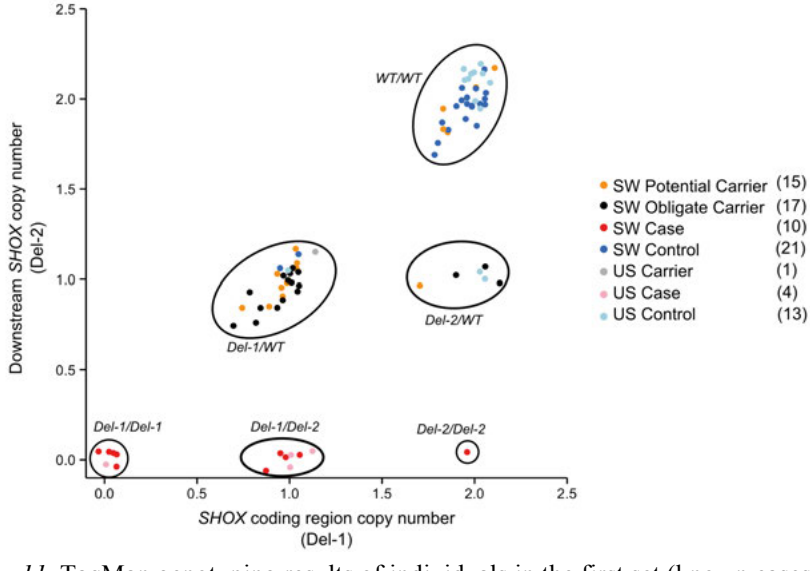

Figure 11. TaqMan genotyping results of individuals in the first set (known cases, obligate carrier, and potential carriers from US and Sweden). Numbers in parenthesis present the number of genotyped individuals. Adapted from [17] (Reprinted with permission from Genetics Society of America).

Based on the genotyping results from Swedish samples collected between 1986 and 2000, we estimated the allele frequencies of *Del-1* and *Del-2* to be

[a] Individuals that are close relatives to known carriers or having unconfirmed SA foals.

4.8% and 1.1%, respectively. In addition, we studied the association between heterozygosity for Del-1 or Del-2 and height at the withers in 130 samples, but we did not find any significant association (paper V, Figure 5C).

Our GWAS analysis based on the 50 kb SNP chip failed to reveal any association, because the markers on the chip were only selected from anchored chromosomes and markers from chrUn were not included in this chip. We could overcome SNP chip limitations by performing WGS of six affected individuals and a pool of control resulting in detection of two overlapping deletions (Del-1 and Del-2). Del-1 removes all *SHOX* coding region which determines a transcriptional activator in osteogenic cells during development [128]. Moreover, Del-2 removes the downstream region of *SHOX* and part of the *CRLF2* gene. *SHOX* regulation is very complex and deletion of its regulatory elements has resulted in mis-expression of this gene [129]. Furthermore, malfunction of the gene has resulted in skeletal deformities in humans such as Léri-Weill dyschondrosteosis (LWD) and idiopathic short stature (ISS) [129, 130]. *CRLF2* is a cytokine receptor involved in the JAK-STAT pathway and is active in bone development and metabolism [131, 132]. We cannot exclude that *CRLF2* contributes to an effect on the phenotype, but the similarity of SA to human skeletal anomalies, caused by *SHOX* deficiency, supports *SHOX* mis-expression as the most plausible cause of SA in Shetland ponies.

This study was an example on application of advanced sequencing technology to characterize a complex region of horse genome assembly. Based on our results, we could accurately validate the genotype of known cases and carriers and provide a diagnostic test for breeders to control their breeding program and avoid the conception of affected foals.

Concluding remarks and future prospects

In the work involved in this thesis, we used NGS data to make leaps in the characterization of ecological adaptation and genetic divergence between populations or species.

We characterized the genetic basis for ecological adaptation in Atlantic herring at an unprecedented resolution providing a unique genetic resource for evolutionary studies and the fisheries industry. The herring's biological features (*e.g.* large population size, subordinate role of drift, and widespread distribution across environmentally variable habitat) together with our results (paper I and II) suggest this species could play a role as a model organism to study adaptation across space and time. The comprehensive list of identified genetic markers can be used to monitor stocks, in particular out of spawning season, and to avoid overfishing. In addition, we found strong differentiation between spring and autumn spawners, indicating that seasonal reproduction is not only influenced by environmental factors (*e.g.*, water temperature and nutritional status). We identified strong candidate genes with an established role in reproductive biology for mammals and birds. The genetic basis of photoperiodic regulation in marine organisms has not been well characterized and our findings provide a starting point for follow-up studies.

To further investigate photoperiodic regulation, we can compare gene expression levels between autumn and spring spawners in different regions of the brain, particularly in the saccus vasculosus, which is a sensory organ that may respond to fluctuations in day length [133]. Such research may have implications for the aquaculture industry to manipulate the timing of reproduction in commercial species.

The Baltic Sea has a unique ecology and hosts species with diverse environmental tolerance, including herring. In our studies, we identified loci where allele frequencies showed a close correlation with salinity gradients across the Baltic Sea. Further studies on differential gene expression across a salinity gradient can enhance our understanding of the mechanisms that regulate adaptation to salinity. In addition to salinity, there are many other elements that may have an impact on local adaptation in herring. Pathogens, predators, habitat loss, environmental contaminants, invasion of new species, and climate change are all examples of stressors that may influence the marine environment and biodiversity in the Baltic Sea. International coordination of research using genetic resources generated during our studies (paper I

and II), together with aquatic surveys, offer to clarify the underlying genetic bases of adaptation to the factors described above.

The discovery that both coding and non-coding changes contribute greatly to ecological adaptation in the Atlantic herring open up a new avenue for further investigations. Specifically, we can examine the relative importance of regulatory networks using expression and genotyping data, and we can also locate eQTL underlying gene expression in tissues with a potential role in physiological adaptation to different aquatic environments under varying conditions (*e.g.* salinity).

In addition to the populations included in our studies, herring stocks exist in other geographical regions. In the West Atlantic there are well-characterized spring and autumn spawner populations. We can now explore association of the identified loci in our studies with spawning time in these populations. This will provide new insights into the parallel evolution of seasonal reproduction in geographically decoupled populations. There are examples of such patterns in other taxonomic groups (*e.g.* Drosophila [134]), but they are rare in marine species. In addition to the distribution of Atlantic herring in the Baltic Sea and Atlantic Ocean, remote populations of Pacific herring have been reported in the White Sea [135]. Interestingly, based on mtDNA and allozyme analysis, introgression of Atlantic herring to Pacific herring was reported in northern Europe [136]. Whole genome comparisons of these populations will allow us to study the level of genetic exchange, and identify loci underpinning local adaptation to environmental conditions in the White Sea.

We expanded previous studies on rabbit speciation and domestication to a genome-wide scale by generating a high quality genome assembly and WGS data from different populations (paper III and IV). We uncovered regions likely to be associated with the early stages of speciation between the two rabbit subspecies. Regulatory elements appeared to be the primary driver of incompatibilities compared to coding changes. Misregulation of gene expression due to interactions among regulatory regions can lead to reduced fitness and eventually contribute to reproductive isolation [120]. In a follow-up study, using genotyping and expression data, we can elucidate the regulatory function of differentiated region.

Our expression analysis revealed an extensive misregulation of genes in the testis of hybrid males, but without overrepresentation in regions of reduced introgression. Moreover, we failed to identify causal genes contributing to hybrid subfertility. Mis-expression of such genes may influence hybrid fertility in specific developmental stages. In addition, testis is a heterogeneous tissue, and therefore follow-up studies may need to focus on characterizing expression patterns in different cell-types. Our results have also provided evidence of genetic differentiation in genes associated with chromatin activity. Information regarding chromatin dynamics in the rabbit is

highly limited, and whole-genome chromatin profiling, in particular of the testis, will allow us to study chromatin diversity between the two subspecies.

Changes in behavior are predicted to be an initial target of selection during domestication because animals must become accustomed to human presence and survive in unnatural conditions that would be overly stressful for wild progenitors. Domestic and wild rabbits differ dramatically in their behavior, in particular in their fear response. Behavioral phenotypes are controlled by the brain, usually via complex mechanisms. A comprehensive study on gene expression levels in the brain may allow us to unravel the key region(s) regulating such complex phenotypes. The genomic data from our studies, together with gene expression analyses and detailed phenotyping offer the potential to make the rabbit a major model system for dissecting behavioral phenotypes.

We identified the mutations associated with skeletal atavism in one of the most complex regions of the horse genome. By conducting WGS of cases and a pool of controls, we identified two partially overlapping deletions that removed ~180 kb of sequence, including the *SHOX* gene in the PAR. By sequencing BAC-clones spanning this region we could identify the partial deletion of another gene (*CRLF2*) that was missing from the EquCab2.0 genome assembly. Although we now have a better picture of this region, we are still missing a considerable section of the BAC-clone sequences. The high repeat, high GC content, high recombination rate, and the observed structural changes explain difficulties in sequencing and assembling this region. In humans, ~600 kb of PAR has still not been assembled due to these features [137]. In future analyses, we should perform in-depth sequencing of BAC-clones across this region by utilizing long-read sequencing and probably optical mapping data to stitch together scattered contigs on chrUn.

All six affected individuals carried the Del-2 deletion that is 60-80 kb in size and removes sequence downstream of *SHOX*. In humans, deletion of regulatory elements downstream of *SHOX* has been shown in LWD and ISS patients. Our efforts to characterize these elements failed because this region is evolutionarily poorly conserved. Therefore, important follow-up studies are to improve the annotation of this region by means of comparative analysis with species that share a closer evolutionary relationship to horses than humans, and to conduct further sequencing that is prohibitively expensive at the present time.

Although use of NGS data allowed us to gain a more comprehensive understanding of biological pathways and living organisms, there are still many challenges to overcome in the future. In all papers we used pooled sequencing data to directly estimate allele frequency from read counts. However, the main drawback of this sequencing strategy is the lack of power to infer haplotypes for admixture analyses. Therefore, we can improve our analyses by generating more individual sequencing data that includes haplotype information.

The ongoing decrease in sequencing costs has enabled us to generate data from many individuals of different biological systems. However, there are often limitations associated with studying natural populations. For instance, in the rabbit speciation study we generated expression data from parental and hybrid individuals for which we did not have kinship information to precisely assign cis/trans-regulatory divergence. In a follow-up study, sequencing related individuals of known pedigree may facilitate analysis and interpretation of the data.

The genomes presented in this thesis were annotated by *ab initio* prediction programs and UTRs were annotated by RNA-seq data from just a few tissues. To improve annotations we can generate RNA-seq data from further tissues, and across developmental stages. To accurately identify different isoforms of genes and capture the full transcriptome we can apply long-read sequencing technology.

Today, generating pan-omics data is not as challenging as before, but phenotyping is often the main operational bottleneck in genetic studies. To some extent, phenotyping has lagged behind our capability to generate and study whole genomes. Therefore, we must develop methods for "next generation phenotyping" (NGP) to empower high-resolution mapping and association studies in medical, agricultural, and evolutionary biology. In papers I-IV we identified several loci with significant differentiation between populations that are associated with ecological adaptation. However, we were unable to assess how much of the observed phenotypic variation is explained by these loci. In future, this question can be addressed by conducting experimental studies and additional phenotyping at the individual level.

Our top differentiated sites were located upsrtream or downstream of genes, that may indicate their regulatory role in controlling expression of nearby genes. We can use EMSA to evaluate protein-DNA interaction and identify potential transcription factor binding sites. For example, we used this technique in paper IV for differentiated sites with potential regulatory roles. We can also perform similar experiment to characterize differentiated non-coding regions in herring.

In paper V, although we characterized associated mutations, we did not find any obvious phenotypic differences between *Del-1/Del-1* and *Del-1/Del-2* individuals. In addition, the effect of *CRLF2* deletion on the SA phenotype is unclear, but it may influence unexplored phenotypes in affected individuals. More detailed phenotyping will allow us to evaluate differences between genotypes in atavistic individuals.

Our current results primarily rest upon genome assemblies generated from single individuals that lack some of the unique features related to other populations or breeds. For example, a 1.6 kb region downstream of the *TBX3* gene is missing from the current horse assembly because the reference individual is homozygous for this deletion, which causes the non-dun coat color [138]. Similarly, we may be blind to sequences specific to the Atlantic her-

ring or wild rabbits due to our choice of reference individual. Thus, to further explore the genomes of focal species we should generate new genome assemblies from a range of different ecotypes.

In our studies, we presented draft genome assemblies for a variety of species that were mostly generated from short-read sequencing data. These data have a limited potential to assemble complex features such as repeats and structural changes. Such regions are usually mis-assembled or filled by gaps. These artifacts can influence downstream analysis and subsequent interpretations of results. In future studies, long-read sequencing data will help to overcome these limitations and improve genome assemblies. This also improves detection of structural changes and enables us to study unexplored features (*e.g.* sex-determination in the Atlantic herring).

Targeted genome editing technologies (such as CRISPR/Cas9 techniques) are powerful tools that enable researchers to study the biological function of candidate genes and their association with certain phenotypes. For example, we can use this technique to edit the identified missense mutations (*TTC21B* and *KDM6B*) with marked allele frequency differences between domestic and wild rabbits, and thereby study their effect on rabbit phenotypes. In our speciation study, regions with reduced introgression were enriched for transcription factor binding sites for two transcription factor (TF) genes (*AR* and *NR6A1*). We can knockout these TFs and if the knock-outs are viable we can study consequent phenotypic change and the magnitude of their effect on gene expression in the two rabbit subspecies. The same experiments can be applied to the strongest signals we identified in the herring. Characterizing gene function and the interactions among gene products will allow us to develop a better understanding of genotype-phenotype relationships.

The ultimate goal of biological studies is to fully understand phenotypic features such as health, agricultural production traits, and evolutionary fitness. NGS data provided great scope to characterize genetic diversity, at a genome-wide scale. Fully translating such information into applicable knowledge requires further progress in technology, computational capabilities to handle and analyze big data, and functional validation of the findings.

Acknowledgements

I thought this section would be the easiest part to write. But I soon realized how difficult it is to write your feelings on paper and not to miss anyone. Thus, please forgive me if I have left you out.

Leif, words come short in explaining how grateful I am for having this amazing opportunity to conduct research and work under your supervision. You are not only an outstanding scientist and mentor, but you also have a great personality. You care about the members of your group, like a father. It is quite rare to find these characteristics in one person. Your experience and insights in genetics and experimental design have been a source of motivation. Your aspiring guidance and inputs have always been helpful. Science was always beautiful for me and it became even more beautiful with your advice and help. It was a unique experience working as part of your team. I feel sad that I will now lose you as a mentor, but I will proudly remember what I have gained and the unique time that I had in your group.

Talking about your great group, I should name many people. But one person who I am sincerely thankful to is **Calle**, the king of heatmaps. The first time I met you in your office, you were working on a heatmap and I think you are still doing that. But behind your interest in plotting heatmaps, there is profound knowledge and thought. You are fantastic in teaching and sharing. Not only in analyzing data, but also in designing experiments. I have been very fortunate to have you as a mentor and without your help and constant support I would not have been able to survive the shocks of *SHOX*! Apart from science, I believe we also shared a great time discussing topics other than science.

Another person to whom I am very grateful is **Alvaro**. Your experience and knowledge in computation and bioinformatics is endless. I learnt a lot from you about bioinformatics tricks, programming, and genetics. You helped me to develop the skills needed to conduct bioinformatic analyses. Many thanks for your input and advice during the time I spent at IMBIM, from my Master's degree project and throughout my PhD. Both you and **Calle** made research more fun and exciting to me. **Miguel,** your support has meant a lot to me during these years. It has been inspiring and rewarding to work with you.

I would like to thank **Mats** for guidance on how to "lift-over" information from digit to paper and his inputs on population genetics. Thanks for allow-

ing me to disturb you anytime I needed. **Marcin**, you have been a great friend for me. We both shared similar thoughts and feelings across a range of different topics whether science or politics. It has always been joyful talking with you, and learning from you about R tricks and statistics. I would like to thank **Kerstin** for your broad knowledge and experience in genetics that has always been inspiring to me. **Örjan,** you are statistically one of the most important scientists who I have met. Your knowledge in statistics has always been a great help. Thank you **Matt** for your valuable hints in genetics and Perl programming.

Andreas and Simon, as you know we all share the same interest about "sequences" of music. Although there is slightly high diversity in our taste, I enjoyed talking with you about music and I would like to thank you for introducing me to great (Swedish) bands.

I enjoyed great collaborations with people at **SciLifeLab**, **BILS** annotation group, and **WABI**. In particular I would like to thank **Olga, Jacques**, **Henrik**, **Marc**, **Adam,** and the sequencing platform that helped us to generate high quality data.

I would like to express my appreciation to people at SLU. **Gabriella,** I would like to thank you for organizing the "Genome Analysis" course. I had a fun time and got the chance to learn from outstanding scientists during this course. **Göran Andersson,** you are truly one of the greatest teachers I had during the "Genome Analysis" course. Your endless knowledge in genetics and the extensive list of examples you had in mind, inspired me constantly. **Erik Bongcam-Rudloff,** it was always interesting to discuss with you and your group members about bioinformatics. **Tomas, Sandrine, Sofia, Lisa, and Lars,** thank you for fruitful scientific discussions.

A few people became out of sight, but not out of mind. **Freyja and Jonas,** I never forget the great times and discussions we had about science, politics and other topics. **Markus**, it was a great experience sharing an office and working with you. **Shumaila**, working with you on the rabbit project was a fantastic experience. **Olaf**, it was always fun to talk to you and have some jam sessions. **Miri**, we had a very nice time discussing about science and cultures. It was great chance to work with you and know you.

In the herring project many people were involved who I enjoyed working with. **Görel**, your invaluable knowledge about the evolutionary biology of teleost fish has always amazed me. It was a great experience working with you. **Nils, Linda, Arild,** and **Michele** I had an amazing time working with you in this project.

I have had a great time working with people in our department. Special thanks go to **Jennifer, Freyja, Jonas, Elisabeth, Manfred, Stina, Alex Hayward, and Sam** for your contribution to my thesis. Special thanks to my PCR buddy **Jessica Petterson** and **Chungang** for valued assistance in the atavism project. Special thanks go to people who contributed to this project

with their knowledge and horse samples including **Evan Eichler, Terje, Adam, and John Eberth**.

I think almost everyone in our lab has benefited from the experience of **Ulla and Eva**. Thanks for all your help and instruction in the wet-lab. **Neda Zamani and Ginger**, you have always been full of positive energy and our conversations have been inspirational to me throughout these years.

Shady and Sangeet, we three have had a long journey together since our Master's program. It has been a special experience to study and work with you. I also had an amazing time with **Rakan, Iris, Chao, Fan, Angela, Anna, Ann, Sharda, Matteo, Fabiana, Roni, Erik, Susanne Kerje, Nina, Martin, Gosia, Doreen, Agnes, Behrooz, Behdad, Aida, Amin, and Saber.** I also value the fun times and discussions I had with **Alex Hayward.**

I would like to thank the IMBIM staff who have been very helpful during the time I spent at the department. **Barbro, Rhené, Malin, Eva, Alexis, Susanne, and Veronica,** thank you for all your kindness and help.

I would not have been able to reach this point without the love and support that I received from my **family members.** My lovely **parents, Zohreh and Ali,** you have been the heros of my life, teaching me perseverance, love, and honesty that have helped me to work persistently and passionately. My dear brother, **Sina**, we have shared amazing moments whether at home or abroad. I have learnt a lot from you. Thanks for all your support. **Neda,** my wife, thank you for your support and the great time that we have shared throughout these years. My lovely son, **Nickan,** who cannot read this now, you opened a new chapter in my life.

Jag tror att jag måste säga några ord på svenska, och uttrycka min tacksamhet över att ha fått leva och arbeta i ett härligt land som Sverige.

References

1. Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 1st ed: John Murray; 1859. 386- p.
2. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016;17(7):422-33.doi: 10.1038/nrg.2016.58.
3. Foll M, Gaggiotti O. Identifying the environmental factors that determine the genetic structure of populations. Genetics. 2006;174(2):875-91.doi: 10.1534/genetics.106.059451.
4. Landauer W. Length of survival of homozygous creeper fowl embryos. Science. 1944;100(2607):553.
5. Barton N H. Evolution: Cold Spring Harbor Laboratory Press; 2007.
6. Sung W K. Pan-omics analysis of biological data. Methods. 2016;102:1-2.doi: 10.1016/j.ymeth.2016.05.004.
7. Hake S, Ross-Ibarra J. Genetic, evolutionary and plant breeding insights from the domestication of maize. Elife. 2015;4.doi: 10.7554/eLife.05861.
8. Quintana-Murci L, Clark A G. Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol. 2013;13(4):280-93.doi: 10.1038/nri3421.
9. Vander Wal E, Garant D, Festa-Bianchet M, Pelletier F. Evolutionary rescue in vertebrates: evidence, applications and uncertainty. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013;368(1610):20120090.doi: 10.1098/rstb.2012.0090.
10. Crawford D L, Oleksiak M F. Ecological population genomics in the marine environment. Brief Funct Genomics. 2016;15(5):342-51.doi: 10.1093/bfgp/elw008.
11. Kong A, Gudbjartsson D F, Sainz J, Jonsdottir G M, Gudjonsson S A, Richardsson B, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002;31(3):241-7.doi: http://www.nature.com/ng/journal/v31/n3/suppinfo/ng917_S1.html.
12. de Gortari M J, Freking B A, Cuthbertson R P, Kappes S M, Keele J W, Stone R T, et al. A second-generation linkage map of the sheep genome. Mamm Genome. 1998;9(3):204-9.
13. Rubin C J, Zody M C, Eriksson J, Meadows J R, Sherwood E, Webster M T, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464(7288):587-91.doi: 10.1038/nature08832.
14. Bourret V, Kent M P, Primmer C R, Vasemagi A, Karlsson S, Hindar K, et al. SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). Mol Ecol. 2013;22(3):532-51.doi: 10.1111/mec.12003.
15. Klein R J, Zeiss C, Chew E Y, Tsai J-Y, Sackler R S, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science (New York, N.Y.). 2005;308(5720):385-9.doi: 10.1126/science.1109557.

16. Lamichhaney S, Martinez Barrio A, Rafati N, Sundström G, Rubin C-J, Gilbert E R, et al. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(47):19345-50.doi: 10.1073/pnas.1216128109.
17. Rafati N, Andersson L S, Mikko S, Feng C, Raudsepp T, Pettersson J, et al. Large Deletions at the SHOX Locus in the Pseudoautosomal Region Are Associated with Skeletal Atavism in Shetland Ponies. G3: Genes|Genomes|Genetics. 2016;6(7):2213-23.doi: 10.1534/g3.116.029645.
18. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(48):19529-36.doi: 10.1073/pnas.1217149109.
19. Marguerat S, Bahler J. RNA-seq: from technology to biology. Cell Mol Life Sci. 2010;67(4):569-79.doi: 10.1007/s00018-009-0180-6.
20. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinformatics and Biology Insights. 2015;9(Suppl 1):29-46.doi: 10.4137/BBI.S28991.
21. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. 2013;14(2):R12.doi: 10.1186/gb-2013-14-2-r12.
22. Berger M F, Levin J Z, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, et al. Integrative analysis of the melanoma transcriptome. Genome Research. 2010;20(4):413-27.doi: 10.1101/gr.103697.109.
23. Pritchard V L, Viitaniemi H M, McCairns R J S, Merilä J, Nikinmaa M, Primmer C R, et al. Regulatory Architecture of Gene Expression Variation in the Threespine Stickleback Gasterosteus aculeatus. G3: Genes|Genomes|Genetics. 2017;7(1):165.
24. Wang G, Yang E, Smith K J, Zeng Y, Ji G, Connon R, et al. Gene expression responses of threespine stickleback to salinity: implications for salt-sensitive hypertension. Front Genet. 2014;5:312.doi: 10.3389/fgene.2014.00312.
25. Albert F W, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, et al. A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. PLoS Genet. 2012;8(9):e1002962.
26. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al. Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. PLOS Genetics. 2008;4(2):e1000010.doi: 10.1371/journal.pgen.1000010.
27. Lister R, Gregory B D, Ecker J R. Next is now: new technologies for sequencing of genomes, transcriptomes and beyond. Current opinion in plant biology. 2009;12(2):107-18.doi: 10.1016/j.pbi.2008.11.004.
28. Mack K L, Campbell P, Nachman M W. Gene regulation and speciation in house mice. Genome Res. 2016;26(4):451-61.doi: 10.1101/gr.195743.115.
29. Lehne B, Lewis C M, Schlitt T. From SNPs to Genes: Disease Association at the Gene Level. PLOS ONE. 2011;6(6):e20133.doi: 10.1371/journal.pone.0020133.
30. van Binsbergen R, Calus M P, Bink M C, van Eeuwijk F A, Schrooten C, Veerkamp R F. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2015;47:71.doi: 10.1186/s12711-015-0149-x.

31. Zhang J, Song Q, Cregan P B, Jiang G L. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). Theor Appl Genet. 2016;129(1):117-30.doi: 10.1007/s00122-015-2614-x.
32. Yoder J B, Stanton-Geddes J, Zhou P, Briskine R, Young N D, Tiffin P. Genomic signature of adaptation to climate in Medicago truncatula. Genetics. 2014;196(4):1263-75.doi: 10.1534/genetics.113.159319.
33. Barson N J, Aykanat T, Hindar K, Baranski M, Bolstad G H, Fiske P, et al. Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature. 2015;528(7582):405-8.doi: 10.1038/nature16062.
34. Feuk L, Carson A R, Scherer S W. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85-97.doi: 10.1038/nrg1767.
35. Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann D S, Hoeppner M P, et al. Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nat Genet. 2016;48(1):84-8.doi: 10.1038/ng.3430
http://www.nature.com/ng/journal/v48/n1/abs/ng.3430.html - supplementary-information.
36. Jones F C, Grabherr M G, Chan Y F, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55-61.doi: 10.1038/nature10944.
37. Pang A W, MacDonald J R, Pinto D, Wei J, Rafiq M A, Conrad D F, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biology. 2010;11(5):R52.doi: 10.1186/gb-2010-11-5-r52.
38. Maynard-Smith J, Haigh J. Hitch-Hiking Effect of a Favorable Gene. Genetics Research. 1974;23(1):23-35.doi: 10.1017/S0016672300014634.
39. Vasemägi A, Gross R, Palm D, Paaver T, Primmer C R. Discovery and application of insertion-deletion (INDEL) polymorphisms for QTL mapping of early life-history traits in Atlantic salmon. BMC Genomics. 2010;11(1):156.doi: 10.1186/1471-2164-11-156.
40. Bush W S, Moore J H. Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology. 2012;8(12):e1002822.doi: 10.1371/journal.pcbi.1002822.
41. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29-.doi: 10.1186/1746-4811-9-29.
42. Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195(1):205-20.doi: 10.1534/genetics.113.152462.
43. Coop G, Witonsky D, Di Rienzo A, Pritchard J K. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics. 2010; 10.1534/genetics.110.114819.doi: 10.1534/genetics.110.114819.
44. Andersson L, Rymani N, Rosenberg R, Stahli G. description of protein loci and population data. 1981;78:69-78.
45. Ryman N, Lagercrantz U, Andersson L, Chakraborty R, Rosenberg R. Lack of correspondence between genetic and morphologic variability patterns in Atlantic herring (Clupea harengus). Heredity. 1984;53(3):687-704.
46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
47. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012:1-9.
48. Koboldt D C, Chen K, Wylie T, Larson D E, McLellan M D, Mardis E R, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283-5.doi: 10.1093/bioinformatics/btp373.

49. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297-303.
50. Sudmant P H, Rausch T, Gardner E J, Handsaker R E, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.doi: 10.1038/nature15394 http://www.nature.com/nature/journal/v526/n7571/abs/nature15394.html - supplementary-information.
51. Kidd J M, Graves T, Newman T L, Fulton R, Hayden H S, Malig M, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell. 2010;143(5):837-47.doi: 10.1016/j.cell.2010.10.027.
52. Rausch T, Zichner T, Schlattl A, Stutz A M, Benes V, Korbel J O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333-i9.doi: 10.1093/bioinformatics/bts378.
53. Chen K, Wallis J W, McLellan M D, Larson D E, Kalicki J M, Pohl C S, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677-81.doi: 10.1038/nmeth.1363.
54. Sayyab S, Rafati N, Carneiro M, Garreau H, Andersson G, Andersson L, et al. A computational method for detection of structural variants using Deviant Reads and read pair Orientation: DevRO. bioRxiv. 2016.
55. Ashburner M, Ball C A, Blake J A, Botstein D, Butler H, Cherry J M. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25.doi: 10.1038/75556.
56. Schlotterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014;15(11):749-63.doi: 10.1038/nrg3803.
57. FAO. Fishery and Aquaculture Statistics. Production Statistics 1950-2014 (online query panel). FAO. Rome. 2016.
58. Bager M. Scandinavian Herring Fishery 2010. Available from: Scandinacian Herring Fishery.
59. Iles T D, Sinclair M. Atlantic Herring: Stock Discreteness and Abundance. Science. 1982;215(4533):627-33.doi: 10.1126/science.215.4533.627.
60. Kennedy J, Nash R D M, Slotte A, Kjesbu O S. The role of fecundity regulation and abortive maturation in the reproductive strategy of Norwegian spring-spawning herring (Clupea harengus). Marine Biology. 2011;158(6):1287-99.doi: 10.1007/s00227-011-1648-0.
61. McQuinn I H. Metapopulations and the Atlantic herring. Reviews in Fish Biology and Fisheries. 1997;7(3):297-329.doi: 10.1023/A:1018491828875.
62. Alves P C, Ferrand N, Hackländer K. Lagomorph Biology: Evolution, Ecology, and Conservation: Springer Berlin Heidelberg; 2007.
63. Lopez-Martinez N. The Lagomorph Fossil Record and the Origin of the European Rabbit. In: Alves PC, Ferrand N, Hackländer K, editors. Lagomorph Biology: Evolution, Ecology, and Conservation; 10.1007/978-3-540-72446-9_3. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 27-46.
64. Branco M, Ferrand N, Monnerot M. Phylogeography of the European rabbit (Oryctolagus cuniculus) in the Iberian Peninsula inferred from RFLP analysis of the cytochrome b gene. Heredity (Edinb). 2000;85 Pt 4:307-17.
65. Miller I, Rogel-Gaillard C, Spina D, Fontanesi L, de Almeida A M. The rabbit as an experimental and production animal: from genomics to proteomics. Curr Protein Pept Sci. 2014;15(2):134-45.

66. Peng X, Knouse J A, Hernon K M. Rabbit Models for Studying Human Infectious Diseases. Comp Med. 2015;65(6):499-507.
67. Phadnis N, Orr H A. A single gene causes both male sterility and segregation distortion in Drosophila hybrids. Science. 2009;323(5912):376-9.doi: 10.1126/science.1163934.
68. Mihola O, Trachtulec Z, Vlcek C, Schimenti J C, Forejt J. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science. 2009;323(5912):373-5.doi: 10.1126/science.1163601.
69. Bayes J J, Malik H S. Altered Heterochromatin Binding by a Hybrid Sterility Protein in Drosophila Sibling Species. Science. 2009;326(5959):1538-41.doi: 10.1126/science.1181756.
70. Larson G, Piperno D R, Allaby R G, Purugganan M D, Andersson L, Arroyo-Kalin M, et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci U S A. 2014;111(17):6139-46.doi: 10.1073/pnas.1323964111.
71. Darwin C. The Variation of Animals and Plants under Domestication. First ed. London: John Murray; 1868.
72. Wilkins A S, Wrangham R W, Fitch W T. The "domestication syndrome" in mammals: a unified explanation based on neural crest cell behavior and genetics. Genetics. 2014;197(3):795-808.doi: 10.1534/genetics.114.165423.
73. Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet. 2004;5(3):202-12.doi: 10.1038/nrg1294.
74. de Simoni Gouveia J J, da Silva M V G B, Paiva S R, de Oliveira S M P. Identification of selection signatures in livestock species. Genetics and molecular biology. 2014;37(2):330-42.
75. Carneiro M, Rubin C J, Di Palma F, Albert F W, Alfoldi J, Barrio A M, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science. 2014;345(6200):1074-9.doi: 10.1126/science.1253714.
76. Clutton-Brock J. A Natural History of Domesticated Mammals: Cambridge University Press; 1999.
77. Carneiro M, Afonso S, Geraldes A, Garreau H, Bolet G, Boucher S, et al. The Genetic Structure of Domestic Rabbits. Molecular Biology and Evolution. 2011;28(6):1801-16.doi: 10.1093/molbev/msr003.
78. Hall B K. Atavisms and atavistic mutations. Nat Genet. 1995;10(2):126-7.
79. Verhulst J. Atavisms in homo sapiens: A bolkian heterodoxy revisited. Acta Biotheoretica. 1996;44(1):59-73.doi: 10.1007/bf00046435.
80. Gingerich P D, Smith B H, Simons E L. Hind Limbs of Eocene Basilosaurus - Evidence of Feet in Whales. Science. 1990;249(4965):154-7.doi: DOI 10.1126/science.249.4965.154.
81. Hermans W A. A hereditary anomaly in Shetland ponies. Netherlands Journal of Veterinary Science. 1970;3:55-63.
82. Hermans W A, Kersjes A W, van der Mey G J, Dik K J. Investigation into the heredity of congenital lateral patellar (sub)luxation in the Shetland pony. Vet Q. 1987;9(1):1-8.doi: 10.1080/01652176.1987.9694070.
83. Andrén T, Björck S, Andrén E, Conley D, Zillén L, Anjar J. The Development of the Baltic Sea Basin During the Last 130 ka. In: Harff J, Björck S, Hoth P, editors. The Baltic Sea Basin; 10.1007/978-3-642-17220-5_4. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 75-97.
84. Andersson L, Ryman N, Rosenberg R, Stahl G. Genetic-Variability in Atlantic Herring (Clupea-Harengus-Harengus) - Description of Protein Loci and Population-Data. Hereditas. 1981;95(1):69-78.

85. Larsson L C, Laikre L, Palm S, Andre C, Carvalho G R, Ryman N. Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. Mol Ecol. 2007;16(6):1135-47.doi: 10.1111/j.1365-294X.2006.03217.x.
86. Larsson L C, Laikre L, Andre C, Dahlgren T G, Ryman N. Temporally stable genetic structure of heavily exploited Atlantic herring (Clupea harengus) in Swedish waters. Heredity. 2010;104(1):40-51.doi: 10.1038/hdy.2009.98.
87. Limborg M T, Helyar S J, De Bruyn M, Taylor M I, Nielsen E E, Ogden R, et al. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). Mol Ecol. 2012;21(15):3686-703.doi: 10.1111/j.1365-294X.2012.05639.x.
88. Knudsen S. Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics. 1999;15(5):356-61.doi: 10.1093/bioinformatics/15.5.356.
89. Hinegardner R, Rosen D E. Cellular DNA Content and the Evolution of Teleostean Fishes. The American Naturalist. 1972;106(951):621-44.doi: doi:10.1086/282801.
90. Ida H, Oka N, Hayashigaki K-i. Karyotypes and cellular DNA contents of three species of the subfamily Clupeinae. Japanese Journal of Ichthyology. 1991;38(3):289-94.doi: 10.1007/bf02905574.
91. Kawaguchi M, Fujita H, Yoshizaki N, Hiroi J, Okouchi H, Nagakura Y, et al. Different hatching strategies in embryos of two species, pacific herring Clupea pallasii and Japanese anchovy Engraulis japonicus, that belong to the same order Clupeiformes, and their environmental adaptation. J Exp Zool B Mol Dev Evol. 2009;312(2):95-107.doi: 10.1002/jez.b.21247.
92. Kawaguchi M, Yasumasu S, Shimizu A, Kudo N, Sano K, Iuchi I, et al. Adaptive evolution of fish hatching enzyme: one amino acid substitution results in differential salt dependency of the enzyme. J Exp Biol. 2013;216(Pt 9):1609-15.doi: 10.1242/jeb.069716.
93. Aneer G. Some Speculations About the Baltic Herring (Clupea-Harengus-Membras) in Connection with the Eutrophication of the Baltic Sea. Canadian Journal of Fisheries and Aquatic Sciences. 1985;42:83-90.
94. Hanon E A, Lincoln G A, Fustin J-M, Dardente H, Masson-Pévet M, Morgan P J, et al. Ancestral TSH Mechanism Signals Summer in a Photoperiodic Mammal. Current Biology. 2008;18(15):1147-52.doi: http://dx.doi.org/10.1016/j.cub.2008.06.076.
95. Nakao N, Ono H, Yamamura T, Anraku T, Takagi T, Higashi K, et al. Thyrotrophin in the pars tuberalis triggers photoperiodic response. Nature. 2008;452(7185):317-U1.doi: 10.1038/nature06738.
96. Ono H, Hoshino Y, Yasuo S, Watanabe M, Nakane Y, Murai A, et al. Involvement of thyrotropin in photoperiodic signal transduction in mice. Proceedings of the National Academy of Sciences. 2008;105(47):18238-42.doi: 10.1073/pnas.0808952105.
97. Kim H D, Choe H K, Chung S, Kim M, Seong J Y, Son G H, et al. Class-C SOX transcription factors control GnRH gene expression via the intronic transcriptional enhancer. Mol Endocrinol. 2011;25(7):1184-96.doi: 10.1210/me.2010-0332.
98. Melamed P, Savulescu D, Lim S, Wijeweera A, Luo Z, Luo M, et al. Gonadotrophin-releasing hormone signalling downstream of calmodulin. J Neuroendocrinol. 2012;24(12):1463-75.doi: 10.1111/j.1365-2826.2012.02359.x.
99. King M C, Wilson A C. Evolution at two levels in humans and chimpanzees. Science. 1975;188(4184):107-16.

100. Geraldes A, Carneiro M, Delibes-Mateos M, Villafuerte R, Nachman M W, Ferrand N. Reduced introgression of the Y chromosome between subspecies of the European rabbit (Oryctolagus cuniculus) in the Iberian Peninsula. Mol Ecol. 2008;17(20):4489-99.doi: 10.1111/j.1365-294X.2008.03943.x.
101. Geraldes A, Ferrand N, Nachman M W. Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (Oryctolagus cuniculus). Genetics. 2006;173(2):919-33.doi: 10.1534/genetics.105.054106.
102. Carneiro M, Ferrand N, Nachman M W. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (Oryctolagus cuniculus). Genetics. 2009;181(2):593-606.doi: 10.1534/genetics.108.096826.
103. Yan Z H, Medvedev A, Hirose T, Gotoh H, Jetten A M. Characterization of the response element and DNA binding properties of the nuclear orphan receptor germ cell nuclear factor/retinoid receptor-related testis-associated receptor. J Biol Chem. 1997;272(16):10565-72.
104. Lan Z J, Gu P, Xu X, Jackson K J, DeMayo F J, O'Malley B W, et al. GCNF-dependent repression of BMP-15 and GDF-9 mediates gamete regulation of female fertility. EMBO J. 2003;22(16):4070-81.doi: 10.1093/emboj/cdg405.
105. Walters K A, Simanainen U, Handelsman D J. Molecular insights into androgen actions in male and female reproductive function from androgen receptor knockout models. Human Reproduction Update. 2010;16(5):543-58.doi: 10.1093/humupd/dmq003.
106. Rossin E J, Lage K, Raychaudhuri S, Xavier R J, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 2011;7(1):e1001273.doi: 10.1371/journal.pgen.1001273.
107. Bateson W. Heredity and variation in modern lights. In: Seward AC, editor. Darwin and Modern Science. Cambridge: Cambridge University Press; 1909. p. 85-101.
108. Dobzhansky T. Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. Genetics. 1936;21(2):113-35.
109. Muller H J. Isolating mechanisms, evolution, and temperature. Biol. Symp. 1942;6:71-125.
110. Noor M A, Grams K L, Bertucci L A, Reiland J. Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci U S A. 2001;98(21):12084-8.doi: 10.1073/pnas.221274498.
111. Rieseberg L H. Chromosomal rearrangements and speciation. Trends Ecol Evol. 2001;16(7):351-8.
112. Leader B, Lim H, Carabatsos M J, Harrington A, Ecsedy J, Pellman D, et al. Formin-2, polyploidy, hypofertility and positioning of the meiotic spindle in mouse oocytes. Nat Cell Biol. 2002;4(12):921-8.doi: 10.1038/ncb880.
113. Matsui M, Motomura D, Karasawa H, Fujikawa T, Jiang J, Komiya Y, et al. Multiple functional defects in peripheral autonomic organs in mice lacking muscarinic acetylcholine receptor gene for the M3 subtype. Proc Natl Acad Sci U S A. 2000;97(17):9579-84.
114. Kimura M, Ishida K, Kashiwabara S I, Baba T. Characterization of Two Cytoplasmic Poly(A)-Binding Proteins, PABPC1 and PABPC2, in Mouse Spermatogenic Cells. Biology of Reproduction. 2009;80(3):545-54.doi: 10.1095/biolreprod.108.072553.
115. José Blanco-Aguiar et al. Intrinsic postzygotic barriers in an incipient species: implications on the conservation of European wild rabbit. Manuscript. 2017.

116. Orr H A. Genetics of male and female sterility in hybrids of Drosophila pseudoobscura and D. persimilis. Genetics. 1987;116(4):555-63.
117. Coyne J A, Orr H A. Speciation. Sunderland, MA: Sinauer Associates, Inc.; 2004. 545 p.
118. Dobzhansky T, Beadle G W. Studies on Hybrid Sterility IV. Transplanted Testes in Drosophila Pseudoobscura. Genetics. 1936;21(6):832-40.
119. Good J M, Giger T, Dean M D, Nachman M W. Widespread over-expression of the X chromosome in sterile Fhybrid mice. PLoS Genet. 2010;6(9).
120. Turner L M, White M A, Tautz D, Payseur B A. Genomic networks of hybrid sterility. PLoS Genet. 2014;10(2):e1004162.doi: 10.1371/journal.pgen.1004162.
121. Nielsen R, Williamson S, Kim Y, Hubisz M J, Clark A G, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15(11):1566-75.doi: 10.1101/gr.4252305.
122. Motazacker M M, Rost B R, Hucho T, Garshasbi M, Kahrizi K, Ullmann R, et al. A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. Am J Hum Genet. 2007;81(4):792-8.doi: 10.1086/521275.
123. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126(4):663-76.doi: 10.1016/j.cell.2006.07.024.
124. Raudsepp T, Das P J, Avila F, Chowdhary B P. The pseudoautosomal region and sex chromosome aneuploidies in domestic species. Sex Dev. 2012;6(1-3):72-83.doi: 10.1159/000330627.
125. Benito-Sanz S, Royo J L, Barroso E, Paumard-Hernandez B, Barreda-Bonis A C, Liu P, et al. Identification of the first recurrent PAR1 deletion in Leri-Weill dyschondrosteosis and idiopathic short stature reveals the presence of a novel SHOX enhancer. J Med Genet. 2012;49(7):442-50.doi: 10.1136/jmedgenet-2011-100678.
126. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78-94.doi: 10.1006/jmbi.1997.0951.
127. Kim H, Lee T, Park W, Lee J W, Kim J, Lee B Y, et al. Peeling back the evolutionary layers of molecular mechanisms responsive to exercise-stress in the skeletal muscle of the racing horse. DNA Res. 2013;20(3):287-98.doi: 10.1093/dnares/dst010.
128. Rao E, Blaschke R J, Marchini A, Niesler B, Burnett M, Rappold G A. The Leri-Weill and Turner syndrome homeobox gene SHOX encodes a cell-type specific transcriptional activator. Hum Mol Genet. 2001;10(26):3083-91.
129. Chen J, Wildhardt G, Zhong Z, Roth R, Weiss B, Steinberger D, et al. Enhancer deletions of the SHOX gene as a frequent cause of short stature: the essential role of a 250 kb downstream regulatory domain. J Med Genet. 2009;46(12):834-9.doi: 10.1136/jmg.2009.067785.
130. Jorge A A, Souza S C, Nishi M Y, Billerbeck A E, Liborio D C, Kim C A, et al. SHOX mutations in idiopathic short stature and Leri-Weill dyschondrosteosis: frequency and phenotypic variability. Clin Endocrinol (Oxf). 2007;66(1):130-5.doi: 10.1111/j.1365-2265.2006.02698.x.
131. Al-Shami A, Spolski R, Kelly J, Fry T, Schwartzberg P L, Pandey A, et al. A role for thymic stromal lymphopoietin in CD4(+) T cell development. J Exp Med. 2004;200(2):159-68.doi: 10.1084/jem.20031975.
132. Li J. JAK-STAT and bone metabolism. JAKSTAT. 2013;2(3):e23930.doi: 10.4161/jkst.23930.

133. Nakane Y, Ikegami K, Iigo M, Ono H, Takeda K, Takahashi D, et al. The saccus vasculosus of fish is a sensor of seasonal changes in day length. Nat Commun. 2013;4:2108.doi: 10.1038/ncomms3108.
134. Pool J E, Braun D T, Lack J B. Parallel Evolution of Cold Tolerance Within Drosophila melanogaster. Mol Biol Evol. 2016; 10.1093/molbev/msw232.doi: 10.1093/molbev/msw232.
135. Laakkonen H M, Lajus D L, Strelkov P, Vainola R. Phylogeography of amphi-boreal fish: tracing the history of the Pacific herring Clupea pallasii in North-East European seas. BMC Evol Biol. 2013;13:67.doi: 10.1186/1471-2148-13-67.
136. Laakkonen H M, Strelkov P, Lajus D L, Väinölä R. Introgressive hybridization between the Atlantic and Pacific herrings (Clupea harengus and C. pallasii) in the north of Europe. Marine Biology. 2015;162(1):39-54.doi: 10.1007/s00227-014-2564-x.
137. Blaschke R J, Rappold G. The pseudoautosomal regions, SHOX and disease. Curr Opin Genet Dev. 2006;16(3):233-9.doi: 10.1016/j.gde.2006.04.004.
138. Imsland F, McGowan K, Rubin C J, Henegar C, Sundstrom E, Berglund J, et al. Regulatory mutations in TBX3 disrupt asymmetric hair pigmentation that underlies Dun camouflage color in horses. Nat Genet. 2016;48(2):152-8.doi: 10.1038/ng.3475.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1301

Editor: The Dean of the Faculty of Medicine

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".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-315032

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2017