Using whole-genome sequencing data for demographic and functional evaluations of small managed populations

Chiara Bortoluzzi

Propositions:

1. Local chicken breeds harbour a wealth of genetic variation that can only be effectively conserved if characterised in terms of demography and function (this thesis)

2. Sequence conservation needs to be coupled to functional data to become the next genomics-based approach to identify functionally important variation (this thesis)

3. Science and technology are an illusionary hope to overcome the world's instability

4. Conservation is not neutral to a species' popularity

5. Projecting the flaws of your own country onto another is an indirect way of not accepting its flaws

6. Cultural appropriation should be acknowledged as an opportunity to oppose racism

Propositions belonging to the thesis, entitled:
Using whole-genome sequencing data for demographic and functional evaluations of small managed populations
Chiara Bortoluzzi, 14 October 2020, Wageningen

Using whole-genome sequencing data for demographic and functional evaluations of small managed populations

Chiara Bortoluzzi

Thesis committee

Promotor:

Prof. dr. M.A.M Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Co-promotors:

Dr. H-J.W.C Megens
Assistant Professor, Animal Breeding and Genomics
Wageningen University & Research

Dr. M Bosse
Postdoctoral researcher, Animal Breeding and Genomics
Wageningen University & Research

Other members:

Prof. dr. K van Oers, Netherlands Institute of Ecology
Dr. F Leenstra, Wageningen Livestock Research
Prof. dr. C van Oosterhout, University of East Anglia, United Kingdom
Dr. N Duijvesteijn, Hendrix Genetics, the Netherlands

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS)

Using whole-genome sequencing data for demographic and functional evaluations of small managed populations

Chiara Bortoluzzi

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on 14 October 2020
at 16:00 in the Aula.

Bortoluzzi, C.
Using whole-genome sequencing data for demopraphic and functional evaluations
of small managed populations

PhD thesis, Wageningen University, the Netherlands (2020)
With references, with summary in English

ISBN: 978-94-6395-482-2
DOI: https://doi.org/10.18174/528193

Abstract

Bortoluzzi, C. (2020). Using whole-genome sequencing data for demographic and functional evaluations of small managed populations. PhD thesis, Wageningen University & Research, the Netherlands.

Genetic diversity is the foundation for selection to act upon a population: without diversity, evolution can not occur and species can not adapt to changing environments. In this thesis, I provide an in-depth analysis of the genome-wide patterns of diversity in local chicken breeds of small effective population size. Since their domestication in Southeast Asia around 9,500 years ago, domestic chickens have undergone human-driven selection resulting in the creation of hundreds of breeds, each described by a precise set of morphological features. Domesticated chicken breeds are therefore an excellent study system to investigate the effects of demography on genomic variation. Using a combination of the latest genomics tools, I show that, besides signs of recent inbreeding and declined diversity, changes in breeding preferences generated novel and identifiable variation. Interestingly, the genetic basis of some of this variation has evolved in other bird species through parallel evolution despite their divergence from chicken millions of years ago. However, as I emphasize, the outstanding diversity harboured by local chicken breeds can only be preserved in the near future if conservation programmes become genomics-informed. The rationale is the ability of genomic data to provide additional information on the functional relevance of such variation, which has important implications for conservation. Therefore, by means of genomic data, we can better control for deleterious alleles, while increasing genetic diversity. The identification of functionally important mutations have for a long time been limited to protein-coding genes. In this thesis, I demonstrate through the development of the ch(icken)CADD model that sub-regions within conserved non-coding elements also harbour variants with a negative fitness effect, as their association with known disease genes in other vertebrate species demonstrate. Overall, the findings of this thesis show that genetic diversity should be characterized from both a demographic and functional perspective to best manage populations and genetic resources.

Contents

The future is unwritten

Joe Strummer

1.

Introduction

1.1 Introduction

World biodiversity is declining globally at rates unprecedented in human history. Temporal analysis of biodiversity loss in the last century estimated the current extinction rates to vastly exceed natural average background rates [44], supporting the claim that Earth's biodiversity is on a trajectory for a sixth mass extinction event [14]. According to the recent Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES), at least 680 vertebrate species have been driven to extinction since the 16^{th} century [34]. The picture is even more bleak if we consider that 32% of known vertebrate species show substantial population declines [44], while for a great many the risk status is unknown.

Although wild species are emblematic to biodiversity, it is important to realize that also livestock breeds are an important component of world biodiversity [107]. Livestock breeds harbour genes and genetic variation of value for agriculture in the future, including adaptability and resilience in the face of climate change, emerging diseases, and shifting market demands [262]. Similarly to wild diversity, farm diversity is currently under threat: 10% of domesticated breeds of mammals and 3.5% domesticated breeds of birds became extinct by 2016 [34]. Most importantly, 67% of domesticated breeds still have an unknown risk status [262].

In the last decades, efficient breeding programmes and genomic selection for few, highly productive transboundary breeds [216, 95, 268] have accelerated the extinction rate of many local breeds. The rapid decline in local livestock diversity calls for immediate and effective conservation measures to prevent future breeds – the biological unit relevant for management – to go extinct. Conservation measures are needed now more than ever because many local livestock breeds are facing population decline. If unrestrained, population decline may shortly threaten the ability of local breeds to adapt and thus survive. However, parameters related to, for instance, genetic diversity and inbreeding should be estimated before setting up a breeding or conservation programme, in order to successfully intervene. Whole-genome sequencing (WGS) data are an invaluable tool to more precisely estimate population parameters [8]. Nonetheless, the use of WGS data to characterize local breeds diversity has only recently become a feasible option in conservation, as world-wide projects illustrate (e.g. IMAGE).

In this thesis, I explore the use of whole-genome sequencing data to characterize livestock population and functional diversity to guide conservation efforts. I illustrate this using local chicken breeds of divergent demographic and selection history as a case study. As I show, the long-lasting human-driven phenotypic selection has created, with some genetic costs, a wide diversity of local breeds displaying a wide range of within and among breeds phenotypic variation. However, the rich phenotypic diversity can only be preserved if well-informed conservation programmes are in place.

1.2 Population size and genetic diversity

Genetic variation provides the foundation for natural selection to act upon a population. From a genetic standpoint, genetic diversity, also known as genetic polymorphism, is the variation in a DNA sequence (i.e. alleles, genotypes) between distinct individuals of a given population (or species). Therefore, individuals in a population have underlying different combinations of variants, which are key to their evolution. The distribution of genetic variation in a population is constantly changing as a result of intra-genomic features and external forces influenced by an individual's demography. In this section, I discuss the role of demography in determining the level of genetic variation in a population. The functional relevance of genetic variation is discussed in **section 1.3**.

1.2.1 N_e is key to understand current genetic diversity of a population

To fully understand and appreciate the genetic diversity of species, the first key step is to unravel a species demographic history [275, 93]. The neutral theory of molecular evolution predicts that, in a population of constant size, genetic diversity is proportional to the product of the effective population size, N_e (**BOX 3**), and the mutation rate per site per generation, μ [164]. The effective size of a population, N_e, is an evolutionary parameter that determines how fast the composition of a population is expected to change as a result of genetic drift (**BOX 3**). N_e is crucial in determining the level of variability in a population [47]. Hence, controlling demographic processes is critical to prevent future loss of diversity.

The direct estimation of N_e is generally impractical, because knowledge on, for e.g., the total number of breeding individuals of a species (i.e. census population size or N_c), is difficult to obtain [93]. However, in commercial breeding, the direct estimation of N_e has become possible since the development of medium and high density SNP arrays. Next-generation sequencing technologies now allow an even more precise estimation of N_e, which, in commercial breeding, is further improved by the availability of rather complete pedigree information. Although possible in a commercial setting, for local breeds and wild species N_e can only be indirectly estimated from genome-wide information. Therefore, it is not surprising to see that N_e values are often lower than those of N_c [47].

Demography is a major determinant of the current levels of genetic diversity. Knowledge about past demographic processes is key to understand the relationship between population size and diversity [219, 311, 108]. However, past trajectories of changes in N_e are less informative for planning conservation actions [305]. The current N_e is a crucial parameter in conservation genetics because it provides insights into the rate at which a population is losing genetic diversity. For population management, N_e estimated from all current breeding individuals is the best measure for informing conservation or breeding programmes. However, as I describe, current N_e is not sufficient to define the conservation status of a species.

1.2.2 Genetic drift

Another important aspect of N_e is that it quantifies the rate of change in the composition of a population caused by genetic drift (**BOX 3**) and the rate of inbreeding (**BOX 3**) [47, 301]. Genetic drift is one of the basic mechanisms of evolution along with mutation and natural selection. The survival and reproduction of some, but not all, individuals in a population is a common phenomenon that occurs in each generation. In fact, just by chance, these individuals, which are not necessarily the fittest, can leave behind a few more descendants (and genes), contributing more to the gene pool of the next generation.

Genetic drift is an inevitable evolutionary force. However, in small and isolated populations genetic drift can accelerate the loss of genetic diversity, as alleles can randomly either drift to fixation or being lost. Over time, if unrestrained, the amount of genetic variation necessary for adaptation is expected to erode, taking a population (or species) to the brink of extinction [162]. Although the effects of genetic drift on individual genomes have been thought to be gradual, recent studies have shown that the response of a population to genetic drift depends on the type and time-frame of the bottleneck [25] (**Chapter 3**), the (long-term) persistence of a population small in size [174, 256], and the presence of a conservation programme [1, 28] (**Chapter 4**).

1.2.3 Inbreeding and runs of homozygosity (ROHs)

Inbreeding is an important threat that can affect population persistence, making it a key concern for conservation biologists [162]. Compared to genetic drift, inbreeding can act swiftly causing, in few generations, a severe decline in population variation. Inbreeding does not change the frequency of alleles in a population, but it redistributes the frequency of genotypes: individuals that are homozygous for alleles that are identical-by-descent (IBD) (**BOX 3**) increase in frequency over heterozygotes [162]. Inbreeding generally causes offspring to have lowered fitness, a phenomenon known as inbreeding depression [160] (**BOX 3**). Inbreeding depression can be caused by increased homozygosity at loci with deleterious recessive alleles, or decreased heterozygosity at loci displaying heterozygous advantage [159]. To date, evidence has been reported for both mechanisms.

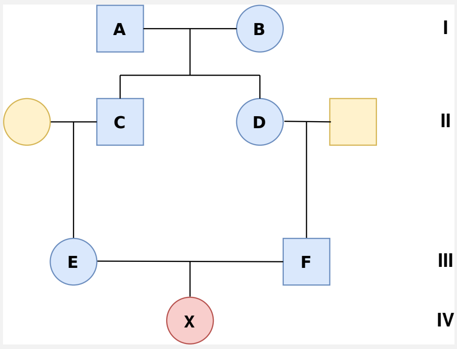
Charles Darwin was among the first to note the effects of inbreeding depression. In his experiments in 57 species of plants from 52 genera and 30 families, Darwin observed that offspring of inbred plants were on average shorter, flowered later, and produced fewer seeds than offspring of outbred plants [66, 67]. Studies in wild and domesticated species have shown that inbreeding depression can be a common phenomenon also in animals, as seen in numerous fitness-related traits [147, 152]. Although inbreeding depression may play a role in a population collapse [257], it is important to realize that stochastic demographic and environmental events may have a more profound acute outcome.

The degree of individual inbreeding has traditionally been defined by the pedigree inbreeding coefficient (F_{ped}) (**BOX 1**). However, large-scale molecular data have dramatically improved our ability to estimate individual inbreeding and inbreeding depression. Since the discovery in the mid-1990s, uninterrupted autozygous segments, also called runs of homozygosity (ROHs) (**BOX 3**), have repeatedly been used to quantify individual autozygosity [43]. Moreover, by characterizing ROHs important insights into a population demographic history can be derived (e.g. historical fluctuations in N_e). Different population histories give rise to different ROHs size distribution, ranging from short (a few kilobases) indicating mating between distantly related individuals (background relatedness) to long (a few megabases) resulting from mating between recent common ancestors [204, 43]. Long ROHs are often abundant in small populations, because mating among relatives is an inevitable process even if species have evolved inbreeding avoidance mechanisms. However, in domesticated species, mating between family members can be intentionally applied for trait selection [26] **(Chapter 2**).

Studies in humans and livestock have generally found that the abundance of ROHs within populations follows a nonuniform distribution along the genome. In fact, regions of high ROH frequency (i.e. ROH hotspots) are alternated with regions of low ROH frequency (i.e. ROH coldspots). ROH hotspots and coldspots directly affect the distribution of genetic diversity and their existence is explained, almost entirely, by individual recombination rate [236]. Because of the nonuniform ROHs distribution, ROH hotspots are frequently used to identify regions under selection, since also selection increases overall autozygosity [163]. However, only the sharing of ROHs among individuals in a population can be viewed as a candidate selective sweep. Interestingly, autozygosity mapping has also been used to identify loci harbouring mutations underlying diseases or genetic defects in cattle [52], pig [78, 77], and human [303].

BOX 1

Individual inbreeding is often quantified by two major inbreeding coefficients based either on the pedigree (F_{ped}) or molecular data (F_{ROH}). The pedigree inbreeding coefficient predicts the probability of a locus being IBD based on a known pedigree where the founders are assumed to be unrelated and non-inbred [162]. F_{ped} has traditionally been the cornerstone of studies on individual inbreeding. However, F_{ped} can only predict the *expected* proportion of the genome that is IBD, because of Mendelian sampling, linkage disequilibrium (LD), limited pedigree depth, and biased assumption on founders unrelatedness [160]. In wild and local livestock breeds, estimates of $F ped$ are often challenged by the difficulties in recording individuals' relatedness.



In this example, E and F are related because their common ancestors are A and B, which are assumed to be unrelated. To calculate the $F_{ped,X}$, we need to go through all possible routes in the pedigree between E and F over each common ancestor:

Route 1 (4 generations): E-C-A-D-F = $(1/2)^4$ = 1 / (2*2*2*2) = 1/16
Route 2 (4 generations): E-C-B-D-F = $(1/2)^4$ = 1 / (2*2*2*2) = 1/16
$a_{E,F}$ = 1/16 + 1/16 = 1/8, from which $F_{ped,X}$ = $a_{E,F}/2$ = 1/16 = 0.0625 = 6.25%

Since the advent of next-generation sequencing, molecular markers have increasingly been used to estimate *F*. With the reduction in sequencing costs, F_{ped} is often replaced by the genomic inbreeding coefficient F_{ROH}, which is the proportion of the genome that is in ROHs [204]. Therefore, F_{ROH} can be used to measure individual inbreeding due to ancient or recent ancestors based on the ROH size considered. Compared to F_{ped}, F_{ROH} measures the *actual* proportion of the autosomal genome that is autozygous. Recent studies in wild [257, 159] and domesticated species [28] have shown that F_{ped} positively correlates with F_{ROH}. However, all studies consistently report an underestimation of the individual's inbreeding coefficient by F_{ped}, arguing that, where possible, genomic-based inbreeding should be preferred over pedigree-based inbreeding. However, the use of F_{ROH} over F_{ped} in the context of small populations not under a conservation programme strongly depends on management practices and available funding.

1.3 Functional evaluation of genetic diversity

1.3.1 Deleterious alleles

From a functional perspective, genetic variation can be neutral, adaptive [253, 116], and deleterious, with important conceptual implications for conservation [136, 170] (**BOX 3**). According to the population genetics theory, deleterious alleles are maintained in a population by an equilibrium between mutation, selection, and N_e (mutation-selection balance). The mutation-selection balance illustrates two important aspects. First, that the level of deleterious variation could depend, to a certain extent, on the demographic history of the population. And second, that small populations are more prone to accumulate deleterious alleles, because their N_e is not large enough to allow purifying selection to remove weakly deleterious alleles. Hence, in small populations, deleterious alleles are likely to accumulate and rise in frequency owing to genetic drift [47, 170].

Most deleterious alleles are present in a population at low allele frequency and have a mild deleterious effect. However, in small-sized populations inbreeding can amplify the harmful effects on fitness of deleterious variants in homozygous state [29]. Some of these deleterious alleles are lethal recessive as they cause pre- or postnatal mortality in homozygous individuals by affecting genes linked to fertility, development and growth [76]. Nevertheless, studies have now shown that recessive lethals can also exert their effects later in gestation or postnatally [74]. Therefore, if not purged, harmful homozygous alleles can lead in the worst case scenario to a 'mutational meltdown' and a population (or species) extinction [162].

Population bottlenecks can remarkably shape the deleterious variation landscape of a species. In humans, Out-of-Africa populations display a shifted allele frequency spectrum of deleterious variants that has been shaped by genetic drift accompanying the bottleneck [140]. Similar shifts in deleterious variation were also observed in, among others, horse [263], dog [200], and farm species [197], a phenomenon we commonly refer to as the genetic 'cost of domestication' [62, 214]. However, since the first, and most severe, bottleneck, the demographic history of many domesticated species has been characterized by repeated declines and/or expansions, accompanied by strong artificial selection for desirable traits that may have been deleterious in the ancestral population. Interestingly, studies have also shown that artificial selection is indirectly responsible for part of an individual's deleterious variation landscape, as deleterious variants can hitchhike with nearby positively selected alleles increasing in frequency [200].

In the context of small populations, recent studies in wild and livestock species have shown that deleterious alleles relate to N_e in a complex fashion. In their study on the Channel island fox, Robinson and colleagues (2016) observed that the long-term small population size of the island fox has enabled the purging of strongly deleterious alleles, resulting in no signs of inbreeding depression [255]. The hypothesis that the type and time-frame of a population decline have important consequences on a population's genetic variation was more formally analysed in local chicken breeds by Bortoluzzi et al. (2020) (**Chapter 3**). In fact, the authors showed that populations that remained small for a long period of time have been able to minimize the effects of genetic drift on the deleterious variation contrary to recently bottlenecked populations that are more susceptible to genetic stochasticity [25]. These conclusions have important implications for management, as also illustrated in recent simulations of Kyriazis et al. (2019).

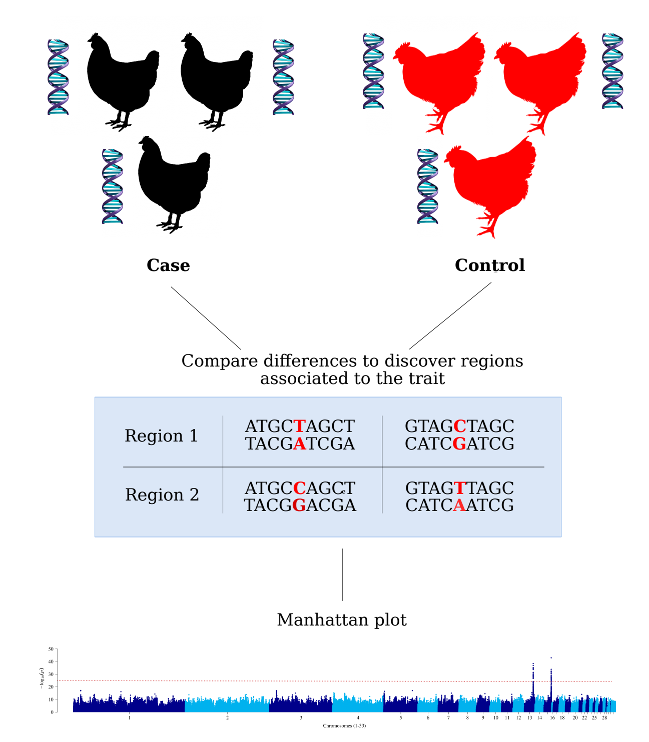
The first important step in the study of deleterious variation is the identification of variants (SNPs and InDels) with a likely detrimental effect. From a genome-wide perspective, this step can now easily be carried out using several prediction tools, such as SIFT [203], PolyPhen-2 [2], and PROVEAN [54]. However, the major constraint of these methods is their limitation to variants altering the DNA sequence of protein-coding genes, including loss-of-function (LoF) mutations and point missense mutations. Although protein-coding mutations are extremely important, they account for only a small fraction of an individual's genome. Moreover, there is increasing evidence that phenotypic changes are not necessarily due to deleterious changes in the protein sequence, suggesting that variants that lie outside coding regions (i.e. regulatory mutations) can also alter gene regulation and expression. Therefore, to fully understand the astonishing phenotypic diversity of our farm animal populations, it is necessary to extend the research focus to each variant, independently of its coding potential. As I will discuss in section 2.3 new methods that go beyond the protein-coding definition of a mutation are already on the way.

1.3.2 Gene-to-phenotype mapping

Farm animal populations harbour a rich collection of mutations with phenotypic effects that have been purposefully selected by humans for thousands of years since the domestication bottleneck. Human-driven selection has resulted in a remarkable variety of phenotypic changes in morphology, physiology, and behaviour, making domestic animals an excellent model to study the genetic basis of phenotypic evolution [260]. Some of these traits have a simple monogenetic basis [217], but most have a multi-factorial complex inheritance [135]. Even though most mutations underlying specific traits do not have pathological phenotypic consequences [10], it is of extreme importance to identify and characterize genomic loci associated with various phenotypes. In fact, it is only by uncovering the genetic basis of phenotypes that we can fully preserve farm animal genetic diversity. Whole-genome sequencing data have revolutionized our understanding of phenotypes. At present, genome-wide association studies (GWAS) are routinely performed to associate genomic loci to the target trait (**BOX 2**). GWAS have enabled the identification of genes regulating body size in cattle [32], number of teats (NTE), number of vertebrae (NVE), and number of ribs (RIB) in pig [298], and a variety of complex diseases in dog [135] and human [300]. Interestingly, GWAS have also elucidated the evolutionary history of similar traits evolved by different genetic changes (e.g. feathering rate in turkey and chicken [76]), or similar genetic changes (e.g. foot feathering in chicken and pigeon [27] (**Chapter 5**). Although GWAS have considerably advanced our understanding of the genetic basis of (complex) traits, they are characterized by several limitations that have only recently been addressed. A major limitation derives from the prevalence of non-protein-coding variants in most GWA studies, whose function is very often difficult to unravel as appropriate species-specific methods are lacking. Hence, many studies often report either the genomic loci, assuming that it affects the closest neighbouring gene(s), or the protein-coding variants whose function is relatively easy to prove. Another limitation is the correlation between neighbouring variants in linkage disequilibrium (LD) (**BOX 3**). High LD limits our ability to fine map candidate loci (often megabases in size) to few base pairs, because LD results in potentially thousands of variants to be similarly associated with a phenotype. Genomic distance can reasonably play a role in determining the phenotype. However, there is an increasing number of examples that show that regulatory elements and non-protein-coding variants have important consequences on an individual's phenotype, making the assumption on the genomic distance simplistic [7]. As I describe in the next section, studies at the interface of population genomics and comparative genomics are the future genomic resources that will allow us to pinpoint and unravel the evolutionary and functional importance of genetic variation in a species genome.

BOX 2

A genome-wide association study (GWAS) is an hypothesis-free approach for identifying associations between genetic regions (loci) and traits (including diseases). A typical GWA study is shown in the figure below. Genomic information coming from either SNP arrays or WGS are collected from individuals that, based on the presence or absence of the trait of interest, are grouped under the case or control group, respectively.



Region 1	ATGCTAGCT TACGATCGA	GTAGCTAGC CATCGATCG
Region 2	ATGCCAGCT TACGGACGA	GTAGTTAGC CATCAATCG

Variants (SNPs and InDels) are then compared across the complete sets of genomes of cases and controls to identify variants associated with a particular trait. Statistical analyses are thus carried out to calculate the probability that an allele is likely to be associated with the trait. The probability is captured by the p-value expressed as corrected -$\log_{10}$(*p-value*) after multiple testing.

Although it has long been known that genetic variation between individuals can cause differences in phenotypes, genome-wide association studies do not generally identify the actual causal variants. In most cases, these studies identify common variants that, because of LD, tag along a region containing the causal variant(s). Therefore, follow-up studies are usually required to further narrow down the candidate region associated with the trait. Once identified, causal variant(s) can be targeted by artificial selection in future breeding programmes to increase their frequency in a population, or, if linked to a deleterious phenotype, be used to identify carriers that can thus be excluded from the mating system.

1.3.3 Non-protein-coding variants

One of the biggest challenge of current genomic research is understanding the evolutionary variation observed within and across species and the genetic mechanisms underlying it. Comparative genomics has proven itself to be an invaluable approach for dissecting the genetic basis of phenotypic variation. As shown in Figure 1.1, comparative genomics relies on the comparison of several species genome. Species included in the alignment are chosen based on a specific phylogenetic scope. By comparing genomic sequences among species, functionally important regions can be identified and be targeted by subsequent functional analysis. In the framework of comparative genomics, genomic regions that display high sequence similarity (conservation) among species are more likely to be functional [7]. This is because such conserved regions are highly constraint, meaning that purifying selection acts against the rise of new mutations. Programmes, such as phastCons [266], phyloP, and GERP [69], are among the most widely used to predict the level of conservation of each aligned base in the multiple sequence alignment to subsequently identify conserved elements (CEs). Although conservation detected by a comparative genomic approach may provide important clues on the functional relevance of non-protein-coding mutations, constraint alone does not reveal what that function is. To circumvent this, one must instead rely on combining constraint with other types of annotation data coming from functional studies, such as RNA-seq, ChIP-seq, DNA methylation, histone modifications, and Hi-C [7]. In livestock, functional data, which are often complementary to sequence conservation analysis, are rather limited, although significant progress is currently on the way [118].

At present, new opportunities are coming from human genetics, where the development of the Combined Annotation-Dependent Depletion (CADD) framework [168] provides a better functional prediction of (non-protein-coding) variation. In fact, the advantage of such annotation is its integrative nature, as it combines a diverse set of genomic features, including sequence context, gene model annotations, evolutionary constraint, epigenetic measurements, and functional predictions [251]. In human genetics, the primary use of CADD has been to identify variants that are most likely deleterious and potentially pathogenic (e.g. Mendelian disorders) [251]. For instance, CADD has allowed the discovery of *de novo* dominant and recessive variants in family-based sequencing [293, 294, 269] and population-based studies [309]. Despite its importance in human genetics, the CADD framework has recently been extended to non-human species, including mouse [127] and pig [128], with promising results. As I show in **Chapter 6**, the application of CADD to chicken can accelerate the discovery of mutations with a phenotypic effect, deleterious or not, making prioritization of variants in future population genomics, GWAS, and functional studies possible.

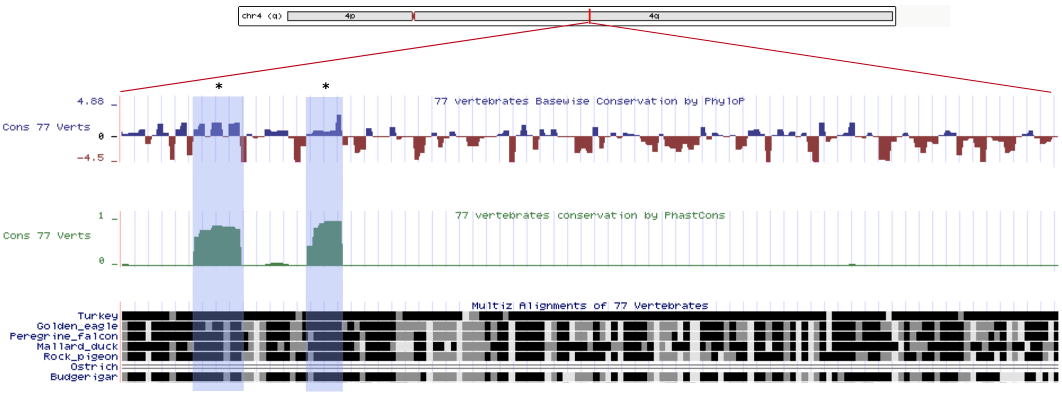


Figure 1.1: Distribution of conserved elements in a 157 bp region on chicken chromosome 4 (chr4: 45,669,249-45,669,405). The identification of conserved elements (highlighted in blue) relies on the availability of a multi-species genome alignment, where species are aligned against each other or to a reference. The phylogenetic distance among species in the alignment is an important factor that can affect the identification of conserved elements.

1.4 Chicken: a model species in demographic and functional studies

Understanding genetic diversity from a demographic and functional perspective is a challenging task. Domesticated species have been, and still are, important models for unravelling the consequences of selection and demography on phenotypes [205]. This is due to their very often well documented demographic history and availability of data, ranging from a good reference genome to a multitude of SNP assays to functional data (i.e. Functional Annotation of ANimal Genomes - FAANG) [9]. In the next paragraphs, I explain the rationale of using chicken as a model species, providing a brief description of the main milestones that signed its evolutionary history.

1.4.1 Suitability of chicken as a model species

Aside from a suitable history and documentation that dates back to *The Variation of Animals and Plants Under Domestication* [66], the availability of detailed genetic information is crucial to study the intra-genomic and external forces shaping genomic variation. Domestic chicken (*Gallus gallus domesticus*) was the first livestock species to have its genome sequenced [145]. The generation of a first draft sequence of the chicken genome represented a milestone in the field of animal genetics contributing to the genomics revolution. Since then, several improved genome assemblies have been published with the latest - GRCg6a - submitted by the Genome Reference Consortium in April 2018. The publication of the first genome assembly was immediately followed by the first genetic variation map (2.8 million SNPs) [58], the first high-density SNP-based linkage map [124], and several SNP arrays, ranging from a medium (60K) [123] to high (600K) density [173]. Although the design of SNP arrays has allowed for an efficient and cheap genome-wide characterization of variation across a wide range of domesticated and wild forms, the advent of next-generation sequencing (NGS) technologies has opened up even greater windows of opportunities to study genetic diversity at a resolution that scientists only dreamed of. Hence, the combination of genomic data with the evolutionary history of chicken offers an unprecedented model to characterize genetic diversity from a demographic and functional perspective.

1.4.2 It all began in the wild Asian forests

The world-wide distribution of chicken is a fascinating example of a successful story of cross-species relationship with human. Although chicken domestication has been a controversial topic for decades, a recent study published in *Cell Research* by Wang and colleagues (2020) seems to have taken the field a step forward in answering key questions on the geographic and temporal origins of chicken domestication [302]. Erasmus, Darwin's grandfather, was one of the first to suggest that the red junglefowl (*Gallus gallus*) was the wild ancestor of the domestic chicken. Darwin backed this theory up while formulating his theory of evolution, postulating a monophyletic origin of the domestic chicken identifying in the Indus valley the centre of domestication [66]. Darwin's idea was initially supported by genetic studies on mitochondrial DNA [112, 113] and retrovirus insertions [111]. However, with the advent of SNP arrays and next-generation sequencing technologies, chicken domestication was proposed to have occurred via multiple independent events in different areas of Asia between 7,000 and 10,000 AC [189, 158] from at least two wild progenitors [94]. The 'multiple origins' hypothesis has recently been questioned by Wang et al. (2020). In their study on 863 chicken genomes from a variety of domestic breeds and wild (sub)species, the authors showed that domestic chickens were initially derived from the wild red junglefowl subspecies *G.g. spadiceus*, which is endemic to Southwestern China, Thailand, and Myanmar. According to their molecular clock analysis, domestic chickens diverged from *G.g. spadiceus* 9,500±3300 years ago, after which they were translocated across Southeast and South Asia where they interbred locally with other subspecies and junglefowl species [302]. Although this study provides new insights into the domestication process of chicken, relevant questions on how many *G.g. spadiceus* lineages were involved in the initial domestication process remain unanswered. Whole-genome sequencing data holds a great promise for clarifying the chicken domestication process in the near future.

1.4.3 Chicken genome: a still largely unknown chest of wonders

Domestication and thousands of (chicken) generations of selection have left clear signatures on the genome. Unlike its ancestors, modern-day chickens display an impressive phenotypic diversity that is comparable to only few other domesticated species, such as dogs [228] and pigeons [83]. The remarkable phenotypic diversity displayed by the domestic chicken owes its origin to the standing genetic variation existing in the ancestral population (the effective population size must, in fact, have been quite large) [260]. This is consistent with the rather important sequence diversity (5 SNPs per kilobase) [145], which is six- to seven-fold larger than that of humans and domestic dog [232] and three-fold larger than gorillas [319]. Nevertheless, mutations arose during domestication and (repeated) hybridization between wild and domestic populations are also thought to have played a major role [177]. For instance, the haplotype responsible for the yellow skin phenotype found in modern commercial lines and several local breeds was acquired through introgression from the grey junglefowl (*Gallus sonneratii*), which is endemic to South Asia [94].

Archaeological findings agree with the hypothesis that leisure, cockfighting, and religion were the initial drivers of domestication [286]. Birds were selectively bred for specific physical traits (particularly colours and/or morphological features such as comb size), which have often a selective disadvantage in the wild. It is also likely that cultural values attributed to certain features further contributed to the large morphological diversity within chickens. The subsequent geographical dispersal of chickens from their centres of domestication resulted in the creation of many local breeds [242], which already existed in the 18^{th} century when livestock diversity started to be documented [262]. During this time, many local breeds underwent selection for subsets of morphological variants, resulting in the creation of breed standardized descriptions [242]. On top of that, a breed can be subdivided into different varieties, which are usually (but not always) identified by a specific plumage colouration or patterning. As an example, according to the Hollandse Kriel Club, over 20 different colour varieties are recognised for the Dutch bantam, a small variety of fowl that does not have any large size counterpart (http://www.hollandsekriel.nl).

1.4.4 Local chicken breeds

The development of the quantitative genetics theory in the mid 20^{th} century had a major impact on traditional chicken breeds diversity. On one side, the creation of commercial lines highly specialized for either meat (broiler lines) or egg production (layer lines) remarkably improved in few generations the capacity of the newly established industrialized poultry farming to answer increasing market demands [286, 242]. On the other had, as only a handful of standardized breeds started to be intensely selected for either growth (e.g. Cornish, Plymouth Rock) or reproductive traits (e.g. White Leghorn, Rhode Island Red) [216], numerous local chicken breeds around the world have gone extinct or are at critical population size. The decline in local livestock diversity was also enhanced by the consolidation of the breeding industry into a limited number of companies. According to The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture of the Food and Agriculture Organization (FAO), among avian species, chickens have by far the highest number of breeds at risk (particularly in Europe and North America), although still a large number of breeds lack population figures (61.7%) and up to 3.4% of the local breeds have gone extinct [262].

Conservation of the genetic diversity is an important issue for local chicken breeds, because their small population size poses them at higher and immediate risks of demographic and environmental stochasticity, with severe implications for their long-term survival, as extensively described in the previous sections. The small population size and often lack of conservation programmes call for immediate and effective management measures to prevent more local chicken breeds to go extinct. However, the demographic history and genetic forces shaping an individual's genomic variation should be identified in order to successfully intervene. Moreover, once characterized, the functional components of such variation should also be investigated, including the prediction and monitoring of major genes and genetic defects.

1.5 Thesis outline

The main goal of the research presented in this thesis is to investigate which genetic factors influence genomic variation and how these factors are shaped by an individual demography and selection history. By means of whole-genome sequencing data from multiple breeds coming from several European countries (i.e. the Netherlands, Germany, France) I relate patterns of breed-specific variation to their demographic and selection history. Moreover, by an interdisciplinary approach at the interface of comparative genomics and transcriptomic I elucidate the importance of such variation in terms of functionality. In this thesis I also provide a set of approaches and methods that are novel for local breeds and that I expect to benefit future demographic and functional studies in local livestock breeds and wild species.

In **Chapter 2** I provide a comprehensive characterization of the genetic diversity, demographic history, and level of inbreeding of local Dutch chicken breeds using the medium density 60K SNP array. In particular, I show that the bantamisation trend started a few decades ago has generated unique genetic diversity today observable in the many bantamized large fowl breeds (i.e. neo-bantams). By comparing local Dutch chicken breeds with commercial white egg-layers, I also illustrate the positive role that management and breeding programmes can have on the genetic diversity of local breeds suffering from high levels of inbreeding. In **Chapter 3** I use whole-genome sequencing data to investigate the impact that varying population bottlenecks occurring at different time points may have on the genetic and deleterious variation of populations of small size, using local chicken breeds as a case study. In this chapter, I show that breeds that since the ancient domestication bottleneck have been kept small in size seem to have eliminated harmful mutations of very strong effect, contrary to recently bottlenecked populations in which the accumulation of weakly harmful mutations is mainly driven by genetic drift. In **Chapter 4** I analyse temporal sequencing data to quantify genomic erosion in small populations under a recently established conservation programme. I show that management can control genetic drift, allowing purging of deleterious alleles. However, populations can benefit from a conservation programme in the long-term only if management practices become genomics-informed. **Chapter 5** illustrates the genetic basis of ptilopody, a trait observed in domesticated and wild bird species. As I show, ptilopody is a polygenetic trait caused by a deletion and a unique haplotype in foot feathered birds affecting genes involved in forelimb and hindlimb development. Interestingly, both causal mutations and genes have evolved by parallel evolution in chicken and pigeon despite their divergence million of years ago. In **Chapter 6** I analyse the non-protein-coding component of the chicken genome in the form of conserved non-coding elements. In this chapter I also present the ch(icken)CADD model, a promising approach that is capable of identifying and prioritizing protein-coding and non-protein-coding variants along the chicken genome. Finally, in **Chapter 7** I discuss the relevance of my findings and place them in a broader context.

BOX 3

Adaptive genetic variation: variation in a DNA sequence (i.e. alleles, genotypes) between distinct individuals of a given population (or species) that affects fitness.

Bottleneck: a severe decline in population size over a short or long time.

Deleterious genetic variation: variation that has a negative effect on fitness, often brought into the population by mutation, gene flow, or genetic drift.

Effective population size (N_e): the size of an idealized population that would show the same amount of, for example, genetic drift as the population under consideration.

Genetic drift: random fluctuation in the number and frequency of alleles among generations in a population owing to stochastic events (e.g. random survival and reproduction).

Identity-by-descent (IBD) segments: chromosomal segments that are identical because they have been inherited from the same common ancestor without recombination.

Inbreeding (or consanguinity): mating among relatives that favours the inheritance of segments IBD.

Inbreeding coefficient (*F*): the probability that two alleles in an individual are identical-by-descent at a given locus.

Inbreeding depression: the reduction in evolutionary fitness of a population or individual due to an increased homozygosity arising from inbreeding.

Linkage disequilibrium: the non-random association of alleles at two loci.

Runs of homozygosity (ROHs): uninterrupted regions of the genome where an individual is homozygous across all sites.

2.
The effects of recent changes in breeding preferences on maintaining traditional Dutch chicken genomic diversity

The chapter is based on:

Bortoluzzi, C., Crooijmans, R. P., Bosse, M., Hiemstra, S. J., Groenen, M. A., Megens, H. J. (2018). *Heredity*, 121(6), 564-578.

Abstract

Traditional Dutch chicken breeds are marginalized breeds of ornamental and cultural-historical importance. In the last decades, miniaturizing of existing breeds (so called neo-bantam) has become popular and resulted in alternatives to original large breeds. However, while backcrossing is increasing the neo-bantams homozygosity, genetic exchange between breeders may increase their genetic diversity. We use the 60K SNP array to characterize the genetic diversity, demographic history, and level of inbreeding of Dutch heritage breeds, and particularly of neo-bantams. Commercial white layers are used to contrast the impact of management strategy on genetic diversity and demography. A high proportion of alleles was found to be shared between large fowls and neo-bantams, suggesting gene flow during neo-bantams development. Population admixture analysis supports these findings, in addition to revealing introgression from neo-bantams of the same breed and of phenotypically similar breeds. The prevalence of long runs of homozygosity (ROH) confirms the importance of recent inbreeding. A high diversity in management, carried out in small breeding units explains the high heterogeneity in diversity and ROH profile displayed by traditional breeds compared to commercial lines. Population bottlenecks may explain the long ROHs in large fowls, while repetitive backcrossing for phenotype selection may account for them in neo-bantams. Our results highlight the importance of using markers to inform breeding programmes on potentially harmful homozygosity to prevent loss of genetic diversity. We conclude that bantamisation has generated unique and identifiable genetic diversity. However, this diversity can only be preserved in the near future through structured breeding programmes.

2.1 Introduction

Since the time of multiple, independent domestication events in South and Southeast Asia [189, 158, 209], domestic chicken (*Gallus gallus domesticus*) populations have experienced intensive human-induced evolution. As a result of domestication and selection for a variety of purposes [189], domesticated chicken breeds have developed an exceptional diversity in morphology, physiology, and behavior [260]. However, demographic events, such as population bottlenecks, admixture of populations, founder effects, genetic drift, and inbreeding, have also contributed to shaping most of the novel genetic variation within the domesticated chicken genome [125, 64, 90].

Chicken populations began to differentiate into breeds after domestication. Preferential breeding of traditional populations exhibiting subsets of specific morphological variants gave rise to a wide range of standardized fancy breeds fixed for a few morphological traits and subjected to low-selection intensity for diverse purposes [125, 286]. However, it was with the increased interest in more efficient selection programmes especially since the second half of the 20th century [125, 286], that a handful of standardized breeds started to be intensively selected for either growth (meat production) or reproductive (egg-laying) traits [40, 216, 125]. The development of experimental and commercial lines led to the replacement of local breeds across the world. As a result, numerous traditional breeds have gone extinct [120], while those that survived are nowadays used for either backyard hobby farming, ornamental and (competitive) fancy breeding [314, 280, 312], or cultural-historical heritage conservation [314, 320, 238]. In rare cases, traditional chicken breeds are used for high-value niche market products [295, 296, 238].

The recent history of traditional breeds in The Netherlands provides numerous genotypes to characterize. Furthermore, the results of such genetic characterization studies can be used as a tool in on-going efforts to preserve chicken diversity nationally. Although the majority of the Dutch breeds have originally been bred for production traits, breeding for egg or meat production has usually ceased. As a result, traditional breeds have become marginalized and have now an almost exclusively ornamental or cultural historical significance. Hobby breeders are the most important stakeholders involved in the conservation of specific varieties or breeds [314], but their number is limited and often getting smaller over the years.

Until recently, Dutch chicken genetic resources comprised mostly large fowls and bantam breeds, whose origin can be dated back to the 16th and 18th century [64]. Traditional Dutch breeds have their origin in Europe, although some East- and Southeast Asian influences have been found in a few breeds, as a result of occasional or repetitive introgression. Introgression from European and Mediterranean breeds has also been observed to a lesser extent [64]. Of the traditional breeds with past productive significance, the North Holland blue derived from the Belgian Malines, whereas Asian chickens, such as the Cochin, Brahma, and Langshan, were involved in the formation of Barnevelders and Welsummers. Malays, Japanese bantams, and Sumatras were involved in the formation of several old traditional breeds (the so-called country fowls) including Frisian fowls, Dutch bantams, Breda fowls, Dutch booted bantams, Dutch fowls, and, Schijndelaars whereas no recorded history of genetic influence from Asiatic chickens was recorded for other country fowls, including the Assendelft fowls, Drenthe fowls, and Groninger Mews. Ornamental breeds, including the Dutch Owl bearded, Dutch Polish bearded and Dutch Polish non bearded, are thought to derive from Polish bearded chickens firstly introduced in the Italian peninsula from Asia through Greece and, despite their ancient origin, are still kept by hobby breeders for ornamental and (competitive) fancy showing. As for some of the country fowls, also the Lakenvelder does not have a recorded history of genetic influence from Asiatic chickens [64].

In the last decades, fancy breeders have become interested in the development of bantam forms of large breeds. Neo-bantams have become popular among hobby breeders for their captivating and petite appearance and because they are more easily housed in a hobby setting. For these reasons, it is likely that neo-bantams will soon replace the large fowl counterparts. The aim of the bantamising trend is to obtain a small-sized individual exhibiting all of the standard (large) breed characteristics, but in a smaller size. Invariably, the bantam forms of large breeds are made by crossing the original large breed with a small breed, such as the Dutch bantam (a 'true' bantam), or more recently with other more recently created bantam breeds. Breeders have repetitively crossed the first generation of neo-bantams to the parental generation of large fowls. Although backcrossing has contributed to the creation of bantam forms of almost all of the standard large breeds, crossing of related animals may pose a threat to the long-term existence of neo-bantams due to the accumulation of harmful and deleterious variants and inbreeding depression. However, introgression from local and imported stocks from Asia and neighboring European countries, along with the occasional genetic exchange between farmers, are important sources of increased genetic diversity. Therefore, informing management and conservation programmes based on genetic data may prevent future loss of genetic diversity of traditional breeds.

The study of the genetic diversity harbored by commercial chicken lines may provide insights into and new perspectives on the genetic management and conservation priority of traditional populations. Such insights are possible since the genetic management of populations, along with demographic and selection history, influence the extent of genetic diversity and breed identity [120, 216, 215, 90]. In particular, effective management has shown to be critically important when the target population shows reduced genetic diversity and high level of inbreeding. In commercial lines, this is due to the decreased number of active breeding stocks, restricted within-line selection, and absence of genetic introgression from non-commercial populations [216].

Informing management of heritage breeds with genetic data has become feasible due to the development of cheap, chicken SNP panels [123]. Moreover, by applying SNP genotype data, major questions in conservation genetics can finally be addressed [31]. In the absence of pedigree data, which is the norm for non-commercial chicken populations, of immediate importance is the assessment of the degree of relatedness between populations, their genetic uniqueness, and degree of inbreeding. SNP arrays provide an alternative approach to estimate the traditional inbreeding coefficient, F_{ped}, by detecting continuous segments of homozygous SNPs (runs of homozygosity - ROH) [163, 279]. Studying ROHs provides insights into past and present population history, selection pressure, and management [31, 245, 142, 163].

Numerous diversity studies of traditional chicken populations from different countries and continents are reported in the literature. However, these studies have been based on a limited number of genetic markers of lower resolution and genome coverage than SNP arrays. Regarding the traditional Dutch chicken breeds, only few studies have focused on the assessment of the breed genetic diversity and contribution to conservation [88, 144]. Invariably, such characterization studies were incomplete, since neo-bantam breeds were not considered.

Here, we use the 60K SNP array to characterize the genetic diversity and inbreeding of all recognized Dutch heritage breeds and most of the bantam forms. In particular, we investigate the process by which the neo-bantams are formed, their degree of inbreeding due to presumed small founder size, and their potential contribution to the total Dutch chicken genetic diversity. Finally, we study the effects of the structured management experienced by commercial lines on their genetic diversity and demographic history to better inform genetic management and conservation of Dutch heritage breeds.

2.2 Materials and methods

2.2.1 Chicken populations

A total of 674 individuals from forty-one chicken populations originating from the Netherlands and resulting from different demographies and management strategies were included in the study (Table 1). The complete set of chicken populations included 37 traditional fancy breeds (480 individuals), comprising true bantams, large fowls, and bantam counterparts (neo-bantam), two commercial white egg layer sire lines (R01 and W1) (80 individuals), and two commercial white egg layer dam lines (WA and WD) (114 individuals). Among the traditional breeds, 476 individuals were sampled from part-time, hobby, and fancy breeders of known provenance, while sperm of four individuals of the breed North Holland Blue was provided by the Center for Genetic Resources (CGN), The Netherlands. The total number of individuals per breed varied from 1 to 66, with a maximum of 24 individuals for the fancy breeds, whereas the total number of fancy breeders contributing to the total sample size ranged from 1 to 10. Sample collection of fancy breeds took place over 13 years (1998-2011) (Table 1). Due to the important variation in sample size over the time frame considered, changes in genetic diversity over time were not analyzed. Since pedigrees are generally not recorded by fancy breeders or breeding organizations, this information was not available to this study. In fact, the absence of such information that is vital for effective management was a major incentive for the comprehensive genotyping effort detailed in this study. Phenotypic information was collected in the form of feather color only for those breeds sampled between 2009 and 2011.

2.2.2 Sampling and genotyping

DNA was extracted from blood of 191 samples (1998), sperm of 4 individuals (2007), and from fertile hatching eggs of 287 samples (2007-2011). Genotyping and quality control (QC) were performed separately using the standard protocols for the Illumina Infinium iSelect 60K BeadChip. Raw data were analysed using the Genome Studio software package (Illumina Inc.) [123]. The 60K SNP chip contained 52 232 SNPs uniformly distributed across the Gallus_gallus5.0 chicken genome, comprising 29 autosomes (Gga 1-28 and Gga 33), two sex chromosomes (that is, 2 577 SNPs on the Z chromosome and 7 on the W chromosome), and one linkage group (that is, 49 SNPs on LGE64). The array also included several variants of unknown mapping position (n = 507), whereas no variants were mapped to the mitochondrial genome. Variants on the two sex chromosomes, linkage group, and variants of unknown physical position were all removed from both traditional and commercial breeds, separately. A total of 49 092 variants were retained in both datasets. Genotype filtering was applied to the merged dataset (traditional breeds and commercial lines) after removing individuals mislabeled.

2.2.3 SNP quality control and marker selection

We used PLINK v1.9 [45] for genotyping data quality control. Samples genotyped for less than 90% of markers were excluded, along with SNPs genotyped for less than 90% of the animals. Monomorphic variants were also discarded. Using these criteria, 2 121 SNPs were excluded because of the low genotyping rates (Table S1), while 38 samples were removed due to a low genotyping rate (Table S2). Although only one individual of the Assendelft fowl bantam, Drenthe fowl bantam, North Holland Blue bantam, and Schijndelaar bantam passed the quality

Table 1: Summary details of the 37 traditional Dutch chicken breeds and 4 commercial white egg layer lines. *N* represents the sample size, *Flocks* the number of fancy breeders that contributed to the total sample size of a breed, and *Types* the number of morphological varieties (feather color) present in the total sample size of a breed. There is no correspondence between *Flocks* and *Types*, as a single breeder can have contributed to the total sample size with different morphological varieties. The number in parenthesis reported after the year in the column *Sampling year* identifies the number of individuals within a breed sampled in that specific year. Abbreviations under the column *Management* represent the subdivision of chicken populations into clusters based on their genetic management (LF, large fowl; B, bantam; NB, neo-bantam; C, commercial). Cluster identifies the group the breed belongs to according to the principal component analysis (CL1: past-productive; CL2: ornamental; CL3: country fowls; CL4: Lakenvelder)

Population	Abbreviated name	Management	Cluster	N	Flocks	Types	Sampling country	Sampling year
Assendelft fowl	AssFw	LF	CL3	15	4	4	Netherlands	2011 (10); 1998(5)
Assendelft fowl bantam	AssFwB	NB	CL3	2	2	1	Netherlands	2011 (2)
Barnevelder	Barnev	LF	CL1	24	10	2	Netherlands	2009 (10); 1998 (14)
Barnevelder bantam	BarnevB	NB	CL1	7	6	3	Netherlands	2011 (7)
Brabanter	Brab	LF	CL2	20	4	6	Netherlands	2011 (5); 2009 (5); 1998 (10)
Brabanter bantam	BrabB	NB	CL2	10	3	6	Netherlands	2011 (10)
Breda fowl	BreFw	LF	CL3	20	5	7	Netherlands	2011 (10); 1998 (10)
Breda fowl bantam	BreFwB	NB	CL3	10	3	7	Netherlands	2011 (10)
Chaam fowl	ChaFw	LF	CL3	10	2	3	Netherlands	2011 (29; 2009 (8)
Dutch bantam	DB	B	CL3	20	9	7	Netherlands	2011 (1); 2009 (9); 1998 (10)
Dutch booted bantam	DBdB	B	CL3	19	3	8	Netherlands	2011 (2); 2009 (7); 1998 (10)
Dutch fowl	DFw	LF	CL3	20	3	6	Netherlands	2011 (10); 1998 (10)
Dutch fowl bantam	DFwB	NB	CL3	4	3	4	Netherlands	2011 (4)
Dutch owl bearded	DOwBd	LF	CL2	24	5	8	Netherlands	2011 (5); 2009 (5); 1998 (14)
Dutch owl bearded bantam	DOwBdB	NB	CL2	11	4	6	Netherlands	2011 (11)
Dutch Polish bearded	DPBd	LF	CL2	13	2	2	Netherlands	2011 (1); 2009 (2); 1998 (10)
Dutch Polish bearded bantam	DPBdB	NB	CL2	10	5	7	Netherlands	2011 (9); 2009 (1)
Dutch Polish non bearded	DPnBd	LF	CL2	20	3	8	Netherlands	2011 (4); 2009 (6); 1998 (10)
Dutch Polish non bearded bantam	DPnBdB	NB	CL2	10	3	9	Netherlands	2011 (5); 2009 (5)
Drenthe fowl	DrFw	LF	CL3	20	2	7	Netherlands	2009 (10); 1998 (10)
Drenthe fowl bantam	DrFwB	NB	CL3	2	2	1	Netherlands	2011 (2)
Eikenburger bantam	EikenbB	B	CL3	4	1	1	Netherlands	2011 (4)
Frisian fowl	FriFw	LF	CL3	24	4	6	Netherlands	2011 (4); 2009 (6); 1998 (14)
Frisian fowl bantam	FriFwB	NB	CL3	7	5	6	Netherlands	2011 (7)
Groninger Mew	GrMw	LF	CL3	19	7	3	Netherlands	2009 (9); 1998 (10)
Groninger Mew bantam	GrMwB	NB	CL3	10	7	3	Netherlands	2009 (10)
Kraienkoppe	KraiK	LF	CL2	20	3	7	Netherlands	2011 (2); 2009 (8); 1998 (10)
Kraienkoppe fowl bantam	KraiKFwB	NB	CL2	5	2	4	Netherlands	2011 (5)
Lakenvelder	LakVe	LF	CL4	20	8	-	Netherlands	2011 (10); 1998 (10)
Lakenvelder bantam	LakVeB	NB	CL4	6	4	1	Netherlands	2011 (6)
North Holland Blue	NHBl	LF	CL1	20	4	1	Netherlands	2016 (1); 2011 (5); 2009 (3); 2007 (1); 1998 (10)
North Holland Blue bantam	NHBlB	NB	CL1	1	1	-	Netherlands	2011 (1)
Schijndelaar	Schijd	LF	CL3	10	1	4	Netherlands	2009 (10)
Schijndelaar bantam	SchijdB	NB	CL3	1	1	1	Netherlands	2009 (1)
Sumatra	Sumt	B	CL3	10	-	-	Netherlands	1998 (10)
Welsummer	Welsum	LF	CL1	24	6	1	Netherlands	2011 (10); 1998 (14)
Welsummer bantam	WelsumB	NB	CL1	8	6	1	Netherlands	2011 (8)
White egg layers – line R01	White_R01	C	-	29	1	-	Netherlands	-
White egg layers – line W1	White_W1	C	-	51	1	-	Netherlands	-
White egg layers – line WA	White_WA	C	-	66	1	-	Netherlands	-

control, we excluded these breeds only from the calculation of the population genetic diversity estimates, because the extremely small sample size precluded such calculations. No additional filtering for minor allele frequencies was carried out, because removal of rare alleles could lead to overestimated results in under sampled populations [287]. Similarly, no filtering for linkage disequilibrium, Mendelian error, and deviation from Hardy-Weinberg equilibrium (HWE) were performed, owing to the lack of pedigree information for the traditional populations and the interest in investigating deviations from HWE. The final data set consisted of 632 individuals from 33 traditional Dutch breeds and 4 commercial lines genotyped for 46 971 SNPs.

2.2.4 Population genetic diversity

Mean expected (H_E) and observed (H_O) heterozygosity, mean minor allele frequency (MAF), and mean inbreeding coefficient (F_{IS}) were averaged across all loci and individuals within a population, respectively. Measures of molecular diversity were estimated using PLINK v1.9. As a result of the wide range of sample size across traditional breeds (from 4 to 23 after QC), we decided to test the influence of a variable sample size on the population genetic diversity estimates by randomly sample a different number of individuals within each breed. Individuals were randomly selected from those that passed the quality control. Mean expected (H_E) and observed (H_O) heterozygosity, mean minor allele frequency (MAF), and mean inbreeding coefficient (F_{IS}) were then estimated for each newly sampled breed. A t-test was used to statistically test whether differences in population statistics were attributed to the sample size. To further investigate the consequences of divergent management practices to the genetic diversity, we followed the genetic cluster analysis carried out by Hillel *et al.* (2003) [144] dividing our populations into four groups, as follow: (1) LARGE FOWL, including large fowls of fancy breeds selected for specific morphological traits; (2) BANTAM, which included all the true bantams for which no large fowl counterparts exist; (3) NEO BANTAM, consisting of recently established bantams of large fowls; and (4) COMMERCIAL, which included the commercial lines intensively selected for quantitative traits related to egg production. Genetic relationships between traditional and commercial populations and among fancy breeds were investigated through the Principal Component Analysis (PCA), which was performed on the genotype data using the R package SNPRelate [327] for R v3.2.0 [283]. Pairwise genetic distance (D) for all pairwise combinations of individuals of traditional breeds was calculated on unpruned data as , where is the average proportion of alleles shared among individuals. The 1-IBS matrix was afterward used for phylogenetic reconstruction producing a Neighbor-Joining (NJ) tree in PHYLIP v3.696 (Felsenstein, 2004) with random input order. The unrooted tree was then visualized with Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

2.2.5 Population genetic admixture

Population genetic admixture was analyzed using the model based clustering method ADMIXTURE v1.3.0 [4]. Although the software assumes that all populations share the same ancestral group, a K value identifying the number of ancestral components needs to be provided to perform the analysis. Due to the unknown genetic structure of our chicken populations, we decided to perform an unsupervised admixture analysis by carrying out a cross-validation (CV) procedure for K values ranging from 1 to 40. The CV procedure aimed to select the K value exhibiting the lowest cross-validation error estimate, which represents the most parsimonious number of clusters. Following the genetic diversity analysis, we resolved to restrict the admixture analysis to the clusters identified by the PCA of traditional breeds, thus reducing the likelihood of biased results owing to the considerable different sample size. Results were visualized with Pong [16]. To formally test whether admixture occurred across our traditional chicken populations, and to additionally measure its extent, we calculated three-population test estimates (*f3* statistics) [249] and their corresponding normalized value (z-score), calling the *threepop* module implemented in the TREEMIX software package v1.13 [239]. We decided to restrict the three-population test to the traditional populations, because we did not expect genetic admixture between commercial and non-commercial breeds. In the *f3* statistics, we considered the triplet of the populations (C; A, B), where C is the target, or test, population, and A and B are the source, or reference, populations. The normalized z-scores were calculated by jack-knifing in blocks of 500 SNPs. A significant negative value of the f3 statistic (z $\leq$ -3.80) indicated an admixture event between the test and the two ancestral populations. We performed all possible triplet combinations, considering only breeds with more than one sample as test population.

2.2.6 Runs of homozygosity

Population demographic history was investigated through the detection of homozygous stretches along an individual's SNP data, as implemented in the *-homozyg* option in PLINK [148]. We defined a run of homozygosity (ROH) as a tract of homozygous genotypes that was greater than 10 Kb in length, and identified in a genome-sliding window of 30 SNPs. To ensure that the entire observed stretch from the first SNP to the last SNP was homozygous (true ROH), we excluded stretches with a mean tract density $>$1Mb/SNP, and with a maximum gap between two consecutive homozygous SNPs of 1 000 Kb. To lower the underestimation of ROHs due to genotyping errors and/or missing genotypes, we allowed only one heterozygous SNP and one missing call per window. The detected ROHs were then classified into three categories intended to correspond to different demographic processes: short ROH ($<$ 1 000 Kb) reflected homozygosity of ancient haplotypes if not founder effects; medium (1-3 Mb) background relatedness within populations; and long (> 3 Mb) recent parental relatedness [279] A genomic measure of individual autozygosity, F_{ROH}, was calculated as the proportion (0-1) of the autosomal genome covered by stretches of consecutive homozygous SNPs following McQuillan et al. [204],

$$F_{ROH} = \frac{\sum L_{ROH}}{L_{auto}}, \tag{2.1}$$

where $\sum L_{ROH}$ is the total length of all of an individual's runs, and is the total genome length across the autosomes covered by SNPs [204]. In calculating the genetic map containing markers not filtered for low genotype calls was used to reduce the likelihood of underestimating the total autosomal genome length. According to our SNP panel, was approximately 906 Mb. Individual and population mean values of F_{ROH} were estimated for all ROHs and for the three ROH-length threshold classes. The correlation between the genomic measure of autozygosity (F_{ROH}) and the inbreeding coefficient estimated from genotype frequency (F_{IS}) was calculated for all homozygous stretches and for the three ROH classes, respectively. All plots were generated with the R package ggplot2 for R v3.2.0.

2.3 Results

2.3.1 Population genetic diversity

Table 2 shows the results of the population and management-based genetic diversity analysis of the traditional breeds and commercial lines that passed genotyping data quality control. The management-based analysis also included the Assendelft fowl bantam, Drenthe fowl bantam, North Holland blue bantam, and Schijndelaar bantam breed, from which the total number of 37 populations. Average minor allele frequency across traditional chicken populations ranged from 0.165±0.16 (Dutch Polish bearded) to 0.256±0.15 (Barnevelder). Average observed (H_O) and expected heterozygosity (H_E) varied between 0.116±0.22 (Eikenburger bantam) and 0.327±0.33 (Sumatra) and between 0.108±0.18 (Eikenburger bantam) and 0.335±0.16 (Barnevelder), respectively. Average inbreeding coefficient (F_{IS}) ranged from -0.311±0.01 (Sumatra) to 0.61±0.04 (Eikenburger bantam) (Table 2). Traditional breeds that showed signatures of outbreeding were also the Dutch Polish bearded (-0.067±0.24), Frisian fowl (-0.063±0.17), and Groninger Mew (-0.011±0.15), whereas high inbreeding coefficient estimates were also reported for the neo-bantams of Brabanter (0.537±012) and Kraienkoppe fowl bantam (0.520±0.03) (Table 2). Overall, higher within-breed inbreeding coefficient estimates were displayed by neo-bantams than large fowl counterparts. Moreover, compared to commercial lines, more heterogeneous molecular diversity estimates were observed for traditional breeds, supporting differences in selective breeding and demographic history. As expected, level of excess homozygosity (F_{IS}) negatively correlated with the observed heterozygote frequencies (r = -0.54, p-value = 0.001).

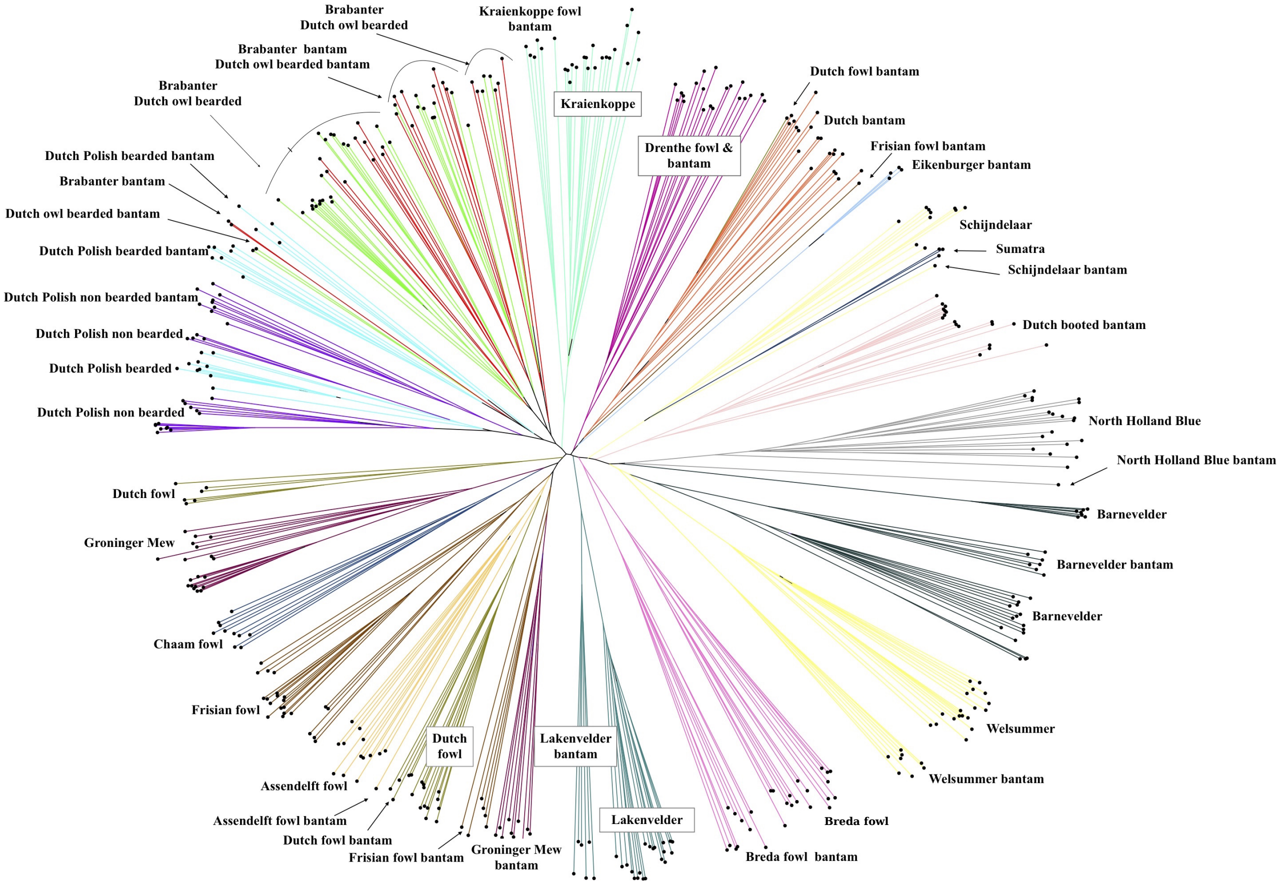
At the management level, breeds selected for a specific morphological standard (LARGE FOWL) were the most polymorphic, followed by neo-bantams, which showed slightly similar genetic diversity estimates, and true bantams (that are, Dutch bantam, Dutch booted bantam, Eikenburger bantam, and Sumatra). Polymorphism measures of the COMMERCIAL cluster displayed intermediate values, supporting the reduced genetic diversity reported in previous studies [120, 216]. Average inbreeding coefficient showed an opposite pattern, with the NEO-BANTAM cluster being the most inbred of the traditional breeds cluster, followed by LARGE FOWL and BANTAM (Table 2).

Population genetic estimates calculated for each population after randomly select a different number of individuals that passed the quality control are reported in Table S3 of the Supplementary Material. Population genetic diversity estimates of the same population calculated on a different sample size did not significantly differ from those reported in Table 2, except for the expected heterozygosity, whose estimates were significantly different in both random sampling scenarios (Table S4). The standard deviation of all genetic estimates were considerably high, especially that of the inbreeding coefficient, which decreased when increasing the sample size, except for the Frisian fowl bantam and Dutch Polish bearded bantam (Table S3).

Table 2: Molecular diversity statistics of traditional chicken breeds and commercial lines and of populations within the four management-based clusters. Population genetic diversity statistics are averaged across loci and individuals within each population. Abbreviations: MAF, minor allele frequency; H_O, observed heterozygosity; H_E, expected heterozygosity; F_{IS}, inbreeding coefficient; SD, standard deviation. Abbreviations under the column *Management* represent the subdivision of the chicken populations into clusters based on their genetic management (LF, large fowl; B, bantam; NB, neo-bantam; C, commercial)

Population	**Management**	**Average MAF ± SD**	**Average H_O ± SD**	**Average H_E ± SD**	F_{IS} **± SD**
Assendelft fowl	LF	0.221 ± 0.15	0.272 ± 0.18	0.295 ± 0.17	0.076 ± 0.21
Barnevelder	LF	0.256 ± 0.15	0.204 ± 0.14	0.335 ± 0.16	0.369 ± 0.16
Barnevelder bantam	NB	0.184 ± 0.16	0.222 ± 0.21	0.246 ± 0.19	0.113 ± 0.12
Brabanter	LF	0.252 ± 0.15	0.283 ± 0.17	0.331 ± 0.16	0.083 ± 0.15
Brabanter bantam	NB	0.223 ± 0.16	0.204 ± 0.17	0.294 ± 0.18	0.537 ± 0.12
Breda fowl	LF	0.235 ± 0.15	0.258 ± 0.17	0.309 ± 0.17	0.154 ± 0.15
Breda fowl bantam	NB	0.217 ± 0.15	0.234 ± 0.18	0.289 ± 0.17	0.482 ± 0.12
Chaam fowl	LF	0.240 ± 0.15	0.316 ± 0.20	0.316 ± 0.16	0.332 ± 0.06
Dutch bantam	B	0.222 ± 0.15	0.210 ± 0.14	0.298 ± 0.16	0.263 ± 0.18
Dutch booted bantam	B	0.220 ± 0.15	0.247 ± 0.18	0.296 ± 0.16	0.091 ± 0.21
Dutch fowl	LF	0.213 ± 0.15	0.217 ± 0.16	0.287 ± 0.17	0.104 ± 0.24
Dutch fowl bantam	NB	0.208 ± 0.16	0.216 ± 0.22	0.275 ± 0.19	0.453 ± 0.06
Dutch owl bearded	LF	0.249 ± 0.15	0.293 ± 0.17	0.328 ± 0.16	0.033 ± 0.25
Dutch owl bearded bantam	NB	0.225 ± 0.15	0.280 ± 0.19	0.297 ± 0.17	0.162 ± 0.41
Dutch Polish bearded	LF	0.165 ± 0.16	0.219 ± 0.22	0.222 ± 0.19	-0.067 ± 0.24
Dutch Polish bearded bantam	NB	0.205 ± 0.16	0.181 ± 0.15	0.274 ± 0.18	0.329 ± 0.49
Dutch Polish non bearded	LF	0.166 ± 0.15	0.162 ± 0.15	0.232 ± 0.17	0.237 ± 0.25
Dutch Polish non bearded bantam	NB	0.215 ± 0.15	0.212 ± 0.17	0.288 ± 0.17	0.280 ± 0.25
Drenthe fowl	LF	0.236 ± 0.15	0.250 ± 0.15	0.314 ± 0.16	0.156 ± 0.15
Eikenburger bantam	B	0.083 ± 0.14	0.116 ± 0.22	0.108 ± 0.18	0.613 ± 0.04
Frisian fowl	LF	0.211 ± 0.15	0.245 ± 0.18	0.284 ± 0.17	-0.063 ± 0.17
Frisian fowl bantam	NB	0.229 ± 0.15	0.238 ± 0.19	0.302 ± 0.17	0.208 ± 0.14
Groninger Mew	LF	0.177 ± 0.15	0.197 ± 0.18	0.242 ± 0.18	-0.011 ± 0.15
Groninger Mew bantam	NB	0.203 ± 0.16	0.225 ± 0.19	0.268 ± 0.19	0.498 ± 0.05
Kraienkoppe	LF	0.232 ± 0.15	0.288 ± 0.17	0.311 ± 0.16	0.010 ± 0.22
Kraienkoppe fowl bantam	NB	0.183 ± 0.16	0.188 ± 0.20	0.242 ± 0.19	0.520 ± 0.03
Lakenvelder	LF	0.171 ± 0.16	0.214 ± 0.19	0.230 ± 0.19	0.054 ± 0.17
Lakenvelder bantam	NB	0.207 ± 0.16	0.266 ± 0.22	0.272 ± 0.19	0.008 ± 0.26
North Holland Blue	LF	0.245 ± 0.14	0.317 ± 0.17	0.326 ± 0.15	0.001 ± 0.20
Schijndelaar	LF	0.213 ± 0.15	0.266 ± 0.19	0.287 ± 0.17	0.425 ± 0.12
Sumatra	B	0.179 ± 0.17	0.327 ± 0.33	0.230 ± 0.20	-0.311 ± 0.01
Welsummer	LF	0.198 ± 0.16	0.237 ± 0.18	0.266 ± 0.18	0.036 ± 0.20
Welsummer bantam	NB	0.171 ± 0.16	0.210 ± 0.21	0.228 ± 0.19	0.137 ± 0.48
White egg layers – line R01	C	0.181 ± 0.16	0.250 ± 0.21	0.241 ± 0.19	0.482 ± 0.02
White egg layers – line W1	C	0.183 ± 0.16	0.247 ± 0.20	0.243 ± 0.19	0.495 ± 0.02
White egg layers – line WA	C	0.153 ± 0.16	0.209 ± 0.21	0.209 ± 0.20	0.572 ± 0.02
White egg layers – line WD	C	0.154 ± 0.16	0.212 ± 0.20	0.207 ± 0.19	0.563 ± 0.02
Management group (N. populations)					
LARGE FOWL (n=17)	LF	0.324 ± 0.11	0.247 ± 0.07	0.410 ± 0.10	0.390 ± 0.16
BANTAM (n=4)	B	0.285 ± 0.13	0.224 ± 0.11	0.368 ± 0.13	0.370 ± 0.18
NEO BANTAM (n=16)	NB	0.311 ± 0.12	0.224 ± 0.08	0.397 ± 0.11	0.427 ± 0.14
COMMERCIAL (n=4)	C	0.260 ± 0.15	0.227 ± 0.13	0.337 ± 0.16	0.542 ± 0.04

Figure 1: Unrooted phylogenetic tree showing genetic relatedness of the 37 traditional chicken breeds. The phylogeny tree was constructed using the Neighbor-Joining method with random input orders and the pairwise 1-IBS-distance matrix. Large fowls and bantam counterparts of each breed are reported with the same color. The name of each breed is indicated in the figure with, in some cases, the use of an arrow. For abbreviations, refer to Table 1

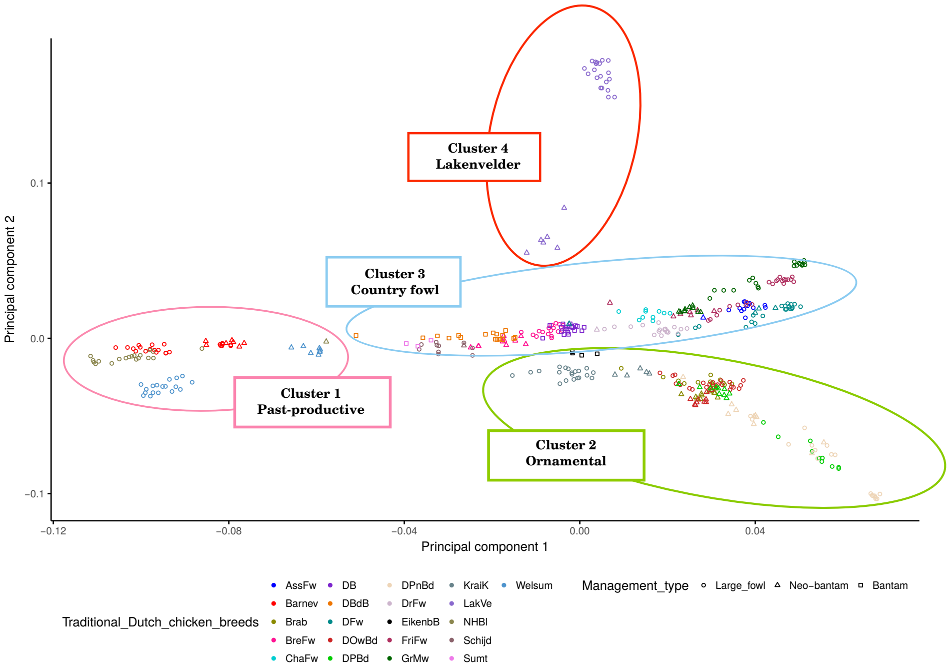


2.3.2 Population genetic structure and admixture

Results on the breed genetic differentiation were consistent in the Neighbor-Joining (NJ) tree and principal component analysis. Moreover, the principal component analysis of traditional breeds revealed a more complex population structure and higher genetic similarities between traditional breeds than with commercial lines (Figure S1). The NJ tree showed an average high proportion of alleles identical-by-state (IBS) shared between the neo-bantams and large fowl counterparts. As a result, large fowls and neo-bantams were separately grouped within the same cluster, as shown, for example, by the Breda fowl and Breda fowl bantam (Figure 1). The NJ tree also identified several subdivided breeds, including the Barnevelder, Frisian fowl, Groninger Mew, and Dutch bantam, for which some individuals were separately grouped within the same cluster, and the Dutch fowl, for which individuals were grouped in two separate clusters, one closer to the Groninger Mew while the other between the Assendelft fowl and Frisian fowl bantam (Figure 1). The NJ tree also captured recent gene flow, as shown by the Dutch fowl bantam and Frisian fowl bantam (indicated with an arrow in Figure 1), which clustered together with the Dutch bantam, a breed that has been used in the bantamisation of the large fowl counterparts. Similar pattern was observed for the Schijndelaar and Schijndelaar bantam, which both clustered together with their source population represented by the Sumatra (indicated with an arrow in Figure 1).

The heterogeneity showed by the subdivided breeds and the recent gene flow reported for some individuals were well captured in the PCA under cluster 1, which identified large fowls and neo-bantams of breeds with past productive significance, and cluster 3, which was defined by the oldest breeds of The Netherlands, the so-called country fowls (Figure 2). In both PCA and NJ tree, the ornamental breeds (cluster 2 in the PCA) showed a distinctive clustering pattern, as displayed by the intermingled breeds, including the Brabanter, Dutch Owl bearded, Dutch Polish bearded, Dutch Polish non-bearded, and bantam counterparts. The NJ tree and PCA identified two main sub-clusters: the first represented by the Dutch Owl bearded and Brabanter, and the second by the Dutch Polish bearded and Dutch Polish non-bearded (Figure 1-2). We did not observe a clear separation between the large fowls and neo-bantams, which may indicate a complex on-going gene flow among the ornamental breeds. The high similarity in phenotypes displayed by the fancy breeds may also support the genetic exchange, as well as question their genetic identity. The results of the ADMIXTURE analysis (Figure S2-S4) performed on the clusters identified by the PCA (except for the Lakenvelder cluster which was combined with cluster 3) were consistent with what reported in the PCA and NJ tree and complied with the breeds' development history. However, we also observed divergent admixture patterns across samples based on their year of sampling, with the recently sampled individuals showing a less unique genetic make-up, as a result of on-going gene flow. The genetic origin of the neo-bantams already captured in the NJ and PCA were better represented in the ADMIXTURE analysis, in which neo-bantams are the result of introgression from original, large-sized fowls and true bantams of morphologically analogous breeds, and, more recently, of neo-bantams of the same or of different breeds. For instance, the Dutch owl bearded bantam showed introgression from the large fowl counterpart, along with the Dutch Polish bearded bantam, Dutch Polish non bearded, and Dutch Polish non bearded bantam (Figure S3). A high number of significant negative *f3* statistics (Table S5) was observed with the Dutch fowl bantam as admixed population and most of the remaining breeds as source populations, confirming introgression from varying populations. Of the 27 282 *f3* statistics, 1 125 were found to be significant (z-score < -3.80)

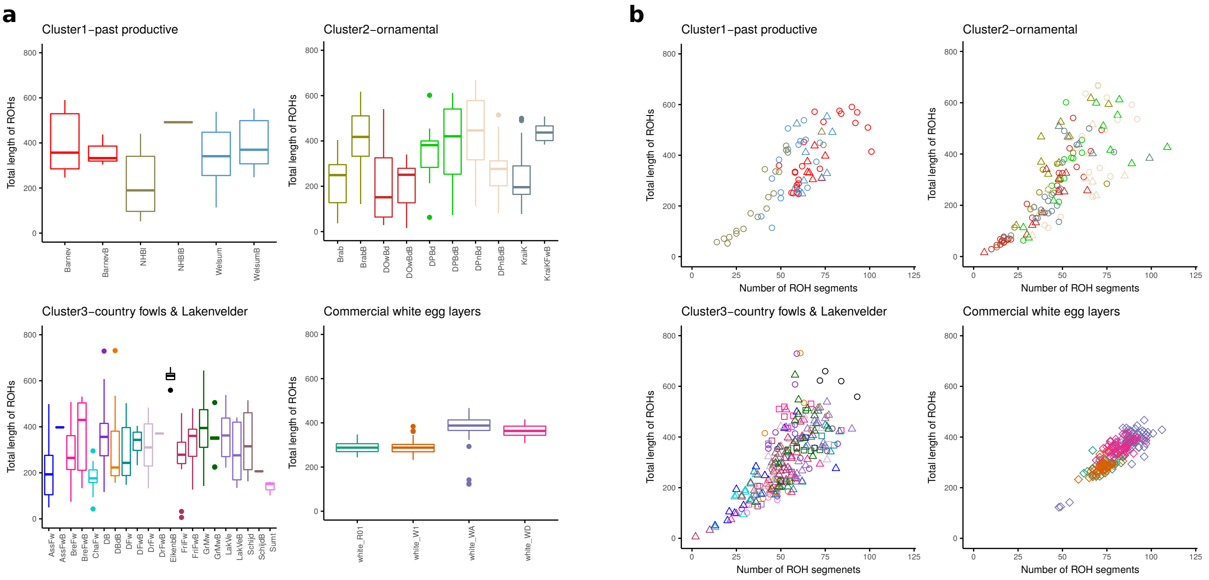
Figure 2: Principal component analysis plot for PC1 and PC2 of traditional Dutch chicken breeds. The principal component analysis was performed on all individuals that passed the genotyping data quality control, for a total of 442 samples. Large fowls and bantam counterparts are represented with the same color but in two different shapes (circle and triangle, respectively, as reported in the legend Management_type). For simplicity, only the abbreviated name of the large fowl is reported in the legend. True bantams, that are the Dutch bantam, Dutch booted bantam, Eikenburger bantam, and Sumatra, are represented with the same square shape (Management_type), since they do not have any large fowl counterpart, but have different colors since they are distinctive breeds. The four colored circles represent the first (purple), second (green), third (light blue), and fourth (red) cluster described in the main text. For abbreviations, refer to Table 1



2.3.3 Runs of homozygosity

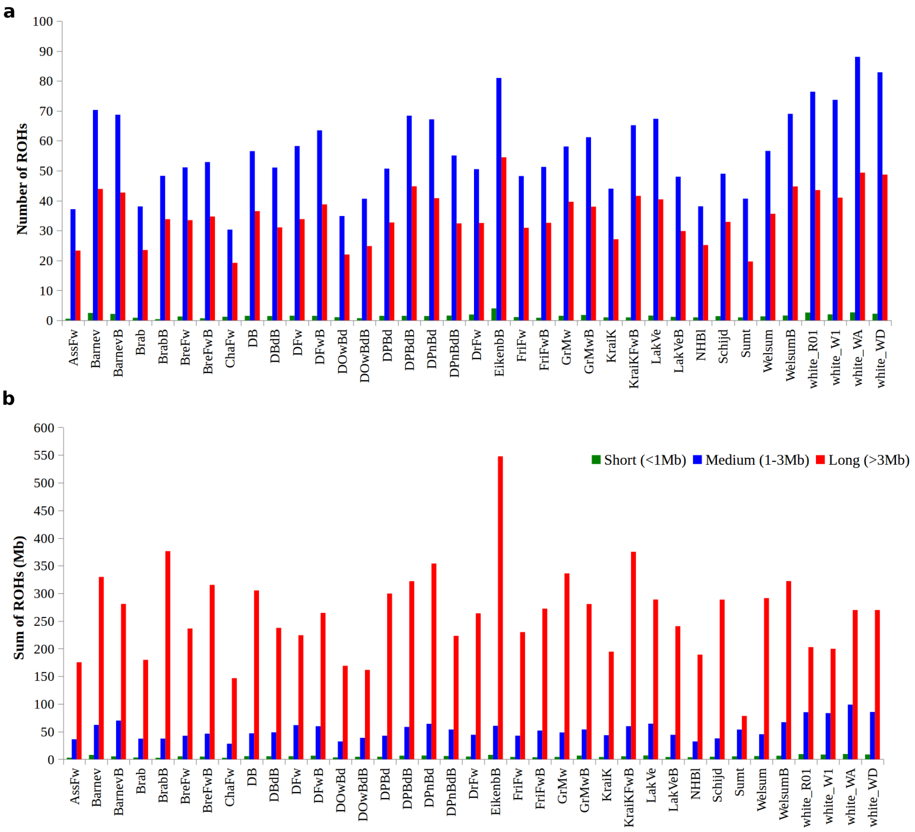
The average proportion of the genome covered by ROHs reflected the genetic diversity and demographic history of traditional breeds, along with the degree of inbreeding. Compared to the homogeneous values reported for the commercial lines (Figure 3a, cluster commercial white egg layers), traditional breeds showed important variation in the total length of ROHs (cluster 1-

Figure 3: Distribution of ROHs across the genome and ROH profile. a. The total length of homozygous segments (ROHs) detected across the genome is reported for each breed and expressed in megabases (Mb). **b.** ROH profile given by the total number of homozygous segments and total segment size in megabases of all 37 traditional chicken breeds. In both figures, breeds were grouped according to the clusters identified in the principal component analysis, except for cluster 3, which also included the Lakenvelder and Lakenvelder bantam. Commercial lines where grouped in another cluster called Commercial white egg layers. For the North Holland blue bantam, Assendelft fowl bantam, Drenthe fowl bantam, and Schijndelaar bantam, the distribution of ROH segments across the genome is represented by a straight line, because only one individual was left after quality control. Colors and shapes were the same of Figure 1 (cycle: large fowl; triangle: bantam counterpart; square: true bantam), except for the commercial white egg layers cluster, which were represented with a different shape



3, Figure 3a). The ROH profile (cluster 1-3, Figure 3b) of the traditional breeds displayed more differences in pattern compared to the commercial lines (Figure 3b), which exhibited similar average cumulative size as well as average ROH number. Although white egg layers displayed a similar ROH profile, individuals from the dam line (WA, WD) had a higher average cumulative size and average ROH number compared to the sire line (R01, W1), confirming the higher homozygosity reported in Table 2. To investigate the effects of specific demographic processes on the distribution of ROHs across the genome, we divided the detected segments into three ROH-length threshold classes: short (< 1Mb), medium (1-3Mb), and long (> 3Mb). Medium and long ROHs were the most abundant classes in both traditional breeds and commercial lines (Figure 4a). Compared to the large fowl counterparts, neo-bantams showed a higher number of medium and long ROHs, with the highest number found in the Eikenburger bantam, followed by Barnevelder bantam and Dutch Polish bearded bantam (Figure 4a). Although traditional breeds had a lower number of ROHs than commercial lines (Figure 4a), homozygous segments covered a significant proportion of their genome, and particularly of neo-bantams, as shown by the Brabanter bantam and Kraienkoppe fowl bantam (Figure 4b). To assess ROH as an indicator of inbreeding, we compared frequency-based estimates of inbreeding with genomic measures of individual autozygosity. Results confirmed the higher degree of inbreeding of neo-bantams compared to the large fowl counterparts for all three ROH-length threshold classes (Table S6). Also based on the genomic measure of autozygosity, the Eikenburger bantam was the most inbred breed, with a genome-wide F_{ROH} of 0.66. Moreover, we reported a positive, significant correlation (r = 0.57, p-value = < 2.2e-6) between F_{ROH} and F_{IS} estimates.

Figure 4: Total number of ROHs and proportion of the genome covered by ROHs. a. The average number of ROHs belonging to the three size classes short (<1Mb), medium (1-3Mb), and long (>3Mb) for the 33 traditional breeds and 4 commercial lines. **b.** The total size of the genome covered by a particular class of ROH in one individual averaged per breed. In both figures, the Assendelft fowl bantam, Drenthe fowl bantam, North Holland Blue bantam, and Schijndelaar bantam were excluded because of the extremely small sample size (N=1), which precludes the calculation of the breed-averaged parameters represented in the figure. For abbreviations, refer to Table 1



2.4 Discussion

The availability of a large number of SNPs resulting from the development of high-density SNP assays has considerably improved the accuracy to assess population structure and relationship among populations, along with the genetic diversity either within or between populations [142]. High-density SNP arrays are also used to assess the effects of inbreeding through the occurrence of runs of homozygosity (ROH), which are increasingly used to infer past and present demography [31, 245, 142]. The application of SNP chip data to assess population genetic diversity and genetic management of traditional breeds is, although still scarce, improving. Here, we present the first comprehensive study on the population genetic diversity, population relationship, and demography of all traditional chicken breeds of the Netherlands recognized by the poultry community, to provide recommendations on their conservation and genetic management. The breeds studied are part of the Dutch poultry genetic resources and are also included in the FAO Domestic Animal Diversity Information System (DAD-IS). All traditional breeds of chicken are described by breed-specific morphological standards, including, among others, plumage color and pattern, feather structure and pigmentation, comb morphology, skin color, and eggshell pigmentation. Despite the exceptional diversity both in qualitative and quantitative traits, most of the traditional breeds are rare breeds or varieties having the status of endangered or critically endangered [314].

Assessing the genetic diversity and understanding the relationships among and within populations are the first necessary steps to establish conservation priorities and strategies [20, 87]. Large sample sizes are usually recommended to accurately estimate population statistics. However, large sample sizes are often difficult or impossible to achieve in genetic diversity studies of threatened and endangered populations [211]. In this study, sample size was a major limitation, as shown by the wide range of sample size across traditional breeds. Such limitation was mainly caused by the small effective population size of most of the breeds here considered and by the limited number of breeders involved in their conservation. For instance, at the time of sampling, the Eikenburger bantam was kept by only one breeder, leading to a sample size of only 4 individuals, while the Barnevelder was kept by roughly 10 farmers, which contributed to a sample size of 24 individuals (Table 1). It is therefore clear that in this study the number of hobby breeders and the popularity of the breed played a major role. We showed that, although genetic diversity estimates showed a varying degree of bias (Table S3-S4), population statistics can nevertheless provide revealing insights into the genetic diversity and, more importantly, the inbreeding history of a breed. Therefore, by keeping in mind potential bias in the results, conclusions and recommendations on the conservation and genetic management of traditional breeds can, but above all, should be attempted; genomic data is the only reliable source to estimate inbreeding and relatedness of marginalized populations in absence of other data sources such as pedigree data.

Practical conservation breeding should also be aware that the ascertainment of the SNP chip results in some bias in terms of detectability of unique genetic variation. Such bias is particularly important if the breed of interest was not involved in the development of the SNP array. Although the traditional Dutch breeds here considered were not used in the development of the 60K SNP chip, we did not observe a systematic difference between the neo-bantams and the other breeds. Therefore, we expect the ascertainment bias to be minimal. Moreover, we focused on the relatedness and runs of homozygosity (ROHs), both statistics that are less sensitive to missing such breed-specific variation. A meta-population-like structure can explain the variable genetic diversity estimates observed across traditional breeds, which are therefore subdivided into small breeder-based breeding units. The genetic diversity within each sub-population is strongly influenced by the breeder's breeding practices and selection preferences. However, results also show that a more or less restricted gene flow and a small local flock size have also divergent consequences on the breeds' genetic diversity, leading to the distinction between large fowls and neo-bantams.

The low genetic diversity observed in some of the large fowls, such as the Groninger Mew, may be explained by the drastic reduction in size of the breeding population occurred in the last century, which further suggests that some of the large fowls may have been close to extinction at some point or points in their history. Despite such population bottleneck, genetic data suggest that diversity was usually restored by crossing surviving individuals with other breeds showing complementary traits. However, the incorporation of genes from phenotypically similar breeds has decreased in popularity, because breeders prefer to use their prized cocks within their own farm. Currently, breeders use backcrossing to obtain a new generation of individuals sharing the same number of traits of the parental generation. As a result, the degree of inbreeding of the sub-population has increased, although the consequences on the long-term genetic diversity vary depending on the farmer's selection preferences and intensity, and flock size. Compared with the large fowls, neo-bantams have a recent historical origin and because of that, the demographic history and within-population genetic relationships are largely unknown. According to our analysis, different evolutionary and human-induced processes have and are now contributing to their meta-population structure. First of all, the creation of miniaturized fowls has been achieved using breeding strategies different from those of large fowl counterparts. And secondly, breeding strategies to create neo-bantams have rapidly changed over the past decades.

Neo-bantams initially resulted from a cross between large fowls of the same breed and true bantams of a distinct breed, such as Dutch bantams and Sumatras. However, our results indicate that, in the last decades, the bantamising trend has seen a significant change in the number of breed/varieties used. Such changes may have been driven either by the development of new phenotypes in large fowls that breeders want to have in a smaller sized individual or by the breeder's initiative to develop new small size varieties. Changes in the bantamising trend were well captured in the admixture analysis, which highlighted the use of even neo-bantams of the same sub-population and of phenotypically similar breeds to bantamize large-sized individuals. Although the genetic exchange between breeders is at the basis of the neo-bantam genetic diversity, backcrossing pursued for phenotype selection has, over time, considerably increased the degree of inbreeding, which, in our analysis, was the highest across the traditional breed clusters. However, the higher inbreeding coefficient may also result from the small effective flock size.

The analysis of runs of homozygosity (ROHs) can be used to address major concerns in conservation genetics, including inbreeding and population demography [31, 245, 142]. Although the 60K SNP panel allows an appropriate estimation of ROHs, ascertainment bias may underestimate the number of small ROHs [31], while amplifying the total length of the medium and longest ROHs [245]. Our results confirm the accuracy of the SNP panel for the analysis of medium and large ROHs and the ability of ROHs to reflect past and present population history, validating previous studies[31].

The analysis of ROH highlights the importance of novel marker-based information to prevent future loss of diversity. The prevalence of long ROHs across traditional chicken breeds is consistent with the limits to effective genetic management resulting from the absence of pedigree data and breed registry, and clearly shows the importance of recent inbreeding for the long-term viability of the populations. The ROH profile confirms such conclusions, in addition to suggest a major effect of individual breeders and breed associations practices on inbreeding. Historic and severe bottlenecks reported in some of the traditional large fowl breeds may further explain the greater proportion of long ROHs, since populations that have already experienced a drastic reduction in the effective population size tend to show a slow recovery, despite the potential increase in population size following the bottleneck [47]. On the other hand, the higher proportion of the genome covered by long ROHs displayed by neo-bantams can be explained by repetitive (sequential) backcrossing pursued for phenotype selection, since strong selection for breed-specific morphological standards or novel phenotypes acts to maintain long homozygous tracts [245]. However, the limited number of founders from which neo-bantams originated also supports their ROH content. The ROH profile of traditional breeds significantly differs from that of modern white-egg layers. White-egg layers experienced a strong population bottleneck in the early second half of the 20th century, which makes them interesting models of inbred populations. In this study, the white-egg layers demonstrate the strong influence of management strategy on inbreeding level. In fact, despite the high number of ROHs, the relative low proportion of genome covered by homozygous segments supports effective genetic management, which is meant to pursue intense, directional selection allowing recessive deleterious alleles to be purged with inbreeding. Furthermore, the higher number of long ROHs confirms the closed population history resulting from the absence of genetic exchange with other breeds/lines, resulting in continuous stretches of homozygosity. In commercial lines we also observed a relative small number of short ROHs (Figure 4a), which may indicate founder effects and distant inbreeding. In fact, it is likely that some relatedness was already present in the founders. However, recombination deriving from directional selection may also have contributed to break down ROHs in short segments. The analysis presented here confirms the importance of using genotype data to set up structured breeding programmes and inform genetic conservation of traditional breeds of chicken of the Netherlands. The selection and demographic history that we here reconstructed allow us to provide recommendations on how to effectively conserve genetic diversity of traditional breeds. According to our results, we recommend the national gene bank to consider traditional breeds separately. Moreover, to capture the available genetic diversity, a sufficient number of representative individuals within each breed should be sampled to preferably embrace all breed-specific morphological standards. A major drawback of current conservation efforts of traditional Dutch chicken breeds is the lack of inventories of their genetic resources. The present study demonstrates the use of high-density genotype data to investigate diversity in marginalized populations that can be used to guide sampling for in situ and ex situ conservation.

Genotype data can, in part, make up for the lack of traditional sources of information that inform breeding programs, such as pedigree data and phenotype information, It can also provide insight in the origin and consequences of strong demographic discontinuities, such as population bottlenecks and introgression, as in the case of the bantamized breeds. As such, the Dutch chicken breeds are a good model for other marginalized populations, and a good example for how genomic data can guide conservation efforts in populations that have little other information to go from. We conclude that bantamisation has generated novel genetic diversity. However, this outstanding diversity can only be preserved in the near future by applying structured breeding programmes that are either informed by pedigree data or genomic variation information.

2.5 Acknowledgements

We would like to acknowledge the owners and breed associations of the chicken populations used in this study for their help and cooperation during sampling. We would also like to thank the Centre for Genetic Resources for providing some of the samples used in this study. The research leading to some of these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353.

2.6 Data archiving

Genotype data are available from Dryad: 10.5061/dryad.1d832h3

2.7 Compliance with ethical standards

Conflict of interest. The authors declare that they have no conflict of interest.

2.8 Additional data

The online version of this article (https://www.nature.com/articles/s41437-018-0072-3) contains supplementary material, which is available to all users.

3.

The type of bottleneck matters: insights into the deleterious variation landscape of small managed populations

The chapter is based on:

Bortoluzzi, C., Bosse, M., Derks, M. F., Crooijmans, R. P., Groenen, M. A., Megens, H. J. (2020). *Evolutionary Applications*, 13(2), 330-341.

Abstract

Predictions about the consequences of a small population size on genetic and deleterious variation are fundamental to population genetics. As small populations are more affected by genetic drift, purifying selection acting against deleterious alleles is predicted to be less efficient, therefore increasing the risk of inbreeding depression. However, the extent to which small populations are subjected to genetic drift depends on the nature and time frame in which the bottleneck occurs. Domesticated species are an excellent model to investigate the consequences of population bottlenecks on genetic and deleterious variation in small populations. This is because their history is dominated by known bottlenecks associated with domestication, breed formation, and intense selective breeding. Here, we use whole-genome sequencing data from 97 chickens representing 39 traditional fancy breeds to directly examine the consequences of two types of bottlenecks for deleterious variation: the severe domestication bottleneck and the recent population decline accompanying breed formation. We find that recently bottlenecked populations have a higher proportion of deleterious variants relative to populations that have been kept at small population sizes since domestication. We also observe that long tracts of homozygous genotypes (runs of homozygosity) are proportionally more enriched in deleterious variants than the rest of the genome. This enrichment is particularly evident in recently bottlenecked populations, suggesting that homozygosity of these variants is likely to occur due to genetic drift and recent inbreeding. Our results indicate that the timing and nature of population bottlenecks can substantially shape the deleterious variation landscape in small populations.

3.1 Introduction

Deleterious mutations are expected to be held at low frequency by an equilibrium between the rate at which they arise by mutation and the efficacy of purifying selection at removing them from the population (mutation-selection balance) [224, 196]. However, the number and frequency of deleterious genetic variants segregating in a population are affected by many evolutionary forces, including artificial selection, genetic drift, and genetic hitchhiking, which is the change in allele frequency of a variant that is passed along together with another variant under positive selection [50, 270].

In small populations, the mutation-selection balance is challenged by population contractions, which reduce the efficacy of purifying selection to remove harmful mutations [223]. As a result, the genetic load, defined as the reduction in mean fitness in a population caused by deleterious variation relative to a mutation-free population [166], is predicted to be larger. In the long-term, the high genetic load and the rapid increase in frequency of harmful mutations could impact population survival and genetic diversity, increasing the risk of inbreeding depression [166].

Genetic drift, or the random fluctuations in the number and frequency of alleles, is mostly responsible for the deleterious genetic landscape in small populations. However, as studies in plant and animal species have suggested, the extent to which small populations are subjected to genetic drift considerably varies depending on the nature and time frame in which the bottleneck occurs [324, 200, 188]. For instance, a long-term population decline is expected to result in a lower proportion of amino-acid changing variants, along with a reduction in the additive genetic load, due to purifying selection acting against deleterious variants [200]. However, if populations have undergone recent and sudden declines, deleterious variation is predicted to be mainly shaped by genetic drift [223].

Domesticated species are an excellent model to investigate the consequences of population bottlenecks on genetic and deleterious variation. This is because their demographic history is characterized by multiple population contractions associated with domestication, breed formation, and intense selective breeding [200, 29, 197, 213]. Domestication involves the (partial or complete) isolation of a number of individuals from a wild progenitor population and entails drastic changes in the nature and strength of selective forces acting on the population, as well as its size [178]. The domestication bottleneck is usually followed by a long period of relatively weak and varying artificial selection, during which the reduced Ne may either be stable or fluctuate depending on human-driven selection. Contrary to the long-term domestication process, breed formation is a more recent event that often entails intense selection over short time periods and is coupled with limited recombination and an additional reduction in N_e [213]. In this study, traditional fancy breeds of chicken were used as a model species to investigate the consequences of two types of bottlenecks on deleterious variation: the severe domestication bottleneck occurred some thousands of years ago and the recent population decline accompanying breed formation in the last decades.

Since their development in the 16^{th} and 18^{th} century [64], traditional fancy breeds have persisted at small population sizes and comprised normal-sized (large fowl) and miniature (bantam) breeds. These traditional breeds experienced domestication only, which was based upon preferential breeding of birds exhibiting specific morphological features. The subsequent long period of weak and varying artificial selection resulted in the foundation of numerous breeds that are nowadays identified by an accurate phenotypic description [286]. In the last decades, hobby breeders have become interested in miniature forms of historical large breeds, which are called neo-bantams, and were initially created by crossing a large fowl with a bantam individual. Even though mating between neo-bantams has recently started to become very popular among hobby breeders, the selection purpose of obtaining an individual exhibiting all of the standard large fowl characteristics still remains [26]. The recent creation of neo-bantam breeds involved, on top of domestication, an additional population bottleneck. As we showed in our previous study, the reduced Ne and parent-offspring mating pursued within a neo-bantam breed to consolidate favorable traits considerably increased the level of inbreeding [26]. Although we expect the recent and sudden bottleneck to have acted differently on the accumulation of deleterious variants relative to the domestication bottleneck experienced by historical breeds, its effect on genome-wide patterns of deleterious variation remains unclear.

Accurate predictions of deleterious variants are essential when assessing their contribution to phenotypic variation [171]. To date, numerous approaches have been developed and applied to non-human species [188, 250, 324, 197, 200, 256], of which the Sort Intolerant From Tolerant (SIFT) approach is among the most widely used. However, as shown in Kono et al. (2016) and Derks et al. (2018), additional filtering criteria should be applied to the set of deleterious mutations to improve the reliability of the prediction. These criteria should include orthologous genes to minimize the effect of off-site mapping of sequence reads, RNA expression of protein-coding variants, and the use of different prediction approaches [79, 171]. We here expanded the approach of Derks et al. (2018) to predict deleterious mutations in domestic chickens by addressing a potential source of bias not previously investigated. That is reference bias, which is the higher probability of calling a variant as reference. We corrected for that by polarizing all protein-coding variants by ancestral and derived state, rather than reference and non-reference, to not underestimate the inferred number of nonsynonymous and deleterious variants.

Whole-genome sequencing data from 97 chickens representing 39 traditional fancy breeds were here used to directly examine the impact of different population bottlenecks on patterns of deleterious variation in small populations. Overall, we find that the recent population bottleneck associated with the creation of neo-bantams has resulted in a higher proportion of deleterious variants relative to large fowl and bantam counterparts, as genetic drift has reduced the efficacy of purifying selection to eliminate harmful mutations. We also observe that most deleterious variants are found in long tracts of homozygous genotypes, suggesting that homozygosity of these variants is likely to occur due to genetic drift and recent inbreeding. Our results indicate that the time frame and nature of the bottleneck can substantially shape the deleterious variation landscape in small populations.

3.2 Material and methods

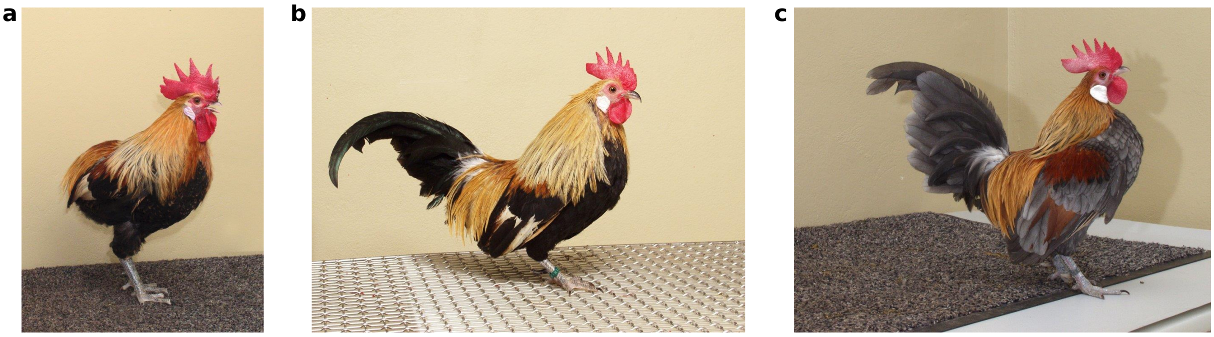
Samples and sequencing

DNA of 97 individuals from 39 traditional chicken breeds from the Netherlands was used for whole-genome sequencing on an Illumina HiSeq 3000. Four samples from each of the known living *Gallus* species were also sequenced for this study (Table S1). Based on their demographic and selection history, samples were classified into large fowls (n=51) (Figure 1a), neo-bantams (n=39) (Figure 1b), and bantams (n=7) (Figure 1c). Sequence reads were processed using standard bioinformatic pipelines (Supplementary Information, S1 Text), including alignment to the chicken reference genome (GenBank Accession: GCA_000002315.3) [308] using the Burrows Wheeler Aligner (BWA) [183], indel realignment, variant calling, and filtering of variants with quality <30. As two samples were discarded from further analyses because of low genome coverage (< 5x), the final dataset comprised 99 individuals (95 samples from traditional breeds and 4 samples from the *Gallus* species).

Principal component analysis

Genetic relationships among the 95 sequenced individuals (*Gallus* species excluded) were investigated through a principal component analysis (PCA) performed on the filtered vcf files in PLINK v1.9 [244]. First and second principal component were plotted using R v3.2.0 [283].

Figure 1: Traditional Dutch chicken breeds. **a.** Large fowl. The bird shown here is a Drenthe fowl boolstat, a breed of chicken whose main trait under selection is the absence of the tail. **b.** Neo-bantam. The individual is a bantamised bird of the Dutch fowl breed, and thus called Dutch fowl bantam. Neo-bantams are usually 2/3 the size of the original large fowl counterpart. **c.** Bantam. The bird shown is the Dutch bantam, one of the few true bantam breeds that exists only small in size. Bantam chickens are usually about a third to half the size of a regular large fowl chicken.



Heterozygosity analysis

Individual heterozygosity was estimated for each of the 95 chicken on a genome-wide scale by dividing the genome into non-overlapping windows of 10 kb (Additional file 1, Table S2). Within each window, heterozygous variants were called only if their depth of coverage met the minimum and maximum threshold, which were set at 4 and 2 times the average coverage, respectively (Figure S2). The total number of heterozygous sites called in a 10-kb window was corrected for the total number of sites not called because of low coverage, following Bosse et al. (2012). Insufficiently covered bins (10-kb windows with less than 1,000 well-covered sites) were excluded from the genome-wide autosomal heterozygosity analysis.

Runs of homozygosity (ROHs)

Runs of homozygosity were extracted from the genome of the 95 sequenced individuals implementing the method developed by Bosse et al. (2012). Information on heterozygosity was used to identify autosomal ROHs, which are here defined as genomic regions showing lower heterozygosity than expected based on the genome-wide average. To identify ROHs, we considered 10 consecutive bins at a time (100,000 bps) in which we calculated the average window heterozygosity that was then compared to the average genome-wide heterozygosity. The 10 consecutive bins were retained as candidate ROHs only if their level of heterozygosity was below 0.25 the average genomic diversity. All 10 consecutive bins that did not meet these criteria were considered to contribute to the genome-wide heterozygosity level outside ROHs. Local assembly or alignment errors were avoided as much as possible by relaxing the threshold within candidate homozygous stretches allowing for maximally twice the average heterozygosity in a bin, only if the heterozygosity within the candidate ROH did not exceed 1/3 the average genomic diversity (for more information refer to Bosse et al. (2012). Insufficiently covered bins were not considered in the actual size of each ROH but were considered in the calculation of the actual ROH length (assuming that all bins were highly covered). ROHs with insufficient coverage (less than 2/3) were removed from our calculations. From the final list of individual ROHs (Additional file 1, Table S2), we classified ROHs into three size classes, each of them corresponding to a specific demographic event, including past relatedness (short ROHs: <100 kb), background relatedness (medium: 0.1-3 Mb), and recent relatedness (long: ≥ 3 Mb).

Population history estimation

SMC++ was used on unphased whole-genome sequencing data to estimate population history [284]. Only samples with an average genome coverage >10x and percentage of missing sites $<10\%$ were considered (Table S5). Population history was estimated for each breed separately setting the mutation rate to 1.9x10-9 site-1 year-1 [220] and generation time at 1 year.

Inferring the ancestral state

Sequencing data of the three wild *Gallus* species included in the dataset (i.e. *G. varius, G. sonneratii, G. lafayetii*) were used as an outgroup to predict the ancestral and derived allelic state of all polymorphic sites. A variant was categorized as ancestral if the three wild samples had the same genotype (homozygous reference or homozygous alternative). The 11 706 316 identified variants were extracted from each sample with a minimum average genome coverage >10x and classified as homozygous ancestral, heterozygous, or homozygous derived.

Functional annotation of variants

Variant annotation was performed with the Variant Effect Predictor (VEP) [203] running the Sort Intolerant From Tolerate (SIFT) algorithm, using the Ensembl *G. gallus* annotation database (release 90). Protein-coding variants were defined based on their SIFT score as synonymous, nonsynonymous tolerated (SIFT score ≥ 0.05), and nonsynonymous deleterious (SIFT score <0.05). We also catalogued mutations that disrupt the generation of a a fully functional protein either by introducing a stop codon or by truncating the protein reading frame as loss of function (LoF) variants. To be more confident on the detection of deleterious variants, we implemented the approach developed by Derks et al. (2018) considering only variants annotated in genes that were 1:1 orthologous in Ensembl with zebra finch and for which the RNA-seq expression coverage was at least 200 in the Ensembl merged RNA-seq dataset (release 86). To increase our confidence in the deleteriousness of our set of putatively deleterious variants predicted by SIFT, we used the GERP scores computed for the 7-sauropsids multiple whole-genome alignment as an additional approach (ftp://ftp.ensembl.org/pub/release-94/compara/). Variants were considered truly deleterious if SIFT score <0.05 and GERP score >1.0.

Test for elevated homozygosity of derived genotypes

Following Robinson et al. (2016), we used likelihood ratio tests to evaluate whether the number of homozygous derived genotypes per individual differed between large fowls and neo-bantams at synonymous, tolerated, deleterious and LoF variant sites (Supplementary Information, S1 Text). We focused on these two management groups only as they were previously found to be genetically more similar than to bantam breeds [26]. Therefore, by testing for differences in homozygosity of derived genotypes we wanted to see whether these two groups might also share similar proportion of deleterious variants. Briefly, the test compared the likelihoods under two models. Under the null model, we assumed a similar proportion of homozygous derived alleles between neo-bantams and large fowls ($p_{lf} = p_{nb}$), whereas under the alternative model differences in the number of homozygous derived genotypes are expected between large fowls and neo-bantams ($p_{lf} \neq p_{nb}$). The log-likelihood values of both the null and alternative models were used to calculate the likelihood ratio test (LRT) as,

$$\Lambda = -2(\text{loglikelihood}_{null} - \text{loglikelihood}_{alternative})$$

Genetic load

Genetic load was calculated as the ratio of nonsynonymous deleterious to synonymous sites in each individual and averaged across individuals within each of the three management groups. Genetic load was separately estimated for heterozygous and homozygous derived alleles.

Site-frequency spectrum (SFS)

The derived allele frequency (DAF) spectrum was calculated for synonymous, tolerated, and deleterious variants, considering only bi-allelic SNPs. We then generated a histogram with 10 bins (with steps of 0.1 allele frequency) starting from a very low (0-0.10) to a very high (0.90-1.0) derived allele frequency.

Enrichment of ROHs for deleterious variants

The distribution of putatively deleterious mutations inside and outside of ROHs was investigated following the method proposed by Szpiech et al. (2013). Homozygous derived variants were grouped into non-damaging or putatively neutral (e.g. synonymous and tolerated) and damaging (e.g. deleterious and LoF). The occurrence of damaging and non-damaging variants was investigated inside and outside each ROH size class. Coordinates of ROHs were used to calculate the fraction of the genome covered by any ROH and by each ROH size class as:

$$G_{i,j} = \frac{L_{ROH}}{L_g} \tag{3.1}$$

where L_g is the total length of the genome, L_{ROH} the total length of ROHs, i is the individual, and j is the ROH class $j \in$ (S,M,L), representing small, medium, long, and any ROH respectively (Additional file 1, Table S3).

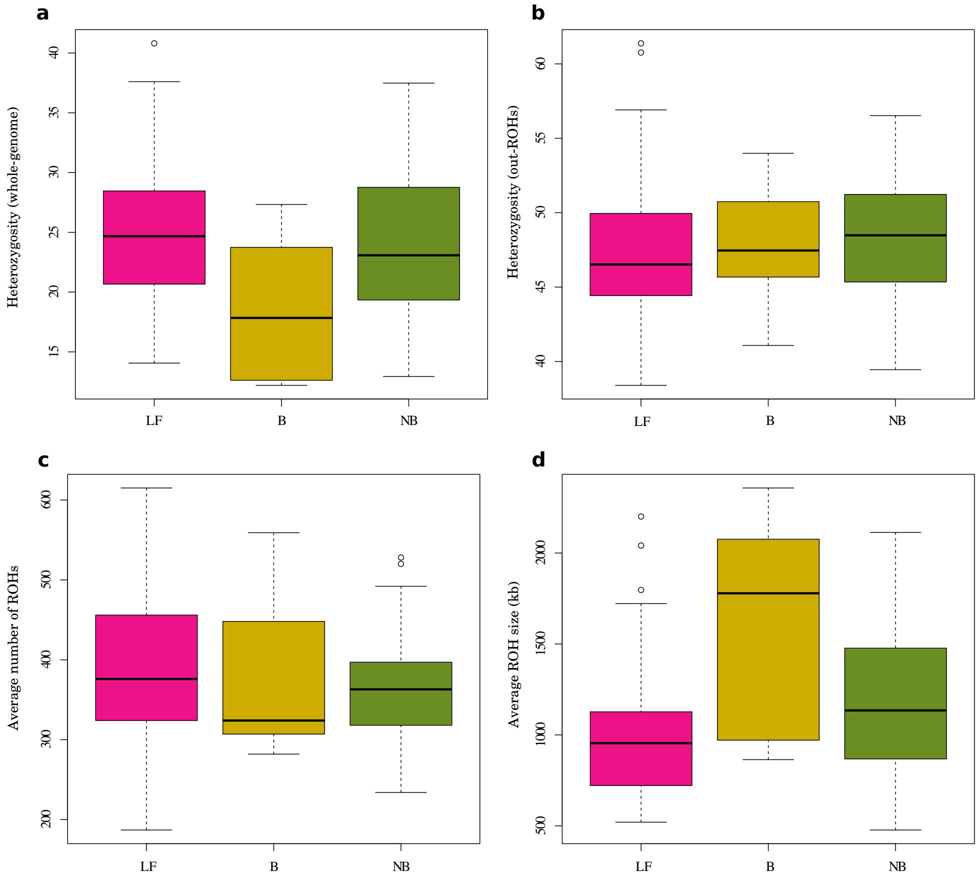
3.3 Results

Patterns of deleterious variation were investigated using whole-genome sequencing (WGS) of 39 traditional chicken breeds (Table S1). On average, 13.4x coverage was generated for each individual (Additional file 1, Table S1). The population-based variant calling approach identified 17 million SNPs and 1.2 million insertions/deletions (indels) (Table S2). Variants were distributed with an average density of 20 SNPs/100-kb, ranging from 0 to 82. The average transition to transversion (Ts/Tv) ratio was 2.58 (Table S2), which is in line with previous findings in commercial chicken populations [79]. Samples origin was validated with the principal component analysis (Figure S1).

Population history is responsible for the current autozygosity landscape

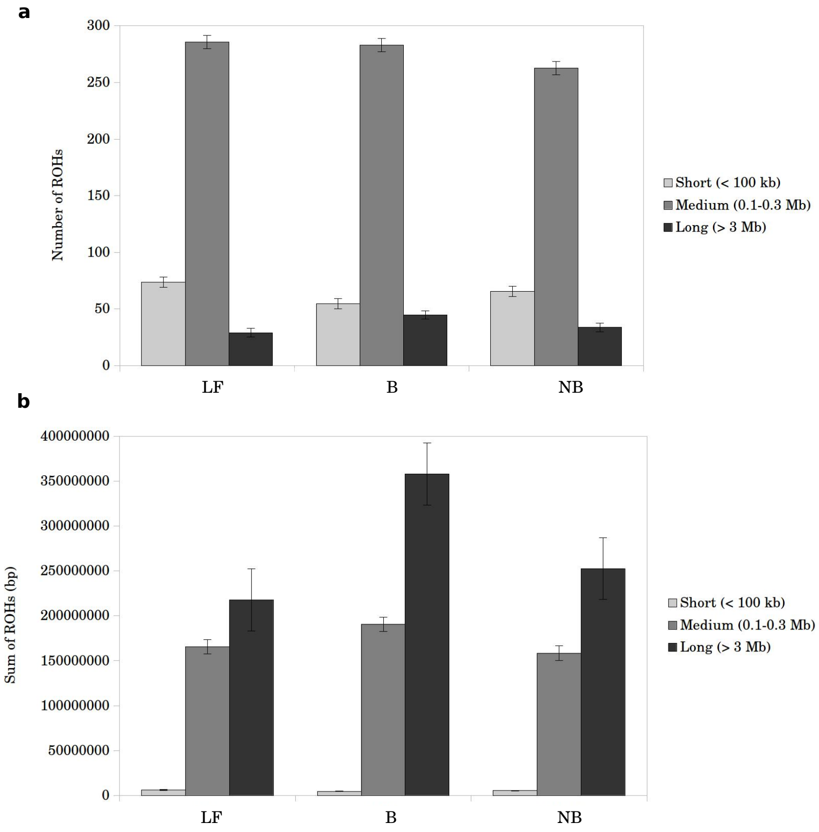
Genome-wide autosomal heterozygosity and ROHs were used to investigate the extent and nature of genetic variation in our populations. Whole-genome heterozygosity ranged from 12.2 to 40.8 SNPs/10-kb (Additional file 1, Table S2). On average, neo-bantams showed slightly lower heterozygosity than their original large fowl counterparts, though the level of heterozygosity was considerably higher than that observed in the bantam breeds (Figure 2a). The level of heterozygosity increased almost two-fold in all breeds when excluding ROHs, with neo-bantams showing slightly higher heterozygosity than both source populations (large fowls and bantams)

Figure 2: Heterozygosity and runs of homozygosity. **a.** Average heterozygosity including ROHs. **b.** Average heterozygosity outside ROHs. **c.** Average number of ROHs along the genome. **d.** Average ROH size in kb. Abbreviations: LF, large fowls (n=49); B, true bantams (n=7); NB, neo-bantams (n=39)



(Additional file 1, Table S2; Figure 2b). The lower genome-wide heterozygosity observed in neo-bantam and bantam breeds is therefore explained by their higher average ROH size (Figure 2d). In fact, the genome of neo-bantams and bantams is mostly covered by (few) long ROHs (>3 Mb) rather than by small (<100 kb) and medium ROHs (0.1-3 Mb) (Figure 3) (Additional file 1, Table S2). On average, 25% of the genome in neo-bantams was covered by long ROHs, 0.6% by short, and 17% by medium ROHs (Additional file 1, Table S3). Of the bantam breeds, the Eikenburger bantam was the most inbred, with up to 70% of the genome covered by ROHs (Additional file 1, Table S3). To further investigate how historical demographic changes have shaped the genomic patterns of homozygosity observed in our populations we decided to infer past effective population size (Ne). According to our results, the chicken ancestral population size remained stable up to approximately 10,000 years, after which it dropped from an initial Ne of 106 to 104-103 (Figure S3). Both management groups showed a constant flattening population size which has hardly recovered since the bottleneck.

Figure 3: Runs of homozygosity. **a.** Average number of short (< 100 kb), medium (0.1-3 Mb), and long (> 3 Mb) ROHs. **b.** Average ROHs size for the three ROH size classes (short, medium, long). Abbreviations: LF, large fowls (n=49); B, true bantams (n=7); NB, neo-bantams (n=39)



More rare than fixed deleterious variants

The role of genetic drift and purifying selection was investigated by annotating variants with respect to their effects on the amino-acid sequence (Table S3). We also annotated alleles as ancestral and derived using three wild *Gallus* species as an outgroup. After filtering for RNA-seq coverage and 1:1 orthologues, the final set of variants comprised 61 567 synonymous, 16,840 nonsynonymous tolerated, 3 833 nonsynonymous deleterious, and 755 loss of function (LoF) mutations. Of the initial set of deleterious variants, 1 674 were classified as deleterious by both SIFT and GERP++. The efficacy of selection at removing deleterious variants from a population was investigated by looking at the distribution of the derived allele frequency (DAF) spectrum. The frequency spectrum showed more rare (DAF <0.1) derived alleles than nearly fixed or fixed deleterious alleles (DAF ≥0.9) (Figure 4). We observed similar DAF spectra for large fowls (Figure S4A) and neo-bantams (Figure S4B), with neo-bantams showing slightly higher derived allele frequency than large fowl counterparts, even up to a DAF of 0.5. Moreover, neo-bantams showed fewer nearly fixed or fixed deleterious variants than large fowl counterparts.

Figure 4: Derived allele frequency spectra of the 39 traditional chicken breeds. Derived allele frequency was inferred for synonymous, nonsynonymous tolerated (SIFT score ≥ 0.05), and nonsynonymous deleterious (SIFT score < 0.05) mutations

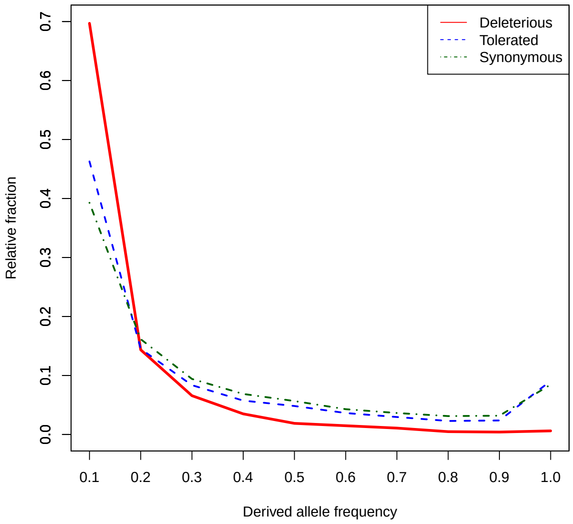
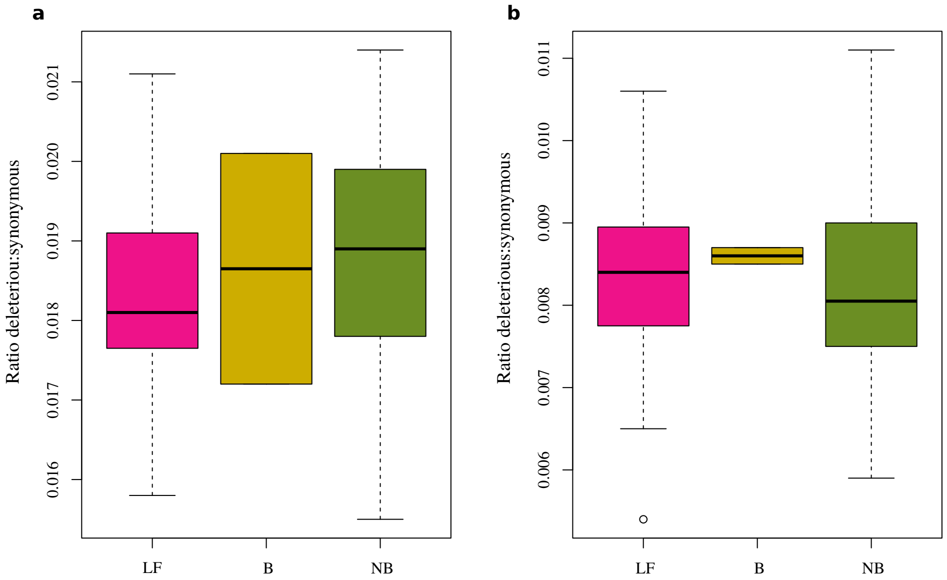


Figure 5: Genetic load expressed as deleterious to synonymous ratio. **a.** Genetic load for heterozygous sites. **b.** Genetic load for homozygous derived sites. Abbreviations: LF, large fowls (n=35); B, true bantams (n=3); NB, neo-bantams (n=22)



The effects of genetic drift on deleterious variation

In traditional fancy breeds the total number of derived alleles (heterozygous and homozygous derived) was lower for deleterious and LoF variants relative to putatively neutral ones (synonymous and nonsynonymous tolerated) (Figure S5). Compared to large fowls, neo-bantams were slightly more enriched in the total number of deleterious and LoF homozygous derived mutations (Figure S5D). Despite these differences in the total number of homozygous derived genotypes, we decided to perform a likelihood ratio test (LRT) to formally test for individual differences between neo-bantams and large fowls. According to the likelihood ratio test, the number of homozygous derived genotypes was not significantly different between large fowl and neo-bantam counterparts for deleterious (*p-value:* 0.730) and LoF (*p-value*: 0.272) variants (Table 1). Significant were the differences for synonymous (*p-value:* 3.442e-08) and nonsynonymous tolerated (*p-value:* 0.015) variants. We also investigated in each of the three management groups the total genetic burden resulting from the accumulation of deleterious mutations (genetic load) (Figure 5). The deleterious to synonymous ratio of heterozygous variants was, on average, lower in large fowls compared to bantam and neo-bantams. The same ratio when considering homozygous derived variants was, on average, slightly higher in large fowls than neo-bantams, which, on the other hand, showed extensive variation (Figure 5). Contrary, bantam breeds showed little variation with an average higher deleterious to synonymous ratio than that of large fowls and neo-bantams.

Table 1: Test for elevated homozygosity of derived genotypes per individual between large fowls and neo-bantams

	Null model[1]		**Alternative model**[2]		
Functional category	**MLE**	**Log-likelihood**	**MLE**	**Log-likelihood**	**Likelihood ratio test (LRT)**[3]
Synonymous	$p_{lf} = p_{nb} = 0.230$	-2903.70	$p_{lf} = 0.229$; $p_{nb} = 0.232$	-2888.48	$\Lambda = 30.44; \textit{p-value}: 3.44e-08$
Nonsynonymous tolerated	$p_{lf} = p_{nb} = 0.209$	-902.61	$p_{lf} = 0.208$; $p_{nb} = 0.210$	-899.69	$\Lambda = 5.843; \textit{p-value}: 0.015$
Nonsynonymous deleterious	$p_{lf} = p_{nb} = 0.070$	-323.85	$p_{lf} = 0.070$; $p_{nb} = 0.070$	-323.79	$\Lambda = 0.118; \textit{p-value}: 0.730$
Loss of function (LoF)	$p_{lf} = p_{nb} = 0.169$	-229.89	$p_{lf} = 0.167$; $p_{nb} = 0.171$	-229.28	$\Lambda = 1.202; \textit{p-value}: 0.272$

[1] The null model states that the proportion of homozygous derived genotypes that a large fowl carries is equal to that of the same genotypes carried by a neo-bantam, so that $p_{lf} = p_{nb}$

[2] Contrary to the null model, in the alternative model the proportion of homozygous derived genotypes carried by a large fowl individual is different from that of a neo-bantam ($p_{lf} \neq p_{nb}$)

[3] Likelihood ratio test for differences in the number of homozygous derived genotypes per individual between large fowl and neo-bantams. The χ^2 distribution with one degree of freedom of Λ was used to calculate p-values.

Deleterious variation and demographic history

The study of ROHs offers a new basis for assessing the mechanisms by which demography and selection produce patterns of deleterious variation. The total number of putatively neutral homozygous derived sites was higher than that of damaging sites (Figure S6). Moreover, with the increasing proportion of the genome covered by longer ROHs, the number of homozygous derived variants within ROHs increased for both damaging *(Pearson's correlation*: 0.844, *p-value:* <2.2e-16) and non-damaging sites (*Pearson's correlation:* 0.844, *p-value:* <2.2e-16). As expected, the number of homozygotes occurring outside ROHs decreased with the fraction of the genome in any ROH, as the genome simply contains fewer ROH-free regions. A negative correlation between homozygous sites and genome not covered by ROHs confirmed our expectations for damaging (*Pearson's correlation:* -0.624 *p-value:* 1.261e-07) (Figure S6B) and non-damaging sites (*Pearson's correlation:* -0.644 , *p-value:* 3.524e-08) (Figure S6A). The fraction of damaging and non-damaging homozygous derived genotypes in ROHs positively correlates with the total genomic ROH coverage (*Pearson correlation:* 0.956, *p-value:* <2.2e-16 for damaging; *Pearson correlation:* 0.981, *p-value:* <2.2e-16 for non-damaging (Figure 6). We also observed that each ROH size class (short, medium, long, and any) is e more enriched for deleterious homozygous derived variants than for non-damaging homozygotes. However, this excess in deleterious mutations is particularly evident for any (Figure 6a) and long ROHs (*Pearson correlation:* 0.972, *p-value:* <2.2e-16 for damaging; *Pearson correlation:* 0.987, *p-value:* <2.2e-16 for non-damaging) (Figure 6d).

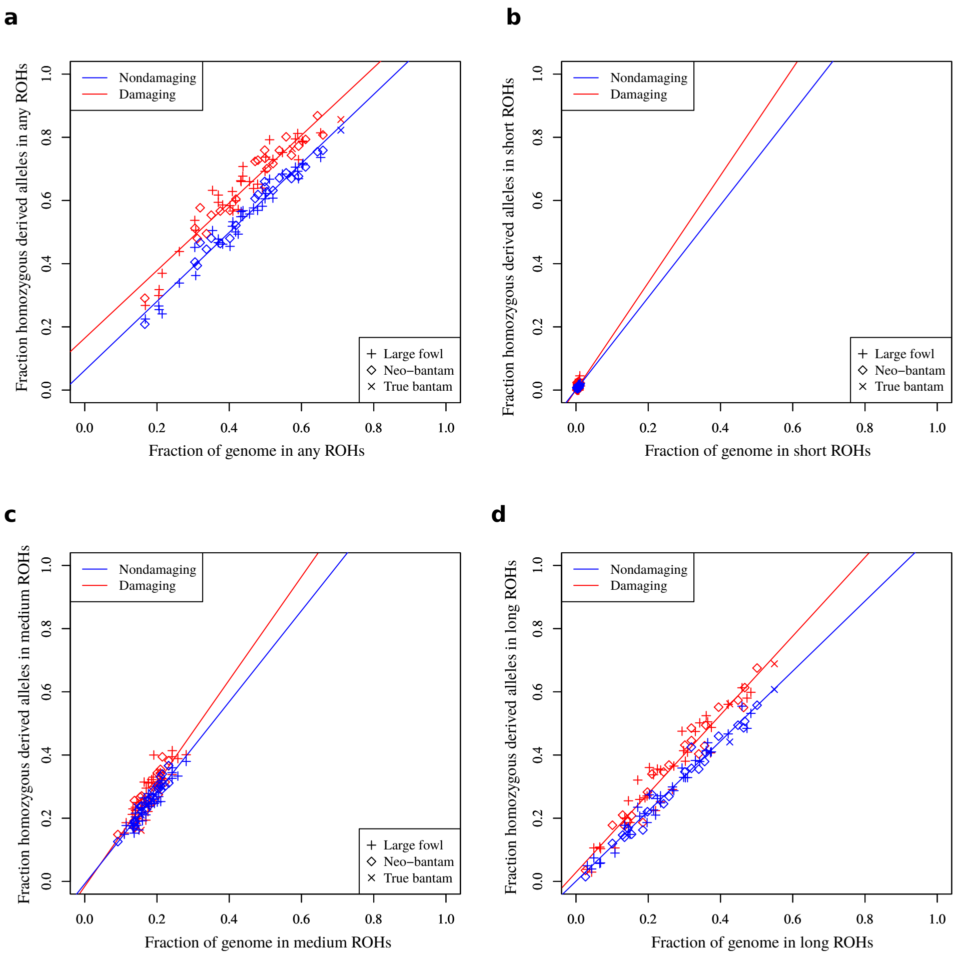
3.4 Discussion

In this study, we used whole-genome sequencing data from traditional fancy breeds of chicken to investigate the consequences of population bottlenecks on genome-wide patterns of deleterious variation in small populations. To do so, we combined individuals from multiple breeds with similar demographic history into one population to better estimate genetic and deleterious variation. Such approach was most suitable give the low number of individuals per breed (between 1 and 4), which is the direct consequence of the threatened population size of most of these breeds (http://edepot.wur.nl/424249). Even though (slightly) different breeds were grouped into the same population, we expect potential bias to be minimal, as already in Bortoluzzi et al. (2018) when using genome-wide SNP chip data.

In line with the small-population paradigm, we showed that the size of a population (N_e) is an important evolutionary factor in determining the level of genetic variability and the effectiveness of purifying selection at removing harmful mutations [42, 47]. In fact, as we showed, populations of small Ne have a lower genetic diversity and high level of inbreeding, along with being more affected by genetic drift.

In traditional large fowls, the accumulation of deleterious alleles is characteristic of a population that since the domestication bottleneck has persisted for a long period of time at small

Figure 6: Fraction of the genome covered by ROHs versus the fraction of damaging and non-damaging sites. **a.** Any ROH. **b.** Short ROHs. **c.** Medium ROHs. **d.** Long ROHs. Damaging homozygotes are shown for each individual belonging to the three management group, which are identified by different shapes



size. Therefore, in these populations deleterious mutations of especially small effects are expected to have risen in frequency and become fixed [136]. As several studies have shown, domestication substantially decrease the effective population size and efficacy of purifying selection, which in turn reduces the genetic diversity and increases the mutational load [213]. These major effects have been observed in many species despite the multiple domestication centres and large population size of the ancestor (in the case of chicken, the red jungle fowl) [158, 209, 197]. For example, Mardsen et al. (2016) recently observed that the dog genome harbours more amino acid changing variants than that of the wild wolf ancestor. The higher proportion of deleterious variants has also led to an increase in the additive genetic load in many dog breeds, which clearly indicates that the efficacy of purifying selection is lowered by strong population contractions accompanying domestication [200, 62]. Similar conclusions have been reached in other domestic animal species [197, 263] and plants [194, 188]. Even though demographic contractions associated with domestication have a major impact on the genome-wide genetic and deleterious variation, processes that co-occurred during domestication have recently questioned the role of domestication itself in increasing the mutation load. For example, the shift in mating system from outcrossing to predominantly selfing rice has been suggested to have substantially influenced the occurrence of deleterious mutation in domesticated rice [188]. It is, however, not to exclude that also the long period of weak and varying artificial selection for desirable traits following domestication could have further reduced N_e [213]. Despite the small population size and high genetic load, large fowl breeds retain substantial genetic variation, mainly because of crossing with other breeds performed in the past to maintain a viable population size and nowadays for phenotypic selection [26]. The favourable consequences of genetic exchange (gene flow) observed in our large fowl populations find support in wild species affected by similar drastic population bottlenecks. For example, in the case of the Iberian lynx, the promotion of admixture with Eurasian populations has resulted in less inbred and more genetically diverse populations, potentially more adapted to environmental changes [1].

Contrary to their large fowl counterparts, the extent and nature of deleterious variation in neo-bantams is characteristic of a population that went through a more recent and severe population decline. In the case of neo-bantams, the bottleneck is associated with their creation in the last decades [26] Because of the very small number of founder individuals, neo-bantams may not yet have had the time to adapt to stochastic demographic and genetic events. This is mostly because the small Ne of neo-bantams is not large enough to discount the effects of genetic drift. As a result, weakly deleterious mutations accumulate in the genome due to genetic drift, as purifying selection does not have the time to purge these harmful mutations [195]. The central role of genetic drift observed in our recently bottlenecked populations on deleterious variation is also observed in small populations under natural selection [1, 139, 240, 256]. In their study, Robinson et al. (2016) showed that a severe bottleneck occurred 30 generations ago has substantially reduced the already small effective size of the island fox from originally 64 individuals to fewer than a dozen. This recent population contraction has severely affected not only the genetic variation of the species, but also the genetic load. As a result, all island populations show more homozygous deleterious mutations relative to the heterozygous, which, as the authors suggest, have become homozygous likely through strong genetic drift [256].

Although we have shown that the demographic history accompanying the bottleneck and genetic drift are important factors in shaping deleterious variation, inbreeding and artificial selection can also affect the mutational load by increasing the probability of harmful mutations to become homozygous. If in homozygous state, these recessive deleterious mutations can potentially lower an individual fitness (inbreeding depression) [29]. Small populations are more prone to suffer from inbreeding depression, as the probability of mating between relatives is high. To test whether the high level of inbreeding observed in our populations affects the deleterious variation landscape, we looked at the distribution of deleterious mutations in- and outside ROHs following Szpiech et al. (2013). In line with results in humans [279], domesticated pig [29], commercial chicken [29] and cattle [325], we found ROHs to be proportionally more enriched in homozygous deleterious alleles than the rest of the genome. However, when looking at the ROH size classes, long ROHs, which are an indication of recent inbreeding, were significantly more enriched than any other size class. This pattern was particularly clear in neo-bantams, which supports the role of both inbreeding and genetic drift in increasing the occurrence of deleterious mutations in homozygous state.

In a recent study on commercial chicken lines, Derks et al. (2018) observed that putative highly deleterious variants can be rare in populations of small effective size if specific breeding programmes aiming to select individuals against inbreeding depression are in place. Therefore, the presence of a breeding program counterbalances the effects of inbreeding and strong artificial selection. In traditional breeds, as well as in small populations under natural selection, the risk of increasing an individual mutational load is considerably higher, because breeding and conservation programmes are often not in place to genetically manage these populations [26]. Moreover, as mating between family members is intentionally pursued to select for specific traits, the proportion of homozygous segments in individual genomes is expected to substantially increase along with that of slightly deleterious mutations. Therefore, we expect the viability of the traditional breeds investigated in this study to strongly depend on future breeding preferences, which, if not genetically managed, are likely to limit the full exploitation of their genetic potential.

3.5 Conclusions

In this study, we showed that the timing and nature of a population bottleneck can substantially shape the deleterious variation landscape in small populations. In particular, we showed that populations kept at small size for long period of time since the bottleneck have a reduced burden of deleterious alleles compared to recently bottlenecked populations. The reduced deleterious burden in these populations, which is also linked to a reduced number and total length of ROHs across the genome, is likely responsible for their genetic success. According to our study, facilitating purging of deleterious mutations through inbreeding avoidance should be at the core of future breeding and conservation programmes in small populations [70]. However, genomic information on deleterious variation can, and should, be incorporated and used in the development of conservation programmes that assure the long-term survival and enhance the genetic diversity of small populations. Fitness-related traits should also be considered to better measure potential fitness consequences at the individual and population level associated with recessive deleterious mutations.

3.6 Acknowledgements

We would like to acknowledge the poultry breeding clubs and hobby breeders for their collaboration during sampling. A special thanks to Miguel Correa-Marrero and Carlos de Lannoy for their useful suggestions and input on improving the runs of homozygosity pipeline used in the study. The research leading to some of these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353. The research was also partly funded by the STW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality.

3.7 Data availability statement

Whole-genome sequencing data from this study have been submitted to the European Nucleotide Archive (ENA) under accession number PRJEB34245. Source codes for running the ROH pipeline have been deposited in https://github.com/cbortoluzzi/ROHs.

3.8 Additional data

The online version of this article (https://onlinelibrary.wiley.com/doi/full/10.1111/eva.12872) contains supplementary material, which is available to all users.

3.9 Supplementary Information

SI Text

Alignment and variant calling

Reads were trimmed using sickle v1.33 [157] before being mapped to the chicken genome (GenBank Accession: GCA_000002315.3) with the Burrows Wheeler Aligner (BWA) v0.7.15 [183]. Mapping was performed with the default settings. Duplicate reads were removed using the rmdup function of samtools v1.19 [184]. The GATK IndelRealigner was then used to perform realignment of reads around indels [202]. Genome-wide coverage and mapping quality were evaluated with Qualimap v2.2 [114]. We performed a population-based variant calling using Freebayes v0.9.10 [115]. The following criteria were used: (1) –min-base-quality 10 (a support base quality 10), (2) –min-mapping-quality 20 (a support mapping quality 20), (3) –min-alternate-fraction 0.2 (at least 20% of reads supporting the alternative allele), and a –min-alternate-count 2 (at least 2 reads supporting the alternative allele). To reduce the number of missed variants, variants of each individual were filtered out if they did not meet the coverage requirements, which were a minimum read depth of 4X and a maximum of 2.5 times the average individual genome-wide coverage. The false discovery rate was reduced by performing additional post-processing using Bcftools v1.4.1 [184], setting (1) a phred quality score > 30, (2) allele count supporting the alternative allele > 2, (3) maximum number of 10 alleles, (4) variants located within 3 bp of an indel, and (5) call rate < 0.70.

Likelihood ratio test for elevated proportion of homozygosity of derived genotypes

We tested whether the number of homozygous derived genotypes per individual differed between large fowls and neo-bantams, implementing the approach developed in Robinson et al. (2016). For a given set of genotypes *g*, the number of homozygous derived alleles that an individual carries follows a binomial distribution with parameters *p* and *g*, where *p* is the proportion of SNPs for which an individual is homozygous derived. According to the null model, the proportion of homozygous derived genotypes that a large fowl individual carries is equal to the proportion of the same genotypes carried by a neo-bantam individual, as expressed by the formula $p_{lf} = p_{nb}$. Therefore, the null model was:

$$l(p|x,g) = \sum_{j=1}^{57} x_j log(p) + (g_j - x_j) log(1-p) \tag{3.2}$$

where x$_j$ is the number of homozygous derived genotypes for individual j and g$_j$ the total number of called genotypes in individual j . Since we assumed similar proportion of homozygous derived alleles between neo-bantams and large fowls, the null model was calculated over the 57 individuals (35 large fowl and 22 neo-bantam individuals). Under the alternative model, differences in the number of homozygous derived genotypes were expected between large fowls and neo-bantams, so that p$_{lf}$ $\neq$ p$_{nb}$:

$$l(p_{lf}, p_{nb}|x, g) = \sum_{j=1}^{35} x_{lf} log(p_{lf}) + (g_{lf} - x_{lf}) log(1 - p_{lf}) + \sum_{j=1}^{22} x_{nb} log(p_{nb}) + (g_{nb} - x_{nb}) log(1 - p_{nb}) \quad (3.3)$$

where x$_{lf}$ and x$_{nb}$ is the number of homozygous derived genotyped of the large fowls (n=35) and neo-bantams (n=22), respectively, and g$_{lf}$ and g$_{nb}$ the total number of called genotypes in the large fowls and neo-bantams. We then tested whether the alternative model fit the data better than the null model using a likelihood ratio test (LRT) as:

$$\Lambda = -2(l(p|x, g) - l(p_{lf}, p_{nb}|x, g)) \quad (3.4)$$

where Λ is χ^2 distribution with one degree of freedom, from which we calculated p-values. The likelihood ratio test was separately calculated for synonymous, missense tolerated, missense deleterious, and loss-of-function variants.

4.
Quantifying temporal genomic erosion in small managed populations under a recently established conservation programme

The chapter is based on:

Bortoluzzi, C., Restoux, G., Rougé, R., Desnoues, B., Petitjean, F., Bosse, M., Tixier-Boichard, M. (2020). *In preparation*

Abstract

Livestock biodiversity is declining globally at rates unprecedented in human history. In chickens, local breeds are the most affected, because their small population size makes them more susceptible to demographic stochasticity and genetic drift. The maintenance of genetic diversity and control over genetic drift by conservation programmes are well understood, but often overlooked. We here used temporal whole-genome sequencing data to assess the consequences of a conservation programme on the genetic diversity, deleterious variation, and inbreeding of two local French chicken breeds. We show that, despite the small population size, conservation programmes can maintain genetic diversity while limiting the negative effects of genetic drift. However, breeds can benefit from management only if conservation practices are consistent over time. Our results reinforce the imperative to establish and sustain existing conservation programmes that aim to keep local livestock breeds from the brink of extinction.

Introduction

Livestock breeds are recognized as important components of world biodiversity since they harbour genes and genetic variants that are and will be useful to agriculture in the future [107]. Nevertheless, livestock diversity is declining globally at rates unprecedented in human history, resulting in an accelerated rate of breeds extinction. According to the recent report on *The State of the World's Biodiversity for Food and Agriculture* of the Food and Agriculture Organization (FAO), only 7,745 local breeds of livestock are still in existence, 26% of which are at risk of extinction, while 67% are of unknown risk status [18]. Many of the threats that are affecting livestock diversity have been identified, such as indiscriminate cross-breeding, production system intensification, and introduction/increased use of exotic breeds. However, the assessment of these threats still needs to be improved, particularly with respect to local animal genetic resources (AnGR) [262].

Avian species, and particularly chicken, are among the livestock species with the highest number of breeds with critical status (Figure 1). The establishment in the mid 20^{th} century of few, specialized breeding industries that rely on few selected lines for egg (layer) or meat (broiler) production has been partially responsible for the decline in local chicken diversity in Europe and North America (Figure 1) [216]. However, the large number of chicken breeds at risk is also due to the often unclear and problematic definition of a breed, which makes any direct risk assessment rather challenging. From a genetic perspective, local chicken breeds are at major risk of extinction because their small population size makes them more susceptible to stochastic demographic and genetic events. The risk of genetic erosion is often enhanced by the lack of conservation programmes, either *in situ* (i.e. conservation of live animals by livestock keepers in the production system) or *ex situ in vivo* (i.e. conservation through maintenance of live animals in, for instance, zoo, ark farms or experimental facilities) [26].

Genetic drift, or the random fluctuation in allele frequencies, is the main stochastic event responsible for the loss of genetic diversity in small populations [102]. In fact, genetic drift can reduce the viability and adaptive potential of a population [170]. Recent studies in wild and domesticated species [257, 318, 292, 25, 1] have shown that the risk of extinction in small populations is also a consequence of harmful mutations that can lower the fitness of an individual carrying them. The rationale is the reduced efficiency of natural selection at purging harmful mutations because of genetic drift [223, 165]. Therefore, deleterious alleles can accumulate and reach fixation in the genome. Additionally, as small populations suffer from inbreeding resulting from mating between close relatives [160], (recessive) deleterious mutations in homozygous state can express their harmful nature.

Conservation programmes are well-known to maintain genetic diversity while controlling for genetic drift [262]. However, the consequences of a conservation programme on a population in terms of genetic diversity, deleterious variation, and inbreeding have rarely been investigated in local livestock breeds. Such assessment is nonetheless of particular relevance today, as the maintenance of high genetic diversity alone is not sufficient to ensure the long-term survival of populations of small size [174, 257]. Recent advances in sequencing technologies can help us in the task of evaluating current conservation programmes with the aim of providing objective recommendations to effective management practices for small local populations [80, 129]. Temporally sampled genomic data are a powerful tool to monitor changes in genetic parameters, including genetic diversity ($\Delta\pi$), inbreeding level (ΔF), deleterious variation (ΔL), and, if applicable, selection (ΔS). Hence, when possible, temporal genomic indices should be quantified to evaluate and guide existing and future conservation programmes [80].

In this study, we assessed the consequences of a conservation programme on the genetic and deleterious variation of two local French chicken breeds, the Barbezieux and Gasconne, by means of whole-genome sequencing data. For each breed, a conservation programme was established in 2003 and is presently under the coordination of the French Poultry and Aquaculture Breeders Technical Center (SYSAAF). The SYSAAF is a professional union that, by close collaboration with several research institutes, provides genetic, genomic and reproductive services to preserve, exploit, and value the genetic diversity of the breeds concerned. The ultimate objective is to produce original products under quality labels at a local or national scale.

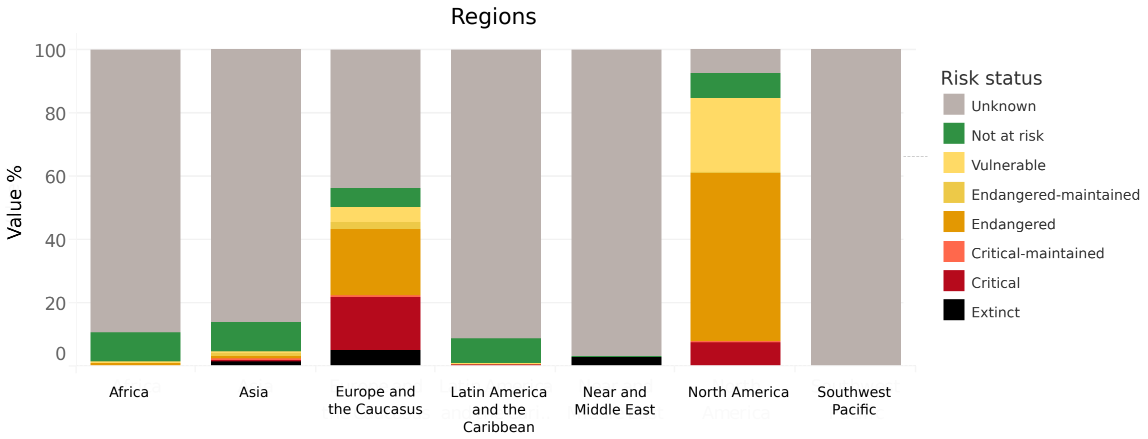
We here sampled for each breed an equal number of individuals, taking into account pedigree data, at the start (i.e. founder population) of the conservation programme and 10 generations after. To assess the effectiveness of maintaining genetic variation in the presence of a conservation programme, temporal genomic erosion was analyzed by quantifying delta indices related to genetic diversity ($\Delta\pi$), inbreeding (ΔF), and deleterious variation (ΔL), which were ultimately used as reference to provide recommendations for future management practices.

Materials and Methods

Samples and sequencing

Blood of individuals sequenced in this study was collected by the Institut National de la Recherche Agronomique (INRA). In 2003, sampling of 10 local chicken breeds, 30 animals each, was performed to study the factors conditioning the sustainability of valorisation programmes of heritage breeds. In 2013, sampling of 22 breeds, 60 animals each, was performed to assess the overall diversity of local chicken breeds under a conservation programme. The Barbezieux and Gasconne breeds were included in each study and chosen for the present

Figure 1: Risk status of local chicken breeds by regions. Risk status per region. The risk status categories are the ones of the Domestic Animal Diversity Information System (DAD-IS). Data were obtained from the Food and Agriculture Organization (FAO) website (last access: 14/07/2020)



temporal analysis because their conservation programme was just starting in 2003. It has been implemented by the Béchanne Breeding Center, under the supervision of the breeder's association set up for each breed, with the methodological support of SYSAAF. Whole-genome sequencing (WGS) data were generated for two local French chicken breeds, the Barbezieux and Gasconne (details are given in Table S1). The origin of the two breeds dates back to the 19^{th} century in south-west of France in the city of Barbezieux-Saint-Hilaire for the Barbezieux and Masseube for the Gasconne (Figure 2a). For the time series analysis, 15 founder individuals were included from the set sampled in 2003, while 14 individuals were included from the set sampled in 2013, i.e. 10 generations after. Semen of a Barbezieux male sampled in 2015 for the *ex situ* gene bank was subsequently added to the dataset for a total of 59 individuals. Birds were chosen based on their pedigree inbreeding coefficient (F_{ped}) to minimize relatedness in the dataset. Sequencing was carried out on a NovaSeq 6000 sequencing machine using standard library preparation protocols. Sequencing statistics are given for each individual in table S2. Whole pedigree data, individual performance for body weight, and reproduction were provided by the SYSAAF under a data transfer agreement established with the breeders' association for each breed.

Read processing and alignment

All analyses were based on an alignment of sequence data from all samples to the chicken GRCg6a genome reference sequence (GenBank assembly accession: GCA_000002315.5). The alignment and variant calling (Figure S1) pipelines were developed under the IMAGE (Innovative Management of Animal GEenetic resources) project and are publicly available (see Data availability). In the first bioinformatic workflow step, sequence data were mapped to the chicken reference genome with the BWA-mem v0.7.17 algorithm [183] using default options. Local realignment around indels and base quality recalibration were subsequently carried out in GATK v3.7 [202] to improve variant concordance and to correct for sequencing errors. SNPs and InDels calling was performed independently for each sample and by each caller (i.e. Mpileup [185], Freebayes [115], and GATK GenotypeGVCFs [202]) retaining only variants with a mapping quality > 30 and base quality > 10 (Figure S1A). In the second workflow step, GATK variants were filtered using the Variant Quality Score Recalibration (VQSR), which takes as positive training set the set of variants called by the three callers and as negative training set the unfiltered variants uniquely called by one of the three callers (Figure S1B). Additional filtering was performed on the final VCF file, retaining only genotypes whose coverage was between 4x and 2.5 the individual mean genome-wide coverage.

Principal component analysis

A principal component analysis (PCA) of genetic variation was carried out in SNPRelate [327] for R v3.2.0 to detect any existing structure within and between the two populations. The first PCA was performed on all samples considering as input only bi-allelic SNPs with a missing rate $< 10\%$ (n = 15,191,755 SNPs). We did not perform any linkage disequilibrium (LD) pruning to avoid excluding sites corresponding to fixed differences between the two populations. Population differentiation was further analyzed by estimating the fixation index (F_{st}) in consecutive non-overlapping 50-kb windows considering only windows with at least 300 variants (Figure S3). In addition to the all-samples PCA, we performed a population-specific PCA, in which bi-allelic SNPs were also pruned for an LD threshold of 0.5. After LD pruning, 108,403 and 84,930 SNPs remained for the Gasconne and Barbezieux, respectively.

Linkage disequilibrium

Genome-wide patterns of linkage disequilibrium (LD) were analyzed for each population and time point separately. Sites with missing data were discarded, as well as sites with a minor allele frequency (MAF) < 0.10 and > 0.80 and an exact test p-value for Hardy-Weinberg equilibrium of 1.0^{-07}. Following this initial filtering step performed in VCFTools v0.1.13 [65], we used PLINK v1.9 [244] to thin each chromosome to include only a random 20% subset of all sites (Xue et al., 2015). This step was performed to reduce the number of pairwise r^2 calculation and avoid an uneven contribution of macro-chromosomes over intermediate and micro-chromosomes in the calculation of LD. After filtering and thinning approximately 1.5 million and 1.6 million SNPs remained per individual for the Barbezieux and Gasconne, respectively. LD was calculated as the mean r^2 correlation coefficient between pairs of SNPs separated by at most 2,000 SNPs and within a 1 Mb SNP interval. Each chromosome was analyzed separately and results were then combined to obtain an overall LD decay profile for each breed and time point. LD decay was plot as average r^2 in consecutive 10-kb windows.

Genome-wide heterozygosity

Heterozygosity was calculated for each individual separately as the corrected number of heterozygous genotypes in consecutive non-overlapping windows of 100-kb, following the approach of Bortoluzzi et al. (2020) based on Bosse et al. (2012). Heterozygosity was calculated for the entire autosomal genome (InDels excluded), though only windows where at least 80% of sites that met the coverage criteria (i.e. 4x $\leq$ coverage $\leq$ 2*genome-wide coverage) were considered for the individual genome-wide heterozygosity estimation.

Within-individual runs of homozygosity

Runs of homozygosity (ROHs), here defined as genomic regions showing lower heterozygosity than expected based on the genome-wide average heterozygosity, were identified using the approach of Bortoluzzi et al. (2020) based on Bosse et al. (2012). To identify ROHs, we used the autosomal heterozygosity calculated in consecutive non-overlapping 10-kb windows along the genome of each individual. We considered 10 consecutive windows at a time in which the average window heterozygosity was calculated for only those windows with at least 80% of sites meeting the coverage criteria. In the first filtering step we retained only the 10 consecutive windows with a level of heterozygosity below 0.25 the average genomic diversity. In the second step we tried to avoid local assembly or alignment errors as much as possible by relaxing the threshold within the candidate homozygous stretches allowing for maximum twice the average heterozygosity in a bin, only if the heterozygosity within the candidate ROH did not exceed 1/5 the average genomic diversity [25, 31]. For each ROH, we calculated its size (i.e. the number of 10-kb windows that make up the ROH) and length (i.e. the total length of the ROH including windows that did not meet the coverage criteria). Even though, ideally, these two measures are the same, we excluded from further analyses 187 ROHs with insufficient coverage (size:length ratio $<$ 2/3). The total retained 4,658 ROHs were further classified into short ($\leq$ 100-kb), medium (0.1-3 Mb), and long ($\geq$ 3 Mb).

Runs of homozygosity validation

We validated our set of ROHs by comparing them with the within homozygous-by-descent (HBD) segments identified by the refinedIBD program [36] (see below) and with ROHs identified in VCFTools v0.1.13 [65]. The refinedIBD program in Beagle v5.0 was particularly good at detecting short ROHs (Person's r = 0.5642; *p-value* = 1.617^{-05}) (Figure S4a), whereas VCFTools mostly identified medium size ROHs (Pearson's r = 0.8648; *p-value* = $<2.2^{-16}$) (Figure S4b). The issue with detecting medium and long ROHs in Beagle is mostly due to the fact that the algorithm for detecting HBD is quite sensitive to clusters of heterozygous genotype calls [37]. In fact, when the number of heterozygous genotypes is larger than 3, the method inserts a non-HBD gap, therefore splitting a potential long segment into smaller chunks. Since clustering of many heterozygous genotypes is a typical issue of next-generation sequencing data, it is likely that the length distribution of HBD segments was subsequently affected. A different methodological constraint may have affected the ROHs identification when using VCFTools, which relies on detecting long ROHs with an autozygosity probability of about 0.99 [12]. VCFTools incorporates assumptions about recombination rate based on a human reference, which is very different from, in our case, the chicken recombination rate. Because of these methodological constraints, we decided to continue our analyses using the set of ROHs identified by the method of Bosse et al. (2012) [31].

Pedigree and genomic measure of inbreeding

We used pedigree information provided by the Syndicat des Selectionneurs Avicoles et Aquacoles Francais (SYSAAF) following the approval of the breeders' associations to determine the inbreeding coefficient (F_{ped}) of our 59 samples. We also calculated the proportion of the genome within ROH (F_{ROH}) following [204] as ratio between the total length of ROHs within an individual (L_{ROH}) and the actual length of the genome (L_{auto}) covered in our dataset (n = 960,268,821 nucleotides). Sex chromosomes and mitochondrial genome were excluded in the calculation of L_{auto}.

Between-individual sequence identity

To identify genetic sequences shared between individuals (identity-by-descent segments or IBD), we first resolved the phase of the distinct haplotypes within each sample using Beagle v5.0 [36]. Phasing was performed on the all-samples dataset of filtered variants using 10 burnin iterations, 12 phasing iterations, a window length of 20 cM, a window overlap of 2.0 cM, and an effective population size of 100000. Phasing was performed on each chromosome separately, providing each time a genetic map with information on variants positions in cM units using the linkage map of Elferink et al. (2010) [92]. Identity-by-descent (IBD) segments between individuals and homozygous-by-descent (HBD) segments within each individual were detected using the refinedIBD program [35]. Parameters used were: a 20 cM window length, a minimum length of 1.0 cM to report an IBD segment, and a LOD score of 3.0.

Polarization and annotation of variants

In order to reduce the reference bias, alleles were polarized as ancestral or derived with respect to the ancestral chicken sequence reconstructed from the 4-sauropsids multiple whole-genome alignment (Ensembl release 95) [126]. The species included in the alignment were: chicken, turkey, zebra finch (*Taeniopygia guttata*; taeGut3.2.4) [307], and green anole lizard (*Anolis carolinensis*; AnoCar2.0) [6]. We retained only SNPs for which either the reference or alternative allele matched the ancestral allele, while ancestral alleles that did not match either chicken allele were discarded.

We assigned consequences to each of the polarized variant using the Ensembl Variant Effect Predictor (VEP) [203] (release 95). Although most studies often consider protein-coding variants only, we here retained protein-coding and non-protein-coding variants, after applying a combination of filtering steps to improve the reliability of the prediction [25, 85]. Filtering criteria included: (1) bi-allelic variants with a call rate $>$ 70%; (2) genes 1:1 orthologues between chicken and zebra finch (release 97) to reduce the effect of off-site mapping of sequence reads; (3) variants outside repetitive elements as these genomic regions are often difficult to sequence and are thus prone to errors; and (4) intronic and intergenic variants outside intronic EST. Protein-coding variants were also discarded if were found outside coding sequences (CDS).

Functional classes

Protein-coding variants retained after filtering were classified, following the VEP annotation, into synonymous, missense tolerated (SIFT $>$ 0.05), missense deleterious (SIFT $\leq$ 0.05), and loss of function (LoF) (i.e. splice donor, splice acceptor, start lost, stop gained, and stop loss). GERP conservation scores [69] computed from a multiple whole-genome alignment of 34 sauropsids available in Ensembl (release 97) were used to validate the potential impact of deleterious mutations. This is because mutations at sites that remain highly constrained during evolution are likely to be deleterious and are thus excellent predictors of fitness effects. Therefore, of the initial set of putative missense deleterious and LoF mutations, only those with a GERP score $>$ 1.0 were considered to be truly deleterious. We further assigned the chCADD score to all filtered variants independently of their coding potential following Groß, Bortoluzzi et al. (2020).

Genetic load

Estimating an individual's genetic load based on genomic data is challenging. We therefore expressed the genetic load using three different approaches. We initially measured genetic load in each individual as the ratio of homozygous derived damaging (i.e. missense deleterious, LoF) mutations over synonymous mutations. We also expressed genetic load using evolutionary constraints as predictor for the fitness consequences of a deleterious mutation. This second measure of genetic load, here called GERP load, was calculated for each individual considering only damaging mutations with a GERP score > 1.0 [227]. We finally estimated the genetic load using the chCADD score assigned to protein-coding and non-protein-coding mutations that belong to functional classes with an average chCADD score > 10, as:

$$chCADD\_load = \frac{\sum ichCADD_i}{NHomozygous} \tag{4.1}$$

where $chCADD_i$ is the chCADD score of an homozygous mutation at genomic position i and $NHomozygous$ is the total number of homozygous mutations identified in each individual's genome.

Mutation burden

The mutation burden was estimated based on counts of derived damaging alleles [247] following three models: 1) homozygous-mutation burden, 2) heterozygous-mutation burden, and 3) total mutation burden (i.e 2*homozygous burden + heterozygous burden).

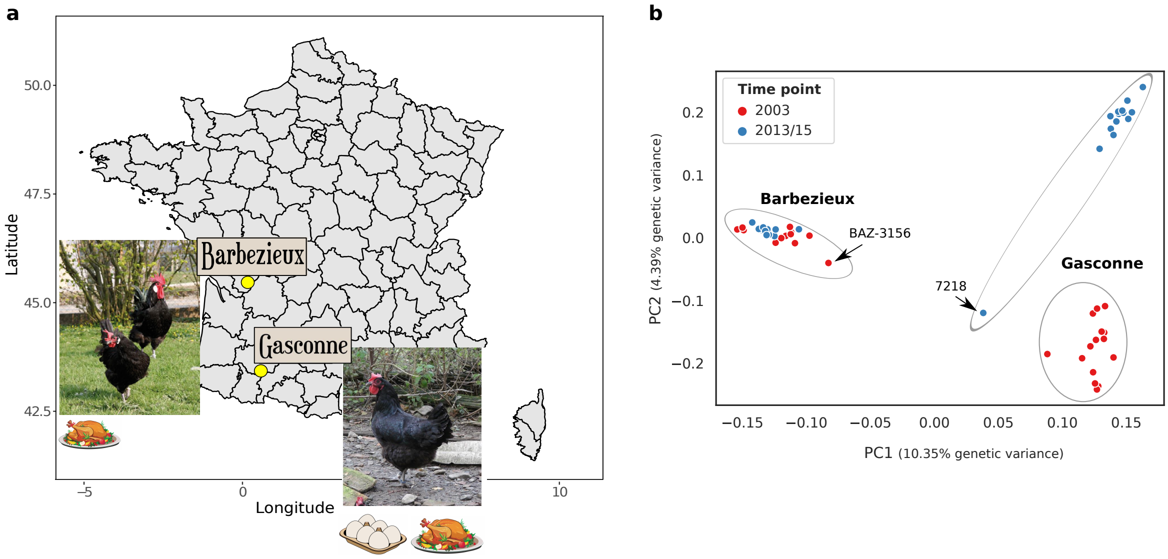
Signatures of selection

Genomic regions under selection were identified using the new generic Hidden Markov Model (HMM) likelihood calculator developed by Paris et al. (2019). This HMM model approximate the Wright-Fisher model implementing a Beta with spikes approximation, which combines discrete fixation probabilities with a continuous Beta distribution [230]. The advantage of this model over existing ones is its applicability to time series genomic data. Prior to detect regions under selection, we estimated the effective population size (N_e) in each breed separately using the NB R package [151], whose underlying model is an HMM with Beta transitions similar to the framework of Paris et al. (2019). To estimate N_e we removed SNPs with an allele frequency <20% or >80% as recommended in Paris et al. (2019). The estimated N_e was 153.46 (CI: 145.84-161.73) in the Barbezieux and 50.05 (CI: 50.00-50.08) in the Gasconne. We then applied the HMM model after which we removed SNPs with a false discovery rate (FDR) threshold of 5% estimated using the q-value R package [276].

Results

We generated whole-genome sequencing data from 29 Barbezieux individuals and 30 Gasconne individuals collected between 2003 and 2015 (Figure 2a, Table S1). All genomes were aligned, genotyped, and annotated with respect to the chicken reference genome (GRCg6a), yielding a per-individual mean genome-wide depth >10x and mapping quality >30 (Table S2). Following variant calling and additional post-filtering steps, we identified 2 million InDels and 19 million SNPs uniformly distributed along the genome (Table S3). Because of the limited number of SNPs on chromosomes 30 to 33, we decided to focus our analyses on the first 28 autosomes.

Figure 2: Samples and population structure. a. Geographic origin of the Barbezieux and Gasconne breed, with relative breeding objective (meat or meat/egg). **b.** Principal component analysis (PCA) performed using 15,191,755 bi-allelic SNPs after filtering for a missing rate of 10%. Individuals from each breed are colored with respect to their sampling year



Temporal changes in genetic diversity ($\Delta\pi$) and inbreeding ($\Delta\mathbf{F}_{ROH}$, $\Delta\mathbf{F}_{ped}$)

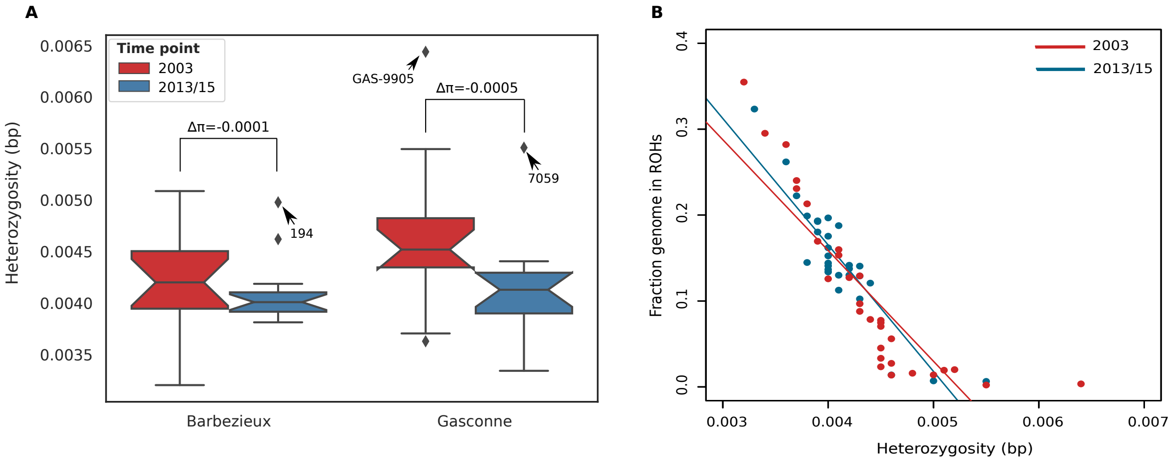
The separation between the Barbezieux and Gasconne samples in the principal component analysis (PCA) confirms them as genetically distinct populations (weighted Fst: 0.107± 0.05) (Figure 2b). At the same time, in the all-samples PCA and population-specific PCA (Figure S4), we observed a separation of Gasconne individuals into two subgroups following sampling time (weighted Fst: 0.060± 0.05), which suggests either structure within the breed or a distinct genetic makeup of individuals sampled in 2013 as compared to the founders of 2003. The result was also confirmed by the Neighbour-Joining (NJ) analysis on the identity-by-state distance relationship matrix (Figure S5). By contrast, very little separation was observed for the Barbezieux breed (weighted Fst: 0.015± 0.03) in both PCA analyses and NJ tree.

To assess how demographic history has shaped genome-wide diversity, we computed the per-site heterozygosity in consecutive 100 kb non-overlapping windows along each individual's genome (Figure S6). We found that, compared to the founding nucleus of the Barbezieux (π: 4.17x10^{-3}) and Gasconne (π: 4.61x10^{-3}), genetic diversity decreased by 2.5% (π: 4.08x10^{-3}) and 10.5% (π: 4.12x10^{-3}), respectively, over the 10 years (Figure 3a). Despite the faster decrease in heterozygosity, the Gasconne still exhibited higher within-breed diversity than the Barbezieux. The reduction in genetic diversity observed in recent samples is the result of a fragmented heterozygosity distribution, where regions of high heterozygosity are interspersed by regions enriched for homozygous genotypes (Figure S6). These regions, otherwise defined as runs of homozygosity (ROH), arise when identical haplotypes are inherited from a recent common ancestor [43]. Therefore, their number and length reflect an individual demographic history and level of inbreeding.

Although mean genome-wide heterozygosity (negatively) correlates with the total fraction of the genome covered by ROHs (Pearson's *r*: -0.90, *p-value*: 3.081^{-11}) (Figure 3b), the correlation does not capture the abundance and size distribution of ROHs (Figure 4a) [257, 80]. Of all ROH size classes, long ROHs ($\geq$ 3 Mb) are of major concern as they result from recent close inbreeding. Barbezieux individuals sampled in 2013/15 contain 1 to 20 long ROHs (0.6-13% of the genome), whereas Gasconne individuals at the same time point contain 1 to 26 long ROHs that cover up to 29% of the genome (Table S4). Although the total number of (short, medium, and long) ROHs increased over the 10 generations in the two breeds (Figure S7), we observed more and longer ROHs in the Gasconne than in the Barbezieux, resulting in a larger fraction of the genome in homozygous state in the former breed (Figure 4a; Figure S8).

We subsequently calculated for each breed the genomic inbreeding coefficient (F$_{ROH}$) using the information on the within-individual ROH (Figure 4b). In line with the heterozygosity and ROH analysis, we observed an increase over time in the genomic inbreeding in the two breeds, with values of delta index 15 times larger in the Gasconne (ΔF_{ROH}: 0.0776) than Barbezieux (ΔF_{ROH}: 0.0051) (Figure 4b). We further calculated the pedigree inbreeding coefficient (F_{ped}) (Table S7) and estimated the accuracy of F_{ped} in capturing individuals' relationships. Although individuals from the founding nucleus are assumed to be unrelated, values of F_{ped} were much more homogeneous in the Barbeziex (F_{ped}: 7%) than in the Gasconne (F_{ped}: 3-15%), confirming the trend in F_{ROH} (Figure 4b; Figure S9). We further correlated F_{ROH} and F_{ped} to verify the usefulness in a conservation programme of the pedigree information. As shown in Figure 4c, we report a significant positive correlation (Pearson's r: 0.425; *p-value*: 4.92^{-03}). Pedigree information provided by the SYSAAF were also used to quantify changes in the number of sires and dames over the 10 generations (Figure S10). We found that the conservation programme was able to increase the number of breeding males and females per generation in both breeds and particularly in the Barbezieux.

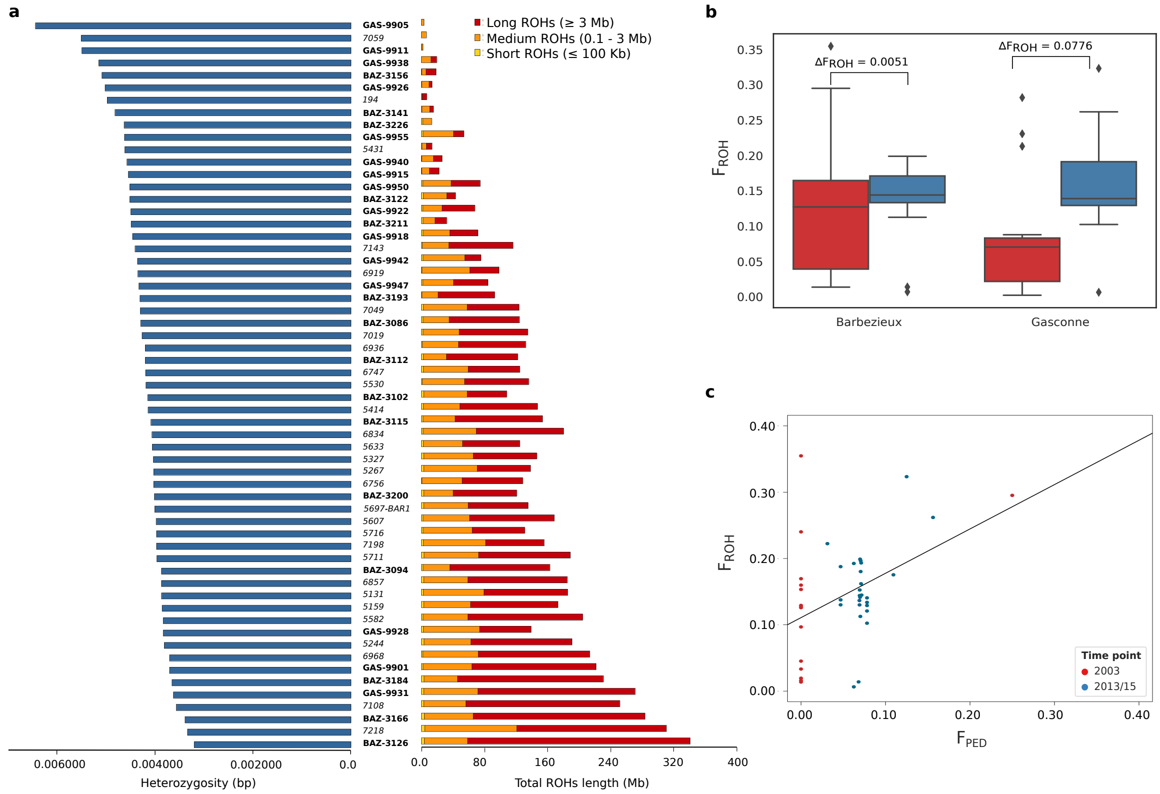
Figure 3: Temporal changes in heterozygosity ($\Delta\pi$). a. Heterozygosity is the mean autosomal heterozygosity calculated for each individual and time point along the genome in consecutive 100 kb non-overlapping windows. **b.** Correlation between individual heterozygosity (bp) and fraction of the genome covered by runs of homozygosity (ROHs)



For each breed we also examined haplotypes shared between individuals (identity-by-descent - IBD) as they also provide information on the levels of recent inbreeding [80, 160]. We found clear differences in the fraction of IBD segments consistent with the ROH and F_{ROH} analysis: individuals from the Gasconne breed (Figure S11c-d), and particularly those sampled in 2013, displayed a 5% higher mean level of sequence sharing than those from the Barbezieux (Figure S11a-b). IBD segments confirm that inbreeding is a major constraint in the management of the Gasconne breed.

To better understand the demographic history of the two breeds, we looked at the genome-wide

Figure 4: Runs of homozygosity (ROHs) and temporal changes in inbreeding level (ΔF). a. Average (autosomal) genome-wide heterozygosity per individual (left) and total length of short ($\leq$ 100 kb), medium (0.1-3 Mb), and long ($\geq$ 3 Mb) ROH per individual (right). Individuals sampled in 2003 are in bold. Samples are ordered by decreasing heterozygosity from top to bottom. **b.** Genomic inbreeding coefficient estimated for each individual as ratio between the total length of ROHs within an individual and the actual length of the genome covered in our dataset. **c.** Correlation between genomic inbreeding coefficient estimated from ROHs (F_{ROH}) and inbreeding coefficient estimated from the pedigree (F_{ped})



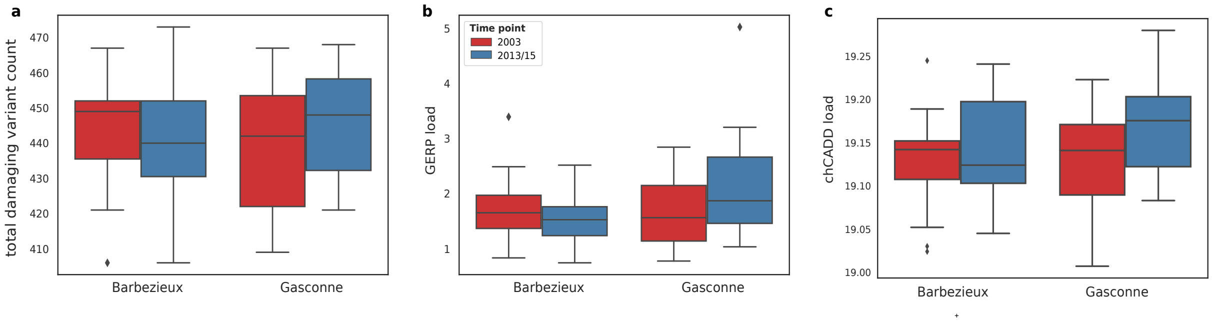
linkage disequilibrium (LD) and found a similar LD decay (Figure S12). In both breeds, the LD decay is consistent with a demographic history marked by a recent population bottleneck [259], as shown by the characteristic L-shape, with the left portion of the curve steeply declining down to an R^2 value of 0.2, after which it remains flat up to a SNP interval of 1 Mb.

Temporal changes in deleterious variation: ΔL

We have shown that since the start of the conservation programme genetic diversity declined ($\Delta\pi$) at the costs of an increase in realized (ΔF_{ROH}) and expected (ΔF_{ped}) inbreeding, resulting from an accumulation of longer ROHs ($\geq$ 3 Mb). To verify whether $\Delta\pi$ and ΔF are associated with changes in deleterious variation, we decided to annotate protein-coding variants with respect to their predicted impact on the encoded amino-acid into synonymous (n = 60,963), nonsynonymous (n = 24,193) and loss-of-function (LoF) (n = 337) (Table S6). We also assigned to all filtered variants the ch(icken)CADD score, a measure of variant deleteriousness that can effectively prioritize variants based on a comprehensive set of functional and evolutionary properties [251, 126].

The derived allele frequency (DAF) spectrum showed that the two breeds have more fixed, high frequency (DAF $>$ 0.90) benign (i.e. synonymous, tolerated) mutations than (putative) damaging (i.e deleterious, LoF) ones at the same frequency (Figure S13). As expected, only 5% of (putative) damaging SNPs were found at very low frequency (DAF $<$ 0.10), as most of these mutations have been purged by purifying selection. To determine how common purging is in the breeds, we looked at three measures of genomic fitness (L) (Figure 5, Figure S15). Individuals sampled in 2013 displayed a net accumulation of homozygous damaging mutations, although this trend was much more pronounced in the Gasconne for which the ΔL index was almost 19 times larger (ΔL: 0.0014) than that of the Barbezieux (ΔL: 0.0001) (Figure S15b). Genomic fitness approximated using the GERP score led to similar results, with the Gasconne exhibiting a 43% increase in GERP load compared to the Barbezieux (Figure 5b). Similar conclusions were also obtained when looking at the chCADD load (Figure 5c).

Figure 5: Genetic load and mutation burden. a. The mutation burden is the total number of damaging alleles existing in an individual at a specific time point. **b.** Genetic load approximated using the GERP score information of each homozygous damaging mutation (GERP $>$ 1.0). **c.** Genetic load approximated from all variants independently of their coding potential using the chCADD score



Signatures of selection and phenotype data

To test whether the two breeds were interested by artificial selection since the start of the conservation programme, we decided to identify genomic regions under selection using the new generic Hidden Markov Model likelihood calculator developed by Paris et al. (2019) [230]. After filtering SNPs for an FDR threshold of 5%, no significant SNPs were retained, meaning that both breeds have not been interested by clear selection objectives. This result was further supported by the unchanged allele frequency distribution in both breeds between 2003 and 2013 (Figure S14).

We subsequently analysed one productive and 6 reproductive traits collected and provided by the SYSAAF. Overall, we did not identify any clear trend for the trait *body weight at 8 weeks* in both Barbezieux (table S8) and Gasconne (table S9). Of the reproductive traits, we found that in the Barbezieux *% of fertile eggs* and *% of hatched eggs* increased in both sire and dam line, leading to a positive selection coefficient in the individuals sampled in 2013 (table S10). The situation in the Gasconne was quite difficult to analyse since reproductive data were only availble for the year 2013 (table S11), making any trend estimate impossible. Hence, trait selection in the Gasconne breed was much less documented than that of the Barbezieux on the whole period considered in this study.

Discussion

In this era of rapid decline in biological diversity, conservation programmes have become critical for preserving the genetic diversity harboured by individual genomes [169]. The importance of a conservation programme on a species genome has extensively been addressed in endangered wild species, such as in the Isle Royale wolf [257], Iberian lynx [169], mountain gorilla [318], and Eastern gorilla [292]. However, for local livestock breeds, the impact and relevance of a conservation programme on an individual's genome has rarely been addressed, as conservation programmes are often not in place. In this study, we used temporal genomic data as a tool to gather critical information on the demographic and genetic processes accompanying a conservation programme to inform management and aid decision-making to keep endangered local breeds from the brink of extinction.

Ideally, a conservation programme should target all breeds at risk of extinction. However, the costs required to conserve all at-risk breeds are often greater than the resources available for conservation. Therefore, determining the conservation value of a breed is a first key step in all conservation programmes of livestock species. For instance, breeds that are genetically distinct should have higher priority, as well as breeds with exceptional economic productivity or unique traits. However, as we here show, conservation programmes can also be established as a mean to preserve a breed genetic diversity through the production of local/niche market products. Once the target breed has been identified, the primary objective is to maintain the highest levels of genetic diversity, while controlling the increase in inbreeding. By doing so, populations will be able to respond to future changes in selection pressures [71].

Conservation programmes often rely on closed practices to maintain un-admixed populations.

For the chicken breeds analysed in this study, this practice implies that introgression of genetic material from other breeds is discouraged to maintain the genetic diversity of the breed unique. However, in closed conservation programmes, inbreeding is an important population parameter that can rapidly increase along with the probability of exposing deleterious alleles in the homozygous state. The increase in inbreeding is also expected as conservation of small populations is often established from a small number of founders (Figure S10). However, as we show, inbreeding can still be controlled, supporting previous successful examples, including the black-footed ferret, California condor, and Przewalski's horse [138]. The control over inbreeding (or coancestry) can be achieved by assuring an optimal contribution of each individual to the next generation in terms of number of offspring. To maximise the effective population size (N_e) [208] and minimise the accumulation and expression of (recessive) deleterious mutations, information on individual relationships is required [71]. Although coancestries can be estimated from pedigree data, high-throughput sequencing data provide more reliable estimates and additional information on the functional relevance of variants [30].

The conservation programme here analyzed is a unique case in Europe for two main reasons. First, management is centrally coordinated by a professional union (SYSAAF). Second, pedigree kinship information is used to select individuals for breeding following an optimal contribution approach. The effects of management can clearly be observed in the F$_{ped}$ of the Barbezieux, which, after an initial steep increase, it stabilizes around 0.07-0.08 (Figure S9). Although inbreeding increased over time, the rate strongly depends on the management strategies. For instance, the strong reduction in heterozygosity observed in the Gasconne (Figure 3a), which mirrors the increase in genomic (F$_{ROH}$) and pedigree inbreeding (F$_{ped}$) (Figure 4), suggests that conservation measures are an issue in this breed. However, pedigree-based estimates tell us a different story. In fact, based on the F$_{ped}$ and pedigree information, it became clear that the population sampled in 2013 does not come from the founder population initially sampled in 2003, but from another, smaller, population sampled in 2009 (Figure S9). Therefore, the change in management explains the genetic distance among individuals observed in the PCA (Figure 2b). These results illustrate two important aspects. First, that in local livestock breeds conservation practices are not always consistent over time, implying that this inconsistency may represent a threat to animal genetic diversity. And second, that pedigree information should always be recorded in any conservation programme as it can provide additional information on management. We however argue that the optimal contribution approach currently applied to the conservation of the Barbezieux and Gasconne breed would considerably benefit from marker-based or molecular coancestry estimates, as F$_{ped}$ estimates expected, and not realized, inbreeding, underestimating the actual proportion of the genome that is IBD [160]. Individual inbreeding estimated from markers information is nowadays possible thanks to the availability of a multitude of SNP arrays at (relative) modest costs. However, such implementation is expected to strongly depend on the resource investments of the stakeholders involved in the conservation programme.

Our findings demonstrate that quantifying temporal erosion is essential for monitoring the evolutionary potential of an endangered species and, in more practical terms, aiding decision-making for conservation efforts. It is, however, important to differentiate the underlying processes leading to genomic erosion, such as inbreeding or genetic drift, from their proximate drivers, which should be directly targeted by management actions [181]. Deleterious mutations have been shown to have important consequences on an individual's survival and genetic potential. However, conservation programmes established without the support of genetic information may indirectly keep deleterious mutations, reducing the mean fitness of the population [70]. According to our estimates of genomic fitness, optimal contribution has been effective at exposing and purging deleterious alleles in the Barbezieux (Figure 5), while maintaining diversity and fitness (Table S8). However, managing contributions seems to have been less effective in the Gasconne (Figure 5), likely due to the limited number of founding individuals in 2009. It is also possible that the genetic diversity and genetic load of the founding individuals were already deviating from the breed average [70], challenging the efficacy of natural selection. However, comparisons with the original population should be carried out in order to validate this hypothesis. Our findings highlight the importance of using molecular information in conservation programmes to maintain heterozygosity, while purging deleterious alleles. Moreover, by investigating the interaction between genetic load and demographic parameters (e.g. census size), molecular data can better inform on which management strategy should be used and how management should change over time to maintain diversity.

In the context of domestic animal diversity, *ex situ* conservation practices are recognized as an essential complementary activity to *in situ* conservation actions for the maintenance of a broader genetic base. In this study, the conservation programme of the Barbezieux and Gasconne relies on the maintenance of live animals (i.e. *in vivo*) rather than on cryoconservation (i.e. *in vitro*). Cryopreservation involves the collection and deep-freezing of semen, ova, embryos or tissues for potential use in breeding [254]. As gene bank collections are stored for an indefinite time, genetic diversity is free from demographic and genetic forces, such as selection and genetic drift [97]. The storage of biological material in gene banks has for long time benefited transboundary breeds, which are currently well represented in many national gene banks, as breeding companies are directly involved in the sampling and financing of the gene bank collections. The interest for cryopreservation has, however, improved over the years also for local livestock breeds, and specifically for poultry, thanks to efforts to enhance the use and exploitation of genetic collections (e.g. IMAGE project). Gene bank collections serve as a pool to support breeding in current populations, including the recover and/or introgression of specific variation to support breeds at risk of extinction, or control breed design in the case of reorientation of the breeding goal [97].

Although a gene bank should in most cases be regarded as a complement to *in situ* and *ex situ in vivo* conservation programmes, stakeholders directly involved in conservation efforts should also take advantage of existing national gene banks to progressively store genetic material for use in the future. This is particularly relevant for local breeds as their small size poses conservation programmes at higher risks of failure if not properly managed.

However, similarly to defining the conservation value of a breed, basic problems for gene bank management should be addressed, including which and how many individuals to store in the collection, what genetic and reproductive material to collect and store, and how stored information is consistently recorded [72]. It is only by addressing these issues that gene banks and conservationists can fully support *in vivo* and *ex situ in vivo* management practices.

Concluding remarks

Conservation programmes will impact individual genomes in terms of diversity and fitness, depending, among others, on the initial genetic diversity, inbreeding load, population size, and management practices. Our study reinforces the imperative to establish and sustain existing conservation programmes that aim to keep local at-risk breeds from the brink of extinction. Moreover, by exploiting the potential of temporal genomic data, our study shows that the incorporation of this information into management practices is now possible and should be of high priority for effective conservation efforts. Although our study focuses on endangered local livestock breeds, our findings and conclusions provide a genomic resource for future conservation and research in domesticated and wild species.

Acknowledgments

The authors would like to thank Cyriel Paris and Simon Boitard for their support on the use of the Hidden Markov Model for the selection signature analysis. We would also like to thank the technicians and scientists from the Centre de Sélection de Béchanne and SYSAAF for the collection and sharing of data. A special thanks to the breeders for their cooperation during data collection. The research leading to some of these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353. Additional funding for C.B.

to join the Génétique Animale et Biologie Intégrative (GABI) group at INRAE was provided by the Wageningen Institute of Animal Sciences (WIAS) as PhD fellowship under project number 4169000100.

Authors contributions

M.T.B conceived the study and organized the data collection with the breeders and SYSAAF. C.B. designed the study, carried out the genomic analyses, and wrote the manuscript. R.R., B.D., and F.P. collected and provided the pedigree and phenotypic data. G.R., M.B., and M.T.B jointly supervised the study and contributed to the writing of the manuscript. All authors revised and approved the final version.

Competing interest

The authors declare that they have no competing interests.

Data and materials availability

The scripts required to run the bioinformatics workflow presented here are available at https://forgemia.inra.fr/bios4biol/workflows/tree/master/Snakemake/. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Supplementary Material

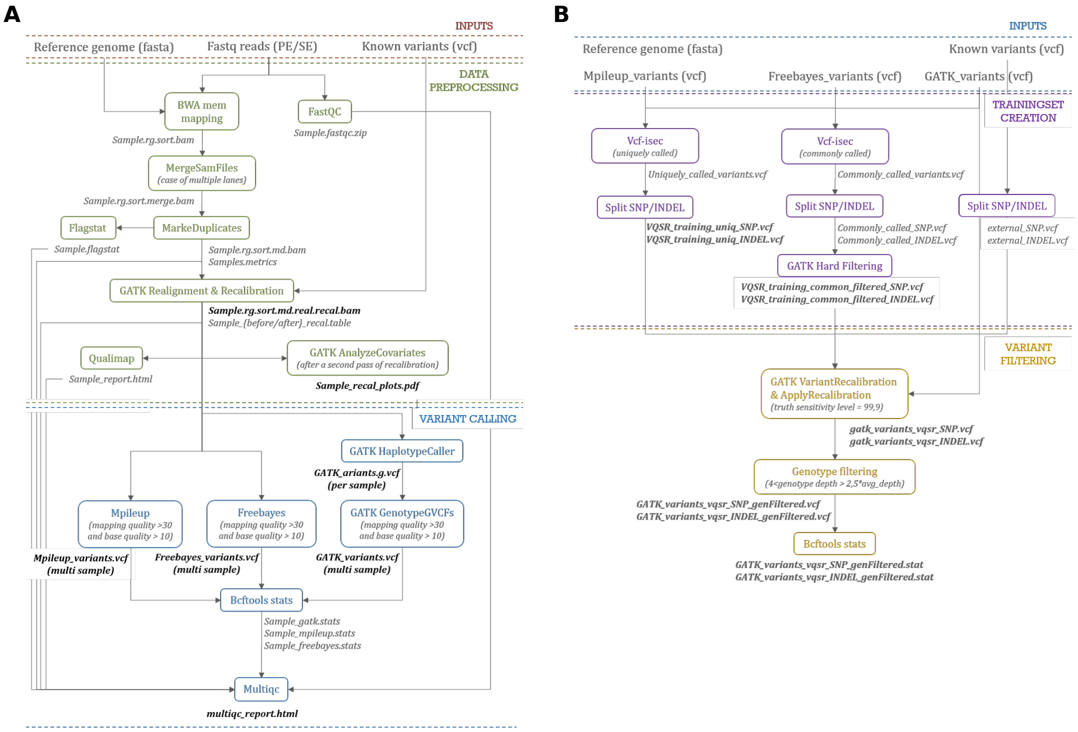


Figure S1: Alignment and variant calling pipeline. A. in the alignment step samples are treated independently and analyzed using the three callers. They are finally merged into 3 files, one per caller. **B.** In the variant calling step, GATK variants (SNPs/InDels) are filtered using information on good quality and poor quality training sets. A final filter on genotypes coverage is applied. The final output of the alignment and calling pipeline is shown in bold

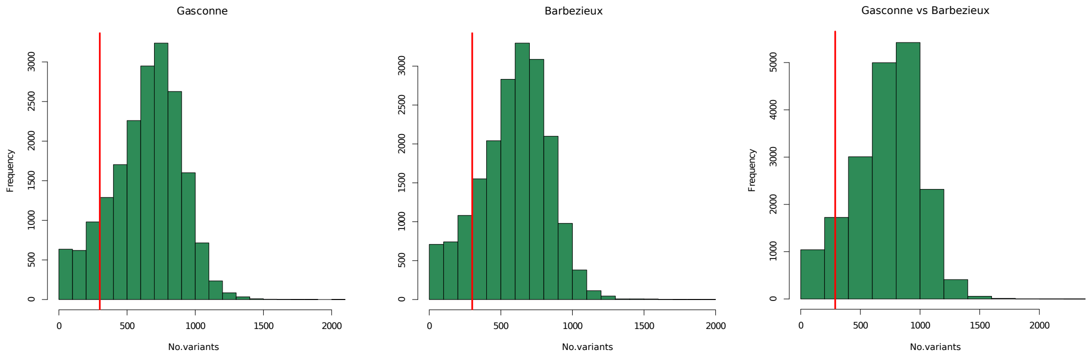


Figure S2: Frequency distribution of variants. The x-axis shows the number of variants found in non-overlapping 50-kb windows that were used for the F_{st} calculation. The vertical red line represents the cut-off (n = 300 variants) for considering windows in the mean genome-wide F_{st} calculation

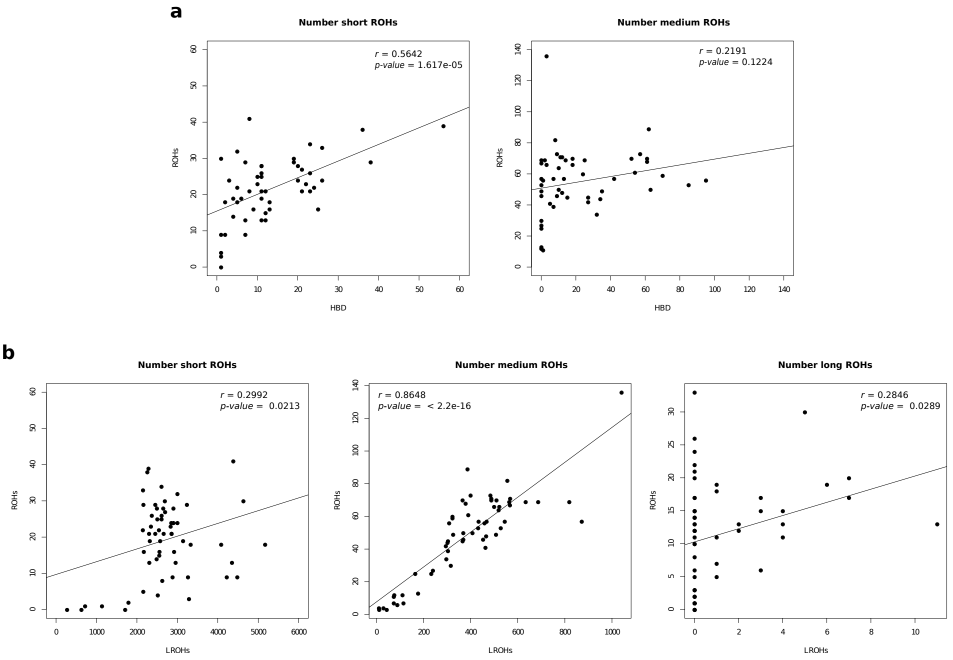


Figure S3: Runs of homozygosity (ROHs) validation. **a.** Validation based on the HBD tracts identified using the refinedIBD program in Beagle v5.0. Results are shown for only short and medium size ROH as these size classes were the only ones identified by Beagle. **b.** Validation based on VCFTools for all ROH size classes. r is the Pearson's correlation test

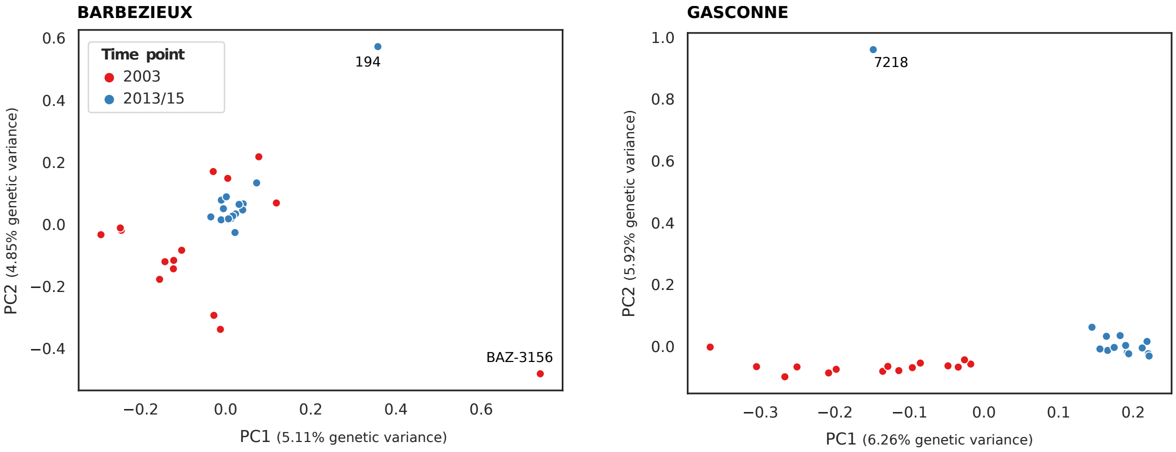


Figure S4: Breed-specific principal component analysis (PCA). The principal component analysis was performed on the two breeds separately after filtering for an LD threshold of 0.5. After LD pruning, a total of 84,930 and 108,403 SNPs were retained for the Barbezieux and Gasconne breed, respectively

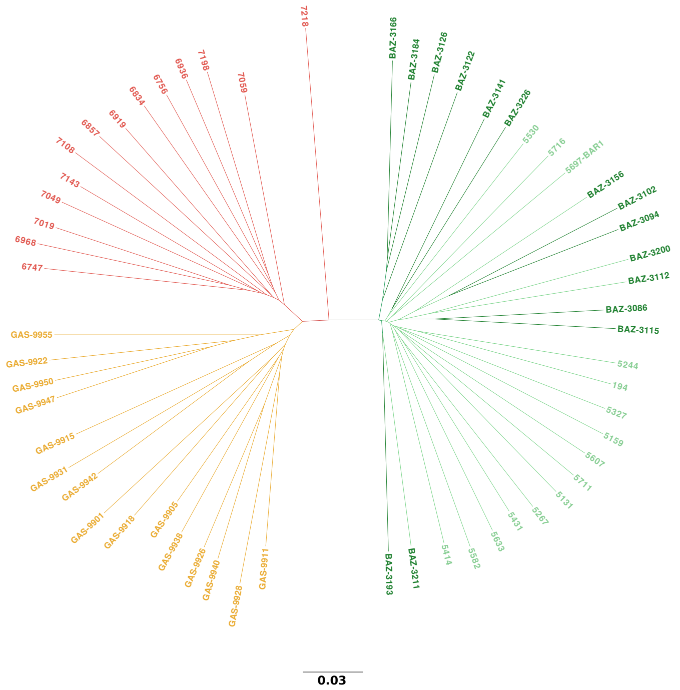
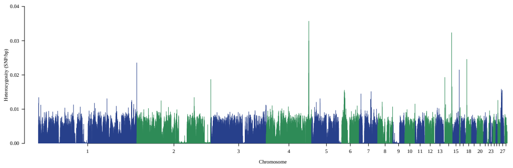
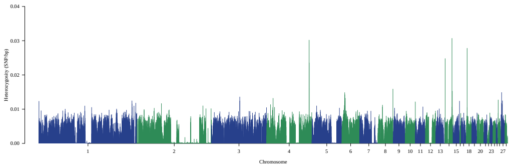
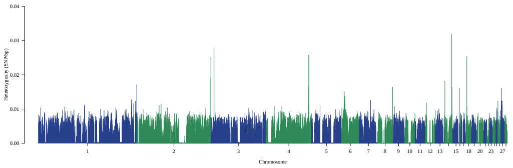
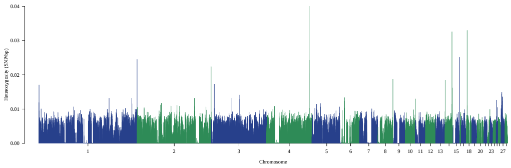


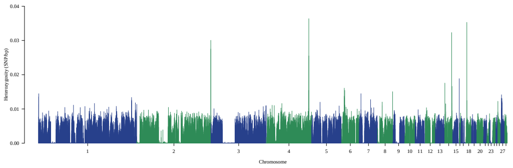
Figure S5: Neighbour-Joining (NJ) tree. The NJ tree was constructed from the identity-by-state (IBS) distance relationship matrix estimated on 12,561,354 SNPs after filtering for a missing rate < 10% and a minor allele frequency (MAF) < 0.05. Orange: Gasconne 2003; Red: Gasconne 2013; Dark green: Barbezieux 2003; Light green: Barbezieux 2013

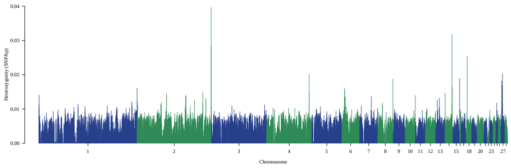

BAZ−3086

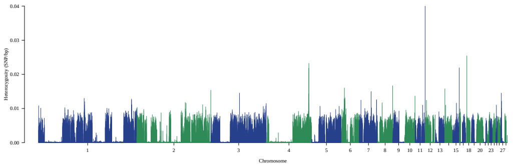

BAZ−3094

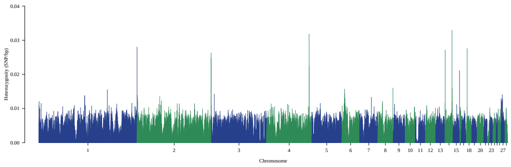

BAZ−3102

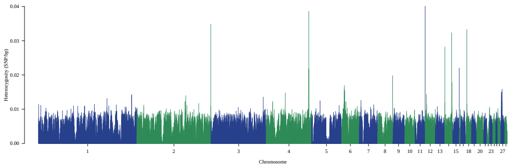

BAZ−3112

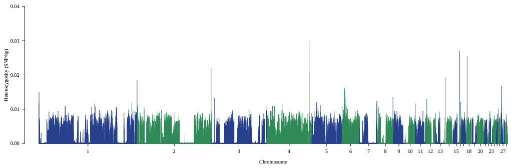

BAZ−3115

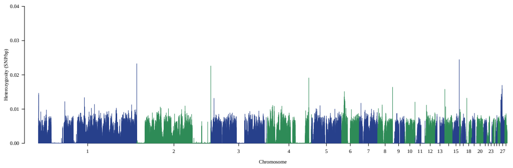

BAZ−3122

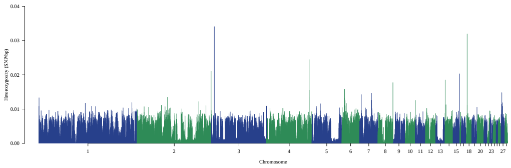

BAZ−3126

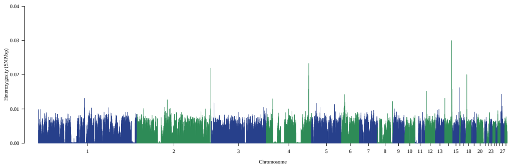

BAZ−3141

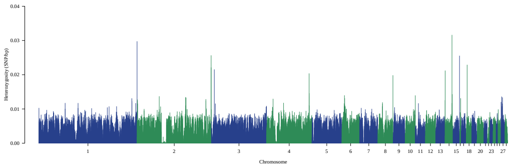

BAZ−3156

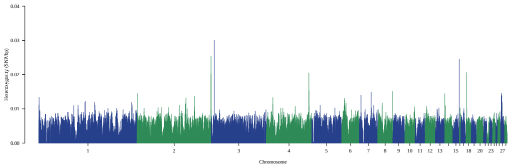

BAZ−3166

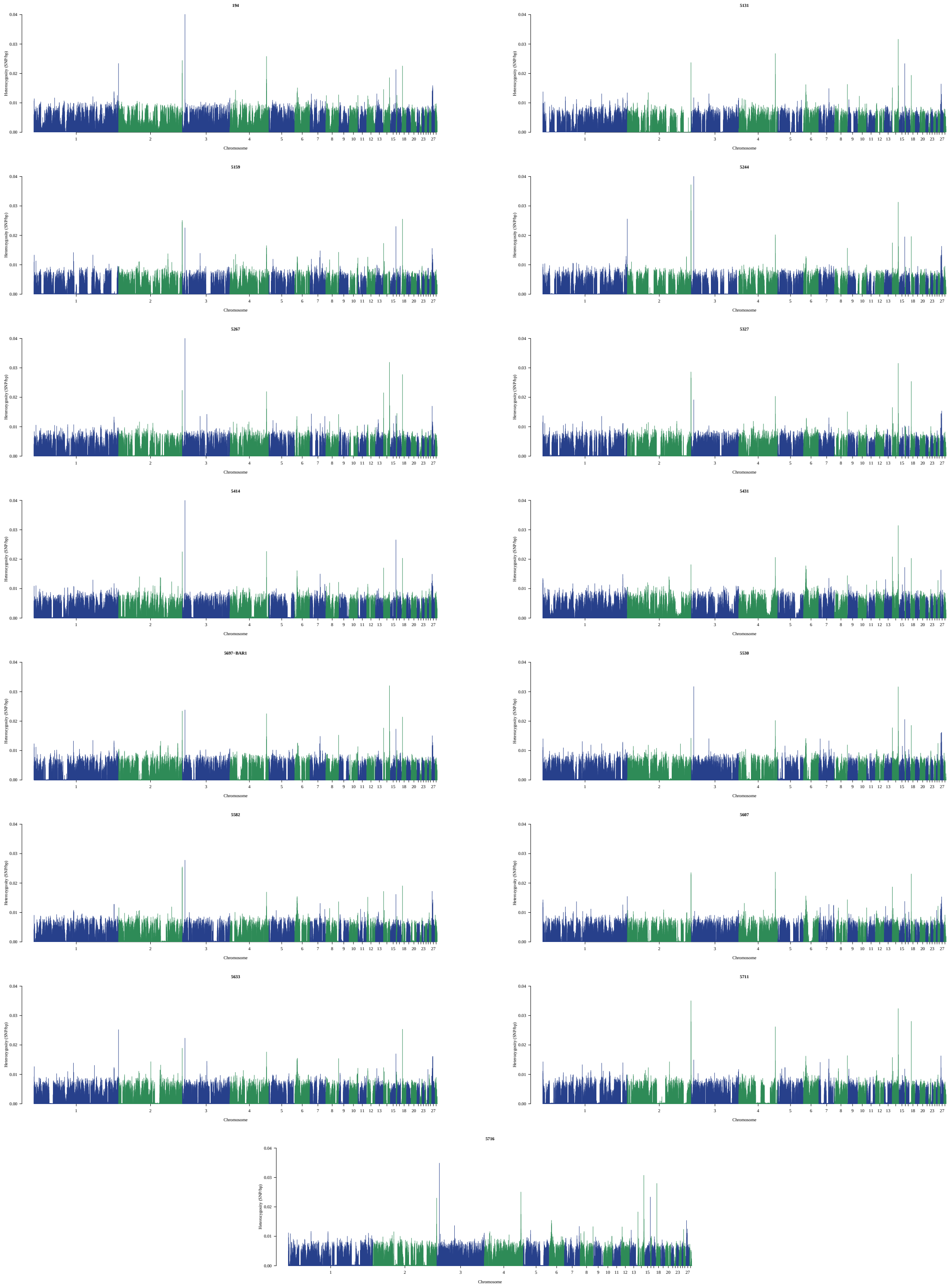

BAZ−3184


BAZ−3193


BAZ−3200


BAZ−3211


BAZ−3226



194

5131

5159

5244

5267

5327

5414

5431

5697−BAR1

5530

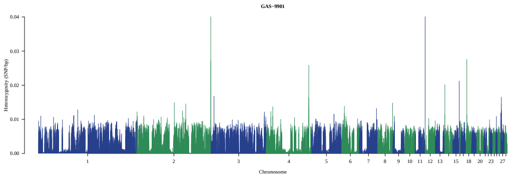
5582

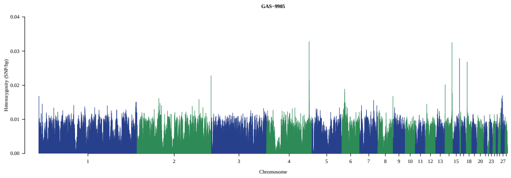
5607

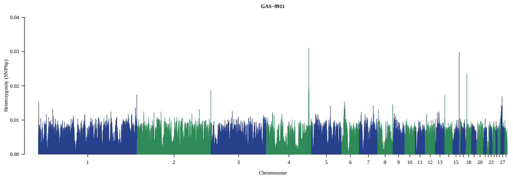
5633

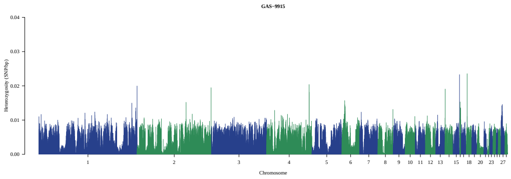
5711

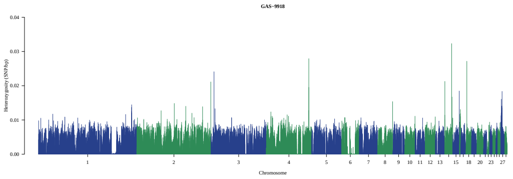
5716

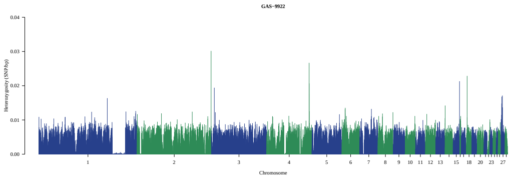

GAS−9901
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

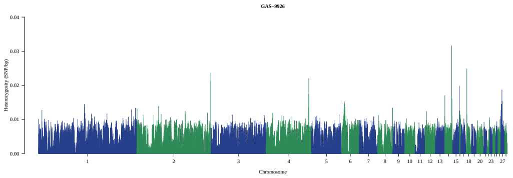

GAS−9905
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

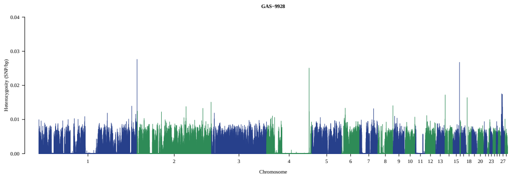

GAS−9911
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

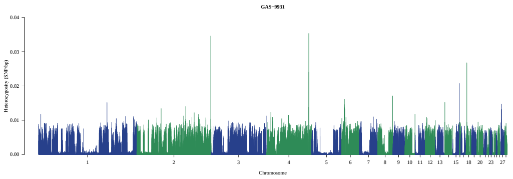

GAS−9915
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

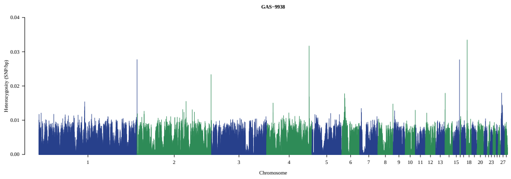

GAS−9918
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

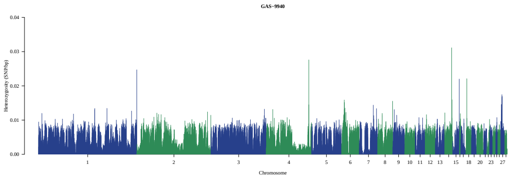

GAS−9922
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

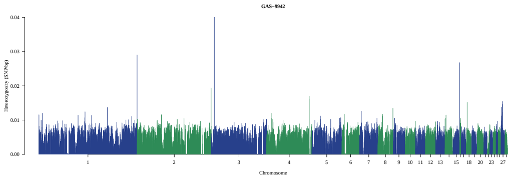

GAS−9926
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

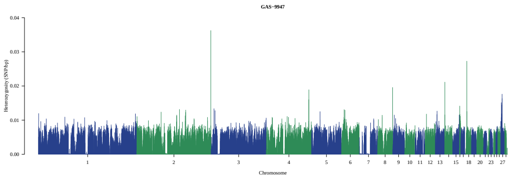

GAS−9928
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

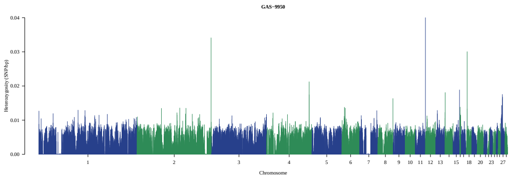

GAS−9931
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

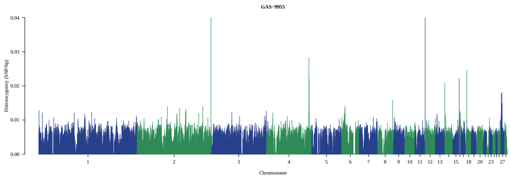

GAS−9938
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome


GAS−9940
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome


GAS−9942
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome


GAS−9947
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome


GAS−9950
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome


GAS−9955
Heterozygosity (SNP/bp)
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 23 27
Chromosome

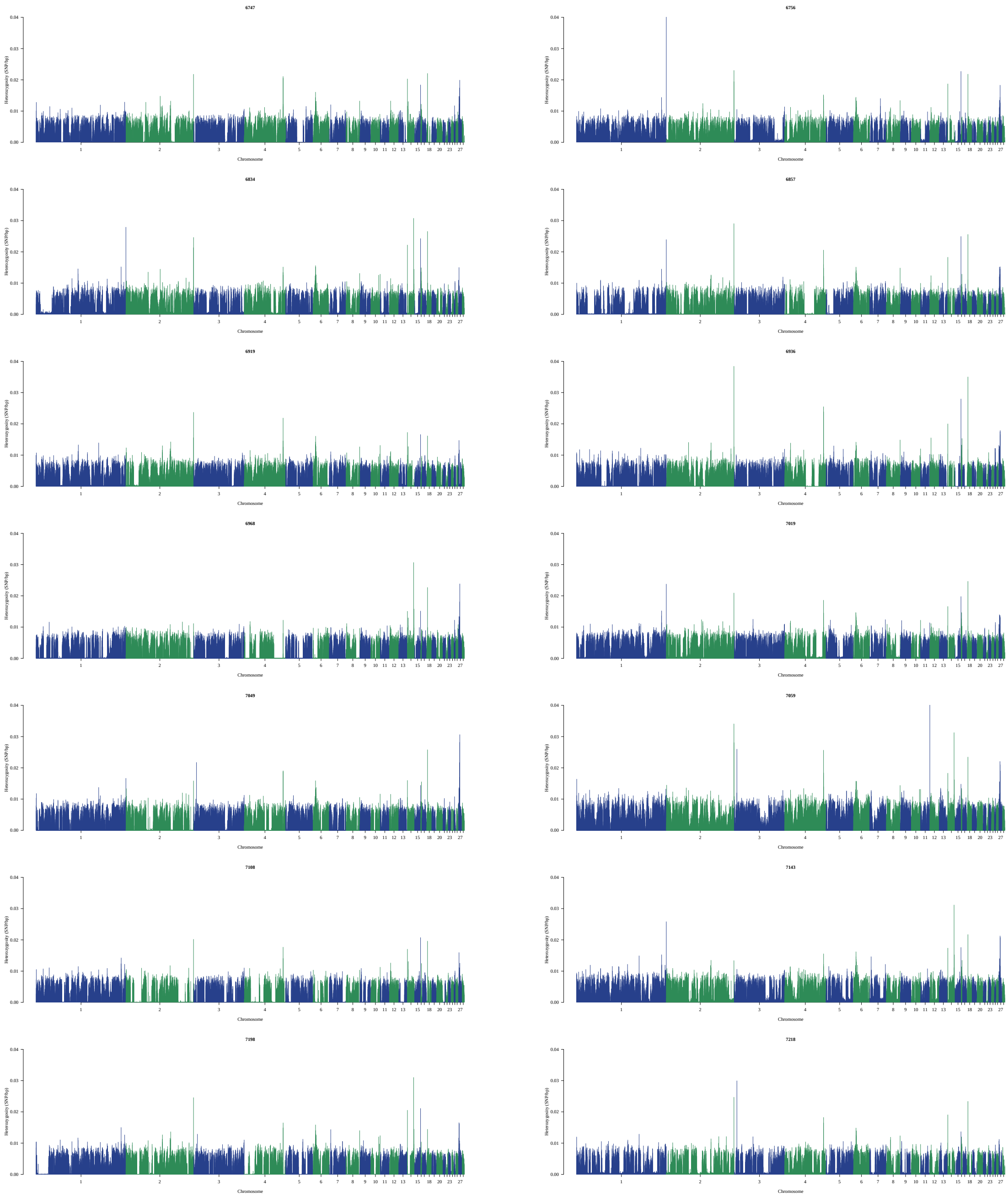


Figure S6: Genome-wide heterozygosity distribution. The x-axis shows the chromosomes (1-28), while the y-axis the corrected number of heterozygous SNPs in each consecutive 100-kb non-overlapping window

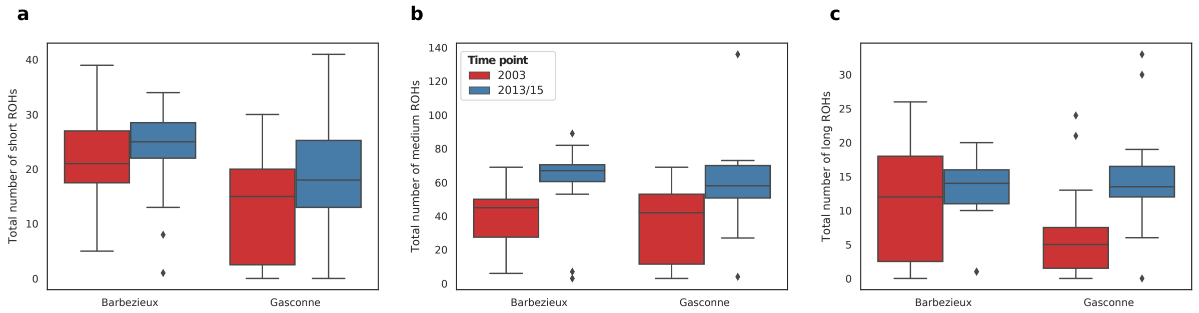


Figure S7: Number of ROH per individual for different ROH size classes. The number of ROH in various size classes is indicative of the individual's demographic history. **a.** Short ROHs (≤ 100 kb) indicate ancient inbreeding. **b.** Medium ROHs (0.1-3 Mb) indicate ancient and historic inbreeding. **c.** Long ROHs (≥ 3 Mb) indicate recent inbreeding

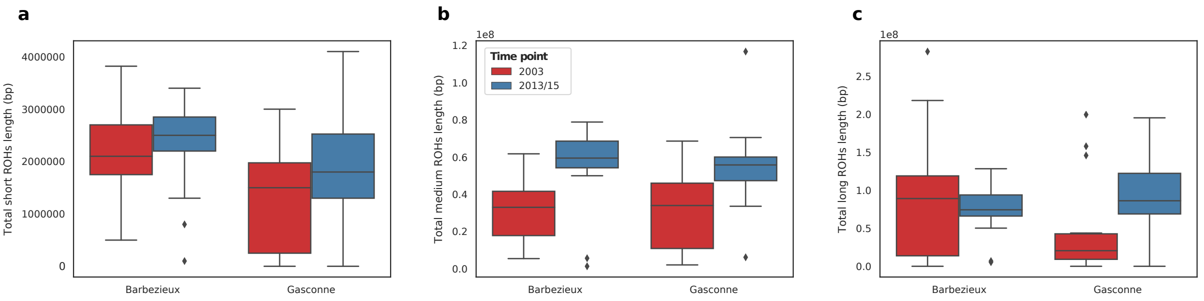


Figure S8: Summed ROH length per individual for different ROH size classes. Individuals for each breed are colored based on their time of sampling

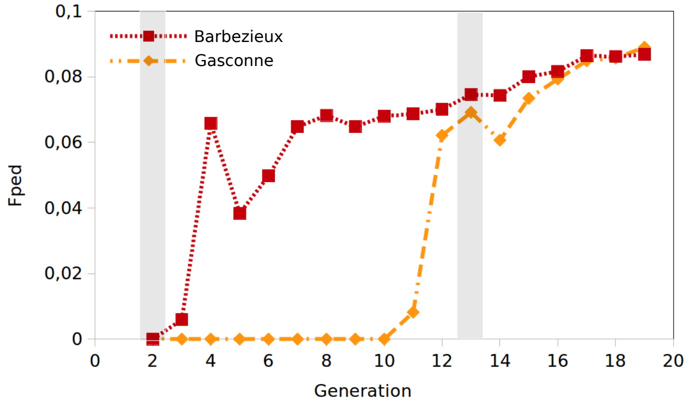


Figure S9: Trend in pedigree-based inbreeding coefficient (F$_{PED}$). Generation 3 (year 2003) and 13 (year 2013) are highlighted in grey

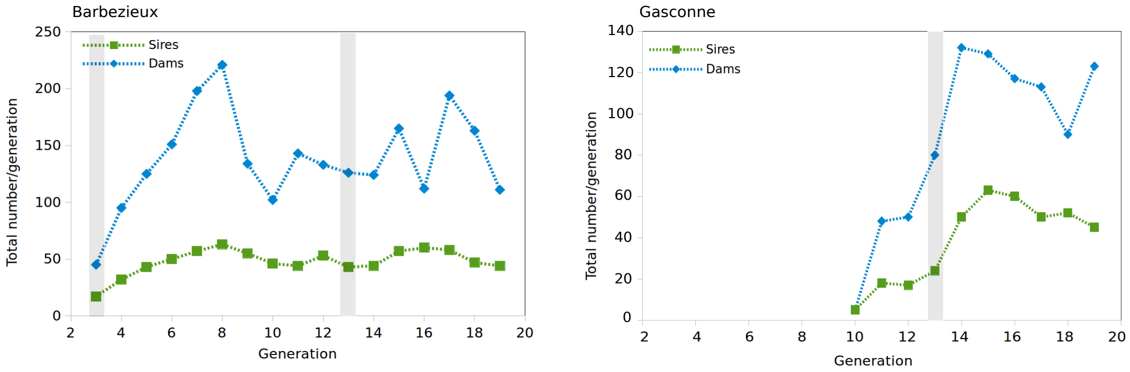


Figure S10: Trend in population size . Generation 3 (year 2003) and 13 (year 2013) are highlighted in grey

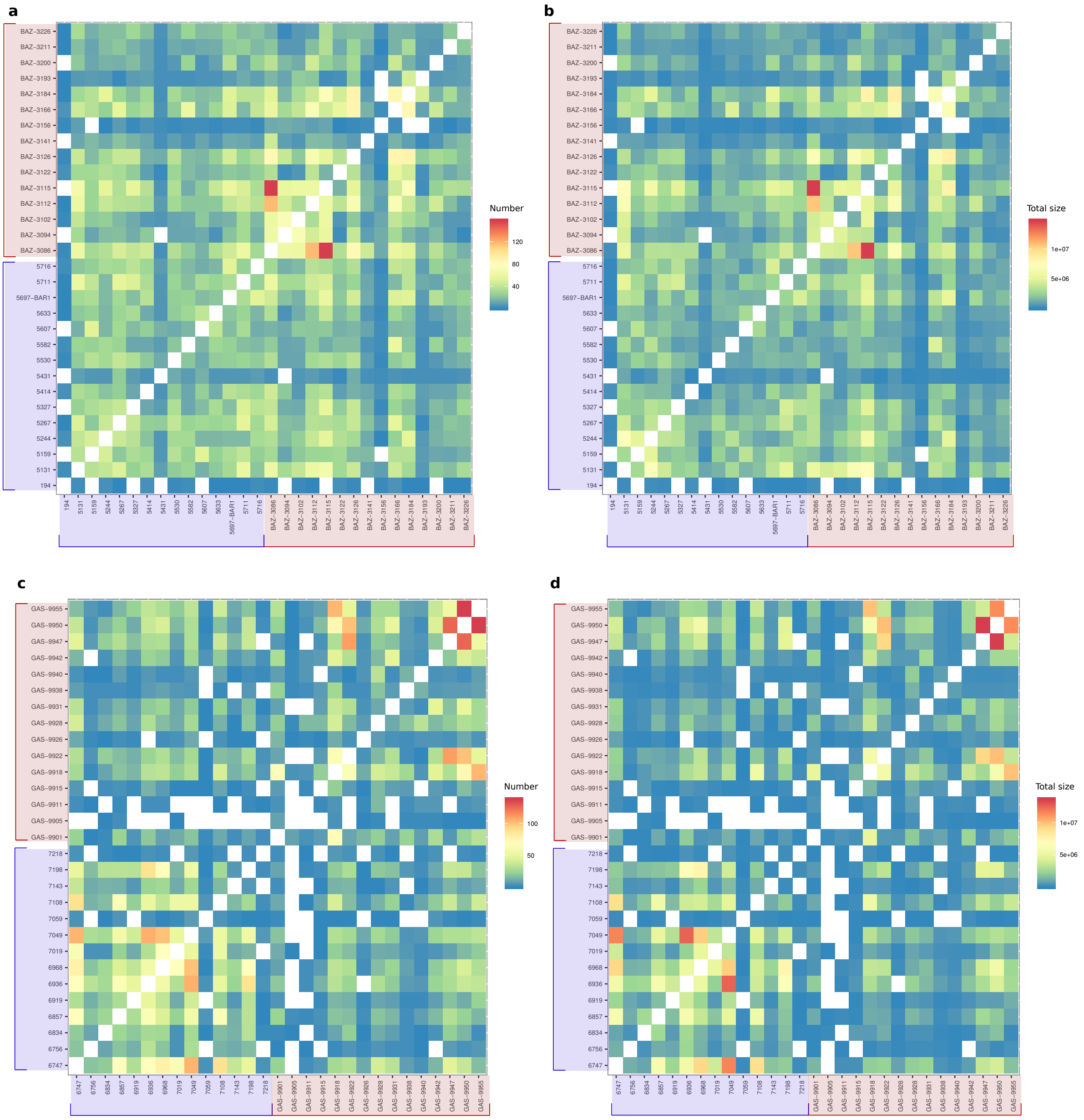


Figure S11: Number of identity-by-descent (IBD) segments and their total size. a.-c. Pairwise number of IBD segments between individuals in the Barbezieux and Gasconne, respectively. **b.-d.** Total size of pairwise IBD segments in the two populations, respectively. Individuals sampled in 2003 are highlighted in red, whereas those sampled in 2013/15 are in blue

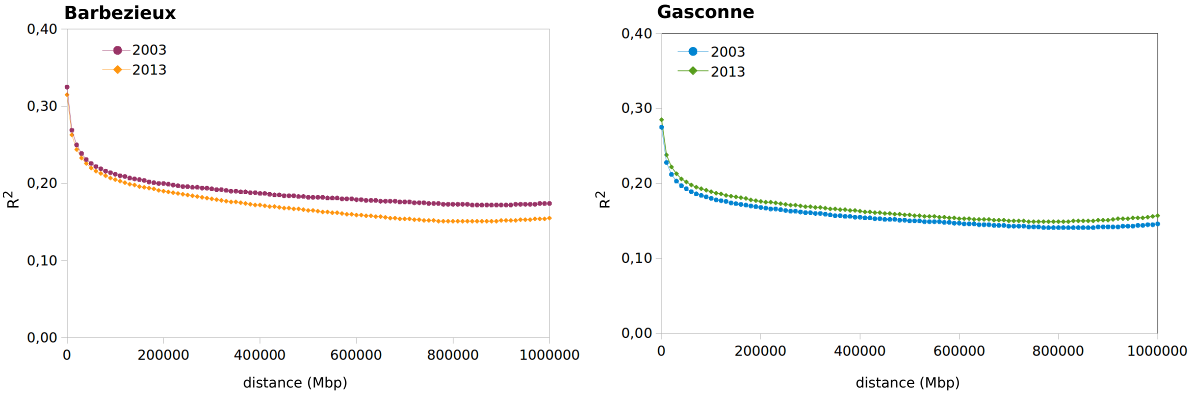


Figure S12: Linkage disequilibrium (LD) decay. LD was calculated as the mean r^2 correlation coefficient between pairs of SNPs separated by at most 2,000 SNPs and within a 1 Mb SNP interval. LD decay, here plotted for each breed and time point, is the average r^2 in consecutive 10-kb windows

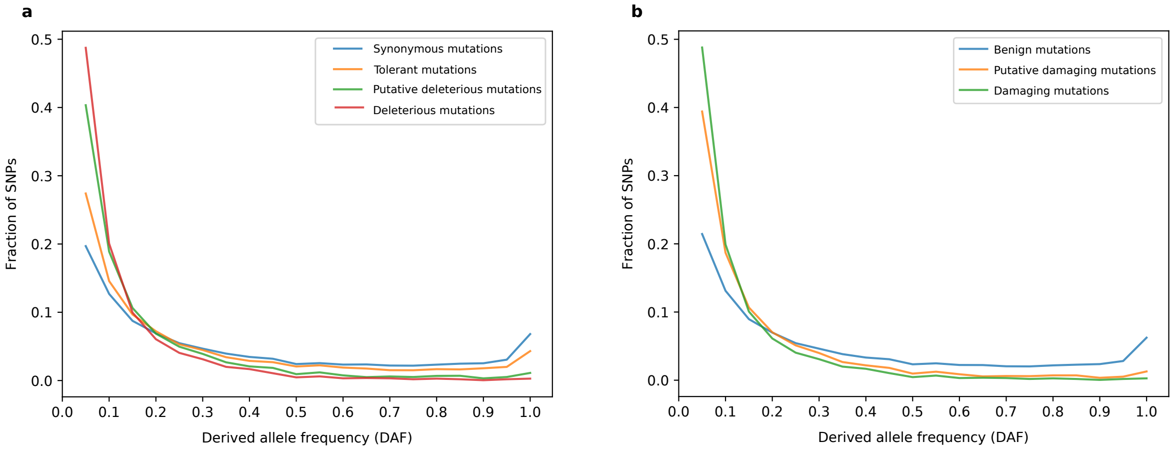


Figure S13: Derived allele frequency distribution. **a.** Derived allele frequency distribution of protein-coding variants, classified into synonymous, nonsynonymous tolerated (SIFT $>$ 0.05), nonsynonymous putative deleterious (SIFT $\leq$ 0.05), and nonsynonymous deleterious (SIFT $\leq$ 0.05; GERP $>$ 1.0). **b.** Derived allele frequency distribution of protein-coding variants classified into benign (i.e. synonymous, tolerated), putative damaging, and truly damaging

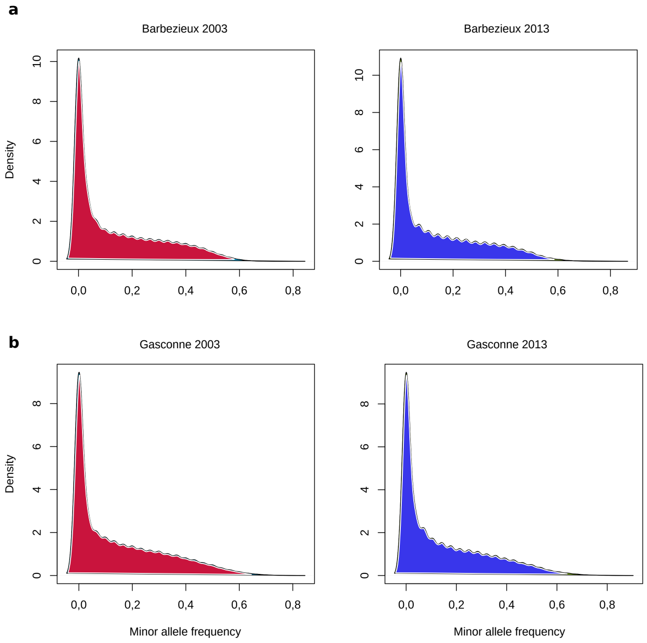


Figure S14: Temporal changes in allele frequency. **a.** Allele frequency distribution of the Barbezieux breed in 2003 (left) and 2013 (right). **b.** Allele frequency distribution of the Gasconne breed in 2003 (right) and 2013 (left)

5.
Parallel genetic origin of foot feathering in birds

The chapter is based on:

Bortoluzzi, C., Megens, H. J., Bosse, M., Derks, M. F., Dibbits, B., Laport, K., Weigend, S., Groenen, M. A., Crooijmans, R. P. (2020). *Molecular Biology and Evolution* (https://doi.org/10.1093/molbev/msaa092)

Abstract

Understanding the genetic basis of similar phenotypes shared between lineages is a long-lasting research interest. Even though animal evolution offers many examples of parallelism, for many phenotypes little is known about the underlying genes and mutations. We here use a combination of whole-genome sequencing, expression analyses, and comparative genomics to study the parallel genetic origin of ptilopody (*Pti*) in chicken. Ptilopody (or foot feathering) is a polygenic trait that can be observed in domesticated and wild avian species and is characterized by the partial or complete development of feathers on the ankle and feet. In domesticated birds, ptilopody is easily selected to fixation, though extensive variation in the type and level of feather development is often observed. By means of a genome-wide association analysis, we identified two genomic regions associated with ptilopody. At one of the loci, we identified a 17 kb deletion affecting *PITX1* expression, a gene known to encode a transcription regulator of hindlimb identity and development. Similarly to pigeon, at the second loci we observed ectopic expression of *TBX5*, a gene involved in forelimb identity and a key determinant of foot feather development. We also observed that the trait evolved only once as foot feathered birds share the same haplotype upstream *TBX5*. Our findings indicate that in chicken and pigeon ptilopody is determined by the same set of genes that affect similar molecular pathways. Our study confirms that ptilopody has evolved through parallel evolution in chicken and pigeon.

5.1 Introduction

Parallel evolution is the independent development of similar phenotypic traits in separate but related lineages [68]. A defining characteristic of parallel evolution is that the trait is absent in the common ancestor. Animal evolution offers many examples of parallelism. These range from seemingly simple phenotypic changes, such as similar female-limited Batesian mimicry in *Papilio* butterflies [154], to more complex ones, such as the reduction of the pelvic complex in threespine and ninespine sticklebacks [264]. Even though studies have shown that parallel evolution is not a rare event in animal evolution [217, 264], the genetic bases are still largely unknown for most traits. One hypothesis is that phenotypes that have evolved in parallel and that are extremely similar must have evolved by applying the same genetic mechanism due to developmental constraints. However, at the other extreme, it can be hypothesised that developmental pathways are so complex, and can be perturbed in so many ways, that essentially an infinite number of combinations of variations could lead to the same outcome.

In vertebrate evolutionary studies, the genomic basis of phenotypic traits that evolved in parallel has not yet been fully understood as appropriate model species are often lacking. Studying the genetic basis of parallelism is further challenged by the fact that the independent evolution of a trait may have occurred many million years ago with very often many genes involved. Domesticated species can provide interesting insights into the genomic architecture of traits that evolved in parallel and the selection that has acted on them [260]. Among vertebrates, domestic chicken (*Gallus gallus domesticu*s) is an excellent model, as many of the domesticated phenotypes are common to other domesticated avian species, such as pigeon and duck, and, in some cases, even to more distantly related species, including dog and cattle, presumably due to the desire of humans for particular traits. Examples of similar traits displayed by chicken and other species include the lack of neck feathers [15, 212], head crest [304, 265], feathering rate [91, 76], and short body stature [278, 21, 316]. Many of these phenotypes are determined by a single mutation affecting a single gene that have evolved in parallel in the two species through artificial selection.

In this study we focus on ptilopody (*Pti*), a trait observed in domesticated and wild avian species in which the epidermis of the ankle and foot are partially or completely covered with feathers [23, 83]. The genetic basis of ptilopody has been extensively studied. Previous classical breeding experiments identified a small number of genetic loci (Pti-1, Pti-2, pti-3) of large effect in chicken [271]. However, only recently several candidate genomic regions were identified. While Dorshorst et al. (2010) identified one quantitative trait locus (QTL) of major effect in Silkie chickens [84], Sun et al. (2015) mapped four QTLs, two of which explain more than 20% of the phenotypic variation [277]. Although recent studies have tried to associate ptilopody to a certain genomic region, it is still unclear which chromosome(s), gene(s), and mutation(s) are directly responsible for the phenotype. A better understanding of the genomic architecture underlying ptilopody comes from a recent study in domestic pigeon, in which two genomic regions containing genes responsible for a partial transformation from hindlimb to forelimb identity were implicated, mainly *PITX1* and *TBX5* [82]. Ptilopody in chicken and pigeon is extremely similar in appearance and this similarity is partly explained by the same genes involved [82]. Even though the same genes are involved, the question is whether a similar underlying mutation has enabled the trait to evolve in both lineages. And, even more intriguing, if indeed a similar mutation was to be involved, it becomes relevant to question how the same pathways are altered by the same regulatory mechanisms in both species. This question has never been addressed directly before, neither from a molecular nor an evolutionary perspective.

In this study, we use a combination of whole-genome sequencing, expression analyses, and comparative genomics to study the genetic basis of foot feathering in chicken. In particular, we identify the underlying causal mutations and affected molecular pathways, while investigating the parallel genetic origin of ptilopody.

5.2 Material and methods

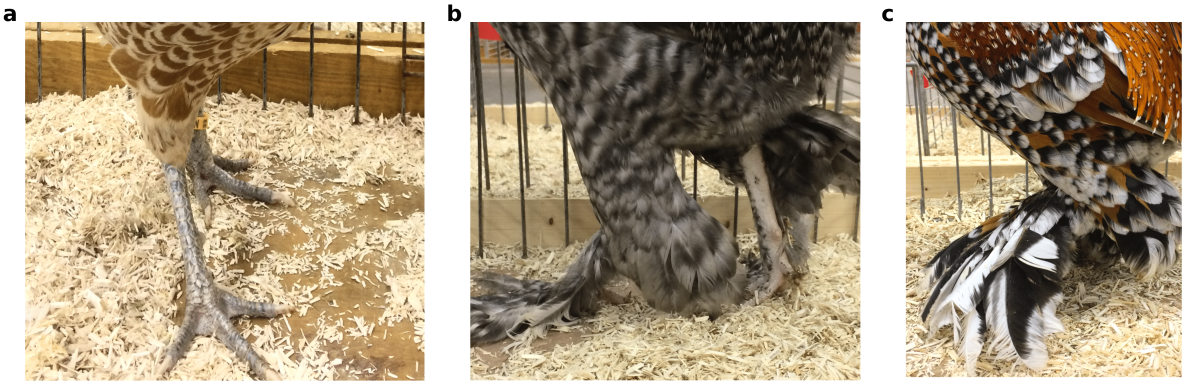
Blood collection and animal experiments

Collections of blood samples was done in accordance with the German Animal Protection Law and was approved by the Committee of Animal Welfare at the Institute of Farm Animal Genetics (Friedrich-Loeffler-Institut) and the Lower Saxony State Office for Consumer Protection and Food Safety (No. 33.9-42502-05-10A064). Sample collection and data recording were also conducted strictly according to the Dutch law on animal protection and welfare (Gezondheids- en welzijnswet voor dieren).

Samples and phenotype

DNA of 169 samples from 87 traditional chicken breeds and one individual from each of the four living wild species of Gallus (i.e. *G. gallus, G. sonneratii, G. lafayetii, G. varius*) was used for whole-genome sequencing on an Illumina HiSeq 3000 (Additional file 1, Table S1). Whole-genome sequencing (WGS) data of the 97 birds sampled in the Netherlands were previously deposited in the European Nucleotide Archive (ENA) under accession number PRJEB34245 (Bortoluzzi et al. 2019). WGS data of the remaining 68 birds sampled in Germany have been deposited in ENA under accession number PRJEB36674. Detailed information on the sequenced reads can be found in Additional file 1, Table S2. Information on foot feathering was collected during sampling and confirmed by the breeding associations. The phenotype was observed in 11 breeds, for a total of 19 samples. Of these, 9 showed short and tight feathers on the metatarsus and digits, while 10 samples had extensive feather development, as well as long flight-like feathers on the posterior toes (Figure 1; Additional file 2, Table S1). The remaining 150 samples, which have scaled epidermis, were used as control.

Figure 1: Whole-genome sequencing of scaled and foot feathered chickens. The scale epidermis is the common phenotype in wild and most chickens (**a**). However, feet of some birds display short and tight feathers on the metatarsus and digits (**b**), which in some cases appear like long flight-like feathers (**c**)



Whole-genome sequencing

Library preparation and sequencing were performed at the Institut national de la recherche agronomique (INRA), France, following their established protocols. Reads were mapped with the Burrows–Wheeler alignment (BWA-MEM) algorithm v0.7.17 [183] to the chicken GRCg6a reference genome (GenBank Accession: GCA_000002315.5) with default settings. Duplicate reads were removed with the *markdup* option in Sambamba v0.6.3 [282]. Sites with mapping quality <30 and base quality <20 were discarded from further analyses (SI Text).

Phasing, imputation, and annotation

Genotypes were imputed and phased into haplotypes with Beagle v4.0 [35] by considering in each of the 10 independent cycles 20,000 markers in each sliding window, allowing 1,000 markers to overlap between sliding windows. Imputation accuracy was estimated by masking 10% of the known sites (SI Text). Imputed variants were annotated to the Ensembl's *G. gallus* annotation database using the Ensembl Variant Effect Predictor (VEP) (release 95) tool [203] (SI Text).

Population stratification

Genetic distances were analysed with a principal component analysis (PCA) and a Neighbour-Joining (NJ) tree, both based on a subset of phased variants filtered for a minor allele frequency < 0.05. Filtering and PCA were performed in PLINK v1.9 [244]. The phylogenetic tree was generated in PHYLIP v3.696 [101] from the distance relationship matrix estimated in PLINK.

Association study and annotation of genes

We performed a standard case/control association analysis using the Fisher's exact test to generate uncorrected and corrected p-values, subsequently applying an adaptive Monte Carlo permutation test with 5,000 replications. Variants with a p-value lower than 1.0e-25 were considered to be significantly associated with the phenotype. Manhattan plots were generated using the qqman library in R v3.2.0 [283, 289]. Genes in genomic regions showing significant association with the phenotype were identified using the Ensembl Genes 95 Database in BioMart [167]. Chicken quantitative trait loci (QTLs) were downloaded from the Animal QTL Database [149]. Genomic coordinates were converted to the GRCg6a assembly in LiftOver [252]).

Signatures of selection

Screening for signatures of selection (SS) was performed on the control (n=150) and case (n=19) group, separately, and only on the autosomes that showed a significant association with the phenotype. For each pool and identified SNP, we determined the number of reads corresponding to the most (n$_{MAJ}$) and least abundant allele (n$_{MIN}$). The pooled heterozygosity (Hp) was calculated in sliding 40 kb windows following Rubin et al. (2010) [260]:

$$H_p = 2 * \frac{\sum n_{MAJ} \sum n_{MIN}}{(\sum n_{MAJ} + \sum n_{MIN})^2} \tag{5.1}$$

To resemble a normal distribution, Hp values were normalized into ZH$_p$ scores as

$$ZH_p = \frac{H_p - \mu H_p}{\sigma H_p} \tag{5.2}$$

Windows with at least 300 SNPs and a ZHp $\leq$ -3 were retained, as windows below this threshold represent the extreme lower end of the distribution (Additional file 3, Figure S3).

Recombination rate

We used the linkage map of Elferink et al. (2010) [92] to estimate the recombination rate, expressed as the genetic length in centimorgans (cM) divided by the physical genomic distance in mega base pairs. Recombination rate was calculated in bins of approximately 100 kb after converting the genomic positions of all SNPs to the GRCg6a genome assembly.

Structural variants

Structural variants (SVs) calling and genotyping were performed using Smoove (`https://github.com/brentp/smoove`). Smoove makes use of various existing tools to call SVs and improves specificity by removing noise from spurious alignment signals. First, discordantly mapped and split reads were extracted from the alignment by Samblaster [100]. Next, Lumpy software [179] was used to call SVs and genotyping was performed by SVtyper [53]. To further filter SV calls, Mosdepth [234] was used to discard reads from regions where the sequence depth of split or discordant reads was greater than 1,000 to remove regions that contribute to spurious calls. Duphold [235] was subsequently used to annotate depth changes within and on the breakpoints of SVs.

PCR-based screens for genomic rearrangement

A set of four PCR primers was designed to amplify two bands around the deletion breakpoints and two bands over the deletion. More information on the primers and PCR protocol are reported in Additional file 4. Gel image with the presence (control, scale-footed samples) or absence (case, feather-footed samples) of PCR product are reported in Additional file 4.

Phylogenetic analysis of haplotypes

For each sample, we extracted and considered the two alternative haplotypes as separate haplotypes, so that haplotypes belonging to the same individual did not necessarily cluster together. No missing alleles were present in the phased haplotypes since missing sites were imputed with Beagle. Haplotypes were reconstructed considering only bi-allelic sites. We then constructed a Neighbour-Joining (NJ) tree based on the distance matrix estimated in PLINK from all haplotypes.

DNA sequence conservation

Conserved elements were predicted from the 23 sauropsids multiple whole-genome alignment generated by [122] (SI Text). Conserved elements (CEs) were predicted using PhastCons [266] from a neutral evolutionary model estimated from 114,709 four-fold degenerate (4D) sites in Phylofit [267, 122]. After filtering for assembly gaps, a total of 1.14 million CEs covering 73 Mb of the chicken genome were retained (SI Text).

Tissue collection, RNA isolation and RNA sequencing

Forelimb and hindlimb buds were harvested from 21 chicken embryos sacrificed at Hamburger Hamilton (HH) stage 35 (n = 11) [131] and HH39 (n = 10) (Additional file 3, Figure S9). Detailed information on the breeds and samples can be found in Additional file 2, Table S8. Sequencing was performed at BGI, China, following the manufacturer's protocol.

RNA-seq analysis

Clean reads were mapped to the chicken GRCg6a reference genome using STAR v2.4.0 [81] with the chicken reference genome and its annotation file as guide, both downloaded from Ensembl (release 95). Quality of mapped RNA-seq data was assessed using Deeptools v3.3.1 [246]. RSEM was used to quantify expression of RNA transcripts and genes(Li and Dewey 2011), while StringTie v2.0.3 [237] was used for gene modelling using the Ensembl gene annotation file as reference. Transcripts of all samples were afterwards combined using the merge option in StringTie and used as input file in the FlExible Extraction of Long noncoding RNA (FEELnc) program [317] to predict and annotate lncRNAs (SI Text).

We used DESeq2 [192] to test whether the genes/lncRNAs identified by the genome-wide association analysis were differentially expressed in foot feathered birds (case) compared to scaled birds (control). The differential expression analysis was performed for each embryonic stage and for the forelimb (F) and hindlimb (H), separately, considering only genes with at least 20 reads. Protein-coding genes and lncRNAs were considered to be significantly differentially expressed only if their adjusted p-value was ¡ 0.05 (Benjamini-Hochberg adjustment).

5.3 Results

Whole-genome sequencing of scaled and foot feathered chickens

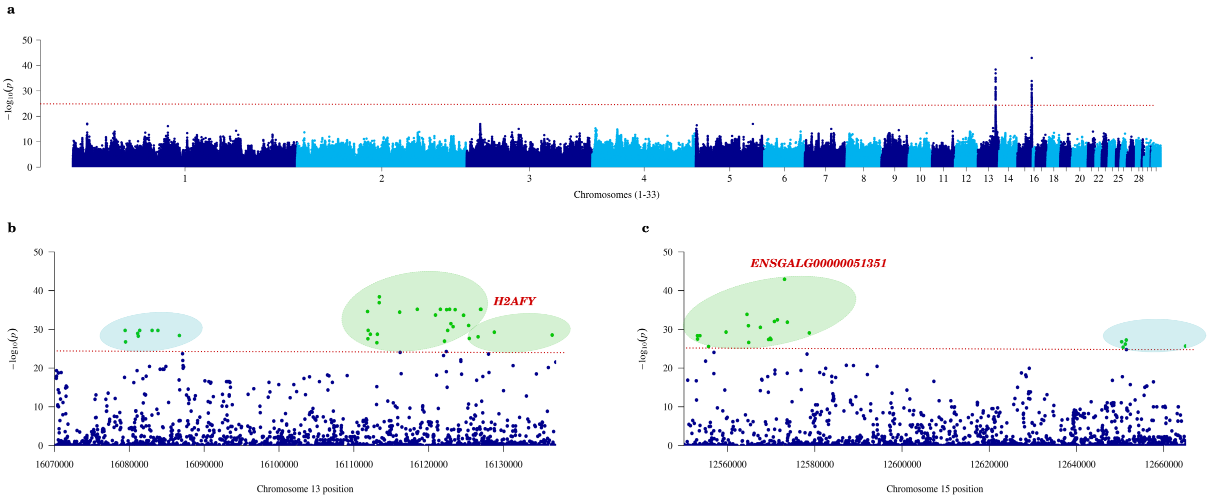
We unravelled the genetic basis and evolutionary history of ptilopody in chicken by whole-genome sequencing (WGS) of 169 samples from a variety of domesticated chicken breeds and wild species of Gallus (i.e. *G. gallus, G. sonneratii, G. lafayetii, G. varius*). On average, 14.8x coverage was generated for each individual after mapping to the chicken reference genome. Mapping quality was, on average, 33.4 with more than 98% of the reads successfully mapped (Additional file 1, Table S3). Since missing data comprised 12% of the total sites, we imputed missing genotypes and phased haplotypes with high accuracy using 21.0 million single-nucleotide polymorphisms (SNPs) and 1.4 million Insertion/Deletions (InDels) (Additional file 2, Table S2). Variants were also assigned to a range of functional classes, though the vast majority were located in introns (58%) and intergenic regions (30%) similarly to Lawal et al. (2018) (Additional file 2, Table S3). Of the 359,176 protein-coding variants, 217,255 were classified as synonymous, 130,147 as missense, and 11,774 as loss-of-function (LoF) (Additional file 2, Table S3).

To assess population stratification we performed a principal component analysis (PCA) and Neighbour-Joining (NJ) analysis on the 19 cases and 150 controls. The PCA did not identify any distinct clustering between the two groups (Additional file 3, Figure S1) aside from clearly separating all traditional breeds from the individuals of the four wild species of *Gallus*. However, the NJ tree based on the distance relationship matrix separated the 19 case individuals into three groups, one of the Breda fowl, one of the Dutch booted bantam, and one of the remaining foot feathered samples (Additional file 3, Figure S2). Despite that, one sample from the Marans (sample 1283), Sundheimer (sample 1769), and German Faverolles (sample 641) breed did not form any specific cluster.

Two genomic regions control foot feathering in chicken

We conducted a genome-wide association study (GWAS) on the 169 samples using a case/control approach through an adaptive Monte Carlo permutation test with 5,000 replications. The GWAS revealed two significant signals on chromosome 13 and 15, respectively (Figure 2a). On chromosome 13, we identified 36 significant variants (8 intergenic, 9 upstream gene, and 19 intron variants), which are located between 16.0 and 16.1 Mb (Additional file 2, Table S4). This 57 kb region contains a protein-coding H2A histone family gene, *H2AFY*, a novel long non-coding RNA, *ENSGALG00000048757*, and one QTL, *QTL127125*, which was previ-

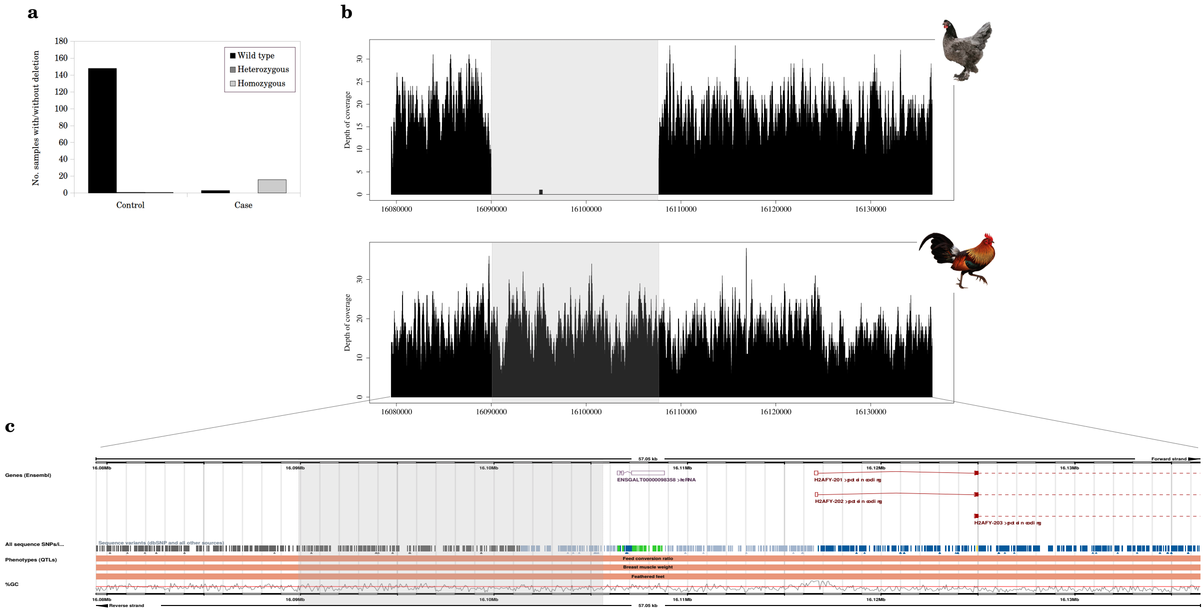
Figure 2: Foot feathering is associated with two genomic regions. **a.** Genome-wide manhattan plot. The –log10(p) for each variant is shown on the y-axis. Two clear signals can be observed on chromosome 13 and 15, respectively. **b.** Manhattan plot of chromosome 13 (16.0-16.2 Mb). **c.** Manhattan plot of chromosome 15 (12.5-12.6 Mb). Significant variants associated with protein-coding genes and lncRNAs are highlighted in green. Intergenic variants are highlighted in light blue. The significant p-value threshold (p-value <1.0e-25) is identified by the red dotted line



ously found to be associated with foot feathering [277] (Figure 2b). The gene *H2AFY* is located 145 kb upstream of *PITX1*, a gene that encodes a homeobox-containing transcription factor that is normally expressed in the vertebrate hindlimb but not the forelimb [190, 191, 258, 281]. For this 57 kb region foot feathered breeds showed elevated levels of homozygosity relative to scaled birds, a clear signature of positive selection as indicated by the low pooled heterozygosity (ZHp = -3.71 vs. ZH$_p$ = 0.80 in control individuals) (Additional file 3, Figure S4b).

On chromosome 15, the 23 significant variants (15 intron non-coding, 3 upstream gene, 5 intergenic) defined a 112 kb region (12.5-12.6 Mb) (Additional file 2, Table S5), in which we identified a T-box 5 protein-coding gene, *TBX5*, a novel lncRNA, *ENSGALG00000052717*, and a previously identified QTL, *QTL127126* [277] (Figure 2c). Most of the variants were found in the lncRNA. Within this 112 kb region we also identified one candidate selective sweep (15: 12,560,000-12,600,000), which had an average ZHp score of -3.41 (Additional file 3, Figure S4c). The ZHp score for the same region was above our threshold (ZHp > -3.0) in scaled samples.

Figure 3: Foot feathered birds share a 17 kb deletion upstream *H2AFY*. a. Total number of samples without deletion (wild type), heterozygous, or homozygous for the deletion. **b.** Genome-wide depth of coverage of chromosome 13 (16.0–16.2 Mb) for a Breda fowl (little feather development) individual and the wild *Gallus gallus* (scale epidermis). The deletion (13:16,089,992-16,107,660) is visible in the foot feathered sample by the absence of coverage. **c.** Location of deletion (highlighted in grey), lncRNA, *H2AFY*, and QTL (*QTL127125*) along the 57 kb significant region on chromosome 13



Foot feathered birds share a 17 kb deletion upstream H2AFY

We performed a copy-number variation (CNV) analysis to test whether a CNV event is associated with foot feathering. We identified a 17 kb deletion on chromosome 13, 9 kb upstream *H2AFY* between 16.08 and 16.10 Mb (Figure 3b). The deletion overlapped two 100 kb bins (13:15,964,681-16,217,433) estimated to have an average recombination rate of 4 cM/Mb (Additional file 3, Figure S7a).

The breakpoints of the deletion are both within the significant peak identified by the GWAS (Figure 3c). In addition to that, the deletion fully overlaps the 44 kb deletion reported in pigeon on scaffold 79 (Additional file 3, Figure S5). Of the 19 foot feathered chickens, 16 were homozygous for the deletion, while no deletion was observed for the Sundheimer, Marans, and German Faverolles breed (Figure 3a; Additional file 1, Table S4). Of the 150 controls, 148 lacked the deletion, while the Phoenix and Toutenkou breed were homozygous and heterozygous, respectively (Figure 3a; Additional file 1, Table S4). Despite that, all four wild species of Gallus did not have the deletion, therefore exhibiting a normal coverage distribution (Figure 3b; Additional file 1, Table S4). The PCR validated our CNV analysis, confirming the presence/absence of a deletion in our samples (Additional file 4). We used the UCSC RepeatMasker track to check for presence of repetitive and transposable elements at the deletion breakpoints in the chicken reference genome and identified four long interspersed nuclear elements (LINEs) (Additional file 2, Table S6). Even though we observed two LINE elements very close to each other in a genome that is not particularly enriched for them, their location (1 kb upstream and 4 kb downstream the deletion breakpoint) suggests that is unlikely that the deletion is caused by transposable elements.

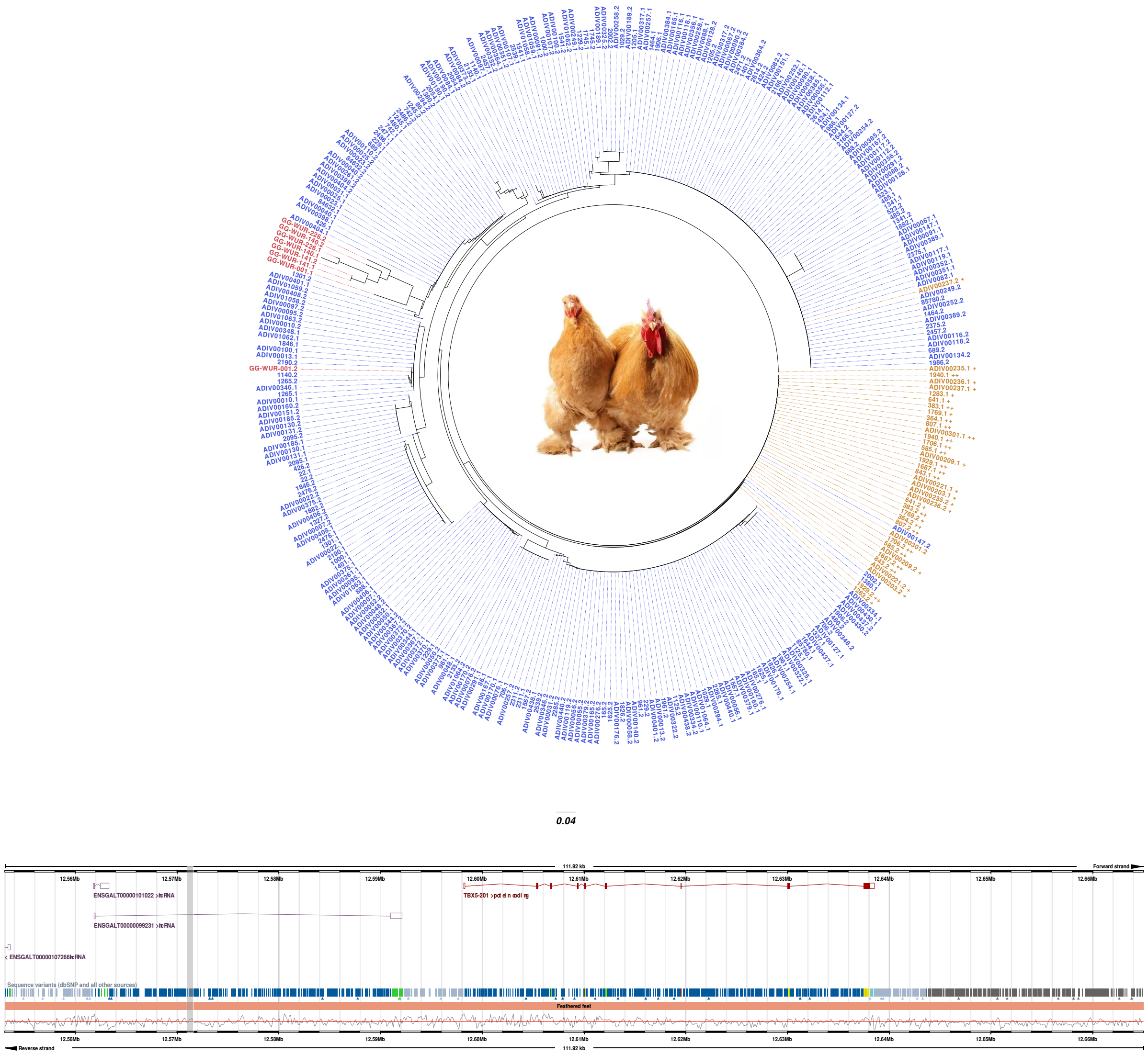
We further identified a 7 bp microhomology at the deletion breakpoint junction (Additional file 3, Figure S6). The nucleotide sequence of the microhomology flanking the first deletion breakpoint (13:16,089,992) is conserved in many other bird species, including duck, pigeon, collared flycatcher, white-throated sparrow, and medium ground finch (Additional file 1, Table S5). On the contrary, the nucleotide sequence of the microhomology adjacent to the second breakpoint (13:16,107,660) is conserved only between chicken and the two other species belonging to the same Galloanserae subgroup, being turkey and duck (Additional file 1, Table S5).

Foot feathering has a single haplotype origin

We also performed a CNV analysis on chromosome 15, but did not observe any CNV potentially associated with the phenotype. We therefore decided to reconstruct haplotypes for each individual by taking 2 kb upstream and 2 kb downstream the intron non-coding variant with the lowest p-value (15:12,573,054) (Additional file 1, Table S6). The total 4 kb region (15:12,571,054-12,575,054) included 44 bi-allelic variants, 3 of which were significantly associated with foot feathering. All 44 bi-allelic variants are located in the lncRNA ENSGALG00000052717. The 4 kb haplotype also overlapped two 100 kb bins (15:12,388,192-12,633,066) estimated to have an average recombination rate of 7 cM/Mb (Additional file 3, Figure S7b).

The phylogenetic analysis of the haplotypes clearly separated scaled from feathered samples, indicating an identical origin of haplotypes for the foot feathered samples (Figure 4). The clear separation between scaled and foot feathered samples was further confirmed by the fixation index (Fst) analysis performed on the 44 bi-allelic variants found in the same 4 kb region. We reported the highest Fst value for the non-coding variant used to reconstruct the haplotypes (Fst = 0.97) (Additional file 2, Table S7).

Figure 4: Foot feathering has a single haplotype origin. Each individual is identified by two haplotypes, one labelled after the name of the individual with the suffix '.1' and the second with the suffix '.2'. Individuals from the control group (i.e. scaled epidermis) are coloured in blue, whereas samples from the case group (i.e. feathered feet) are shown in orange. The individual from the *G. gallus, G. sonneratii, G. lafayetii, and G. varius* is labelled in red. Samples with little feather development are identified by the ' +' symbol following the haplotypes name, otherwise by a '++' symbol if heavily feathered. Haplotypes were defined by taking 2 kb upstream and 2 kb downstream the most significant variant (15:12,571,054). The genomic location of the reconstructed haplotype with respect to the lncRNA and *TBX5* protein-coding gene is highlighted in grey



Variants associated with foot feathering are in highly conserved regions

To understand the evolution of foot feathering, we looked for presence of conserved elements (CEs) in the 23 sauropsids multiple sequence alignment, which includes fifteen birds, three crocodilians, four turtles, and anole lizard (Additional file 3, Figure S8). On chromosome 13, 115 CEs were found within the GWAS peak. Of the 36 significant variants, two (one upstream gene and one intron variant) overlapped a CE (Table 1) and both were associated with the *H2AFY* gene. In the peak region of chromosome 15 we identified 275 CEs, although only the lncRNA intronic variant with the lowest p-value was found in a CE of considerable size (Table 1).

Table 1: Variants associated with foot feathering are highly conserved. The allele associated with the phenotype is underlined in bold. Abbreviations: Chr, chromosome; CE, conserved element; Maj, major allele; Min, minor allele. The p-value is that of the genome-wide association analysis

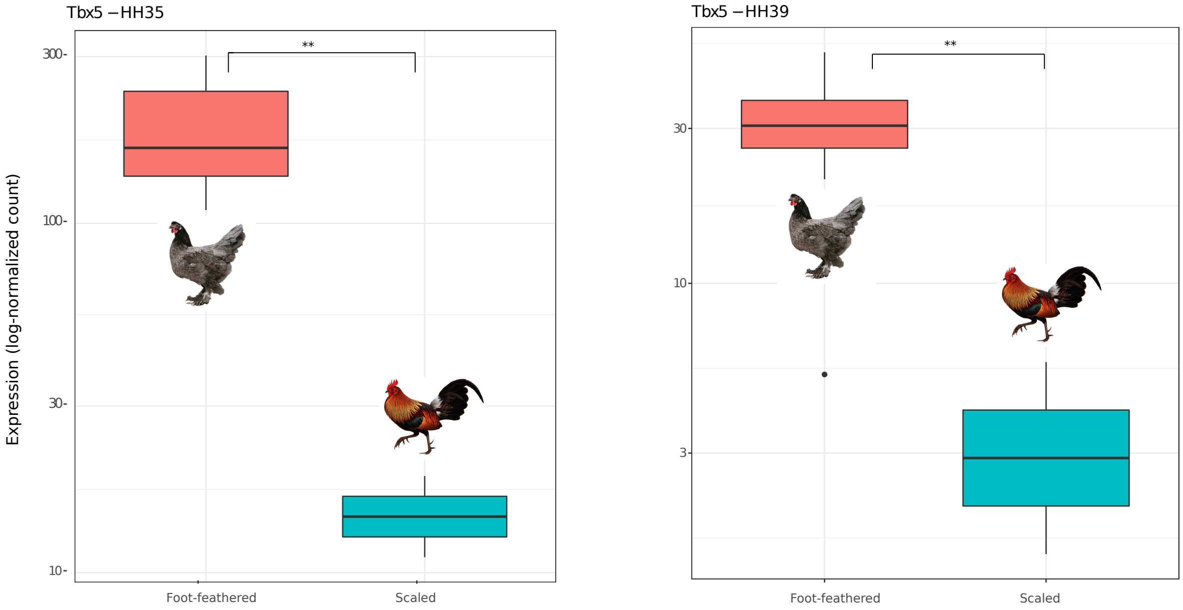
Chr	CE start	CE end	Size	Strand	Variant	Maj/Min allele	p-value
13	16,112,206	16,112,215	9	+	16,112,207	**C**/T	1.822e-29
13	16,128,762	16,128,774	12	+	16,128,768	**G**/A	5.395e-30
15	12,572,741	12,573,218	477	+	12,573,054	**C**/T	1.141e-43

TBX5 is upregulated in the hindlimb of foot feathered birds

We generated high quality RNA-seq data from 21 chicken embryos sacrificed at Hamburger Hamilton (HH) stage 35 and 39 to test whether our candidate genes and lncRNA are significantly differentially expressed in foot feathered samples (Additional file 2, Table S8). Overall, more than 90% of the reads were uniquely mapped with an average size of 300 bp (Additional file 2, Table S9). As expected, clustering of samples based on read counts followed the embryonic HH stage (Additional file 3, Figure S10).

PITX1 was significantly downregulated in the hindlimb of foot feathered birds at HH35 (*q-value*: 1.79e-03), but not at HH39 (*q-value*: 0.38) (Additional file 3, Figure S11). We observed a similar pattern in expression for *H2AFY* at HH35 (*q-value*: 0.016) compared to HH39 (*q-value*: 0.88) (Additional file 3, Figure S12). On the contrary, *TBX5* was always significantly upregulated in foot feathered birds at both embryonic stages (HH35 *q-value*: 2.49e-14; HH39 *q-value*: 6.87e-03) (Figure 5). At HH35, among the first top 10 most significant differentially expressed genes we also identified *ZIC1* (q-value: 2.42e-21), a transcription factor acting as scale-feather converter (P Wu et al. 2018). The FEELnc program classified our candidate lncRNA *ENSGALG00000052717* as belonging to the set of mRNAs, since its coding potential was above the cut-off estimated by a 10-fold cross validation procedure that maximizes both sensitivity and

Figure 5: *TBX5* is upregulated in the hindlimb of foot feathered embryos. Expression values of *TBX5* are shown on the y-axis as log-normalized counts at HH35 (left) and HH39 (right). Differences in expression between foot feathered and scaled birds were significant based on the adjusted p-value at stage HH35 (*q-value*: 2.49e-14) and HH39 (*q-value*: 6.87e-03). ** = q-value < 0.05



specificity [317]. We think that the low lncRNA read counts observed in our samples may have affected the FEELnc classification.

5.4 Discussion

Foot feathering has a parallel genetic origin

Researchers have repeatedly questioned the genetic basis of parallel evolution. Despite this long-lasting interest, for many traits that evolved in parallel little is known whether these are mirrored in underlying genes or mutations. Foot feathering is an interesting case since, although it is a very recognizable trait that can be very easily selected to fixation in breeds, it is in fact not a monogenic trait.

The molecular basis of foot feathering has so far only been studied in detail in domestic pigeon [82, 22]. Chicken and pigeon diverged more than 89 million years ago (Myr) [156] and are currently classified as belonging to two separate subgroups within the Neognathae clade, the Galloanserae and Neoaves, respectively [156, 38]. Since domestication, both species have experienced selection for a variety of traits that are remarkably different or absent in the wild ancestor [286, 83]. Foot feathering has been under artificial selection since ancient times and nowadays extensive variation can be observed among breeds (i.e. groused feet, slippered legs, muffed legs, vulture hocks). As for many other traits, ptilopody has been artificially selected to fixation and has become a breed characteristic in both species [15]. For parallel evolution to occur, loci associated with similar regulatory pathways and likely to generate variation should be targeted by selection. Our findings and those of Domyan et al. (2016) corroborate this hypothesis, as the independent evolution of foot feathering in chicken and pigeon involves a similar genetic basis and set of genes that not only generate outstanding variation, but this variation is also targeted for recurring artificial selection. In this study we showed that artificial selection has left clear signatures of positive selection in the genome of foot feathered birds, as indicated by the negative ZHp scores. These results also illustrate how combining genome-wide association studies and signature of selection analyses forwards our understanding of the genomic basis of traits, providing support for the role of, in this case, chromosome 13 and 15 in foot feather development. Similar signatures of selection were also identified in pigeon by Domyan et al. (2016). However, compared to chicken, pigeons homozygous for the deletion on scaffold 79 showed elevated levels of haplotype homozygosity relative to scaled birds, while positive selection was only observed among heavily feathered birds (i.e. muff phenotype) on scaffold 70 [82].

Interestingly, foot feathering is also observed in avian wild species, including snowy owl, golden eagle, and rock ptarmigan [15, 23]. Even though in raptor and boreal species ptilopody has entirely evolved by natural selection, the occurrence of the phenotype suggests that the same underlying genes and mutations can evolve in different species under different types of selection and selection pressure. However, studies on both wild and domesticated avian species are required to further validate this hypothesis.

Foot feathering is associated with a single, identical haplotype in chicken

As we showed, foot feathering has evolved independently in chicken and pigeon as a result of human-driven selection and this selection pressure has resulted in similar causal mutations. In chicken, foot feathered birds were also found to share an identical 4 kb haplotype on chromosome 15 independently on whether the individual has the 17 kb deletion on chromosome 13. Sharing of an identical haplotype was also reported in pigeon (i.e. scaffold 70), though a clear clustering was only observed among heavily feathered birds [82]. The presence of a single, identical underlying haplotype suggests that foot feathering is caused by mutations that occurred only once in the domestication history of chicken. These causal mutations have then been selected in multiple breeds, in many cases by deliberately crossing foot feathered birds with scaled birds of a different breed. Because of repeated crossing, the causal mutations underlying ptilopody have been recycled many times since domestication and because the genetic basis is strikingly the same among breeds, the underlying genes can easily be detected by an across-breed genome-wide association study. The mutations found in the 4 kb haplotype, which are clearly related to domesticated populations, are likely to have first appeared in Asia and have later been introgressed into Europe through human migration.

Haplotype length can provide important information on the age of the haplotype. This means that longer haplotypes are of more recent origin ("younger"), as recombination events did not break them down into smaller tracks over time. A negative correlation is, therefore, expected between haplotype length and recombination. The relative small size of the haplotype reported on chromosome 15 supports our conclusions on the single, and likely old, occurrence of the mutations underlying ptilopody on chromosome 15. In fact, it is likely that repeated crossing has not only allowed the spreading of the causal mutation, but has also contributed to breaking down the original haplotype at each generation, which, by means of artificial selection, is now fixed in all foot feathered birds considered in this study. The intact haplotype length is also explained by the local recombination rate reported in bins of 100 kb on chromosome 15, if we consider that recombination rate in micro-chromosomes (50-100 kb/cM) is nearly three times higher than that of macro-chromosomes (300 kb/cM) [205]. The limited number of SNPs found in the 4 kb haplotype makes, however, the estimation of the substitution model and mutation rate required to infer the haplotype age a challenging task. A possible solution would be to analyze the same candidate region in ancient samples to better estimate the age of the haplotype.

The long-noncoding variant (15:12,573,054) used for the haplotype analysis is found upstream *TBX5*, a gene encoding a key transcriptional regulator of forelimb identity and development [190, 191, 258, 281]. Interestingly, the same mutation was associated with foot feathering in a parallel study in chicken by Li et al. (2020) [186]. In chicken embryos, *TBX5* is normally expressed in the forelimb, but its mis-expression in the hindlimb at early embryonic stages can induce a partial wing-like transformation, including the formation and development of feathers on the feet [281]. Similarly to pigeon, the absence of fixed non-synonymous coding mutations in *TBX5* confirms the role of expression changes in the determination of feathered versus scaled feet in chicken as well. Interestingly, we could observe mis-expression in the hindlimb of feathered embryos at both HH35 and HH39, meaning that mis-expression starts at a very early embryonic stage (in the study of Domyan similar expression changes were observed at HH25) and is maintained almost up to the end of the embryonic development. Among our most significantly upregulated genes in the hindlimb at HH35 we also found the novel scale-feather converter *ZIC1*, a transcription factor whose overexpression in feather forming regions (i.e. wings and tail feathers) is sufficient to initiate the invagination step required to form the follicle, but not to form mature follicles [315].

The 17 kb deletion on chromosome 13 likely acts as qualitative molecular driver

Even though ectopic expression of *TBX5* is associated with foot feathering in chicken, extensive variation in feather type and distribution is often observed. In our study, the 17 kb deletion on chromosome 13 was homozygous in 16 birds, while absent in three birds (Additional file 1, Table S4). The absence of the deletion suggests that this locus may affect the qualitative variation in epidermal appendages. This means that the deletion is important for an individual to display variation in the type and extent of feathers, such as enlarged feathers on the feet or wing-like feathers on the feet and toes, but is not essential to determine the localized development of feathers on the feet, which seems the function of *TBX5*. Therefore, the qualitative role of the 17 kb deletion reasonably explains the discrepancies observed in the Marans, Sundheimer, and German Faverolles breed.

Strikingly, in pigeon a similar deletion, 44 kb in size, in the exact same region as the 17 kb deletion in chicken, is present [82]. The 7 bp microhomology we identified at the deletion breakpoints in chicken indicates that this structural variant has emerged in both species multiple times independently. However, contrary to chicken, based on their QTL and whole-genome sequencing analyses Domyan et al. (2016) concluded that the deletion is sufficient for the development of small feathers (grouse phenotype), while the development of large feathers (muff phenotype) is mostly driven by the *TBX5* locus.

In chicken, the 17 kb deletion is 9 kb upstream *H2AFY* and 200 kb upstream *PITX1*, a gene encoding a key transcriptional regulators of hindlimb identity and development. *PITX1* is normally expressed in the hindlimb, but not in the forelimb, as an abnormal expression in the forelimb blocks feather development [191]. Interestingly, in pigeon the peak on scaffold 79 is also 200 kb upstream *PITX1*. Compared to *H2AFY*, which was downregulated at HH35 and upregulated at HH39, *PITX1* was always downregulated. The key role of *PITX1* in limb-type morphology determination has been demonstrated across multiple species [191, 281, 73, 229, 172] and in all species investigated the relationship between *PITX1* and *H2AFY* is maintained. In pigeon, the 44 kb deletion spans an element orthologous to a known human limb enhancer, *hs1473*, which shows a strong limb-specific activity [272, 82]. As limb patterning and morphogenesis are regulated by highly conserved networks [22], it is reasonable to assume that, as in pigeon, also in chicken the deletion encompasses this enhancer, causing loss of *PITX1* expression, thus resulting in a partial leg-to-wing homeotic transformation. However, further molecular analysis, including ChiP-seq data, are required to formally confirm this conclusion.

5.5 Conclusions

Foot feathering is an interesting example of a polygenic trait that has evolved by parallel evolution as its parallel evolution is mirrored in almost every detail at the molecular and, most likely, developmental level. In this study, we showed that, although chicken and pigeon diverged more than 89 million years ago (Myr), in both avian species the exact same number of loci containing the exact same set of genes are involved. This similarity is even more striking as a similar deletion at one of the loci has the same outcome in regulating gene expression. Even though genetic variants arose independently millions of years after the species divergence, it is remarkable to see that not only are the exact same genes involved, but they are affected in very similar ways, despite the many ways in which a similar phenotype conceivably could have arisen. Therefore, even under different types of selection and selection pressure, the same genes and causal mutations underlying major phenotypic changes can evolve in different lineages. Our findings provide support for the hypothesis that only a limited number of evolutionary trajectories at the molecular level are open to generate a specific outcome if developmental pathways are sufficiently constrained.

5.6 Acknowledgements

We would like to sincerely thank all hobby breeders and breeders associations for their availability and collaboration during sampling. A special thanks to the breeders and people of the Nederlands Pluimveemuseum for their help, support, and enthusiasm during the sampling of eggs that were used for the RNA-seq analyses. We would also like to sincerely thank Zhou Wu for her help in the lab and for providing us the pictures of the chicken embryos used in this study. The research leading to some of these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353.

5.7 Author Contributions

CB, MB, and HJM conceived the study. CB performed the analysis. MFLD performed the structural variant analysis. BD, KL and SW performed the PCR. SW performed the blood collection of the German samples. CB, BD, KL, and RPMAC performed the sampling of eggs used for the extraction of limb buds. MB, HJM, MAMG, and RPMAC supervised the project. CB, MB and HJM wrote the manuscript. All co-authors read and contributed to the manuscript. All co-authors agreed on the final manuscript.

5.8 Declaration of Interests

The authors declare that they do not have any conflict of interest.

5.9 Data availability

Whole-genome sequencing data of the 97 birds sampled in the Netherlands were previously deposited in the European Nucleotide Archive (ENA) under accession number PRJEB34245 [25]. Whole-genome sequencing data of the 68 birds sampled in Germany and RNA-seq data of the 21 chicken embryos have been deposited under accession number PRJEB36674.

5.10 Additional data

The online version of this article (https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msaa092/5818884) contains supplementary material, which is available to all users.

5.11 Supplementary Information

SI Text

Variant calling and post-filtering

Reads were aligned with the Burrows–Wheeler alignment (BWA-MEM) algorithm v0.7.17 [183] to the chicken GRCg6a reference genome (GenBank Accession: GCA_000002315.5) with default options. Duplicate reads were then removed with the markdup option in sambamba v0.6.3 [282]. Mapping quality and coverage of the aligned sequences were assessed with Qualimap v2.2 [225]. Freebayes v0.9.10 [115] was used to perform a population-based variant calling, applying the following settings: (1) mapping quality >20, (2) base quality >20, (3) at least 20% of observations and 2 reads supporting an alternative allele within an individual, and (4) coverage at SNP position >4 and <2.5*average individual genome-wide coverage. Variant calling was performed on single variants (–haplotype-length 0) assuming diploid organisms (–ploidy 2). We reduced the false discovery rate by performing additional filtering using BCFtools v1.4.1 [183]. The settings were: (1) a phred quality score >30, (2) an allele count supporting the alternative allele >2, (3) maximum number of 10 alleles, (4) variants located within 3 bp of an indel, and (5) call rate <70%.

Imputation accuracy

Imputation accuracy was estimated on three scenarios after randomly assigning samples to a reference and validation set. The size of the reference population was set to 90% (Ref90), 60% (Ref60), and 30% (Ref30) (Table 1). In all three scenarios, 10% of the known variants were masked in the validation set prior to imputation.

Table 1: Overview of scenarios for imputation accuracy. Samples were randomly assigned to a reference (nRef) and validation (nVal) set. The percentage of masked SNPs (SNPsmasked) with a minimum read depth of 4x was set to 10% in all three scenarios

Scenario	nref/nVal	SNPsmasked (%)	Reference/Validation set
Ref90	153/16	10%	Random
Ref60	102/67	10%	Random
Ref30	51/118	10%	Random

Imputation was then performed in Beagle v4.0 [35] with the same criteria previously used to generate the reference set. The percentage of correctly imputed genotypes (% correct) and the correlation between true and imputed genotypes (rs) were computed to assess the imputation accuracy. The % correct is the proportion of correctly imputed genotypes out of all imputed genotypes, whereas r2 is the Pearson's correlation between the true genotypes and the imputed discrete genotypes (0, 1 or 2). Both % corret and r2 were computed for each individual in the validation set and then averaged across all individuals in each scenario (Table 2).

Table 2: Imputation accuracy for each scenario measured as % correct and r2. Imputation accuracy was estimated for each individual and then averaged across all samples in each scenario. Sd is standard deviation

Scenario	Average % correct (sd)	Average r2 (sd)
Ref90	95.88% (1.867)	0.8984 (0.044)
Ref60	95.27% (2.768)	0.8857 (0.066)
Ref30	91.69% (4.065)	0.7897 (0.117)

Annotation of variants

Imputed variants (SNPs and indels) were annotated with the Variant Effect Predictor (VEP) tool [203] using the Ensembl's *G. gallus* annotation database (release 95). Based on the annotation, variants were divided into protein-coding (e.g. synonymous, nonsynonymous, and loss of function) and non-coding (e.g. intronic, intergenic, lncRNAs). Among the nonsynonymous mutations, nonsynonymous tolerated were defined by a SIFT score ≥ 0.05, whereas deleterious by a SIFT score < 0.05. Loss of function mutations were defined as those that econde a premature stop codon and contribute to splicing, stop–gain and stop-loss. These included: splicing acceptor, splicing donor, inframe insertion, inframe deletion, stop gained, stop lost, and start lost variants.

23 sauropsids multiple whole-genome alignment

Analyses on the DNA sequence conservation were performed on the 23 sauropsids multiple whole-genome alignment generated in progressive cactus [233] by [122]. The whole-genome alignment was downloaded in the hierarchical alignment format (HAL) from crocgenomes.org (`ftp://ftp.crocgenomes.org/pub/ICGWG/Supplementary_Materials/main_page/Ancestral_genome_reconstruction_files/crocPaper_reestimated.hal`). The hal tools command hal2maf [143] was used to perform format conversion with the following parameters: -refGenome galGal4 to export the HAL alignment as a multiple alignment format (MAF) referenced to the chicken reference genome (GenBank Accession GCA_000002315.2), -noAncestors to exclude unrequired ancestral reconstructions, -onlyOrthologs to include only sequences orthologous to chicken, and –noDupes to ignore paralogy edges. An in-house python script was then used to perform additional formatting, including splitting the alignment into individual files for the 28 assembled chicken chromosomes and two sex chromosomes (W, Z). A separate file was also created for all remaining unassembled scaffolds. During reformatting, only alignments where chicken aligned to at least two more species were retained.

Prediction of conserved elements

Conserved elements were predicted from the same multiple whole-genome alignment using Phastcons [266]. As the multiple sequence alignment contains very distant species, the estimates of the nonconserved branch lengths to these species will tend to be underestimated, because any ”nonconserved” bases that do align are probably actually at least partially conserved.

For this reason, we first estimated a neutral (nonconserved) evolutionary model from the 114,709 four-fold degenerate (4D) sites downloaded in Phylip format from GigaScience (`https://doi.org/10.5524/100125`) as WholeGenome_4Dsites_filter2.phy. The topology of the tree used in the phast tool Phylofit to run the REV model was identical to that derived by [122]. Phastcons was then run using the standard parameters (SP) defined in the UCSC genome browser to produce the “most conserved” tracks [210]. The parameters are an expected length of 45, a target coverage of 0.3, and an rho of 0.31. We did not perform additional tuning in Phastcons as standard and tuned parameters were found to provide accurate estimates of conserved elements with a small margin of differences [61]. Genomic coordinates of all predicted conserved elements were converted to the chicken GRCg6a reference genome using the pyliftover library in python v3.5.0. All successfully converted CEs were further discarded if they fell or overlapped assembly gaps.

Prediction of long noncoding RNAs

Long noncoding RNA (lncRNA) annotation was performed by the FlExible Extraction of Long noncoding RNAs (FEELnc) program [317]. The first module FEELnc_filter was initially used to filter out transcripts overlapping (in sense) protein-coding exons and pseudogenes of the reference annotation file. Transcripts shorter than 200 bp were also filtered out (default option). The second module FEELnc_codpot was used to compute coding potential score (CPS) for each of the candidate transcripts in order to distinguish putative lncRNAs from protein-coding RNAs. For the training set of protein-coding transcripts, we used the 26,441 (autosomal) known coding transcripts annotated by Ensembl, while for the training set of long noncoding transcripts we used the 19,818 (autosomal) chicken putative transcripts from the NON-

CODEV5 database (v2017) [99] whose genomic coordinates were successfully converted to the latest chicken genome build. The third module FEELnc_classifier was finally used to classify each lncRNA with respect to its location and orientation compared to its closest annotated protein-coding genes.

6.
Prioritizing sequence variants in conserved non-coding elements in the chicken genome using the chCADD model

The chapter is based on:

Groß, C.*, **Bortoluzzi, C.***, de Ridder, D., Megens, H.J., Groenen, M.A.M., Reinders, M., Bosse, M. (2020). *PLOS Genetics (Under review)*

* These authors equally contributed to this work

Abstract

The availability of genomes for many species has advanced our understanding of the non-protein-coding fraction of the genome. Comparative genomics has proven itself to be an invaluable approach for the systematic, genome-wide identification of conserved non-protein-coding elements (CNEs). However, for many non-mammalian model species, including chicken, our capability to interpret the functional importance of variants overlapping CNEs has been limited by current genomic annotations, which rely on a single information type (e.g. conservation). We here studied CNEs in chicken using a combination of population genomics and comparative genomics. To investigate the functional importance of variants found in CNEs we develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD) model, a variant effect prediction tool first introduced for humans and later on for mouse and pig. We show that 73 Mb of the chicken genome has been conserved across more than 280 million years of vertebrate evolution. The vast majority of the conserved elements are in non-protein-coding regions, which display SNP densities and allele frequency distributions characteristic of genomic regions constrained by purifying selection. By annotating SNPs with the chCADD score we are able to pinpoint specific subregions of the CNEs to be of higher functional importance, as supported by SNPs found in these subregions are associated with known disease genes in humans, mice, and rats. Taken together, our findings indicate that CNEs harbor variants of functional significance that should be object of further investigation along with protein-coding mutations. We therefore anticipate chCADD to be of great use to the scientific community and breeding companies in future functional studies in chicken.

Author summary

Chickens are raised worldwide as a livestock species to provide us with their eggs and meat, but besides their huge economical impact their genome remains poorly understood. Here we introduce a variant prioritization tool modeled after the Combined Annotation Dependent Depletion (CADD). CADD is a well-established approach to prioritize variants with respect to their deleteriousness for the interpretation of genetic variation that can substantially impact human phenotypes. We applied the CADD approach to chicken (chCADD) to investigate the functional importance of conserved non-protein-coding elements. The chCADD model assigns a score to all possible variation in the chicken genome. We used these scores to identify subregions within conserved non-coding elements of relative higher importance. The chCADD score and the identified subregions are expected to support our efforts to pinpoint causal genomic variation throughout the chicken genome.

Introduction

The rapidly increasing availability of genomes has considerably advanced our understanding of the non-protein-coding fraction of the genome. With the sequencing of the human genome [59] and the first ENCODE project [56, 57] it was soon realized that protein-coding genes constitute a small fraction of a species functional genome and that the remaining non-protein-coding DNA is not simply ´junk´ DNA as initially thought. Nevertheless, the functional importance of these non-protein-coding regions remained for long time unknown, as determining (molecular) function was far more difficult than for protein-coding genes [199]. A better understanding of the functional importance of these non-protein-coding regions comes from comparative genomics, which has allowed the systematic, genome-wide identification of conserved non-protein-coding elements (CNEs) [5, 134].

Comparative genomics relies on the genome comparison of a group of species related by a narrow or wide time-scale (i.e. phylogenetic scope). Regions in the genome that share some minimum sequence similarity across two or more species are an indication of a selection constraint. Moreover, conservation often implies a biological function [7]. Based on this principle, CNEs can be identified in any species included in the alignment, as reported in recent studies in the collared flycatcher [61], fruit flies [19], and plants [134]. However, the phylogenetic scope [132] and species included in the alignment [33] can have important implications for the identification of CNEs. For instance, by including the spotted gar genome in their alignment, Braasch et al. (2016) were able to identify numerous CNEs previously undetectable in direct human-teleost comparisons, supporting the importance of a bridging species in the alignment [33].

CNEs have been the subject of intense recent interest. The identification of CNEs has had important implications in enhancing genome annotation [187], investigating signatures of adaptive evolution [130, 141, 313], and identifying putative trait loci [198]. CNEs and sequence conservation have also proven crucial in studying the genetic basis of phenotypic diversity. In fact, non-protein-coding SNPs have been linked to traits and diseases in genome-wide association studies [201, 27].

Although the methodological advantages of a comparative genomic approach are well recognized, the functional interpretation of CNEs is incomplete if based on conservation alone, as conservation provides information on restrictions, but not on functionality. A possible solution is combining conservation with other complementary types of data that characterize the biological role of genetic sequences at a genome-wide scale [7]. Such data include, for instance, RNA sequencing (RNA-seq) for the identification of transcriptionally active regions and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for regulatory-factorbinding regions (RFBRs) [231]. In human genetics, integrative annotations such as Combined Annotation-Dependent Depletion (CADD) [168] have been developed. The main advantage of such frameworks is the combination, into a unique score, of diverse genomic features derived from, among others, gene model annotations, evolutionary constraints, epigenetic measurements, and functional predictions [251].

Compared to humans, for many non-mammalian model species, including chicken (*Gallus gallus*), the situation is quite different. First, comparative genomic studies that made use of the very first genome assemblies [323, 207, 193] may have provided an incomplete and biased picture of avian CNEs and avian genome evolution, as recently pointed out by Bornelov et al. (2017) [24]. Second, the lack of species-specific methods that can identify and score functional non-protein-coding mutations throughout the genome has restricted most of the research interest to protein-coding genes. In fact, in the context of protein-coding genes generic predictors such as SIFT [222], PolyPhen2 [2], and Provean [54] can be used.

We here addressed these limitations using a combination of comparative genomic and population genomic approaches to accurately predict CNEs in the chicken genome. Furthermore, we used machine learning to develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD) model, in the tradition of previous CADD models for non-human species, including mouse (mCADD) [127] and pig (pCADD) [128]. As we show, chCADD has the potential of providing new insights into the functional role of non-protein-coding regions of the chicken genome at a single base pair resolution.

Even though deciphering the function of the non-protein-coding portion of a species genome has been a challenging task, we expect our study to provide a new framework for decoding the still largely unknown function of CNEs and their relative variants in chicken, an ideal non-mammalian model and anchor species in evolutionary studies.

Materials and methods

Chicken genomic data

We used a dataset by Bortoluzzi and colleagues available at the European Nucleotide Archive (`http://www.ebi.ac.uk/ena/`) under accession number PRJEB34245 [25] and PRJEB36674 [27]. The dataset comprised a total of 169 chickens sampled from 88 traditional breeds of divergent demographic and selection history. The 169 chicken samples were sequenced at the French Institute of Agricultural Research (INRAe), France, on an Illumina HiSeq 3000. Reads were processed following standard bioinformatics pipelines. Reads were aligned to the chicken GRCg6a reference genome (GenBank Accession: GCA_000002315.5) with the

Burrows-Wheeler alignment (BWA-mem) algorithm v0.7.17 [183]. After removal of duplicate reads with the *markdup* option in sambamba v0.6.3 [282], we performed population-based variant calling in Freebayes [115] using the following settings: (1) mapping quality > 20, (2) base quality > 20, (3) at least 20% of observations and 2 reads supporting an alternative allele within an individual, and (4) coverage at SNP position > 4 and < 2.5*average individual genome-wide coverage. We reduced the false discovery rate by additional filtering using BCFtools v1.4.1 [183]. The settings were: (1) a phred quality score > 30, (2) an allele count supporting the alternative allele > 2, (3) maximum number of 10 alleles, (4) variants located within 3 bp of an indel.

Multiple whole-genome sequence alignment

Conserved elements (CE) were identified using the 23 sauropsids multiple whole-genome sequence alignment (MSA) generated using Progressive Cactus (`https://github.com/glennhickey/progressiveCactus`) [233] by Green et al. (2014) [122]. The MSA downloaded in the hierarchical alignment format (HAL) was converted into multiple alignment format (MAF) using the HAL tools command hal2maf [143] with the following parameters: -refGenome galGal4 (GenBank Accession: GCA_000002315.2) to extract alignments referenced to the chicken genome assembly, -noAncestors to exclude any ancestral sequence reconstruction, -onlyOrthologs to include only sequences orthologous to chicken, and -noDupes to ignore paralogy edges. During reformatting, only blocks of sequences where chicken aligned to at least two other species were considered for a total chicken genome alignability of 90.88%. Genomic coordinates were converted to the GRCg6a genome assembly using the pyliftover library in python v3.6.3.

Prediction of evolutionarily conserved elements

Conserved elements were predicted from the whole-genome alignment using PhastCons [266]. We chose PhastCons because this approach does not use a fixed-size window approach, but can take advantage of the fact that most functional regions involve several consecutive sites [261]. We first generated a neutral evolutionary model from the 114,709 four-fold degenerate (4D) sites previously extracted from the alignment by Green et al. (2014). The topology of the phylogeny was also identical to that derived by Green et al. (2014). PhastCons was run using the set of parameters used by the UCSC genome browser to produce the 'most conserved' tracks (top 5% of the conserved genome): expected length = 45, target coverage = 0.3, and rho = 0.31 [210]. Conserved elements were subsequently excluded if falling in or overlapping assembly gaps and/or if their size was < 4 bp.

Annotation of conserved elements by genomic feature

We use the Ensembl (release 95) chicken genome annotation files to extract sequence coordinates of CDS, exons, 5' and 3' UTRs, pseudogenes, and lncRNAs. Sequence information was extracted from 14,828 genes (out of the 15,636 genes found in the Ensembl annotation), as transcripts of these genes had a properly annotated start and stop codon. For protein-coding genes with an annotated 5' UTR of at least 15 bp, the promoter was defined as the 2-kb region upstream of the transcription start site (TSS) [61]. Sequence coordinates of miRNAs, rRNAs, snoRNAs, snRNAs, ncRNAs, tRNAs, and scRNAs were also extracted from the annotation file. For the identification of intergenic regions we considered all annotated protein-coding genes and defined intergenic regions as DNA regions located between genes that did not overlap any protein-coding genes in either of the DNA strands. The intersection between CEs and the various annotated genomic features was found following the approach of Lindblad-Toh et al. (2011) of assigning a CE overlapping two or more genomic features to a single one in a hierarchical format: CDS, 5' UTR, 3' UTR, promoter, RNA genes, lncRNA, intronic, and intergenic region [187]. Conserved non-protein-coding elements (CNEs) were defined as CEs without any overlap with exon-associated features (CDS, 5' UTR, 3' UTR, promoter, and RNA genes) and include lncRNAs, introns, and intergenic regions.

Gene ontology analysis

Genes in conserved regions overlapping CDS, 5' UTR, 3' UTR, and introns were separately used to perform a Gene Ontology analysis in g:Profiler [248] using *Gallus gallus* as organism. We only considered annotated genes that passed Bonferroni correction for multiple testing with a threshold < 0.05.

Genome-wide distribution and density of conserved non-protein-coding regions

Polymorphic, bi-allelic SNPs belonging to all functional classes predicted by the Variant Effect Predictor (VEP) [203] were considered. However, to improve the reliability of the set of annotated variants, we applied additional filtering steps. SNPs were discarded if they overlapped repetitive elements or if their call rate was $< 70\%$. The rationale for excluding variants found in repetitive elements was to reduce erroneous functional predictions as a result of mapping issues, as regions enriched for repetitive elements are usually difficult to assemble. For intronic and intergenic SNPs, SNPs in exons or that fell within any spliced EST from the UCSC chrN_intronEST tables were discarded [85].

Ancestral allele and derived allele frequency

The sequence of the inferred ancestor between chicken and turkey (*Meleagris gallopavo*; Turkey_2.01) [63] reconstructed from the Ensembl EPO 4 sauropsids alignment (release 95) was used to determine the ancestral and derived state of an allele, along with its derived allele frequency. We considered only SNPs for which either the reference or alternative allele matched the ancestral allele. Ancestral alleles that did not match either chicken allele were discarded. We generated derived allele frequency (DAF) distributions for sets of SNPs based on functional class and whether they were within or outside of CNEs. A derived allele frequency cutoff of 10% was used to distinguish rare from common SNPs.

Chicken Combined Annotation Dependent Depletion (chCADD)

The chicken CADD score is the -10 log relative rank of all possible alternative alleles of all autosomes and Z chromosome of the chicken GRCg6a reference genome, according to the following formula:

$$chCADD_i = -10log_{10}\frac{n_i}{N} \tag{6.1}$$

where N represents the number of all possible alternative alleles (3,073,805,640) on the investigated chromosomes and n is the rank of the ith SNP. The rank is based on the model posteriors of a ridge penalized logistic regression model trained to classify simulated and derived SNPs.

Chicken derived SNPs were defined as those sites where the chicken reference genome differs from the chicken-turkey ancestral genome inferred from the Ensembl EPO 4 sauropsids alignment containing chicken, turkey, zebra finch (*Taeniopygia guttata*; taeGut3.2.4) [307] and green anole lizard (*Anolis carolinensis*; AnoCar2.0) [6]. Sites for which the ancestral allele occurs at a minor allele frequency greater than 5% were excluded. In addition, derived SNPs that are observed with frequency above 90% in our population of 169 individuals were included. In total we identified 17,237,778 SNPs.

The dataset of simulated variants was simulated based on derived nucleotide substitution rates between the different inferred ancestors of the 4 species in the EPO 4 taxa sauropsids alignment. These derived nucleotide substitution rates were obtained for windows of 100 kb and used to simulate de novo variants which have a larger probability to have a deleterious effect than the set of derived variants. All SNPs which have a known ancestral site are retained in the dataset. In total 17,233,727 SNPs were simulated in this way. 17,233,722 SNPs of each dataset were joined and randomly assigned to train and test sets of sizes 15,667,020 and 1,566,702, respectively.

The datasets were annotated with various genomic annotations: among others, PhyloP and PhastCons conservation scores based on three differently deep phylogenies (i.e. 4 sauropsids, 37 amniote/mammalia, 77 vertebrate, all excluding the chicken genome), secondary DNA structure predictions [328], Ensembl Consequence predictions, amino acid substitution scores such as Grantham [121], and amino acid substitution deleterious scores such as SIFT [222]. Further, we utilized RNA expression, ATAC-seq and HI-C [106] data to annotate our data set. An overview is given in Table S1.

Annotations for which values were missing were imputed (Table S1) and categorical values were one hot-encoded [86]. In the one hot-encoding process, an annotation is a series of binary annotations, each indicating the presence of a specific category for a given variant. For scores that are by definition not available for certain parts of the genome, such as SIFT which is found only for missense mutations, columns indicating their availability were introduced.

Combinations of annotations were created of Ensembl Variant Effect Predictor consequences and other annotations, such as distance to transcription start site and conservation scores. The total number of all features used in training was 874. An extensive list of all annotations, combinations of annotations and their learned model weights is shown in Supplementary File 1. Finally, each feature column is scaled by its standard deviation. The logistic regression is trained via the Python Graphlab module. We selected a penalization term of 1 based on results on the test set (Fig. S1).

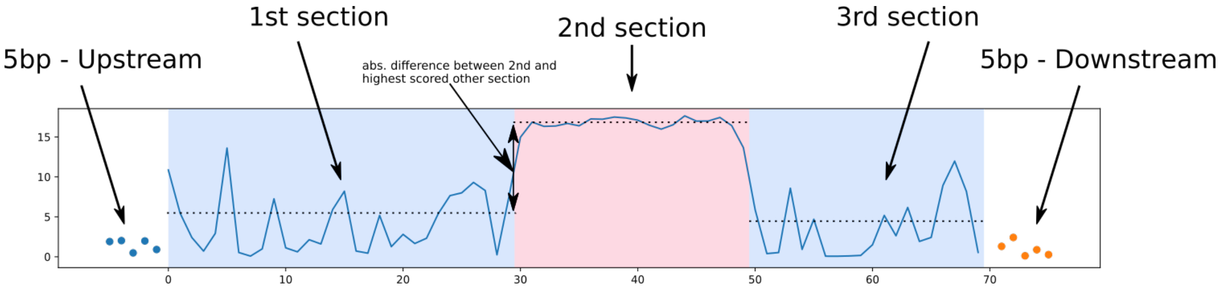
Investigation of likely causal SNPs from the OMIA database

We downloaded the likely causal variants of phenotype changes from the Online Mendelian Inheritance in Animals (OMIA) [180] database (last accessed 25.11.2019). SNPs whose location was reported for older genome assemblies such as Galgal4 and Galgal5 were mapped to the chicken GRCg6a reference genome via CrossMap [326]. We only considered bi-allelic SNPs whose genomic position was successfully mapped to GRCg6a and whose substitution remained the same. In total, 15 SNPs were left and annotated with chCADD.

Change point analysis

To identify subregions of particular importance within each CE, we annotated all CEs with the maximum chCADD score found at each site or the 23 sauropsids PhastCons scores that were used to identify conserved elements in the first place. Our basic assumption was that highly important subregions within a CE are preceded and succeeded by less important sites which would result in a relatively higher score region surrounded by two lower scored regions. Each CE was treated similarly to time series data by conducting an offline change point analysis, once based on maximum chCADD scores and once based on 23 sauropsids PhastCons scores. To this end, we used the Python ruptures module [288] and applied a binary segmentation algorithm with radial basis function (RBF). The algorithm first identifies a single change point. Furthermore, if a change point is detected, the algorithm investigates each sub-sequence independently to identify the next change point We were looking particularly for 2 change points, which would divide the CE into three subregions, numbered from 1 to 3, starting at the 5' end of the sequence. We added 5 bp upstream and downstream of each CE to allow that the borders of the $2nd$ region coincide with the borders of the CE (Fig. 1). After computing the change points, we conducted t-tests between the scores of the $1st$ and $2nd$, as well as $3rd$ and $2nd$ subregions, to identify CEs that have a significantly different score in the $2nd$ section than in the other two. We applied a p-value cutoff of 0.05. We sorted CNEs with respect to the largest difference between the mean chCADD score of the inner and the two outer subregions and selected those with a higher scored $2nd$ section than either of the other two outer ones.

Figure 1: Approach used to identify subregions within CNEs via change point analysis. The scores used to annotate the CE region are displayed on the y-axis. The position in the investigated CE region is shown on the x-axis. In total there are five sections, 5 bp up and downstream, 1st, 2nd and 3rd subregions. The transitions from blue to red background indicate the position of the two identified change points. The up and downstream scores are shown as dots while the scores in the CE region are a continuous blue line.



SNP density distribution within conserved non-protein-coding regions

SNP density was calculated as the number of SNPs identified in the 169 chicken individuals divided by the number of bases found in the sequence. SNP density was computed for conserved coding (CC) and conserved non-protein-coding (CNE) regions, as well as for the subregions identified in the change point analysis of CNEs overlapping lncRNAs, introns, and intergenic regions. We repeated this analysis once for the change points identified using chCADD scores and once for the 23 sauropsids PhastCons based change points.

Homologous phenotypes

We obtained phenotypes from the Ensembl database (release 95) for genes associated with the lncRNA and intronic CNEs. Beside chicken, these phenotypes encompass the observed phenotypes for orthologous genes associated with disease studies in humans (*Homo sapiens*) and gene-knockout studies in mouse (*Mus musculus*) and rat (*Rattus norvegicus*).

Results

Conserved non-protein-coding elements cover a large fraction of the chicken genome

To define CNEs, we first identified conserved elements (CEs) using the UCSC PhastCons most conserved track approach [210]. PhastCons predicted in the 23 sauropsids multiple sequence alignment (MSA) 1.14 million CEs encompassing 8% of the chicken genome for a total of 73 Mb. In line with the density of genes and regulatory features characteristic of the chicken genome [145], we found that most of the predicted CEs are on micro-chromosomes (GGA11-GGA33), followed by intermediate (GGA6-GGA10) and macro-chromosomes (GGA1-GGA5) (Fig. S2). Even though the length of predicted CEs ranged from 4 bp to a maximum of 2,000 bp, the vast majority was short (< 100 bp) (Fig. S3). Therefore, we do not expect any length bias in our final set of CEs.

We annotated CEs by genomic features, considering only genes for which the transcript had a proper annotated start and stop codon, as defined by the Ensembl´s annotation files (n = 14,828 genes). Overall, we found that 23% of the predicted CEs were associated with exonic sequences (i.e. CDS, 5' UTR, 3' UTR, promoter, and RNA genes) spanning 17.14 Mb of the chicken genome (Table 1). The majority of the exon-associated CEs overlapped known coding regions (85% of total exon-associated CEs), followed by 3' UTRs (8% of total), and promoter regions (4% of total). Although we observed conservation in exon sequences, most CEs overlapped non-protein-coding sequences, including lncRNA (15% of total non-exon associated CEs), intronic (36% of total), and intergenic regions (49% of total). We further examined the biological processes and molecular functions of known genes overlapped by CEs in coding regions, 5' UTRs, 3' UTRs, and introns. These genes are associated with basic functions, including cell differentiation and development, anatomical structure development, morphogenesis, and growth (Table S2). Most of these GO categories have also been previously associated with mammalian and vertebrate ultraconserved elements (UCEs) [145, 17].

In total we identified 259,688 CEs in protein-coding regions, leaving 850,920 CNEs spanning over 51 Mb of the chicken genome (Table 1), with a genome-wide distribution of 92.10 CNEs/100-kb. We further observed noticeable differences in the length distribution of CEs associated with different types of annotations. Among the conserved exon-associated CEs, those found in CDSs are, on average, the longest (68 bp), followed by 3' UTRs (61 bp), RNA genes (52 bp), promoters (47 bp), and 5' UTRs (38 bp) (Fig. S4). On the contrary, CEs found in non-protein-coding regions show a homogenous length distribution, ranging from 56 bp in introns to 63 bp in lncRNAs (Fig. S5).

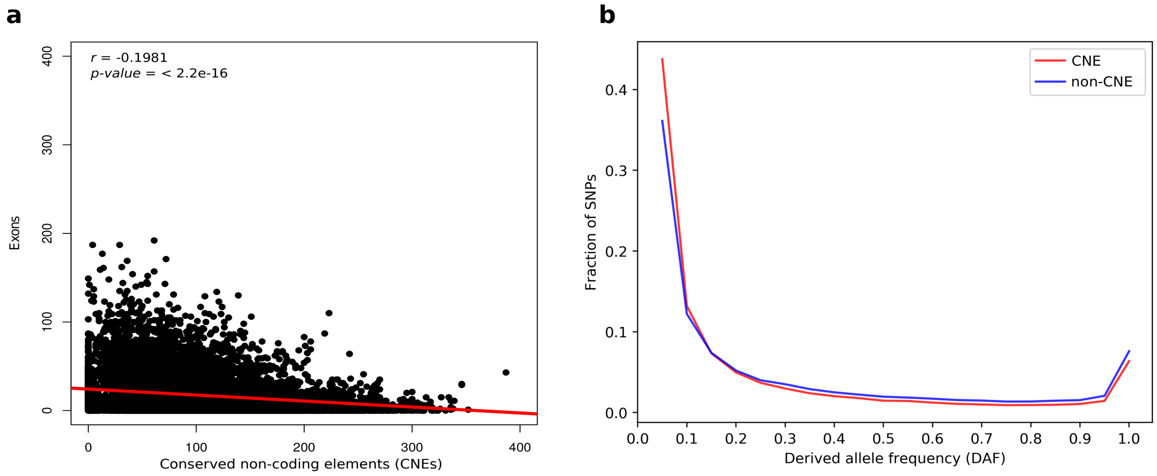
Table 1: Statistics of predicted conserved elements (CEs) based on gene annotation. The fraction of CEs per sites class is presented, for protein-coding gene annotations, in percentages of the exonic CEs (17,148,879 bp). For non-protein-coding gene annotations, the fraction is relative to the non-exonic CEs (51,224,645 bp). Abbreviations: CC, conserved coding; CNE, conserved non-protein-coding elements.

Genomic feature	No. overlapping CEs	Total overlap (bp)	Genome coverage (%)	Fraction of site class conserved (%)
CDS	213,787	14,683,183	1.38	85.62
5'UTRs	5,457	207,320	0.02	1.21
3'UTRs	23,721	1,460,144	0.15	8.51
Promoters	16,022	761,504	0.08	4.44
RNA genes	701	36,728	0.00	0.21
lncRNAs	121,840	7,696,557	0.80	15.03
Introns	328,579	18,520,675	1.93	36.16
Intergenic	400,501	25,007,413	2.60	48.82
Total CC	**259,688**	**17,148,879**	**1.78**	**100.00**
Total CNE	**850,920**	**51,224,645**	**5.33**	**100.00**

CNEs are less common in gene dense regions

We further investigated the genomic location of CNEs as this might provide important clues to their functional role. We found that the distribution of CNEs in windows of 100 kb is significantly negatively correlated (r = -0.20; *p-value*: < 2.2x10-16) with the distribution of exons (Fig. 2a). We subsequently analyzed chicken polymorphism data to address the mutational or evolutionary forces shaping CNEs, following previous studies in humans [85] and *Drosophila* [41, 19]. We used polymorphism densities to investigate whether these forces could still be acting on the chicken genome or they could have acted in other species and may no longer be relevant for chicken. SNP density, which reflects events within the chicken lineage, was calculated in the genomes of 169 chickens from different traditional breeds of divergent demographic and selection history. Specifically, we compared the SNP density found in CNEs with that in non-protein-coding elements that were identified not to be conserved (non-CNEs; i.e. not conserved intronic, lncRNA and intergenic regions), following [85, 41]. Overall, we found that the SNP density in non-CNEs (=0.02) is two-fold higher than CNEs (=0.01).

Figure 2: CNEs are less common in gene dense regions and are selectively constrained. **a** Correlation between exons and conserved non-protein-coding elements (CNEs) along the chicken genome. CNEs and exons count per 100 kb windows are shown with the Pearson correlation coefficient r and corresponding p-value in the top left corner. **b** Derived allele frequency (DAF) distribution of SNPs in CNEs and non-CNEs.



CNEs are selectively constrained in chicken

To test whether low local mutation rates in CNEs or purifying selection is responsible for the observed low SNP density, we looked at the derived allele frequency (DAF) distribution in CNEs and non-CNEs. This is because mutation rate differences are not expected to affect the allele frequency spectra. On the contrary, selective constraint is responsible for the shift in allele frequency distribution of constrained alleles towards lower values. Allele frequencies for derived (new) alleles were compiled using the sequence of the inferred ancestor between chicken and turkey. The ancestral allele was determined for a total of 9 million SNPs that passed several filtering criteria (see Methods). We observed an excess of rare ($\leq$ 10%) derived alleles of SNPs within CNEs in all chicken populations (Fig. 2b). Overall, 57% of SNPs within CNEs had a DAF $\leq$ 10%, compared to only 48% in non-CNEs (the same pattern was observed for each SNP functional class; see also Table 2). Non-CNEs displayed on the contrary a higher proportion of common SNPs (DAF $>$ 10%) (52% versus 43% within CNEs) independently of their functional class (Fig. 2b; Table 2). Therefore, the lower proportion of derived alleles in CNEs indicates that evolutionary pressure has suppressed CNE-derived allele frequencies.

Table 2: Derived allele frequency distribution for SNPs in CNEs and non-CNEs. The derived allele frequency was compiled using the sequence of the inferred ancestor between chicken and turkey. A derived allele frequency of 10% is used as a cut-off to define rare versus common variants. Information are reported for each genomic feature that make up CNEs and non-CNEs.

Genomic feature	**DAF**	**Within CNEs**	**Outside CNEs**	**chCADD within CNEs**	**chCADD outside CNEs**
		Number of SNPs (%)	**Number of SNPs** (%)	**Average (± sd)**	**Average (± sd)**
All	≤0.10	137,871 (57%)	482,685 (48.4%)	9.78 (4.18)	3.21 (3.18)
	> 0.10	103,726 (43%)	513,935 (51.5%)	8.81 (4.25)	2.74 (2.83)
lncRNA	≤0.10	24,364 (57.4%)	26,429 (47.6%)	10.02 (4.00)	3.49 (3.33)
	> 0.10	18,081 (42.5%)	29,014 (52.4%)	9.10 (4.13)	3.03 (2.99)
Intron	≤0.10	43,790 (56.8%)	159,203 (47.4%)	9.81 (4.46)	3.00 (3.11)
	> 0.10	33,171 (43.2%)	176,650 (52.6%)	8.71 (4.53)	2.46 (2.74)
Intergenic	≤0.10	69,717 (57%)	297,053 (44.6%)	9.68 (4.05)	3.31 (3.20)
	> 0.10	52,474 (43%)	308,271 (55,4%)	8.78 (4.11)	2.87 (2.86)

chCADD scores for the investigation of CNE and SNP evaluation

To investigate CNEs further, we developed a model that can evaluate individual SNPs or entire sequences based on a per-base score, with respect to its putative deleteriousness. This model is based on the CADD approach, hence it is labeled ch(icken) CADD. chCADD is a linear logistic model that is trained to differentiate between two classes of variants, one being relatively more enriched in potentially deleterious variants than the other. To obtain these two classes, one class is generated from derived variants, alleles that have accumulated since the last ancestor with turkey and became fixed or almost fixed (allele frequency > 90%) in our chicken populations. These are depleted in deleterious variants and can be assumed to be benign or at least neutral in their nature. The set of putative deleterious variants contains simulated *de novo* variants that are not depleted of deleterious variants. The feature weights obtained during training are shown in Supplementary File 1. Performance on a held out test set to determine an optimal penalization term are shown in Fig. S1.

chCADD distinguishes between potentially causal and not causal variants

We evaluated the performance and applicability of chCADD on two different sets of variants before we annotated non-coding SNPs. First, we assigned a chCADD score to all SNPs found in the genomes of the 169 chickens previously used in the SNP density and DAF analysis and compared these to functional predictions as annotated by the Ensembl VEP (Fig. S6). To this end, we categorized VEP predictions into 14 categories (Table S3). The purpose of this was to test whether chCADD correctly scores SNPs with respect to their potential to cause a deleterious or phenotype-changing effect, as indicated (mostly for protein-coding mutations) by the VEP functional predictions. We observed that mutations with a relatively large deleterious potential, such as stop-gained mutations and splice-site altering mutations, were scored higher than regular missense and synonymous mutations (Fig. S6). SNPs in potentially regulatory active regions were also evaluated to be potentially more deleterious than synonymous SNPs (Fig. S6). We performed a similar analysis considering only protein-coding and regulatory mutations found in the Online Mendelian Inheritance in Animals (OMIA) database (Table 3). We annotated only SNPs whose genomic positions were uniquely mapped to the chicken GRCg6a reference genome and the reference/alternative allele matched that in the genome assembly. Of the 15 annotated SNPs associated with a change of phenotype, 5 were reported to cause a deleterious phenotype change in the affected individual, and an average chCADD score of 27.1. These 5 variants (3 stop-gained, 2 missense) have a chCADD score above 20 and are putatively responsible for dwarfism, scaleless, analphalipoproteinaemia, muscular dystrophy, and wingless phenotypes (Table 3). All these phenotypes display a strong severity and may lead to an early death in uncontrolled environments.

Table 3: Annotation of known causative variants with the chCADD score. SNPs were obtained from the Online Mendelian Inheritance in Animals (OMIA) and their genomic position was lifted over to the GRCg6a reference genome.

OMIA ID(s)	Variant Phenotype	Gene	Type of Variant	Deleterious?	g. or m.	chCADD
OMIA 001622-9031	Resistance to avian sarcoma and leukosis viruses, subgroup C	BTN1A1	stop-gain	no	28:g.903289G>T	17.83
OMIA 000889-9031	Scaleless	FGF20	stop-gain	yes	4:g.63270401A>T	33.02

OMIA ID(s)	Variant Phenotype	Gene	Type of Variant	Deleterious?	g. or m.	chCADD
OMIA 001534-9031	Resistance to myxovirus	MX1	missense	no	1:g.110260061G>A	14.27
OMIA 000915-9031	Feather colour, silver	SLC45A2	missense	no	Z:g.10336596G>T	21.72
OMIA 000915-9031	Feather colour, silver	SLC45A2	missense	no	Z:g.10340909T>C	15.69
OMIA 000679-9031	Muscular dystrophy	WWP1	missense	yes	2:g.123014353G>A	26.30
OMIA 000303-9031	Dwarfism, autosomal	C1H12ORF23	stop-gain	yes	1:g.53638233C>T	35.29
OMIA 001302-9031	Resistance to avian sarcoma and leukosis viruses, subgroup B	TNFRSF10B	stop-gain	no	22:g.1418711C>T	17.63
OMIA 000810-9031	Polydactyly	LMBR1	regulatory	yes	2:g.8553470G>T	17.41
OMIA 000913-9031	Silky/Silkie feathering	PDSS2	regulatory	unknown	3:g.67850419C>G	3.88
OMIA 001547-9031	Wingless-2	RAF1	stop-gain	yes	12:g.5374854G>A	23.44
OMIA 000374-9031	Feather colour, extended black	MC1R	missense	no	11:g.18840857T>C	18.06
OMIA 000374-9031	Feather colour, extended black	MC1R	missense	no	11:g.18840919G>A	18.89
OMIA 000374-9031	Feather colour, buttercup	MC1R	missense	no	11:g.18841289A>C	17.41
OMIA 000374-9031	Feather colour, extended black	MC1R	regulatory	no	11:g.18840609C>T	6.74

chCADD detects evolutionary constraints within CNEs

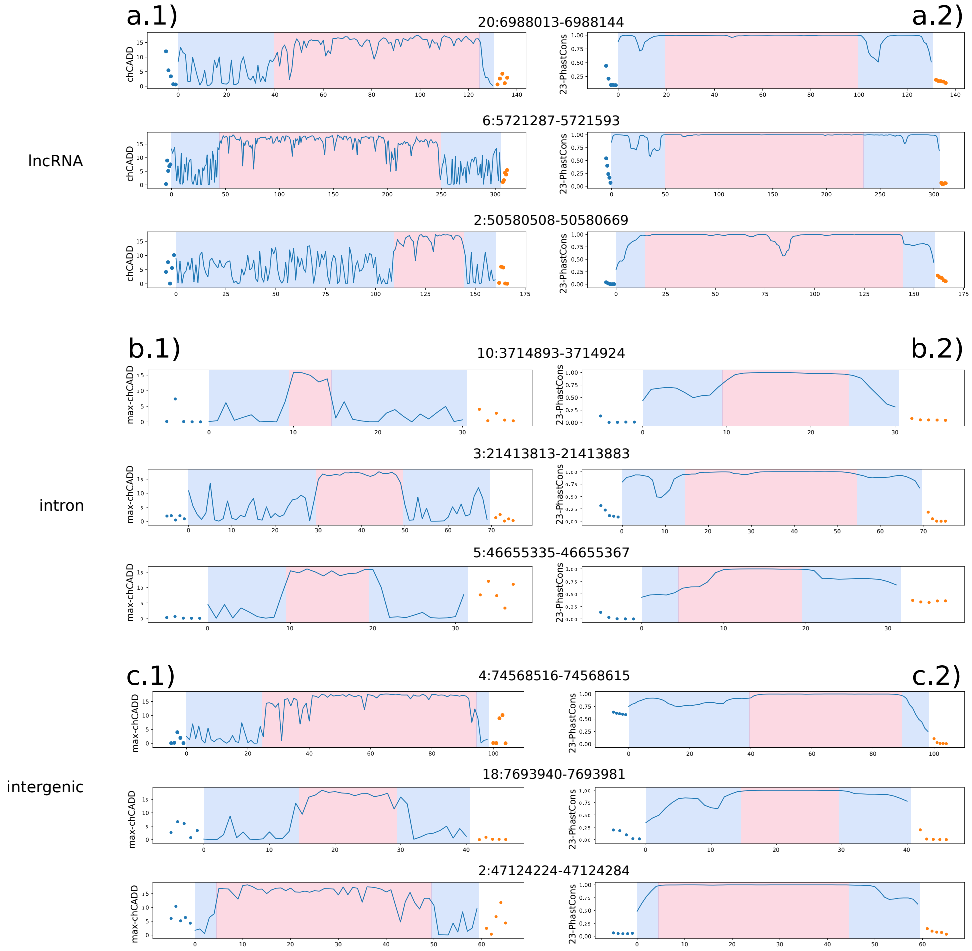
As we showed, chCADD can score functionally important protein-coding variants. We therefore decided to take a step further by annotating SNPs found in CNEs with chCADD to predict their deleteriousness and function (Table 2). We assume that highly scored SNPs can help us to identify truly functionally active regions among CNEs. We observed that rare non-protein-coding variants located within CNEs (DAF $\leq$ 10%) have an overall higher chCADD score compared to rare variants found in non-CNEs (Table 2). This result supports our previous conclusion based on the derived allele frequency spectrum that evolutionarily conserved non-protein-coding variants are likely functional. As expected, this trend was most pronounced in lncRNAs, followed by introns and intergenic regions.

We further used the chCADD score to identify specific subregions of potentially higher functional importance within each CNE, assuming that the high scoring SNPs would indicate that. We applied a change point analysis to search for a center region that has high chCADD scores as opposed to the two outer regions (see Methods). We ranked CNEs based on positive chCADD score differences between the center region and the outer regions and filtered for significant difference (*p-value* $\leq$ 0.05, t-test). The top 3 ranked CNEs that overlap with lncRNAs, intronic and intergenic regions, respectively, are shown in Fig. 3a.1, b.1 and c.1.

Analogous to this subregion analysis based on chCADD score, we performed a subregion analysis based on the 23 sauropsids PhastCons scores. Fig. 3a.2-c.2 show the identified regions for the PhastCons score for the same CNEs as Fig. 3a.1, 4c.1, respectively. These figures indicate that chCADD generates more discriminative subregions than PhastCons. Particularly interesting are the chCADD scores for the top intergenic regions (Fig.. 3c.1). The chCADD score increased from 5 to 15 at the subregion change point. This is equal to an increase of predicted deleteriousness by one magnitude, from the top 33% highest scored sites in the entire genome to the top 3%.

To further investigate the subregion partitioning of the CNEs, we computed the SNP density in each region, for both the chCADD induced regions (Fig. 4, blue bars) as well as the 23 sauropsids PhastCons induced regions (Fig. 4, orange bars). In both bases, the SNP densities of the center region are lower than those of the outer regions. Moreover, all CNE subregions display a lower density than regions up- and downstream the CNE, supporting the functional importance of the CNEs in general. Interestingly, the center regions, as identified by the chCADD score, have in general a 0.07% lower SNP density than the center regions detected using the PhastCons scores. Therefore, our findings suggest that chCADD is more effective in pinpointing potentially regions of interest.

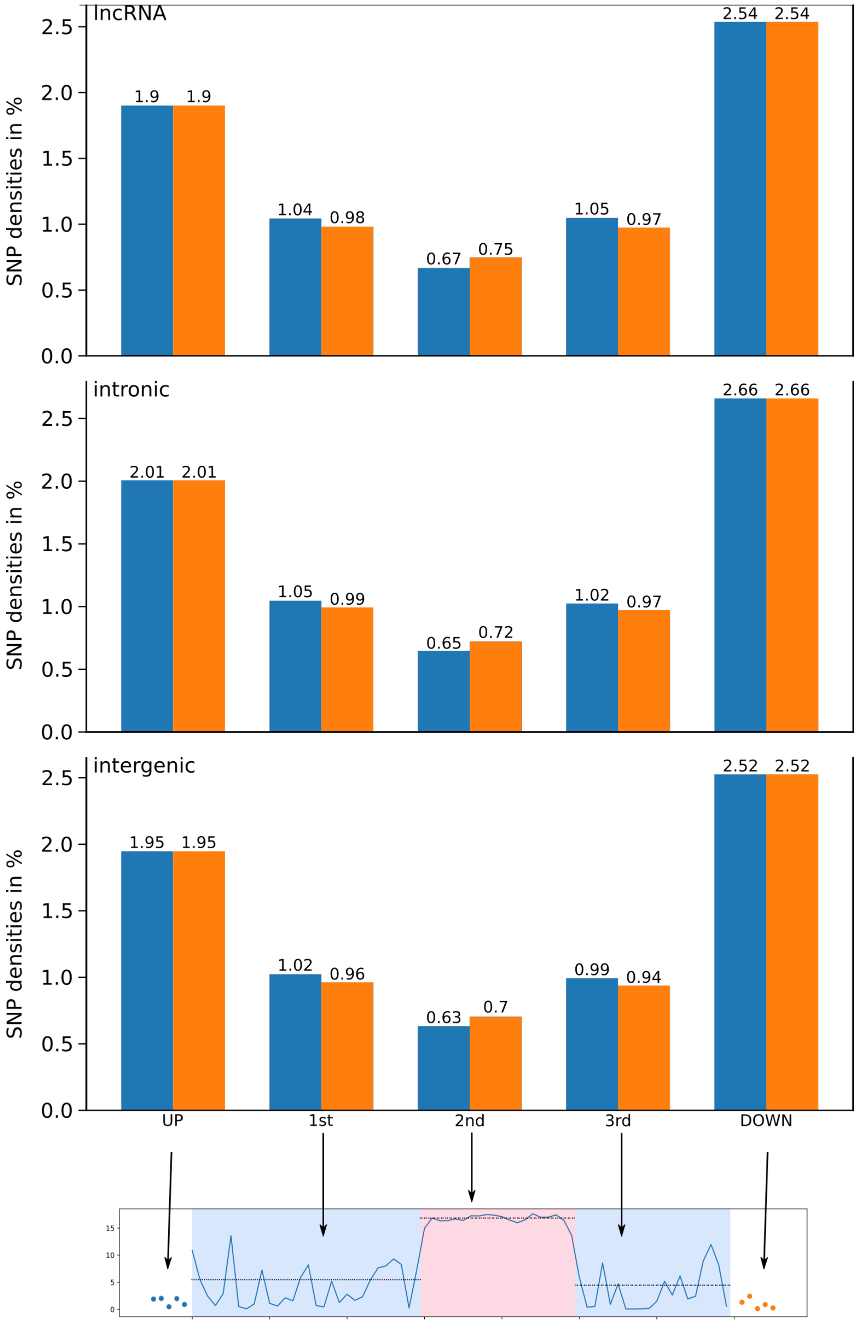
Figure 3: Change point analysis of the top 3 CNEs for each genomic feature, respectively (lncRNA, intronic, intergenic). CNEs are sorted based on the largest difference between the 2nd section and 1st or 3rd section for each of the three CNE classes respectively (lncRNA, intronic, intergenic). Change points were once computed based on maximum chADD score per site (a.1,b.1,c.1) and once on 23 sauropsids PhastCons scores (a.2,b.2,c.2). The dots in each plot display the scores for the 5 bp up- and downstream regions. The transition from blue to red background indicates the identified change points. a.1) lncRNA - maximum chCADD. a.2) lncRNA - PhastCons scores. b.1) intronic - maximum chCADD. b.2) intronic - PhastCons. c.1) intergenic - maximum chCADD. c.2) intergenic - PhastCons.



Conserved non-protein-coding subregions are detected on the basis of a limited number of genomic annotations

As part of the investigation into subregions we identified two change points, splitting each CE into three subregions, starting from 5' to 3', 1*st*-, 2*nd*- and 3*rd* subregion (Fig. 1). Next we were interested how genomic annotations that were used in the creation of chCADD, differ between the three subregions. The model coefficients with the largest weights (Table S4)

Figure 4: SNP densities computed for each section of the three different CNEs (lncRNA, intronic, intergenic). The orange bars represent the SNP densities for that section based on change points derived from 23 sauropsids alignment PhastCons scores, the blue bars represent the SNP densities based on change points identified via chCADD.



point to the importance of the PhastCons conservation scores calculated on the 4 sauropsids alignment. Other important model features are secondary structure predictions and combinations with the intronic identifier from VEP. Over all CNEs, we compared the chCADD model features, especially the conservation scores that are based on different phylogenies, excluding the chicken reference sequence in their computation. For all genomic annotations, we computed absolute Cohen's D values (standardized mean difference) [55]. We observed that the conservation scores based on the largest 77 vertebrate alignments cannot properly distinguish between the $1st$-,$2nd$- and $3rd$ subregions. Conservation scores based on smaller phylogenies (4 sauropsids and 37 amniote/mammalia) are more discriminative between these (Table S5); see columns $1st$-$2nd$, $2nd$-$3rd$).

Considering the three PhastCons scores, based on differently large phylogenies, the average absolute Cohen's D between the $1st$- and $2nd$- and the $2nd$- to the $3rd$- subregions differ less between different genomic features (intergenic, lncRNA and introns) than between genomic annotations (Table S5; see columns $1st$-$2nd$, $2nd$-$3rd$). The average absolute Cohen's D between the three subregions of a CNE ranges from 0.259 to 0.276. In comparison, the average absolute Cohen's D between the same subregions, taking the three conservation scores individually, range from 0.137 to 0.338. The effect sizes between the different multiple sequence alignment PhastCons score (i.e. 4 sauropsids, 37 amniote/mammalia, 77 vertebrates) differ by more than 2-fold.

Intronic CNEs overlap functionally important genes

Intronic CNEs were associated with genes for which we obtained phenotype annotations of their orthologs in human, mouse, and rat. We investigated the top 10 CNEs that are located in introns, with the largest p-value differences between the $1st$ and $3rd$ to the $2nd$ section. In total, 6 CNEs were associated with homologous genes that have annotated phenotypes in other species. Among the phenotypes found for human genes are mental retardation and nonsyndromic male infertility. For mouse, these included neuronal issues and abnormal shape of heart and limbs (Supplementary File 3). The link to highly severe phenotypes in other species highlights the potential importance of regulatory features for orthologous genes in chicken.

Discussion

The prediction of CNEs depend on the phylogenetic scope

Non-protein-coding elements are typically identified by sequence-level similarity across species, which is a generally applicable criterion of conservation and biological function [132]. However, when predicting CEs, and subsequently CNEs, the evolutionary distance among species included in the alignment (or phylogenetic scope) is an important parameter that can considerably affect the prediction and resolution of CEs. If the evolutionary distance among species is too narrow, the specificity of constraint is reduced, but if it is too broad, the number of CEs rapidly declines and lineage-specific conservation is lost [132, 60].

One of the first studies to address the impact of the phylogenetic scope on CEs prediction was that of Lindblad-Toh et al. (2011). In their study on the 29 mammalian multiple sequence alignment, the authors identified 3.6 million conserved elements spanning 4.2% of the genome at a resolution of 12 bp [187]. When comparing these results to a 5 vertebrate alignment, Lindblad-Toh and colleagues observed that only 45% of the 5 taxa CEs were covered by the 29 taxa alignment. The partial overlap indicates that most of the CEs derived from the 29 taxa alignment were mammalian-specific [187]. The issue resulting from a broad phylogenetic scope on CNEs has also recently been reported by Babarinde and Saitou (2016), where authors identified CNEs between chicken and four mammalian species, including human, mouse, dog, and cattle [13]. By applying a minimum length of 100 bp, Babarinde and Saitou (2016) identified 21,584 CNEs in chicken, a small number as expected from the divergence time between human and chicken occurred approximately 310 million years ago [145]. Therefore, CNEs detected among distant species are better predictions of ultraconserved CNEs than CNEs between closely related species (i.e. human-mouse) [241], as they were already present in the ancient common ancestor of the considered species.

In this study we chose the 23 sauropsids multiple sequence alignment for two reasons. First, the phylogenetic distance between crocodilian and bird species (240 million years ago) [122] is large enough to detect likely functional CNEs. Second, the alignment is reference free allowing the identification of lineage-specific CEs. Reference-free alignments should always be preferred over reference-based ones [11]. In fact, genomic regions shared within a certain clade, which would be missed in a reference-based alignment (e.g. MULTIZ), can also be detected. As a result, reference free alignments better enable the study of genome evolution along all phylogenetic branches equally.

Avian genomes have similar genomic characteristics

According to our study, 8% of the chicken genome is covered by CEs for a total of 1.14 million CEs. These results are comparable to those on the collared flycatcher genome (*Ficedula albicollis*) [61]. By means of the same alignment, Craig et al. (2018) identified 1.28 million CEs covering 7% of the flycatcher genome. The genome of many bird species is highly compact and thus small in size. Small genomes are thought to require fewer regulatory sequences involved in the organization of chromatin structure [61]. However, the similarity in genome size between, for example, chicken (i.e. GRCg6a: 1.13 Gb) and flycatcher (i.e. FicAlb1.5: 1.11 Gb), reflects the little cross-species variation characteristic of birds [322].

The limited number of CEs often identified in birds relative to mammals has repeatedly been linked to gene loss [323, 193, 156]. However, the role of gene loss in avian evolution, genome size, and prediction of CEs has recently been questioned. According to Bornelov et al. (2017), gene loss was incorrectly hypothesized from the absence of genes clustering in GC-rich regions in the earlier chicken genome assemblies [24]. In fact, these regions are often difficult to sequence and assemble. The issue is particularly prominent in the GC-rich micro-chromosomes, which, as we show, contribute disproportionately to the total density of functional sequence (Fig. S2). We therefore recommend future comparative genomics studies in chicken to make use of the most recent and complete genome assembly to avoid any erroneous link of CEs to gene loss in chicken.

Conserved non-protein-coding elements are maintained by purifying selection

A fundamental question in the study of CNEs is the role of purifying selection. Purifying selection can be discriminated from a low mutation rate by comparing the derived allele frequency (DAF) spectra in constrained regions (i.e. CNEs) with that of neutral regions (i.e. non-CNEs) [85, 19]. The rationale is that new mutations are unlikely to increase in frequency in constrained regions. Although CNEs are identified using an interspecific comparative genomic approach, the evolution and dynamics of these regions are generally analyzed at an intraspecific scale by looking at polymorphism data [85, 273]. In this study, we showed that the evolutionary constraint acting on the 23 sauropsids is correlated with constraint within the chicken populations, as assessed from chicken polymorphism data. Consistent with studies in humans [85, 187], plants [134], and *Drosophila* [19, 41], the derived allele frequency spectra of our chicken populations is shifted towards an excess of rare variants in CNEs. These results indicate that the conservation of CNEs in the chicken genome is mainly driven by selective constraints, and not by local variation in mutation rate. The role of purifying selection was also confirmed by the reduced SNP density in CNEs compared to non-CNEs and by the reduced SNP density in specific conserved non-protein-coding subregions. The concordance in SNP density is a clear indication of reduced levels of population diversity and functional roles of CNEs as confirmed by the association of subregions within CNEs to highly severe phenotypes in humans, mouse, and rat. However, future population diversity comparisons in terms of nucleotide diversity (π) [221] or Watterson's estimator (θw) [310] between outbred and inbred populations would further elucidate our understanding of purifying selection in CNEs.

Integrating comparative and functional genomics into a single score

We developed a ch(icken) Combined Annotation-Dependent Depletion (chCADD) approach that provides scores for all SNPs throughout the chicken genome. These scores are indicative of putative SNP deleteriousness and can be used to prioritize variants. The annotation of chCADD relies on the combination of a diverse set of genomic features, including evolutionary constraints and functional data [251, 168]. Multiple sequence alignments of distantly related species are better suited to differentiate conserved sites that can reliably be used to identify functionally important regions. However, these regions are often large enough to question the functional role of the entire region. Our findings show that chCADD outperforms any conservation-based method alone (e.g. PhastCons) in the identification of functionally important subregions within CNEs. Therefore, methods, such as chCADD, are required to fine-tune in one step CNEs to identify subregions directly linked to - in some cases deleterious – phenotypes.

According to the authors of the original human CADD [168], SNPs with a score above 20 (i.e. the SNP is among the top 1% highest scored potential SNPs in the genome) could be considered deleterious. This means that the higher the score, the higher the chance the variant has a functional effect or may even be deleterious. When annotating protein-coding and regulatory mutations found in OMIA, we observed that SNPs with a chCADD score of 15 can already be considered functional. Therefore, our findings indicate that by setting an arbitrary threshold of 20 may underestimate the fraction of the genome that is actually functional. This is particularly pronounced when the variants in question are located outside protein-coding regions. Therefore we recommend future chCADD users to evaluate the variants identified in their populations to see if they are particularly highly scored compared to other variants in the same genomic region. Further, the signal to order SNPs of interest is obtained over evolutionary timescale, which means that mutations that would have been deleterious for chicken in the past may not be deleterious for chicken in a commercial environment and vice versa. chCADD is able to support the order SNPs with respect to their potential interest but for final economical evaluations, further information about each investigated SNP may be required.

Future uses of chCADD

The high scoring of non-protein-coding variants in subregions of CNEs has important implications for future functional and genome-wide association studies (GWAS) in chicken. A very large fraction of trait- or disease-associated loci identified in GWAS are intronic or intergenic. This is expected considering the preponderance of non-protein-coding SNPs on genotyping arrays [5] or along the genome. However, because of a lack of understanding of the function of non-protein-coding mutations, most of the causal mutations reported in the OMIA database are coding. Moreover, in the presence of non-protein-coding mutations, many studies stop at the general locus or - understandably - assume that the closest neighboring gene is affected. However, these assumptions on genomic distance are simplistic. Our findings in chicken demonstrate that chCADD can accurately pinpoint non-protein and protein-coding variants associated with important phenotypes in chicken. Therefore, we expect future genome-wide association studies combined with chCADD to identify novel causal mutations or substantially narrow down the list of potential causal variants in large quantitative trait loci (QTLs). We also expect chCADD to accelerate the discovery and understanding of the biology and genetic basis of phenotypes.

Conclusion

Deciphering the function of the non-coding portion of a species genome has been a challenging task. However, the availability of genomes from a great variety of species, along with the development of new computational approaches at the interface of machine learning and bioinformatics, has made this task possible in model and non-model organisms. Our findings indicate that an accurate assessment of selective pressure at individual sites becomes an achievable goal. We have also shown that chCADD is a reliable score for the analysis of non-protein-coding SNPs, which should be targeted along with protein-coding mutations in future genome-wide association studies. We therefore anticipate chCADD to be of great use to the scientific community and breeding companies in future functional studies in chicken.

Data access

Raw sequences of the 169 individuals used in this study are available at the European Nucleotide Archive under accession number PRJEB34245 and PRJEB36674. chCADD scores partitioned per chromosomes can be downloaded from the Open Science Framework project page (https://osf.io/d6wxp/).

Disclosure declaration

The authors declare that they do not have any conflict of interest.

Acknowledgements

We would like to thank Ole Madsen and Martijn Derks for the useful discussion. C.G. was funded by the TTW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality. C.B was funded by the European Union´s Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353 (Innovative Management of Animal Genetic Resources – IMAGE). M.B. is financially supported by the NWO-VENI grant no.016.Veni.181.050.

Supporting information

Figure S1: Model performances measured in Receiver Operator Area under the Curve (ROC-AUC) and log-loss for three different ridge penalization terms (0.1, 1.0, 10.0). The scale is adjusted to make the differences between the models visible. Penalization of 1 was selected due to the lowest log-loss and largest ROC-AUC.

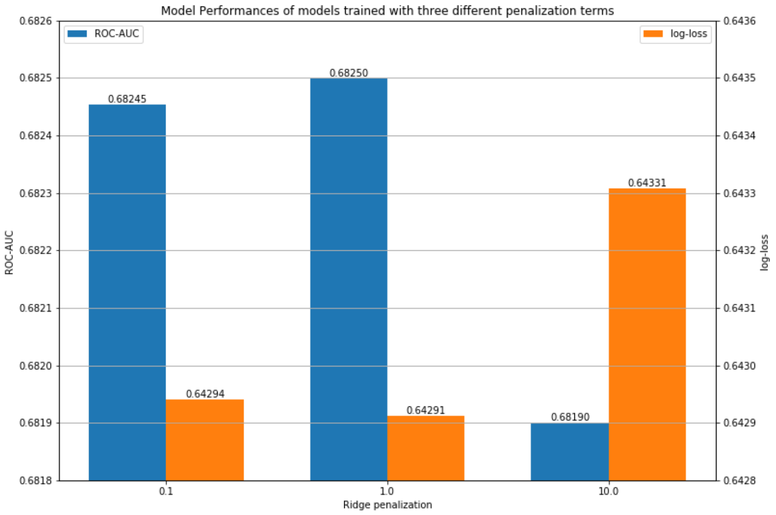


Figure S2: Distribution of conserved elements (CEs) along the chicken genome. The barplot displays the fraction of the genome per chromosome covered by conserved elements.

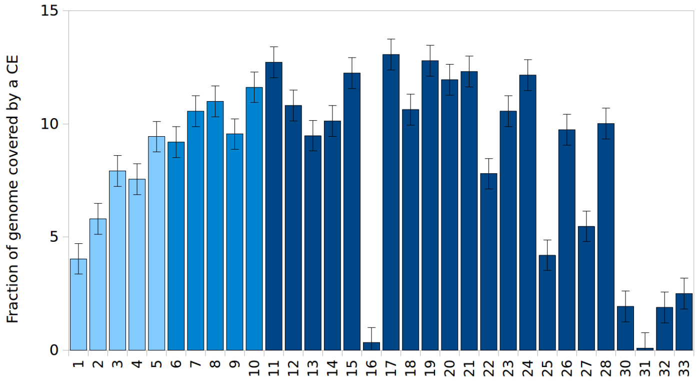


Figure S3: Frequency size distribution of predicted conserved elements. The y-axis shows the frequency, while the x-axis the size in base pairs (bp) of the predicted conserved elements.

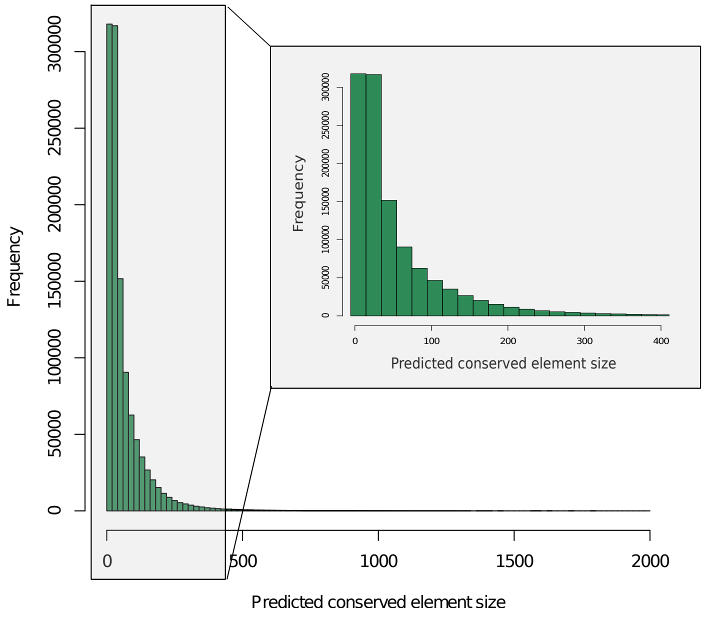


Figure S4: Frequency size distribution of predicted conserved elements overlapping exonic-associated gene annotations. The exonic-associated conserved elements include CDS, 5'UTR, 3'UTR, and promoter regions.

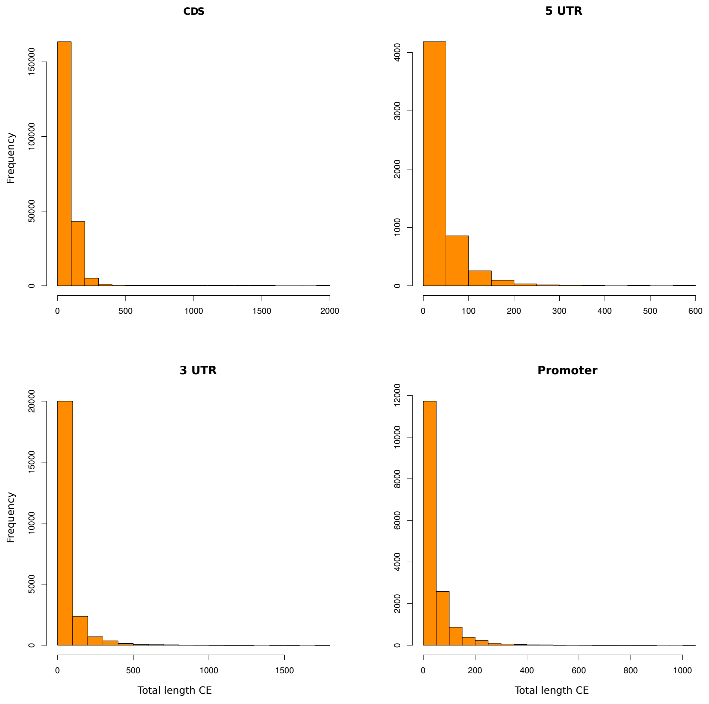


Figure S5: Frequency size distribution of predicted conserved elements overlapping non-protein-coding gene annotations. The non-protein-coding gene annotations include introns, lncRNA, and intergenic regions.

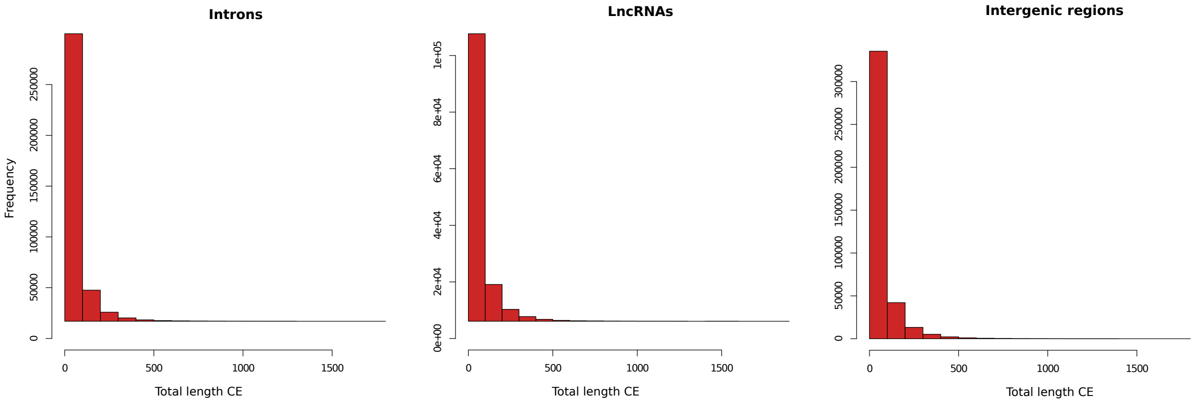


Figure S6: chCADD score distribution of SNPs per VEP cateogy. SNPs from the 169 chickens are categorized based on the VEP categories reported in Table S2. SG: Stop-gained; CS: Canonical Splice; NS: Non-Synonymous; SN: Synonymous; SL: STOP-Lost; S: Splice Site; U5: 5'-UTR; U3: 3'-UTR; IG: Intergenic; NC: Noncoding-change; I: Intronic; UP: Upstream; DN: Downstream; O: Other. The label indicates the category and the number of SNPs falling into that category.

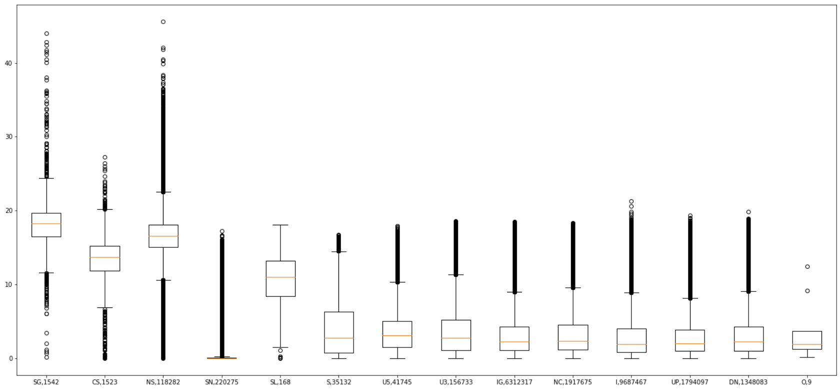


Table S1: List of annotations which form the set of descriptive features for which model weights are learned. Missing values are imputed via the specified values. Annotations of the type (factor) are OneHotEncoded and combinations between annotations form the final feature set.

Annotation label	Data type	Imputed value	Annotation description
Ref	factor	-	Reference allele
Alt	factor	-	Observed allele
isTv	bool	0.5	Is transversion?
Consequence	factor	-	VEP Consequence summaries
GC	num	0.4	Percent GC in a window of +/- 75bp
CpG	num	0.02	Percent CpG in a window of +/- 75bp
motifECount	int	0.0	Total number of overlapping motifs
motifEHIPos	bool	False	Is the position considered highly informative for an overlapping motif by VEP
motifEScoreChng	num	0.0	VEP score change for the overlapping motif site
Domain	factor	UD	Domain annotation inferred from VEP annotation (ncoils, tmhmm, sigp, lcompl, ndomain = "other named domain")
Dst2Splice	int	0.0	Distance to splice site in 20bp; positive: exonic, negative: intronic
Dst2SplType	factor	UD	Closest splice site is ACCEPTOR or DONOR
oAA	factor	UD	Amino acid of observed variant
nAA	factor	UD	Reference amino acid
Grantham	int	0.0	Grantham score: oAA,nAA
SIFTcat	factor	UD	SIFT category of change
SIFTval	num	0.0	SIFT score
cDNApos	int	0.0	Base position from transcription start
relcDNApos	num	0.0	Relative position in transcript
CDSpos	int	0.0	Base position from coding start
relCDSpos	num	0.0	Relative position in coding sequence
protPos	int	0.0	Amino acid position from coding start
relProtPos	num	0.0	Relative position in protein codon
dnaRoll	num	0.23	Predicted local DNA structure effect on dnaRoll
dnaProT	num	0.68	Predicted local DNA structure effect on dnaProT
dnaMGW	num	0.03	Predicted local DNA structure effect on dnaMGW
dnaHelT	num	-0.12	Predicted local DNA structure effect on dnaHelT
GerpS	num	-0.17	Rejected Substitution' score defined by GERP++
GerpN	num	0.64	Neutral evolution score defined by GERP++
GerpRS	num	0.0	Gerp element score
GerpRSpval	num	1.0	Gerp element p-Value
4PhCons_noChick	num	0.17	4-taxa-sauropsids PhastCons score (excl. chicken)
37PhCons_ noChick	num	0.13	37-taxa-Amniota PhastCons score (excl. chicken)
77PhCons_ noChick	num	0.2	77-taxa-Vertebrate PhastCons score (excl. chicken)
4PhyloP_ noChick	num	0.07	4-taxa-sauropsids PhyloP score (excl. chicken)
37PhyloP_ noChick	num	0.04	37-taxa-Amniota PhyloP score (excl. chicken)
77PhyloP_ noChick	num	0.25	77-taxa-Vertebrate PhyloP score (excl. chicken)
minDistTSS	int	10000000	Distance to closest Transcribed Sequence Start (TSS)
minDistTSE	int	10000000	Distance to closest Transcribed Sequence End (TSE)
interaction-score	num	0	Interaction score from Hi-C interaction maps
Exp-score	int	0	RNA expression scores
Exp-pval	num	1	p-Value of RNA expression scores
Exp-logFC	num	0	Log-Fold change of RNA expression
OChrom-Peaknb	Int	0	Read number for open Chromatin; ATAC-seq
OChrom-pval	num	1	p-Value for open chromatin; ATAC-seq
OChrom-logFC	num	0	Log-Fold change for ATAC-seq

Table S2: GO term enrichment analysis of exonic-associated CE and intronic CEs.

Term ID	Term description	Target size	3' UTR				Intron			
			Term size	Query size	Overlap	p-Value	Term size	Query size	Overlap	p-Value
GO:0048856	Anatomical structure development	12,514	3,293	4,736	1,475	1.24x10^{-17}	3,293	6,971	2,128	1.09x10^{-29}
GO:0010646	Regulation of cell communication	12,514	2,038	4,736	917	3.67x10^{-09}	2,038	6,971	1,329	1.33^{-17}
GO:0010604	Positive regulation of macromolecule metabolic process	12,514	2,118	4,736	952	1.49x10^{-09}	2,118	6,971	1,331	2.21x10^{-09}
GO:0023051	Regulating of signaling	12,514	2,056	4,736	926	2x10^{-09}	2,056	6,971	1,339	1.88x10^{-17}
GO:0048583	Regulation of response to stimulus	12,514	2,332	4,736	1,032	1.44x10^{-08}	2,332	6,971	1,477	9.79x10^{-13}
GO:0048468	Cell development	12,514	1,364	4,736	625	1.27x10^{-06}	1,364	6,971	927	1.12x10^{-18}
GO:0031325	Positive regulation of cellular metabolic process	12,514	2,091	4,736	936	9.01x10^{-09}	2,091	6,971	1,304	1.09x10^{-07}

Term ID	Term description	Target size	CDS				5' UTR			
			Term size	Query size	Overlap	p-Value	Term size	Query size	Overlap	p-Value
GO:0048856	Anatomical structure development	12,514	3,293	9,703	2,713	2.06x10^{-11}	3,293	1,896	654	5.13x10^{-14}
GO:0010646	Regulation of cell communication	12,514	2,038	9,703	1,686	2.64x10^{-06}	2,038	1,896	381	9.33x10^{-03}
GO:0010604	Positive regulation of macromolecule metabolic process	12,514	2,118	9,703	1,749	3.53x10^{-06}	2,118	1,896	403	5.06x10^{-04}
GO:0023051	Regulating of signaling	12,514	2,056	9,703	1,699	4.46x10^{-06}	2,056	1,896	384	9.24x10^{-03}
GO:0048583	Regulation of response to stimulus	12,514	2,332	9,703	1918	5.55x10^{-06}	2,332	1,896	424	4.39x10^{-02}
GO:0048468	Cell development	12,514	1,364	9,703	1,142	1.78x10^{-05}	1,364	1,896	282	3.38x10^{-05}
GO:0031325	Positive regulation of cellular metabolic process	12,514	2,091	9,703	1,723	1.91x10^{-05}	2,091	1,896	388	1.60x10^{-02}

Table S3: VEP consequences summarized in 14 categories. If multiple annotations exist for the same variant, the consequence is selected according to the displayed hierarchy, starting at 1 and ending at 14.

Hierarchy	Abbreviation	VEP Consequence
1	SG	Stop Gained
2	CS	Canonical Splice
3	NS	Non-Synonymous
4	SN	Synonymous
5	SL	Stop Lost
6	S	Splice Site
7	U5	5'-UTR
8	U3	3'-UTR
9	IG	Intergenic
10	NC	Noncoding-change
11	I	Intronic
12	UP	Upstream
13	DN	Downstream
14	O	Unknown / Other

Table S4: Top 10 model features with the largest assigned weight and their explanations.

Label	Model weight assigned to feature	Feature explanation
GerpS	0.152568	GERP rejected substitution score
4PhCons_noChick	0.28726	4-sauropsids PhastCons scores (excluding chicken)
I_GerpS	0.109099	GERP rejected substitution score for intronic sites
I_4PhCons_noChick	0.0899441	4-sauropsids PhastCons scores (excluding chicken) for intronic sites
dnaProT	0.083813	DNA secondary structure prediction for ProT
77PhCons_noChick	0.0790709	4-amniota PhastCons scores (excluding chicken
dnaRoll	0.0733429	DNA secondary structure prediction for Roll
IG_4PhCons_noChick	0.067539	4-sauropsids PhastCons scores (excluding chicken) for intergenic sites
I_dnaProT	0.0671401	DNA secondary structure prediction for ProT for intronic sites
IG_GerpS	0.0635293	GERP rejected substitution score for intergenic sites

Table S5: Differences between genomic annotations utilized for the chCADD model. Differences are measured in absolute Cohen's D between the different subregions in which each CNEs was subdivided in the change point analysis.

Intronic	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	594	307	361	609
37PhastCons	446	328	369	448
77PhastCons	1.25	96	195	1.32
4PhyloP	0.43	0.09	126	428
37PhyloP	351	187	214	0.35
77PhyloP	776	186	237	778
GerpS	272	182	196	257
GerpN	212	112	0.11	214
dnaMGW	103	9	7	104
dnaProT	0.08	13	12	0.08
dnaHelT	82	2	2	83
GC	121	45	47	0.12
CpG	34	34	34	34
OChrom-Peaknb	58	1	91	15
OChrom-logFC	62	87	138	17
OChrom-pval	6	13	70	55
LncRNA	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	608	289	338	623
37PhastCons	469	0.31	342	482
77PhastCons	1.29	86	184	1.37
4PhyloP	428	83	117	0.43
37PhyloP	343	161	0.18	348
77PhyloP	788	0.17	0.22	792
GerpS	267	0.17	181	259
GerpN	212	86	98	201
dnaMGW	97	6	8	95
dnaProT	96	9	9	93
dnaHelT	89	3	0.0	86
GC	114	37	41	109
CpG	24	33	29	28
OChrom-Peaknb	59	-0.02	64	23
OChrom-logFC	102	93	137	55
OChrom-pval	12	96	103	5

Intergenic	UP-1st	1st-2nd	2nd-3rd	3rd-Down
4PhastCons	0.61	281	341	619
37PhastCons	474	319	359	481
77PhastCons	1.29	84	179	1.37
4PhyloP	431	84	119	432
37PhyloP	351	162	185	351
77PhyloP	0.79	167	215	795
GerpS	0.29	169	183	274
GerpN	209	91	88	215
dnaMGW	96	8	8	96
dnaProT	97	14	12	96
dnaHelT	86	3	2	84
GC	136	62	62	136
CpG	39	37	36	41
OChrom-Peaknb	17	4	0.02	5
OChrom-logFC	89	5	12	77
OChrom-pval	0.00	5	52	23

7.

Discussion

7.1 Introduction

Our world is currently facing a rapid decline in biodiversity. This trend, which appears to be unstoppable, calls for active transformative changes to overtake the current insufficient global response. However, for that to be sustainable in the long-term, biodiversity must be characterised from a demographic and functional perspective. With this thesis, I aim to provide insights into how demography shapes the genetic variation landscape of a species genome, using chicken as a model. Furthermore, I focus on which genetic factors influence the pattern of variation within individual genomes, and how this variation can be interpreted in terms of functionality. I emphasise how the knowledge on deleterious variation can support appropriate conservation practices that aim to minimise genetic erosion. I also discuss how genomics-based information on demography and functional variation can be incorporated in current and future conservation programmes of endangered populations. In the final part, I highlight how novel genomic approaches are changing the study of functional variation in a species genome. In this concluding chapter, I further elaborate on my findings presented in Chapter 2-6 and put them in a broader perspective.

The work described in this thesis aims to provide a better understanding with regards to *how* demography and selection shape an individual's genome variation. Understanding which factors shape genomic variation is a central question in evolutionary genomics. The generation or elimination of genetic variation from the genome is determined by an interplay between intra-genomic features and external forces. Although important, the intra-genomic determinants of variation were not investigated in this thesis. For a comprehensive overview on the topic, I direct the reader to a review written by Ellegren and Galtier (2016) [93]. In the following paragraph, I mainly discuss the role of external forces, and in particular the effects of population bottlenecks, genetic drift, and inbreeding on genomic variation.

7.2 Demography and diversity inferred from the genome

Genetic diversity is an essential pillar in conservation biology: without diversity, evolution can not occur and species can not adapt to changing environments. From a theoretical perspective, genetic diversity can be viewed as reflecting the balance between the generation and removal of genetic variants (alleles). At each generation, spontaneous mutations resulting from DNA replication errors or mutagen-induced DNA damages contribute to the generation of new alleles [93]. However, the rate of mutation is not constant across the genome [146] and among species [196]. Hence, as new mutations do not equally rise in the genome, genetic diversity is also not expected to be equally distributed. The generation of new alleles can be opposed by external genetic forces, such as genetic drift and selection, which can cause the loss or fixation of alleles. Genetic diversity can nonetheless increase through processes, such as hybridization, although elevated nucleotide diversity is mostly found in regions that represent mixed origins.

Genetic diversity is clearly determined by intrinsic genomic features and external genetic factors that act in unison. The extent to which each factor influences the level of variation in individual genomes is now possible to estimate thanks to the generation of high-throughput sequencing data. As I describe, a new set of questions are on the horizon. Genomics-based conservation research is offering an unparalleled opportunity to preserve the rapidly declining world's biodiversity. However, for it to be effective, conservation itself has to change and be more open to a smarter use of genomic data [297].

For a long time, studying genetic diversity was limited to few neutral molecular markers, including microsatellites and (low-to-medium density) SNP arrays. In livestock, the use of pedigrees was a cornerstone to selective breeding programmes of commercially relevant breeds [98, 96]. However, the use of pedigree rarely caught on in local livestock breeds, as these breeds are not always under coordinated management [26]. Similarly, in wild species detailed and careful monitoring of individuals is rarely carried out because of the practical difficulties involved in inferring genetic relationships in the population [133]. Traditional chicken breeds from the Netherlands are an interesting study system with regards to genetic diversity. Although pedigree information could help in the management of these breeds of small effective population size, pedigree information are not collected. Therefore, in these breeds misinformed population management is an important threat to the long-term conservation of their genetic diversity.

In **Chapter 2**, I use the 60K SNP array to characterise the genetic diversity, demographic history, and level of inbreeding of local Dutch chicken breeds. I show that the genetic diversity displayed by Dutch heritage breeds reflects their divergent demographic history and meta-population structure. However, I also observe that all breeds are facing a dramatic increase in inbreeding level, which mostly results from long ROHs. The recently created neo-bantams are the most affected by high levels of inbreeding because of back-crossing pursued for phenotype selection [26]. Findings presented in Chapter 2 demonstrate that SNP arrays can provide genetic and demographic information that have relevant practical implications for the conservation and effective management of farm animal genetic resources (FAnGR). Therefore, in the absence of better genetic data, SNP arrays can support, with limited costs, immediate management practices, which are beneficial for breeds or populations of small size. In Chapter 2 I also show that mean inbreeding estimated from neutral variation as nonrandom mating (F_{is}) [162] is an important population parameter to identify breeds and populations at higher risk of inbreeding depression. Inbreeding depression is expected in all populations where inbred individuals are found [160]; this would imply that species currently threatened by a small population size are expected to equally suffer from inbreeding depression. However, as I discuss in **section 7.2.1**, the degree to which inbreeding depression is a problem to conservation varies with a population demography.

7.2.1 Genomic data allow to ask qualitatively different questions

Changes in N_e are expected to deeply affect genetic diversity [3]. However, different demographic histories can result in similar genetic variation. For instance, a recently bottlenecked population displays reduced genome-wide heterozygosity [1, 257] similarly to a population that has been small for a long time [256, 159]. Nevertheless, populations that recently experienced a bottleneck are more likely to suffer from stronger inbreeding depression than those that have been small for many generations [47]. Although differences in genetic diversity between small populations are difficult to be detected from SNP arrays, next-generation sequencing technologies can help us in this task. Next-generation sequencing technologies have revolutionised our way of studying and, particularly, looking at genetic diversity. For instance, we now know that in recently bottlenecked populations, genetic diversity is not homogeneously distributed along the genome, while in populations that have been small for a long time genetic diversity is equally distributed.
With the increasing amount of the genome interrogated, researchers can now ask *qualitatively* different questions on, for instance, *which* evolutionary processes and *how* demography shape genetic diversity. Therefore, the simple comparison of mean genome-wide heterozygosity among breeds or populations is no longer sufficient to prevent future losses of genetic variation.

According to the prevailing paradigm in conservation biology, conservation of populations threatened with extinction should focus on the maintenance of high genetic diversity [42], which can be monitored through specific indicators such as N_e [175]. Small and isolated populations are expected to benefit from an increase in genetic diversity, as they face two major threats. First, genetic drift, because the random fixation or loss of alleles indirectly erodes the quantitative genetic variation necessary for adaptation [162]. Second, inbreeding, because mating among relatives increases homozygosity for alleles identical-by-descent (IBD), resulting in fitness reduction [159]. The conservation biology paradigm assumes that fitness is mainly determined by genetic diversity; in other words, as long as genetic variation is present in a population, the effects of inbreeding are expected to be minimal.

Chapter 3 challenges this paradigm. As also observed in the wild [256, 255], populations that have been kept small for a long period of time (i.e. large fowl breeds) have had the time to purge strongly deleterious variants. On the contrary, in recently bottlenecked populations (i.e. neo-bantam breeds) the accumulation of weakly deleterious alleles has been driven by genetic drift, which overruled selection against them. Our study indicates that populations at small size can persist if natural selection is efficient at eliminating harmful mutations. In turn, the ability of natural selection depends on the type and time frame in which the population bottleneck occurs [25].

In their recent simulations, Kyriazis et al. (2019) showed that the extinction risk in small populations often increases when genetic diversity is maximised, while it decreases when deleterious variation is minimised [174]. The key driver of such behaviour is not the *current* population size, but the *ancestral* population size. In plants and animals, domestication had dramatic effects on the ancestral N_e and selection pressure. The domestication bottleneck has been suggested to have substantially increased the mutation load of several species (e.g. dog [62, 200], horse [263], rice [188]), a phenomenon that we commonly refer to as the 'legacy of domestication' [62] or 'cost of domestication' [214]. Of the Dutch heritage breeds, large fowls are the direct descendants of the ancestral population (Figure 1). Therefore, domestication has not only affected the mutation load of these breeds, but has also contributed to their initial genetic erosion. Over time, lethal variants were purged, while the frequency of slightly deleterious mutations increased, as natural selection became less effective. Genetic diversity of large fowls has progressively been restored [26], resulting in an overall higher heterozygosity when excluding ROHs (see Figure 2b, Chapter 3) [25]. Genetic rescue of neo-bantams through large fowls can be viewed as a reasonable practice considering that large fowls are one of the source populations used to create neo-bantam breeds (Figure 1) [26]. However, in line with simulations [174], the deleterious variation landscape of large fowls does not support this management strategy. Conservation programmes established with the support of genomic data are far more promising (**section 7.5**). In fact, neo-bantam individuals with high genetic diversity, but low levels of deleterious variation, can be identified and used for breeding.

The ancestral demography of a species is becoming central to the conservation of small and isolated populations. The availability of an increasing number of sequences from ancient genomes is also assisting us in this task, as temporal data can more accurately quantify demographic changes [80]. The relevance of genomics-based information are well recognised and are increasingly becoming feasible to obtain at reasonable costs. However, global authorities in charge of assessing a species risk status still base their assessment on current population figures. As also recently highlighted by van Oosterhout (2020), biodiversity can only be safeguarded if conservation becomes genomics-informed. For domesticated species under direct human management, the integration of genomic information into the risk status of a breed is 'simplified' by the absence of a strong environmental component. However, for wild species, the fate of a population is highly stochastic, because genetic factors can substantially vary even under the same ecological conditions. Therefore, genetic and environmental stochasticities represent the greatest challenge for genomics-based management efforts in wild species.

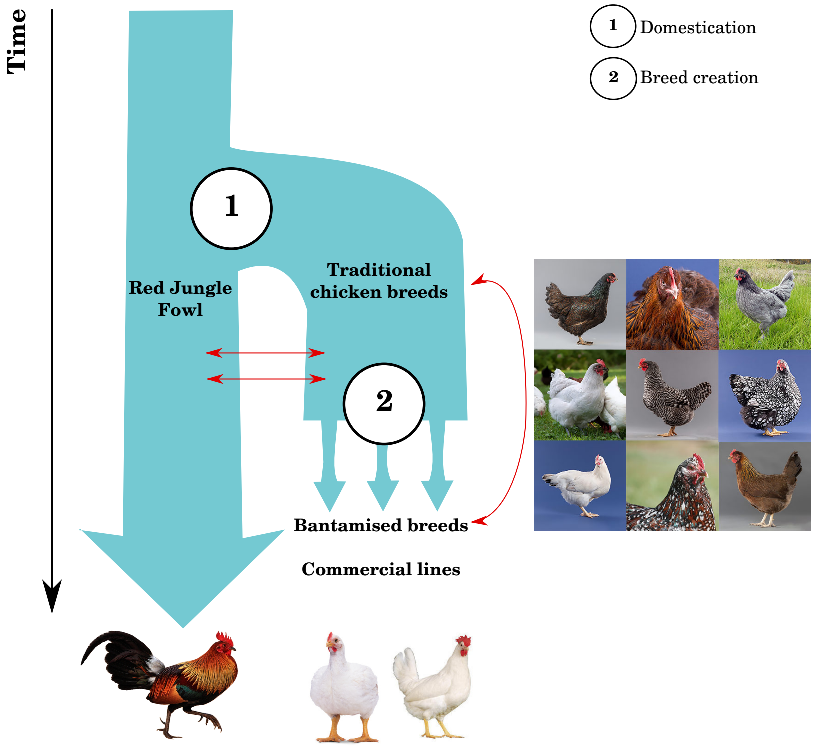


Figure 1: Domestication and breed formation bottleneck. (1) Traditional chicken breeds are the direct descendants of the ancestral wild chicken population here shown, for simplicity, by the red jungle fowl alone. As the red arrows indicate, domestication was a continuous, change process, during which complete isolation rarely occurred, resulting in opportunities for genetic exchange. (2) Commercial lines were created from mainly dual purpose traditional breeds through a second bottleneck (i.e. breed creation). Bantamized breeds were also created through this bottleneck by crossing large fowls with true bantams (red arrow) [26]. Genetic exchange between neo-bantams and the source populations is still on-going.

7.3 Deleterious alleles and inbreeding depression

Individual genomes harbour genetic variation of various fitness effect, meaning that mutations have relevant functional implications for conservation. Deleterious alleles represent the most important threat to a population survival. However, small populations are expected to be the most affected because selection is largely ineffective when N_e is small. In the next paragraphs, I emphasize how deleterious mutations can be used as a proxy to estimate genomic fitness. Moreover, I discuss the role of (recessive) deleterious variants in determining inbreeding depression.

7.3.1 Deleterious alleles as a proxy to estimate genomic fitness

At each generation, deleterious mutations appear with a rate equal to the per genome deleterious mutation rate [48]. At the same time, deleterious mutations are removed from an individual's genome through natural selection. Deleterious mutations are purged because they cause a reduction in the fitness of individuals carrying them. The negative fitness effects of deleterious alleles are dependent on the population evolutionary history and selection history [47]. Functional tests aiming to measure fitness-related traits in individual genomes (e.g. fertility) are the best approach to unambiguously detect and correctly classify deleterious alleles [80]. However, when functional tests are not possible, deleterious alleles identified by large-scale sequencing data can be used as a proxy to accurately estimate individual genomic fitness [140].

Genomic fitness is defined as the ratio between the number of deleterious variants and the number of synonymous variants [29, 79]. Variants found in protein-coding genes have extensively been used to estimate genomic fitness (e.g. chicken [79], pig [29], dog [200]). However, by focusing on protein-coding variants, we interrogate a small fraction of the genome. The advantage of focusing on protein-coding genes is that alterations in the translated protein sequence involve phenotypic changes that, because observable, can be directly assessed. However, not all protein-coding deleterious mutations have an equal fitness effect. In their study on the impact of long-term population decline and inbreeding, Xue et al. (2015) observed fewer loss-of-function mutations in Eastern gorillas (*Gorilla beringei*) than Western gorillas (*Gorilla gorilla*). At the same time, genomes of Eastern gorillas had higher accumulation of missense mutations than those of Western gorillas [318]. Loss-of-function (LoF) mutations, such as stop-gain, stop-loss, and frame-shift, have larger fitness effects than missense mutations [119], because they can disrupt gene function and the generation of a fully functional protein. Hence, because of the larger fitness effects, purging is more efficient against LoF mutations in small and inbred populations (e.g. *G. beringei*) than in populations of larger size (e.g. *G. gorilla*). It is however important to remember that harmful mutations that are recessive are often masked in large populations.

Quantifying genomic fitness, or changes in genomic fitness, is an important first step to determine how common natural selection is against deleterious alleles. However, genomic fitness must always be contextualised within a population demography. In large populations, high genomic fitness does not represent a direct threat, as long as recessive mutations are maintained in a heterozygote state. However, this ceases when historically large populations start to decline, as inbreeding exposes recessive mutations in homozygous state. At this evolutionary stage, key is the ability of a population to survive inbreeding depression. As homozygosity increases, some recessive deleterious mutations are purged because exposed by inbreeding, while others are lost due to genetic drift. Although some bottlenecked populations may appear to have become resistant to inbreeding, genetic drift may continue to affect an individual's fitness [297]. Because of genetic drift, genomic fitness alone may not be a reliable indicator of extinction risk [291].

7.3.2 Evolutionary genomic constraints and genomic fitness

Understanding how much of the genome is under selection and which mutations underlie phenotypic variation in species are long-standing research questions in population genetics. Protein-coding variants have provided relevant information on the role of population history at influencing deleterious variation. However, in our efforts to monitor and conserve genetic diversity through the estimation of genomic fitness, we have ignored a large component of the genome that can also contribute to variation [297].

In **Chapter 6**, I show that 8% of the chicken genome has remained conserved over millions of years of evolution since the species diverged from anole lizard [126]. Highly conserved elements (HCEs) show a deficit of substitutions across evolutionarily distant taxa. Because of the deficit of substitutions, sites are conserved (or identical) across disparate species [7] and are thus thought to be biologically functional. Strong purifying selection is responsible for the level of conservation of HCEs [85, 126, 41]. The rationale is quite simple: mutations arising within HCEs may have a negative fitness effect and to prevent their fixation selection acts against them. Although expected, in practice, mutations may still occur in HCEs [126]. Hence, mutations in HCEs can also contribute to the overall genomic fitness [297].

Evolutionary genomic constraints have been proposed as better predictors for the fitness consequences of mutations [291]. The most used comparative genomic approach is the Genomic Evolutionary Rate Profiling (GERP) score. The GERP score quantifies the reduction in the number of substitutions in the multi-species sequence alignment compared to the neutral expectation [69]. Therefore, the higher the GERP score, the larger the deleterious selection coefficient [140]. The functional relevance of the GERP score has been extensively highlighted in a wide range of mammalian [200, 263, 140, 291] and vertebrate species [25, 28]. However, how conservation relates to the strength of selection (N_es) was only recently addressed by Huber et al. (2020). The authors showed that changes in selection coefficients can strongly affect the GERP distribution, leading to unexpected relationships between GERP and N_es. Interestingly, Huber and colleagues observed a similar pattern when in presence of lineage-specific selection or changes in functional elements over time (i.e. a sequence has a specific regulatory role in one lineage, but does not in another lineage) [150]. Reference-based comparative approaches are known to miss out lineage-specific sequences and selection [11]. As a result, functional turnover, combined with N_es, can affect the power to identify strongly selected sites. Reference-free comparative approaches [233, 11] have all the prerequisites to replace more traditional comparative methods. Nonetheless, similar questions on the GERP distribution and phylogenetic scope are expected to arise.

Comparative genomics is an invaluable tool for the identification of functional sites [7, 69, 251] and accurate estimation of genomic fitness [25, 200, 263, 140]. However, additional improvements can be made. Codon models of evolution combined with the GERP framework are an interesting solution to overcome the weak ability of GERP to distinguish weak selection from strong selection in presence of functional turnover (i.e. coding regions are thought to be functionally more stable). For example, by setting a GERP score threshold of 1.0, I was able to more reliably identify truly deleterious alleles to thus estimate genomic fitness in local chicken breeds (**Chapter 3; Chapter 4**) [25, 28]. However, codon models have a major drawback: functional non-protein-coding mutations are ignored.

7.3.3 Identifying loci responsible for inbreeding depression

Deleterious alleles are the determinants of inbreeding depression [162]. To date, most empirical evidence in a number of species suggests that inbreeding depression is caused by homozygosity at loci with (partially) recessive deleterious mutations (*dominance hypothesis*) [47, 257]. However, in a few species [137], inbreeding depression is linked to decreased heterozygosity at loci displaying heterozygous advantage maintained at intermediate frequencies by balancing selection (*overdominance hypothesis*) [49, 46]. Although empirical studies provide evidence for both hypotheses, the genetics of inbreeding depression remains poorly understood in many species.

Inbreeding depression is the reduced survival and fertility of offspring as a consequence of relatedness. Inbred (or consanguineous) individuals have an increased likelihood of exhibiting major abnormalities, ranging from lethal embryonic phenotypes to lowered fertility, survival and growth rate [51]. Recent studies of inbreeding depression have mostly focused on establishing the link between deleterious alleles and fitness traits (e.g. fertility, reproduction). However, there is increasing evidence that other life history traits, often not recorded, may play a role as well. For example, in several studies on inbreeding depression in the wild, body size was found to indirectly influence reproductive opportunities of smaller individuals by reducing their probability to become territory holders [39, 109]. Another important limitation of recent studies on inbreeding depression is the focus either on the dominance or overdominance hypothesis, indirectly excluding *a priori* the possibility that inbreeding depression is caused by a combination of both mechanisms.

In **Chapter 3** I show that the complete set of deleterious alleles within a genome can provide an indication of the inbreeding depression risk only when variants are identical-by-descent (IBD). However, when we relate ROHs to deleterious alleles, we indirectly assume that the phenotypic effect of deleterious alleles is relatively mild. If the fitness effect results in a lethal phenotype, the deleterious allele would be found in genomic regions with lower than expected ROH frequency rather than in ROHs [159]. As a result, the observed consequence of this early embryonic death is a lower fertility of the parents [78]. Recessive alleles responsible for early lethal phenotypes can also be identified by testing for the statistical absence of haplotypes in homozygous state [299, 78]. The haplotype-depletion approach is a powerful tool that can detect low frequency recessive lethal haplotypes (frequency < 2%) [78]. However, thousands of genotyped or hundreds of sequenced individuals are generally required to detect rare deleterious haplotypes. The small population size of many local livestock breeds and wild species is preventing this approach to be used in the study of inbreeding depression.

The study of inbreeding depression should not be limited to the lethal haplotype (or IBD segment), but should also consider selection at linked loci. In their systematic search for recessive lethal haplotypes in several pig populations, Derks et al. (2017) identified a nearly 1 Mb haplotype on chromosome 18 in the Large White population at a frequency of 5.4% [78]. Although the authors did not report individuals homozygous for the depleted haplotype, carrier-by-carrier matings resulted in a significant reduction in total number of born (19.5%) and live-born piglets (19.3%). Interestingly, the SSC18 haplotype is in complete LD with a 212 kb deletion that, while causing death of homozygous fetuses, increases growth rate and feed intake in heterozygous pigs [77]. Hence, because of LD, the frequency of a recessive lethal haplotype can remain at high or moderate frequency in a population if nearby alleles lead to a fitness advantage in heterozygotes (i.e. *overdominance* hypothesis).

Next-generation sequencing data are clarifying the genetic basis of inbreeding depression through the development of new methods and approaches. However, to go beyond the simple *indication* of inbreeding depression, a comprehensive catalogue of life history traits must be collected and made available to the scientific community.

7.4 The genetics of phenotypic diversity

The genetics underlying phenotypic traits has fascinated scientists for decades. As I extensively discussed in this thesis, livestock species are ideal model systems to study the genetics of phenotypes. For millennia, human-driven selection has resulted in the creation of many phenotypes, some for productive purposes, others for aesthetics. Over the past 10 years, genomic resources and methods have advanced the identification of a number of causative genes and mutations underlying breed-defining characteristics [117]. However, our understanding of a species phenotypic complexity remains patchy. In this section, I discuss the demographic and functional limitations highlighting promising approaches for future gene-to-phenotype studies.

7.4.1 Local chicken breeds are reservoirs of phenotypic variation

Humans, and breeders in particular, have always been interested in animals exhibiting fancy traits. Through phenotype selection, breeders indirectly valued specific mutations that, when desirable, became the key molecular determinants of a breed-defining feature. Artificial selection for specific morphological variants led to the foundation of hundreds of breeds, each described by a precise set of traits [286]. Therefore, local livestock breeds are a rich resource for genomic studies aimed at mapping and identifying causal mutations [286, 10].

In **Chapter 5**, I dissect the molecular basis of a fancy phenotype, called ptilopody, through a combination of several genomic approaches that want to overcome the limitations of frequency-based methods when sample size is relatively small. Ptilopody, or foot feathering can be considered an oligogenic trait, since it is determined by a small number of genes [117], and precisely *PITX1* and *TBX5*. The relatively simple genetic basis of foot feathering has evolved by parallel evolution in pigeon [82] and chicken [27, 176]; hence, similar mutations affecting the exact same genes have evolved in the two species. Although introgression through hybridization may drive the evolution of shared polymorphism in different species, ptilopody clearly owes its origin to domestic-specific variants that are clearly under selection [27], as reported in other phenotypes in domesticates [260, 155, 110].

Parallel evolution, as well as convergent evolution, are not rare events. Genetic and molecular studies have identified many examples of the repeated involvement of the same genes in the evolution of similar traits in lineages that shared a recent (i.e. parallel evolution) or distant (i.e. convergent evolution) ancestor. Examples are the *MC1R*-mediated changes in pigmentation patterns in birds [285, 217] and mammals [218, 89], and albinism in blind Mexican cavefish [243]. The evolution of common genetic mechanisms can also involve major structural changes, as observed in threespine (*Gasterosteus aculeatus*) and ninespine (*Pungitius pungitius*) sticklebacks [264]. Selection, either artificial or natural, is an important force in determining the evolution of similar traits in different lineages. However, an important question remains: why similar phenotypes are determined by the same genes? The answer has likely to do with the finite number of genes required to build a certain structure during development. The constraint in gene number subsequently limits the realm of possible evolutionary changes the species is left with [274].

Ambitious large-scale genome projects, such as the Bird 10K project [321] and Earth BioGenome project [182], have promised the generation of draft genome sequences from thousands of species. The availability of an increasing number of species genomes is expected to boost future research on the genetic basis of phenotypes. Therefore, parallel evolution may not surprise us anymore because of its frequent occurrence.

7.4.2 Linkage disequilibrium can constrain the discovery of functional variants

Genome-wide association studies (GWAS) are widely recognised approaches to bridge the gene-to-phenotype gap. Despite the relative success of identifying loci using GWAS, the identification of the actual causal mutation(s) has been far less successful because of linkage disequilibrium (LD). LD is a consequence of the effective population size (N_e), meaning that different demographic histories result in distinct LD decays [305]. Genetic drift can also affect LD by randomly generate associations between alleles at different loci at a rate inversely proportional to N_e [306]. However, in some species, such as chicken, LD is highly variable along the genome and these differences correlate with the recombination rate [145, 206].

Selection for breed-defining features has contributed to a further reduction in N_e, resulting in extended LD, especially in regions with low recombination rate. LD creates an additional layer of complexity to fine-map GWAS. The complexity is due to the fact that high LD results in thousands of variants to be similarly associated with the phenotype-causing variant. In some species, such as dog, a two-stage mapping strategy is used to pinpoint loci associated with breed-defining features [161]. The two-stage approach exploits the long within breeds LD and short between breeds LD to narrow down the search to the shared ancestral haplotype that carries the causative mutation. The ancestral haplotype, which is often < 100 kb in size, is subsequently screened for functional variants [161]. In livestock, an across-breed GWA is an often used strategy to narrow down the list of candidate variants. The underlying assumption is that the same causal variant is segregating in the breeds, even though the haplotype structure may differ among breeds. However, even after fine-mapping, the number of variants remains relatively high.

In **Chapter 5**, I show that genome-wide association studies combined with comparative genomics and functional genomics provide an interesting and effective solution to the LD constraint. As I show, an *indication* of the potential functional importance of variants can be obtained from their overlap with highly conserved elements along the genome (refer to **section 7.3.2**). Once identified, the *actual* functional role of variants can be assessed through the generation of transcriptomic data and, whenever possible, other functional data (e.g. ATAC-seq, ChIP-seq). The rationale is that conserved sites and elements identified by comparative genomics are not always functional, and vice versa [57]. The approach described in Chapter 5 has invaluable research advantages when sample size is limited, as often the case in local livestock breeds and wild species. However, as described in the next section, new methods coming from the field of human genetics are opening up new and exciting research opportunities.

7.4.3 CADD allows the identification of functional variation

As discussed in the previous section, mutations underlying phenotypic variation should be identified and assessed to understand their genetic value. Methods that seek to combine comparative genomic information with functional data [168, 251] offer new opportunities to pinpoint mutations of functional importance throughout a species genome. The Combined Annotation-Dependent Depletion (CADD) score [168] is a widely used measure of variant deleteriousness. Initially developed to facilitate the interpretation of millions of human genetic variants, the CADD approach has recently been applied to livestock species, including pig [128] and chicken [126]. The main advantage of the CADD approach is its integrative annotation. In fact, instead of relying on a single genomic feature, CADD integrates into a single score various genomic features, whose amount and type strongly depend on the species for which CADD is built for [127].

The functional relevance of CADD for chicken is highlighted in **Chapter 6**. As I show, conservation scores alone are not sufficient to pinpoint phenotypically influential mutations within HCEs. Hence, CADD score could potentially replace comparative genomics-based approaches to identify HCEs or fine-tune HCEs to smaller, functional subregions [126]. The CADD score can also be combined with classic GWAS to prioritize variants. In their search for the causal variant(s) of the trait *number of teats* (NTE), Derks and colleagues used the p(ig)CADD score to identify an interesting candidate SNP overlapping the promoter region of the *VRTN* gene. The variant had a pCADD score of 11.95, which is relatively high considering that the mutation is non-coding [75]. According to this study, the CADD score represents a powerful resource to rank any possible mutation in the genome based on the likelihood of being functional. Although not directly addressed in this thesis, the chCADD score could similarly help prioritising

SNVs in classic gene-to-phenotype mapping studies. It is however important to keep in mind that functional information like, for e.g., epigenetic marks and gene expression data are required to validate the *actual* function of prioritised variants [5].

The Functional Annotation of ANimal Genomes (FAANG) consortium [118] has advanced the functional annotation of many livestock species genome with the ultimate goal of assessing the functional relevance of all genomic variation. Despite the generation and availability of functional annotations, the amount of data remains largely inferior to that of model species and human species. The FAANG consortium is progressively generating new data for the validation of the phenotypic effects of candidate variants [105, 103, 104]. Data include RNA-seq from various tissues, histone modification marks, ATAC-seq, and DNA methylation to assess the chromatin accessibility and architecture of the genome, and HiC-seq to assess the 3D conformation of the genome. The public data generated within the FAANG consortium will certainly improve the precision of the CADD score in the near future, if not promoting the generation of CADD scores for other relevant livestock species.

As this thesis highlights, the usefulness of CADD is not limited to HCEs and GWAS. In **Chapter 4**, I used the ch(icken)CADD score to estimate the genomic fitness (or chCADD load) in two local chicken breeds. Similar to the GERP load, the chCADD load confirmed the importance of conservation programmes at removing deleterious alleles from individual genomes. However, while the GERP load biases genomic fitness towards protein-coding mutations, the chCADD load identifies genome-wide deleterious variants with higher confidence [28]. As discussed in **section 7.3.3**, the true functional relevance of candidate mutations can only be directly proven when deleterious alleles are linked to a phenotype. Unfortunately, the phenotypic effects of the scored mutations were not investigated because individual recording was incomplete [28]. However, I do expect that with the continuation of the conservation programme, recording and monitoring will improve, making future validations possible.

Conservation in the genomics era

A comprehensive demographic and functional characterisation of the diversity within and between populations is key to effective management. Levels of inbreeding and associated effects are important factors that should determine conservation priorities and management practices. However, conservation programmes are not always based on genomic information. In recent years, the implementation of genomics as a tool to inform population management has considerably improved in the context of wild species. However, the use of genomics for the conservation of local livestock breeds remains limited. In the next section, I discuss the benefits of genomics-informed conservation programmes aiming to conserve diversity and alleles relevant for phenotypic traits. Moreover, I highlight the central role that gene bank collections could play in restoring genetic diversity.

Genomics-based conservation programmes

Ideally, genetic diversity should be characterised before applying management. However, when not possible, conservation programmes should nonetheless be established based on, for instance, pedigree management and current population figures [176]. Next-generation sequencing data can inform us in our efforts to improve population management and conservation. In fact, genetic relationships between individuals can be better assessed and populations at higher risk of inbreeding depression can be identified. Also, variants with a negative fitness effect can be identified and potentially removed from the population. In **Chapter 4**, I use whole-genome sequencing data from two local chicken breeds to assess the consequences, at genome-wide level, of starting a conservation programme. To do that, I used temporal genomic data of individuals sampled at two different time points: at the beginning of the conservation programme in 2003 and 10 years later in 2013. Furthermore, the effects of management were analysed through the estimation of delta indices related to genetic diversity ($\Delta\pi$), genomic fitness (ΔL), and inbreeding (ΔF_{ROH}) [28]. Although the French conservation programme analysed is a unique case study in the way it is organised and carried out, its limitations are in line with genetic-based management practices. All forms of population management aim to maximise genetic diversity. However, this decision may lead to an increase of deleterious variants in the population (Figure 5, Chapter 4). Therefore, managing variation means maximising diversity while minimising deleterious variation [70]. Taking as a case study the Barbezieux and Gasconne breed, I show that local chicken breeds can significantly benefit from a conservation programme, even when a relative small number of individuals are used as founding nucleus [28]. A reason is that management becomes centralised.

In order to be preserve diversity in the long-term, conservation programmes have to rely more on genomic data as they can reveal additional and relevant information. In **Chapter 4**, I demonstrate that the usefulness of genomic data relies on their ability to reveal the presence of deleterious mutations that would otherwise pass undetected by simple management practices and genetic approaches. Identifying deleterious variants in a genome has become easier and more reliable. Therefore, implementing estimates of effects of deleterious variants into breeding programmes is no longer unrealistic. On the contrary, it has become a necessity to make breeds more resilient to inbreeding and genetic drift. And last, but not least, mutational load can be controlled through management only if the entire population is screened for deleterious variation.

7.4.4 Gene banks are a solution to the decreasing genetic diversity

The long-term conservation of many livestock species can potentially be boosted through the use of material from animal gene bank collections. Animal gene banks are collections of genetic material in the form of gametes (e.g. sperm, ova) or embryos. By definition, gene banks should store material from key, representative ancestors of the current population and unique variation. Genetic material can be used to introgress old variation in the current population, support the rescue of breeds at risk of extinction, or enable the potential adaptation of breeds to changes in breeding goals [226].

Recent studies on the genetic diversity harboured by gene bank samples have progressively elucidated the variation currently available in several national gene banks [290, 153, 95]. However, while genetic variation has been characterized, deleterious variation has so far been ignored. Much of the progress in the evaluation of gene bank collections has been possible thanks to the support of the Innovative Management of Animal GEnetic resources (IMAGE) project financed by the European Union Horizon 2020 research grant. The objective of IMAGE was to promote the characterisation and use of available data to better exploit animal genetic resources, as well as to improve the accessibility and quality of animal gene bank collections. Transboundary breeds, such as the Dutch Holstein Frisian, have been the target of much genetic studies, leaving unanswered many questions on the genetic contribution of gene bank collections to unmanaged populations. Moreover, the usefulness of gene bank collections as reservoirs of phenotype diversity remains unknown.

Small local breeds are thought to considerably benefit from the use of gene bank collections, because of their higher risk of extinction due to modest economic interest. If we consider that the stored material harbours genetic diversity that since the time of sampling has been lost in the population because of genetic drift, the use of gene bank collections is quite appealing. However, we have to remember that, for local livestock breeds, genetic drift was an important genetic force already before or at the time of sampling. Moreover, as in many local chicken breeds genetic exchange has been - and still is - pursued to generate new varieties as a result of a changing sense of beauty, what is the ancestral and unique variation is a very interesting open question. Studies that aim to elucidate the genetic basis of phenotypes in *in vivo* local breeds (see **Chapter 5**) can provide useful insights into the genetic diversity likely stored in gene bank collections. However, only the direct assessment of gene bank material can provide an accurate description of the stored diversity.

Promising tools that aim to facilitate the assessment of the genetic diversity stored by gene bank collections are under development. The multi-species 10K SNP array aims to achieve this in five iconic livestock species, including cattle, pig, goat, sheep, and chicken (unpublished). The multi-species 10K array represents an important step forward in our efforts to characterise gene bank diversity. However, full genome information will still be required to establish effective management strategies to minimise the risk of driving a species to the brink of extinction.

7.5 Concluding remarks

In this thesis, I performed a comprehensive demographic and functional characterisation of the genetic diversity harboured by local chicken breeds. Moreover, I provided a detailed example of how genomes are affected by demographic bottlenecks, genetic drift, and selection. I expect findings presented in this thesis to be particularly useful to the conservation and management of domesticated species. Findings on domesticated populations can serve as a model for any managed population, including wild endangered species. As this thesis demonstrates, the functional meaning of variation in species can only be investigated through an interdisciplinary approach. Comparative genomics and functional genomics are recognised as invaluable research fields to prove the effect that a single nucleotide polymorphism has on an individual's phenotype. I hope with this thesis to have shown that genetic diversity can and should be measured at multiple levels to best manage populations and genetic resources, while bridging the gene-to-phenotype gap.

References

[1] F. Abascal, A. Corvelo, F. Cruz, J. L. Villanueva-Cañas, A. Vlasova, M. Marcet-Houben, B. Martínez-Cruz, J. Y. Cheng, P. Prieto, V. Quesada, et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered iberian lynx. *Genome Biology*, 17(1):251, 2016.

[2] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1):7–20, 2013.

[3] N. Alcala and S. Vuilleumier. Turnover and accumulation of genetic diversity across large time-scale cycles of isolation and connection of populations. *Proceedings of the Royal Society B: Biological Sciences*, 281(1794):20141369, 2014.

[4] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 2009.

[5] R. P. Alexander, G. Fang, J. Rozowsky, M. Snyder, and M. B. Gerstein. Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11(8):559–571, 2010.

[6] J. Alföldi, F. Di Palma, M. Grabherr, C. Williams, L. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. E. Glor, J. D. Jaffe, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, 477(7366):587–591, 2011.

[7] J. Alföldi and K. Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome Research*, 23(7):1063–1068, 2013.

[8] F. W. Allendorf, P. A. Hohenlohe, and G. Luikart. Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11(10):697–709, 2010.

[9] L. Andersson, A. L. Archibald, C. D. Bottema, R. Brauning, S. C. Burgess, D. W. Burt, E. Casas, H. H. Cheng, L. Clarke, C. Couldrey, et al. Coordinated international action to accelerate genome-to-phenome with faang, the functional annotation of animal genomes project. *Genome Biology*, 16(1):57, 2015.

[10] L. Andersson and M. Georges. Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, 5(3):202–212, 2004.

[11] J. Armstrong, G. Hickey, M. Diekhans, A. Deran, Q. Fang, D. Xie, S. Feng, J. Stiller, D. Genereux, J. Johnson, et al. Progressive alignment with cactus: a multiple-genome aligner for the thousand-genome era. *bioRxiv*, page 730531, 2019.

[12] A. Auton, K. Bryc, A. R. Boyko, K. E. Lohmueller, J. Novembre, A. Reynolds, A. Indap, M. H. Wright, J. D. Degenhardt, R. N. Gutenkunst, et al. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research*, 19(5):795–803, 2009.

[13] I. A. Babarinde and N. Saitou. Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics. *Molecular Biology and Evolution*, 33(7):1807–1817, 2016.

[14] A. D. Barnosky, N. Matzke, S. Tomiya, G. O. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, et al. Has the earth's sixth mass extinction already arrived? *Nature*, 471(7336):51–57, 2011.

[15] T. Bartels. Variations in the morphology, distribution, and arrangement of feathers in domesticated birds. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 298(1):91–108, 2003.

[16] A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran. pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32(18):2817–2823, 2016.

[17] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, 2004.

[18] J. Bélanger and D. Pilling. The state of the world's biodiversity for food and agriculture. *FAO Commission on Genetic Resources for Food and Agriculture Assessments: Rome, Italy*, page 572, 2019.

[19] T. Berr, A. Peticca, and A. Haudry. Evidence for purifying selection on conserved noncoding elements in the genome of drosophila melanogaster. *bioRxiv*, page 623744, 2019.

[20] C. Berthouly, J.-C. Maillard, L. P. Doan, T. N. Van, B. Bed'Hom, G. Leroy, H. H. Thanh, D. Laloë, N. Bruneau, C. V. Chi, et al. Revealing fine scale subpopulation structure in the vietnamese h'mong cattle breed for conservation purposes. *BMC Genetics*, 11(1):45, 2010.

[21] I. J. Boegheim, P. A. Leegwater, H. A. van Lith, and W. Back. Current insights into the molecular genetic basis of dwarfism in livestock. *The Veterinary Journal*, 224:64–75, 2017.

[22] E. F. Boer, H. F. Van Hollebeke, S. Park, C. R. Infante, D. B. Menke, and M. D. Shapiro. Pigeon foot feathering reveals conserved limb identity networks. *Developmental biology*, 454(2):128–144, 2019.

[23] E. F. Boer, H. F. Van Hollebeke, and M. D. Shapiro. Genomic determinants of epidermal appendage patterning and structure in domestic birds. *Developmental biology*, 429(2):409–419, 2017.

[24] S. Bornelöv, E. Seroussi, S. Yosefi, K. Pendavis, S. C. Burgess, M. Grabherr, M. Friedman-Einat, and L. Andersson. Correspondence on lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biology*, 18(1):112, 2017.

[25] C. Bortoluzzi, M. Bosse, M. F. Derks, R. P. Crooijmans, M. A. Groenen, and H.-J. Megens. The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. *Evolutionary applications*, 13(2):330–341, 2020.

[26] C. Bortoluzzi, R. P. Crooijmans, M. Bosse, S. J. Hiemstra, M. A. Groenen, and H.-J. Megens. The effects of recent changes in breeding preferences on maintaining traditional dutch chicken genomic diversity. *Heredity*, 121(6):564, 2018.

[27] C. Bortoluzzi, H.-J. Megens, M. Bosse, M. F. Derks, B. Dibbits, K. Laport, S. Weigend, M. A. Groenen, and R. P. Crooijmans. Parallel genetic origin of foot feathering in birds. *Molecular Biology and Evolution*, 2020.

[28] C. Bortoluzzi, G. Restoux, and M. Tixier-Boichard. Quantifying temporal genomic erosion in small managed populations under a recently established conservation programme. *In preparation*, 2020.

[29] M. Bosse, H.-J. Megens, M. F. Derks, Á. M. de Cara, and M. A. Groenen. Deleterious alleles in the context of domestication, inbreeding, and selection. *Evolutionary applications*, 12(1):6–17, 2019.

[30] M. Bosse, H.-J. Megens, O. Madsen, R. P. Crooijmans, O. A. Ryder, F. Austerlitz, M. A. Groenen, and M. A. R. de Cara. Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research*, 25(7):970–981, 2015.

[31] M. Bosse, H.-J. Megens, O. Madsen, Y. Paudel, L. A. Frantz, L. B. Schook, R. P. Crooijmans, and M. A. Groenen. Regions of homozygosity in the porcine genome: consequence

of demography and the recombination landscape. *PLoS Genetics*, 8(11):e1003100, 2012.

[32] A. C. Bouwman, H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*, 50(3):362–367, 2018.

[33] I. Braasch, A. R. Gehrke, J. J. Smith, K. Kawasaki, T. Manousaki, J. Pasquier, A. Amores, T. Desvignes, P. Batzel, J. Catchen, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics*, 48(4):427–437, 2016.

[34] E. Brondizio, J. Settele, S. Díaz, and H. Ngo. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. *IPBES Secretariat*, 2019.

[35] B. L. Browning and S. R. Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[36] B. L. Browning, Y. Zhou, and S. R. Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[37] S. R. Browning and B. L. Browning. High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics*, 86(4):526–539, 2010.

[38] S. L. Brusatte, J. K. O'Connor, and E. D. Jarvis. The origin and diversification of birds. *Current Biology*, 25(19):R888–R898, 2015.

[39] K. E. Brzeski, D. R. Rabon Jr, M. J. Chamberlain, L. P. Waits, and S. S. Taylor. Inbreeding and inbreeding depression in endangered red wolves (canis rufus). *Molecular Ecology*, 23(17):4241–4255, 2014.

[40] D. W. Burt. Chicken genome: current status and future opportunities. *Genome Research*, 15(12):1692–1698, 2005.

[41] S. Casillas, A. Barbadilla, and C. M. Bergman. Purifying selection maintains highly conserved noncoding sequences in drosophila. *Molecular Biology and Evolution*, 24(10):2222–2234, 2007.

[42] G. Caughley. Directions in conservation biology. *Journal of animal ecology*, pages 215–244, 1994.

[43] F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson. Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics*,

19(4):220, 2018.

[44] G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, 1(5):e1400253, 2015.

[45] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, 2015.

[46] B. Charlesworth. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetics Research*, 63(3):213–227, 1994.

[47] B. Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195, 2009.

[48] B. Charlesworth. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190(1):5–22, 2012.

[49] B. Charlesworth, M. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

[50] D. Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):e64, 2006.

[51] D. Charlesworth and J. H. Willis. The genetics of inbreeding depression. *Nature Reviews Genetics*, 10(11):783–796, 2009.

[52] C. Charlier, W. Coppieters, F. Rollin, D. Desmecht, J. S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, et al. Highly effective snp-based association mapping and management of recessive defects in livestock. *Nature Genetics*, 40(4):449, 2008.

[53] C. Chiang, R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan, and I. M. Hall. Speedseq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10):966, 2015.

[54] Y. Choi and A. P. Chan. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16):2745–2747, 2015.

[55] J. Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[56] E. P. Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[57] E. P. Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799, 2007.

[58] I. C. P. M. Consortium et al. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, 432(7018):717, 2004.

[59] I. H. G. S. Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[60] G. M. Cooper and J. Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640, 2011.

[61] R. J. Craig, A. Suh, M. Wang, and H. Ellegren. Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (ficedula albicollis). *Molecular Ecology*, 27(2):476–492, 2018.

[62] F. Cruz, C. Vilà, and M. T. Webster. The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Molecular Biology and Evolution*, 25(11):2331–2336, 2008.

[63] R. A. Dalloul, J. A. Long, A. V. Zimin, L. Aslam, K. Beal, L. A. Blomberg, P. Bouffard, D. W. Burt, O. Crasta, R. P. Crooijmans, et al. Multi-platform next-generation sequencing of the domestic turkey (meleagris gallopavo): genome assembly and analysis. *PLoS Biology*, 8(9), 2010.

[64] N. Dana, H.-J. Megens, R. P. Crooijmans, O. Hanotte, J. Mwacharo, M. A. Groenen, and J. A. van Arendonk. East asian contributions to dutch traditional and western commercial chickens inferred from mtdna analysis. *Animal Genetics*, 42(2):125–133, 2011.

[65] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[66] C. Darwin. *The variation of animals and plants under domestication.* 1868.

[67] C. Darwin. Cross and self-fertilization of plants. *Murray, London*, 1876.

[68] P. H. Davis, V. H. Heywood, et al. Principles of angiosperm taxonomy. *Principles of angiosperm taxonomy.*, 1963.

[69] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12), 2010.

[70] M. Á. R. De Cara, B. Villanueva, M. Á. Toro, and J. Fernández. Purging deleterious mutations in conservation programmes: combining optimal contributions with inbred matings. *Heredity*, 110(6):530, 2013.

[71] M. Á. R. de Cara, B. Villanueva, M. Á. Toro, and J. Fernández. Using genomic tools to maintain diversity and fitness in conservation programmes. *Molecular Ecology*, 22(24):6091–6099, 2013.

[72] R. De Oliveira Silva, B. V. Ahmadi, S. J. Hiemstra, and D. Moran. Optimizing ex situ genetic resource collections for european livestock conservation. *Journal of Animal Breeding and Genetics*, 136(1):63–73, 2019.

[73] A. DeLaurier, R. Schweitzer, and M. Logan. Pitx1 determines the morphology of muscle, tendon, and bones of the hindlimb. *Developmental biology*, 299(1):22–34, 2006.

[74] M. F. Derks, A. B. Gjuvsland, M. Bosse, M. S. Lopes, M. van Son, B. Harlizius, B. F. Tan, H. Hamland, E. Grindflek, M. A. Groenen, et al. Loss of function mutations in essential genes cause embryonic lethality in pigs. *PLoS Genetics*, 15(3):e1008055, 2019.

[75] M. F. Derks, C. Groß, M. S. Lopes, M. J. Reinders, M. Bosse, A. B. Gjuvsland, D. de Ridder, H.-J. Megens, and M. A. Groenen. Accelerated discovery of functional genomic variation in pigs. *In preparation*, 2020.

[76] M. F. Derks, J. M. Herrero-Medrano, R. P. Crooijmans, A. Vereijken, J. A. Long, H.-J. Megens, and M. A. Groenen. Early and late feathering in turkey and chicken: same gene but different mutations. *Genetics Selection Evolution*, 50(1):7, 2018.

[77] M. F. Derks, M. S. Lopes, M. Bosse, O. Madsen, B. Dibbits, B. Harlizius, M. A. Groenen, and H.-J. Megens. Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. *PLoS Genetics*, 14(9):e1007661, 2018.

[78] M. F. Derks, H.-J. Megens, M. Bosse, M. S. Lopes, B. Harlizius, and M. A. Groenen. A systematic survey to identify lethal recessive variation in highly managed pig populations. *BMC Genomics*, 18(1):858, 2017.

[79] M. F. Derks, H.-J. Megens, M. Bosse, J. Visscher, K. Peeters, M. C. Bink, A. Vereijken, C. Gross, D. De Ridder, M. J. Reinders, et al. A survey of functional genomic variation in domesticated chickens. *Genetics Selection Evolution*, 50(1):17, 2018.

[80] D. Díez-del Molino, F. Sánchez-Barreiro, I. Barnes, M. T. P. Gilbert, and L. Dalén. Quantifying temporal genomic erosion in endangered species. *Trends in Ecology & Evolution*.

[81] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[82] E. T. Domyan, Z. Kronenberg, C. R. Infante, A. I. Vickrey, S. A. Stringham, R. Bruders,

M. W. Guernsey, S. Park, J. Payne, R. B. Beckstead, et al. Molecular shifts in limb identity underlie development of feathered feet in two domestic avian species. *eLife*, 5:e12115, 2016.

[83] E. T. Domyan and M. D. Shapiro. Pigeonetics takes flight: evolution, development, and genetics of intraspecific variation. *Developmental biology*, 427(2):241–250, 2017.

[84] B. Dorshorst, R. Okimoto, and C. Ashwell. Genomic regions associated with dermal hyperpigmentation, polydactyly and other morphological traits in the silkie chicken. *Journal of Heredity*, 101(3):339–350, 2010.

[85] J. A. Drake, C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh, A. Reymond, L. Excoffier, H. Attar, S. E. Antonarakis, E. T. Dermitzakis, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics*, 38(2):223–227, 2006.

[86] N. R. Draper and H. Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[87] T. Druml, K. Salajpal, M. Dikic, M. Urosevic, G. Grilz-Seger, and R. Baumung. Genetic diversity, population structure and subdivision of local balkan pig breeds in austria, croatia, serbia and bosnia-herzegovina and its practical value in conservation programs. *Genetics Selection Evolution*, 44(1):5, 2012.

[88] H. Eding and T. Meuwissen. Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *Journal of Animal Breeding and Genetics*, 118(3):141–159, 2001.

[89] E. Eizirik, N. Yuhki, W. E. Johnson, M. Menotti-Raymond, S. S. Hannah, and S. J. O'Brien. Molecular genetics and evolution of melanism in the cat family. *Current Biology*, 13(5):448–453, 2003.

[90] M. G. Elferink, H.-J. Megens, A. Vereijken, X. Hu, R. P. Crooijmans, and M. A. Groenen. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. *PLoS One*, 7(2):e32720, 2012.

[91] M. G. Elferink, A. A. Vallée, A. P. Jungerius, R. P. Crooijmans, and M. A. Groenen. Partial duplication of the prlr and spef2 genes at the late feathering locus in chicken. *BMC Genomics*, 9(1):391, 2008.

[92] M. G. Elferink, P. van As, T. Veenendaal, R. P. Crooijmans, and M. A. Groenen. Regional differences in recombination hotspots between two chicken populations. *BMC Genetics*, 11(1):11, 2010.

[93] H. Ellegren and N. Galtier. Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422, 2016.

[94] J. Eriksson, G. Larson, U. Gunnarsson, B. Bed'Hom, M. Tixier-Boichard, L. Strömstedt, D. Wright, A. Jungerius, A. Vereijken, E. Randi, et al. Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genetics*, 4(2):e1000010, 2008.

[95] S. E. Eynard. *Using genomic information to conserve genetic diversity in livestock*. Wageningen University, 2018.

[96] S. E. Eynard, J. J. Windig, S. J. Hiemstra, and M. P. Calus. Whole-genome sequence data uncover loss of genetic diversity due to selection. *Genetics Selection Evolution*, 48(1):33, 2016.

[97] S. E. Eynard, J. J. Windig, I. Hulsegge, S.-J. Hiemstra, and M. P. Calus. The impact of using old germplasm on genetic merit and diversity—a cattle breed case study. *Journal of Animal Breeding and Genetics*, 135(4):311–322, 2018.

[98] S. E. Eynard, J. J. Windig, G. Leroy, R. van Binsbergen, and M. P. Calus. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. *BMC Genetics*, 16(1):24, 2015.

[99] S. Fang, L. Zhang, J. Guo, Y. Niu, Y. Wu, H. Li, L. Zhao, X. Li, X. Teng, X. Sun, et al. Noncodev5: a comprehensive annotation database for long non-coding rnas. *Nucleic Acids Research*, 46(D1):D308–D314, 2018.

[100] G. G. Faust and I. M. Hall. Samblaster: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, 2014.

[101] J. Felsenstein. Phylip (phylogeny inference package) version 3.6. 2005.

[102] J. Fernández, M. Toro, A. Mäki-Tanila, et al. Management of genetic diversity in small farm animal populations. *Animal*, 5(11):1684–1698, 2011.

[103] S. Foissac, S. Djebali, K. Munyard, N. Vialaneix, A. Rau, H. Acloque, S. Lagarrigue, and E. Giuffra. 32 functional annotation of livestock genomes: chromatin structure and regulation of gene expression. *Journal of Animal Science*, 97(Supplement_2):15–16, 2019.

[104] S. Foissac, S. Djebali, K. Munyard, N. Vialaneix, A. Rau, K. Muret, D. Esquerre, M. Zytnicki, T. Derrien, P. Bardou, et al. Livestock genome annotation: transcriptome and chromatin structure profiling in cattle, goat, chicken and pig. *bioRxiv*, page 316091, 2018.

[105] S. Foissac, S. Djebali, K. Munyard, N. Vialaneix, A. Rau, K. Muret, D. Esquerré, M. Zytnicki, T. Derrien, P. Bardou, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biology*, 17(1):1–25, 2019.

[106] S. Foissac, S. Djebali, K. Munyard, N. Vialaneix, A. Rau, K. Muret, D. Esquerré, M. Zytnicki, T. Derrien, P. Bardou, F. Blanc, C. Cabau, E. Crisci, S. Dhorne-Pollet, F. Drouet, T. Faraut, I. Gonzalez, A. Goubil, S. Lacroix-Lamandé, F. Laurent, S. Marthey, M. Marti-Marimon, R. Momal-Leisenring, F. Mompart, P. Quéré, D. Robelin, M. Cristobal, G. Tosser-Klopp, S. Vincent-Naulleau, S. Fabre, M. der, Laan, C. Klopp, M. Tixier-Boichard, H. Acloque, S. Lagarrigue, and E. Giuffra. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biology*, 17(108):863–874, 2019.

[107] I. U. for Conservation of Nature, N. Resources, and W. W. Fund. *World conservation strategy: Living resource conservation for sustainable development.* Gland, Switzerland: IUCN, 1980.

[108] L. A. Frantz, J. G. Schraiber, O. Madsen, H.-J. Megens, M. Bosse, Y. Paudel, G. Semiadi, E. Meijaard, N. Li, R. P. Crooijmans, et al. Genome sequencing reveals fine scale diversification and reticulation history during speciation in sus. *Genome Biology*, 14(9):R107, 2013.

[109] R. J. Fredrickson, P. Siminski, M. Woolf, and P. W. Hedrick. Genetic rescue and inbreeding depression in mexican wolves. *Proceedings of the Royal Society B: Biological Sciences*, 274(1623):2365–2371, 2007.

[110] A. H. Freedman, K. E. Lohmueller, and R. K. Wayne. Evolutionary history, selective sweeps, and deleterious variation in the dog. *Annual Review of Ecology, Evolution, and Systematics*, 47:73–96, 2016.

[111] D. P. Frisby, R. A. Weiss, M. Roussel, and D. Stehelin. The distribution of endogenous chicken retrovirus sequences in the dna of galliform birds does not coincide with avian phylogenetic relationships. *Cell*, 17(3):623–634, 1979.

[112] A. Fumihito, T. Miyake, S.-I. Sumi, M. Takada, S. Ohno, and N. Kondo. One subspecies of the red junglefowl (gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds. *Proceedings of the National Academy of Sciences*, 91(26):12505–12509, 1994.

[113] A. Fumihito, T. Miyake, M. Takada, R. Shingu, T. Endo, T. Gojobori, N. Kondo, and S. Ohno. Monophyletic origin and unique dispersal patterns of domestic fowls. *Proceedings of the National Academy of Sciences*, 93(13):6792–6795, 1996.

[114] F. García-Alcalde, K. Okonechnikov, J. Carbonell, L. M. Cruz, S. Götz, S. Tarazona, J. Dopazo, T. F. Meyer, and A. Conesa. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, 28(20):2678–2679, 2012.

[115] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. arxiv. 2012. *arXiv preprint arXiv:1207.3907*, 2012.

[116] P. Gelabert, M. Sandoval-Velasco, A. Serres, M. de Manuel, P. Renom, A. Margaryan, J. Stiller, T. de Dios, Q. Fang, S. Feng, et al. Evolutionary history, genomic adaptation to toxic diet, and extinction of the carolina parakeet. *Current Biology*, 30(1):108–114, 2020.

[117] M. Georges, C. Charlier, and B. Hayes. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*, 20(3):135–156, 2019.

[118] E. Giuffra, C. K. Tuggle, and F. Consortium. Functional annotation of animal genomes (faang): current achievements and roadmap. *Annual review of animal biosciences*, 7:65–88, 2019.

[119] S. Glémin. How are deleterious mutations purged? drift versus nonrandom mating. *Evolution*, 57(12):2678–2687, 2003.

[120] Z. Granevitze, J. Hillel, G. Chen, N. T. K. Cuc, M. Feldman, H. Eding, and S. Weigend. Genetic diversity within chicken populations from different continents and management histories. *Animal Genetics*, 38(6):576–583, 2007.

[121] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974.

[122] R. E. Green, E. L. Braun, J. Armstrong, D. Earl, N. Nguyen, G. Hickey, M. W. Vandewege, J. A. S. John, S. Capella-Gutiérrez, T. A. Castoe, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215):1254449, 2014.

[123] M. A. Groenen, H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, R. P. Crooijmans, A. Vereijken, R. Okimoto, W. M. Muir, and H. H. Cheng. The development and characterization of a 60k snp chip for chicken. *BMC Genomics*, 12(1):274, 2011.

[124] M. A. Groenen, P. Wahlberg, M. Foglio, H. H. Cheng, H.-J. Megens, R. P. Crooijmans, F. Besnier, M. Lathrop, W. M. Muir, G. K.-S. Wong, et al. A high-density snp-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research*, 19(3):510–519, 2009.

[125] L. Groeneveld, J. Lenstra, H. Eding, M. Toro, B. Scherf, D. Pilling, R. Negrini, E. Finlay, H. Jianlin, E. Groeneveld, et al. Genetic diversity in farm animals–a review. *Animal Genetics*, 41:6–31, 2010.

[126] C. Gross, C. Bortoluzzi, D. de Ridder, H.-J. Megens, M. Groenen, M. Reinders, and M. Bosse. Evolutionarily conserved non-protein-coding regions in the chicken genome harbor functionally important variation. *bioRxiv*, 2020.

[127] C. Groß, D. de Ridder, and M. Reinders. Predicting variant deleteriousness in non-human species: applying the cadd approach in mouse. *BMC Bioinformatics*, 19(1):1–10, 2018.

[128] C. Groß, M. Derks, H.-J. Megens, M. Bosse, M. A. Groenen, M. Reinders, and D. De Ridder. pcadd: Snv prioritisation in sus scrofa. *Genetics Selection Evolution*, 52(1):4, 2020.

[129] J. C. Habel, M. Husemann, A. Finger, P. D. Danley, and F. E. Zachos. The relevance of time series in molecular ecology and conservation biology. *Biological Reviews*, 89(2):484–492, 2014.

[130] D. L. Halligan, A. Kousathanas, R. W. Ness, B. Harr, L. Eöry, T. M. Keane, D. J. Adams, and P. D. Keightley. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genetics*, 9(12), 2013.

[131] V. Hamburger and H. L. Hamilton. A series of normal stages in the development of the chick embryo. *Journal of morphology*, 88(1):49–92, 1951.

[132] N. Harmston, A. Barešić, and B. Lenhard. The mystery of extreme non-coding conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632):20130021, 2013.

[133] M. Hasselgren and K. Norén. Inbreeding in natural mammal populations: historical perspectives and future challenges. *Mammal Review*, 49(4):369–383, 2019.

[134] A. Haudry, A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, Z. Joly-Lopez, J. G. Steffen, K. M. Hazzouri, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8):891–898, 2013.

[135] J. J. Hayward, M. G. Castelhano, K. C. Oliveira, E. Corey, C. Balkman, T. L. Baxter, M. L. Casal, S. A. Center, M. Fang, S. J. Garrison, et al. Complex disease and phenotype mapping in the domestic dog. *Nature communications*, 7(1):1–11, 2016.

[136] P. W. Hedrick. Conservation genetics: where are we now? *Trends in Ecology & Evolution*.

[137] P. W. Hedrick. What is the evidence for heterozygote advantage selection? *Trends in Ecology & Evolution*.

[138] P. W. Hedrick and A. Garcia-Dorado. Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology & Evolution*.

[139] P. W. Hedrick, M. Kardos, R. O. Peterson, and J. A. Vucetich. Genomic variation of inbreeding and ancestry in the remaining two isle royale wolves. *Journal of Heredity*, 108(2):120–126, 2016.

[140] B. M. Henn, L. R. Botigué, C. D. Bustamante, A. G. Clark, and S. Gravel. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6):333–343, 2015.

[141] R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, M. Przeworski, et al. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019):920–924, 2011.

[142] J. M. Herrero-Medrano, H.-J. Megens, M. A. Groenen, G. Ramis, M. Bosse, M. Pérez-Enciso, and R. P. Crooijmans. Conservation genomic analysis of domestic and wild pig populations from the iberian peninsula. *BMC Genetics*, 14(1):106, 2013.

[143] G. Hickey, B. Paten, D. Earl, D. Zerbino, and D. Haussler. Hal: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10):1341–1342, 2013.

[144] J. Hillel, M. A. Groenen, M. Tixier-Boichard, A. B. Korol, L. David, V. M. Kirzhner, T. Burke, A. Barre-Dirie, R. P. Crooijmans, K. Elo, et al. Biodiversity of 52 chicken populations assessed by microsatellite typing of dna pools. *Genetics Selection Evolution*, 35(6):533, 2003.

[145] L. W. Hillier, W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. Groenen, M. E. Delany, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 423(10):695–777, 2014.

[146] A. Hodgkinson and A. Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766, 2011.

[147] J. I. Hoffman, F. Simpson, P. David, J. M. Rijks, T. Kuiken, M. A. Thorne, R. C. Lacy, and K. K. Dasmahapatra. High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences*, 111(10):3775–3780, 2014.

[148] D. P. Howrigan, M. A. Simonson, and M. C. Keller. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics*, 12(1):460, 2011.

[149] Z.-L. Hu, C. A. Park, X.-L. Wu, and J. M. Reecy. Animal qtldb: an improved database tool for livestock animal qtl/association data dissemination in the post-genome era. *Nucleic Acids Research*, 41(D1):D871–D879, 2013.

[150] C. D. Huber, B. Y. Kim, and K. E. Lohmueller. Population genetic models of gerp scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS Genetics*, 16(5):e1008827, 2020.

[151] T.-Y. J. Hui and A. Burt. Estimating effective population size from temporally spaced samples with a novel, efficient maximum likelihood algorithm. *Genetics*, 2015.

[152] J. Huisman, L. E. Kruuk, P. A. Ellis, T. Clutton-Brock, and J. M. Pemberton. Inbreeding depression across the lifespan in a wild mammal population. *Proceedings of the National Academy of Sciences*, 113(13):3585–3590, 2016.

[153] I. Hulsegge, M. Calus, R. Hoving-Bolink, M. Lopes, H.-J. Megens, and K. Oldenbroek. Impact of merging commercial breeding lines on the genetic diversity of landrace pigs. *Genetics Selection Evolution*, 51(1):60, 2019.

[154] T. Iijima, R. Kajitani, S. Komata, C.-P. Lin, T. Sota, T. Itoh, and H. Fujiwara. Parallel evolution of batesian mimicry supergene in two papilio butterflies, p. polytes and p. memnon. *Science Advances*, 4(4):eaao5416, 2018.

[155] F. Imsland, C. Feng, H. Boije, B. Bed'Hom, V. Fillon, B. Dorshorst, C.-J. Rubin, R. Liu, Y. Gao, X. Gu, et al. The rose-comb mutation in chickens constitutes a structural rearrangement causing both altered comb morphology and defective sperm motility. *PLoS Genetics*, 8(6), 2012.

[156] E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.

[157] N. Joshi and J. Fass. Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33)[software], 2011.

[158] S. Kanginakudru, M. Metta, R. Jakati, and J. Nagaraju. Genetic evidence from indian red jungle fowl corroborates multiple domestication of modern day chicken. *BMC Evolutionary Biology*, 8(1):174, 2008.

[159] M. Kardos, M. Åkesson, T. Fountain, Ø. Flagstad, O. Liberg, P. Olason, H. Sand, P. Wabakken, C. Wikenros, and H. Ellegren. Genomic consequences of intensive inbreeding in an isolated wolf population. *Nature Ecology & Evolution*, 2(1):124–131, 2018.

[160] M. Kardos, H. R. Taylor, H. Ellegren, G. Luikart, and F. W. Allendorf. Genomics advances the study of inbreeding depression in the wild. *Evolutionary applications*, 9(10):1205–1218, 2016.

[161] E. K. Karlsson and K. Lindblad-Toh. Leader of the pack: gene mapping in dogs and other

model organisms. *Nature Reviews Genetics*, 9(9):713–725, 2008.

[162] L. F. Keller and D. M. Waller. Inbreeding effects in wild populations. *Trends in Ecology & Evolution*.

[163] E.-S. Kim, J. B. Cole, H. Huson, G. R. Wiggans, C. P. Van Tassell, B. A. Crooker, G. Liu, Y. Da, and T. S. Sonstegard. Effect of artificial selection on runs of homozygosity in us holstein cattle. *PLoS One*, 8(11):e80813, 2013.

[164] M. Kimura. Theoretical foundation of population genetics at the molecular level. *Theoretical population biology*, 2(2):174–208, 1971.

[165] M. Kimura et al. Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics*, 28(4):882–901, 1957.

[166] M. Kimura, T. Maruyama, and J. F. Crow. The mutation load in small populations. *Genetics*, 48(10):1303, 1963.

[167] R. J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, et al. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, 2011, 2011.

[168] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310, 2014.

[169] D. Kleinman-Ruiz, L. Soriano, M. Casas-Marce, C. Szychta, I. Sánchez, J. Fernández, and J. A. Godoy. Genetic evaluation of the iberian lynx ex situ conservation programme. *Heredity*, 123(5):647–661, 2019.

[170] M. H. Kohn, W. J. Murphy, E. A. Ostrander, and R. K. Wayne. Genomics and conservation genetics. *Trends in Ecology & Evolution*.

[171] T. J. Kono, F. Fu, M. Mohammadi, P. J. Hoffman, C. Liu, R. M. Stupar, K. P. Smith, P. Tiffin, J. C. Fay, and P. L. Morrell. The role of deleterious substitutions in crop genomes. *Molecular Biology and Evolution*, 33(9):2307–2317, 2016.

[172] B. K. Kragesteen, M. Spielmann, C. Paliou, V. Heinrich, R. Schöpflin, A. Esposito, C. Annunziatella, S. Bianco, A. M. Chiariello, I. Jerković, et al. Dynamic 3d chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nature Genetics*, 50(10):1463–1473, 2018.

[173] A. Kranis, A. A. Gheyas, C. Boschiero, F. Turner, L. Yu, S. Smith, R. Talbot, A. Pirani, F. Brew, P. Kaiser, et al. Development of a high density 600k snp genotyping array for chicken. *BMC Genomics*, 14(1):59, 2013.

[174] C. C. Kyriazis, R. K. Wayne, and K. E. Lohmueller. High genetic diversity can contribute to extinction in small populations. *bioRxiv*, page 678524, 2019.

[175] L. Laikre, S. Hoban, M. W. Bruford, G. Segelbacher, F. W. Allendorf, G. Gajardo, A. G. Rodríguez, P. W. Hedrick, M. Heuertz, P. A. Hohenlohe, et al. Post-2020 goals overlook genetic diversity. *Science*, 367(6482):1083–1085, 2020.

[176] L. Langqing, M. Bosse, H.-J. Megens, M. de Visser, M. A. Groenen, and O. Madsen. Genetic consequences of long-term small effective population size in the critically endangered pygmy hog. *In preparation*, 2020.

[177] G. Larson and J. Burger. A population genetics view of animal domestication. *Trends in Genetics*, 29(4):197–205, 2013.

[178] G. Larson and D. Q. Fuller. The evolution of animal domestication. *Annual Review of Ecology, Evolution, and Systematics*, 45:115–136, 2014.

[179] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. Lumpy: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, 2014.

[180] J. Lenffer, F. W. Nicholas, K. Castle, A. Rao, S. Gregory, M. Poidinger, M. D. Mailman, and S. Ranganathan. Omia (online mendelian inheritance in animals): an enhanced platform and integration into the entrez search interface at ncbi. *Nucleic Acids Research*, 34(suppl_1):D599–D601, 2006.

[181] G. Leroy, E. L. Carroll, M. W. Bruford, J. A. DeWoody, A. Strand, L. Waits, and J. Wang. Next-generation metrics for monitoring genetic erosion within populations of conservation concern. *Evolutionary applications*, 11(7):1066–1083, 2018.

[182] H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, et al. Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, 2018.

[183] H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[184] H. Li and R. Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[185] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[186] J. Li, M. Lee, B. Davis, S. Lamichhaney, B. Dorshorst, P. Siegel, and L. Andersson.

Mutations upstream of the tbx5 and pitx1 transcription factor genes are associated with feathered legs in the domestic chicken. *Molecular Biology and Evolution*, 2020.

[187] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[188] Q. Liu, Y. Zhou, P. L. Morrell, and B. S. Gaut. Deleterious variants in asian rice and the potential cost of domestication. *Molecular Biology and Evolution*, 34(4):908–924, 2017.

[189] Y.-P. Liu, G.-S. Wu, Y.-G. Yao, Y.-W. Miao, G. Luikart, M. Baig, A. Beja-Pereira, Z.-L. Ding, M. G. Palanichamy, and Y.-P. Zhang. Multiple maternal origins of chickens: out of the asian jungles. *Molecular phylogenetics and evolution*, 38(1):12–19, 2006.

[190] M. Logan, H.-G. Simon, and C. Tabin. Differential regulation of t-box and homeobox transcription factors suggests roles in controlling chick limb-type identity. *Development*, 125(15):2825–2835, 1998.

[191] M. Logan and C. J. Tabin. Role of pitx1 upstream of tbx4 in specification of hindlimb identity. *Science*, 283(5408):1736–1739, 1999.

[192] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[193] P. V. Lovell, M. Wirthlin, L. Wilhelm, P. Minx, N. H. Lazar, L. Carbone, W. C. Warren, and C. V. Mello. Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biology*, 15(12):565, 2014.

[194] J. Lu, T. Tang, H. Tang, J. Huang, S. Shi, and C.-I. Wu. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics*, 22(3):126–131, 2006.

[195] G. Luikart, F. Allendorf, J. Cornuet, and W. Sherwin. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity*, 89(3):238–247, 1998.

[196] M. Lynch. Evolution of the mutation rate. *TRENDS in Genetics*, 26(8):345–352, 2010.

[197] T. Makino, C.-J. Rubin, M. Carneiro, E. Axelsson, L. Andersson, and M. T. Webster. Elevated proportions of deleterious genetic variation in domestic animals and plants. *Genome Biology and Evolution*, 10(1):276–290, 2018.

[198] A. Marcovitz, R. Jia, and G. Bejerano. “reverse genomics” predicts function of human conserved noncoding elements. *Molecular Biology and Evolution*, 33(5):1358–1369, 2016.

[199] E. H. Margulies, G. M. Cooper, G. Asimenos, D. J. Thomas, C. N. Dewey, A. Siepel,

E. Birney, D. Keefe, A. S. Schwartz, M. Hou, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Research*, 17(6):760–774, 2007.

[200] C. D. Marsden, D. Ortega-Del Vecchyo, D. P. O'Brien, J. F. Taylor, O. Ramirez, C. Vilà, T. Marques-Bonet, R. D. Schnabel, R. K. Wayne, and K. E. Lohmueller. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences*, 113(1):152–157, 2016.

[201] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[202] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[203] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1):122, 2016.

[204] R. McQuillan, A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, et al. Runs of homozygosity in european populations. *The American Journal of Human Genetics*, 83(3):359–372, 2008.

[205] H. Megens and M. Groenen. Domesticated species form a treasure-trove for molecular characterization of mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. *Heredity*, 109(1):1, 2012.

[206] H.-J. Megens, R. P. Crooijmans, J. W. Bastiaansen, H. H. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, et al. Comparison of linkage disequilibrium and haplotype diversity on macro-and microchromosomes in chicken. *BMC Genetics*, 10(1):86, 2009.

[207] R. W. Meredith, G. Zhang, M. T. P. Gilbert, E. D. Jarvis, and M. S. Springer. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science*, 346(6215):1254390, 2014.

[208] T. Meuwissen. Maximizing the response of selection with a predefined rate of inbreeding. *Journal of Animal Science*, 75(4):934–940, 1997.

[209] Y. Miao, M.-S. Peng, G.-S. Wu, Y. Ouyang, Z. Yang, N. Yu, J. Liang, G. Pianchou, A. Beja-Pereira, B. Mitra, et al. Chicken domestication: an updated perspective based on mito-

chondrial genomes. *Heredity*, 110(3):277–282, 2013.

[210] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, et al. 28-way vertebrate alignment and conservation track in the ucsc genome browser. *Genome Research*, 17(12):1797–1808, 2007.

[211] N. Miyamoto, J. F. Fernández-Manjarrés, M.-E. Morand-Prieur, P. Bertolino, and N. Frascaria-Lacoste. What sampling is needed for reliable estimations of genetic diversity in fraxinus excelsior l.(oleaceae)? *Annals of forest science*, 65(4):1, 2008.

[212] C. Mou, F. Pitel, D. Gourichon, F. Vignoles, A. Tzika, P. Tato, L. Yu, D. W. Burt, B. Bed'Hom, M. Tixier-Boichard, et al. Cryptic patterning of avian skin confers a developmental facility for loss of neck feathering. *PLoS Biology*, 9(3), 2011.

[213] B. T. Moyers, P. L. Morrell, and J. K. McKay. Genetic costs of domestication and improvement. *Journal of Heredity*, 109(2):103–116, 2017.

[214] B. T. Moyers, P. L. Morrell, and J. K. McKay. Genetic costs of domestication and improvement. *Journal of Heredity*, 109(2):103–116, 2018.

[215] B. Mtileni, F. Muchadeyi, A. Maiwashe, E. Groeneveld, L. Groeneveld, K. Dzama, and S. Weigend. Genetic diversity and conservation of south african indigenous chicken populations. *Journal of Animal Breeding and Genetics*, 128(3):209–218, 2011.

[216] W. M. Muir, G. K.-S. Wong, Y. Zhang, J. Wang, M. A. Groenen, R. P. Crooijmans, H.-J. Megens, H. Zhang, R. Okimoto, A. Vereijken, et al. Genome-wide assessment of worldwide chicken snp genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Sciences*, 105(45):17312–17317, 2008.

[217] N. I. Mundy, N. S. Badcock, T. Hart, K. Scribner, K. Janssen, and N. J. Nadeau. Conserved genetic basis of a quantitative plumage trait involved in mate choice. *Science*, 303(5665):1870–1873, 2004.

[218] M. W. Nachman, H. E. Hoekstra, and S. L. D'Agostino. The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences*, 100(9):5268–5273, 2003.

[219] K. Nadachowska-Brzyska, R. Burri, L. Smeds, and H. Ellegren. Psmc analysis of effective population sizes in molecular ecology and its application to black-and-white ficedula flycatchers. *Molecular Ecology*, 25(5):1058–1072, 2016.

[220] K. Nam, C. Mugal, B. Nabholz, H. Schielzeth, J. B. Wolf, N. Backström, A. Künstner, C. N. Balakrishnan, A. Heger, C. P. Ponting, et al. Molecular evolution of genes in avian

genomes. *Genome Biology*, 11(6):R68, 2010.

[221] M. Nei and W.-H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273, 1979.

[222] P. C. Ng and S. Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

[223] T. Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96, 1973.

[224] T. Ohta. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1):263–286, 1992.

[225] K. Okonechnikov, A. Conesa, and F. García-Alcalde. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2):292–294, 2016.

[226] J. Oldenbroek. *Genomic management of animal genetic diversity*. Wageningen Academic Publishers, 2017.

[227] L. Orlando and P. Librado. Origin and evolution of deleterious mutations in horses. *Genes*, 10(9):649, 2019.

[228] E. A. Ostrander and R. K. Wayne. The canine genome. *Genome Research*, 15(12):1706–1716, 2005.

[229] J.-F. Ouimette, M. L. Jolin, A. L'honoré, A. Gifuni, and J. Drouin. Divergent transcriptional activities determine limb identity. *Nature communications*, 1(1):1–9, 2010.

[230] C. Paris, B. Servin, and S. Boitard. Inference of selection from genetic time series using various parametric approximations to the wright-fisher model. *G3: Genes, Genomes, Genetics*, pages g3–400778, 2019.

[231] P. J. Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[232] H. G. Parker, L. V. Kim, N. B. Sutter, S. Carlson, T. D. Lorentzen, T. B. Malek, G. S. Johnson, H. B. DeFrance, E. A. Ostrander, and L. Kruglyak. Genetic structure of the purebred domestic dog. *Science*, 304(5674):1160–1164, 2004.

[233] B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, and D. Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

[234] B. S. Pedersen and A. R. Quinlan. Mosdepth: quick coverage calculation for genomes

and exomes. *Bioinformatics*, 34(5):867–868, 2018.

[235] B. S. Pedersen and A. R. Quinlan. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience*, 8(4):giz040, 2019.

[236] T. J. Pemberton, D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li. Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics*, 91(2):275–292, 2012.

[237] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature Biotechnology*, 33(3):290, 2015.

[238] M. Pham, C. Berthouly-Salazar, X. Tran, W. Chang, R. Crooijmans, D. Lin, V. Hoang, Y. Lee, M. Tixier-Boichard, and C. Chen. Genetic diversity of v ietnamese domestic chicken populations as decision-making support for conservation strategies. *Animal Genetics*, 44(5):509–521, 2013.

[239] J. K. Pickrell and J. K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics*, 8(11):e1002967, 2012.

[240] J. P. Pollinger, K. E. Lohmueller, E. Han, H. G. Parker, P. Quignon, J. D. Degenhardt, A. R. Boyko, D. A. Earl, A. Auton, A. Reynolds, et al. Genome-wide snp and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, 464(7290):898, 2010.

[241] D. Polychronopoulos, J. W. King, A. J. Nash, G. Tan, and B. Lenhard. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Research*, 45(22):12611–12624, 2017.

[242] A. Potts. *Chicken*. Reaktion Books, 2012.

[243] M. E. Protas, C. Hersey, D. Kochanek, Y. Zhou, H. Wilkens, W. R. Jeffery, L. I. Zon, R. Borowsky, and C. J. Tabin. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genetics*, 38(1):107–111, 2006.

[244] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[245] D. C. Purfield, D. P. Berry, S. McParland, and D. G. Bradley. Runs of homozygosity and population history in cattle. *BMC Genetics*, 13(1):70, 2012.

[246] F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, and T. Manke. deeptools: a flexible

platform for exploring deep-sequencing data. *Nucleic Acids Research*, 42(W1):W187–W191, 2014.

[247] P. Ramu, W. Esuma, R. Kawuki, I. Y. Rabbi, C. Egesi, J. V. Bredeson, R. S. Bart, J. Verma, E. S. Buckler, and F. Lu. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature Genetics*, 49(6):959, 2017.

[248] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 2019.

[249] D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. Reconstructing indian population history. *Nature*, 461(7263):489, 2009.

[250] S. Renaut and L. H. Rieseberg. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Molecular Biology and Evolution*, 32(9):2273–2283, 2015.

[251] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

[252] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, M. Diekhans, K. E. Smith, K. R. Rosenbloom, B. J. Raney, et al. The ucsc genome browser database: update 2010. *Nucleic Acids Research*, 38(suppl_1):D613–D619, 2010.

[253] Â. M. Ribeiro, L. Puetz, N. B. Pattinson, L. Dalén, Y. Deng, G. Zhang, R. R. da Fonseca, B. Smit, and M. T. P. Gilbert. 31° south: The physiology of adaptation to arid conditions in a passerine bird. *Molecular Ecology*, 28(16):3709–3721, 2019.

[254] B. Rischkowsky and D. Pilling. *The state of the world's animal genetic resources for food and agriculture*. Food & Agriculture Org., 2007.

[255] J. A. Robinson, C. Brown, B. Y. Kim, K. E. Lohmueller, and R. K. Wayne. Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Current Biology*, 28(21):3487–3494, 2018.

[256] J. A. Robinson, D. Ortega-Del Vecchyo, Z. Fan, B. Y. Kim, C. D. Marsden, K. E. Lohmueller, R. K. Wayne, et al. Genomic flatlining in the endangered island fox. *Current Biology*, 26(9):1183–1189, 2016.

[257] J. A. Robinson, J. Räikkönen, L. M. Vucetich, J. A. Vucetich, R. O. Peterson, K. E. Lohmueller, and R. K. Wayne. Genomic signatures of extensive inbreeding in isle royale wolves, a population on the threshold of extinction. *Science Advances*, 5(5):eaau0757,

2019.

[258] C. Rodriguez-Esteban, T. Tsukui, S. Yonei, J. Magallon, K. Tamura, and J. C. I. Belmonte. The t-box genes tbx4 and tbx5 regulate limb outgrowth and identity. *Nature*, 398(6730):814–818, 1999.

[259] A. R. Rogers. How population growth affects linkage disequilibrium. *Genetics*, 197(4):1329–1341, 2014.

[260] C.-J. Rubin, M. C. Zody, J. Eriksson, J. R. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 464(7288):587, 2010.

[261] J. Sadri, A. B. Diallo, and M. Blanchette. Predicting site-specific human selective pressure using evolutionary signatures. *Bioinformatics*, 27(13):i266–i274, 2011.

[262] B. D. Scherf, D. Pilling, et al. The second report on the state of the world's animal genetic resources for food and agriculture. 2015.

[263] M. Schubert, H. Jónsson, D. Chang, C. Der Sarkissian, L. Ermini, A. Ginolhac, A. Albrechtsen, I. Dupanloup, A. Foucal, B. Petersen, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences*, 111(52):E5661–E5669, 2014.

[264] M. D. Shapiro, M. A. Bell, and D. M. Kingsley. Parallel genetic origins of pelvic reduction in vertebrates. *Proceedings of the National Academy of Sciences*, 103(37):13753–13758, 2006.

[265] M. D. Shapiro, Z. Kronenberg, C. Li, E. T. Domyan, H. Pan, M. Campbell, H. Tan, C. D. Huff, H. Hu, A. I. Vickrey, et al. Genomic diversity and evolution of the head crest in the rock pigeon. *Science*, 339(6123):1063–1067, 2013.

[266] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[267] A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21(3):468–488, 2004.

[268] F. Sitzenstock, F. Ytournel, A. R. Sharifi, D. Cavero, H. Täubert, R. Preisinger, and H. Simianer. Efficiency of genomic selection in an established commercial layer breeding program. *Genetics Selection Evolution*, 45(1):29, 2013.

[269] D. Smedley, M. Schubach, J. O. Jacobsen, S. Köhler, T. Zemojtel, M. Spielmann, M. Jäger, H. Hochheiser, N. L. Washington, J. A. McMurry, et al. A whole-genome analy-

sis framework for effective identification of pathogenic regulatory variants in mendelian disease. *The American Journal of Human Genetics*, 99(3):595–606, 2016.

[270] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35, 1974.

[271] R. Somes Jr. Mutations and major variants of plumage and skin in chickens. *Developments in Animal and Veterinary Sciences (Netherlands)*, 1990.

[272] M. Spielmann, F. Brancati, P. M. Krawitz, P. N. Robinson, D. M. Ibrahim, M. Franke, J. Hecht, S. Lohan, K. Dathe, A. M. Nardone, et al. Homeotic arm-to-leg transformation associated with genomic rearrangements at the pitx1 locus. *The American Journal of Human Genetics*, 91(4):629–635, 2012.

[273] K. A. Steige, B. Laenen, J. Reimegård, D. G. Scofield, and T. Slotte. Genomic analysis reveals major determinants of cis-regulatory variation in capsella grandiflora. *Proceedings of the National Academy of Sciences*, 114(5):1087–1092, 2017.

[274] D. L. Stern. Perspective: evolutionary developmental biology and the problem of variation. *Evolution*, 54(4):1079–1091, 2000.

[275] M. Stoffel, E. Humble, A. Paijmans, K. Acevedo-Whitehouse, B. Chilvers, B. Dickerson, F. Galimberti, N. Gemmell, S. Goldsworthy, H. Nichols, et al. Demographic histories and genetic diversity across pinnipeds are shaped by human exploitation, ecology and life-history. *Nature communications*, 9(1):1–12, 2018.

[276] J. Storey, A. Bass, A. Dabney, and D. Robinson. qvalue: Q-value estimation for false discovery rate control. r package version 2.0. 0. *Available at github. com/jdstorey/qvalue. Accessed April*, 14:2017, 2015.

[277] Y. Sun, R. Liu, G. Zhao, M. Zheng, Y. Sun, X. Yu, P. Li, and J. Wen. Genome-wide linkage analysis identifies loci for physical appearance traits in chickens. *G3: Genes, Genomes, Genetics*, 5(10):2037–2041, 2015.

[278] N. B. Sutter, C. D. Bustamante, K. Chase, M. M. Gray, K. Zhao, L. Zhu, B. Padhukasahasram, E. Karlins, S. Davis, P. G. Jones, et al. A single igf1 allele is a major determinant of small size in dogs. *Science*, 316(5821):112–115, 2007.

[279] Z. A. Szpiech, J. Xu, T. J. Pemberton, W. Peng, S. Zöllner, N. A. Rosenberg, and J. Z. Li. Long runs of homozygosity are enriched for deleterious variation. *The American Journal of Human Genetics*, 93(1):90–102, 2013.

[280] R. Tadano, M. Nishibori, Y. Imamura, M. Matsuzaki, K. Kinoshita, M. Mizutani, T. Namikawa, and M. Tsudzuki. High genetic divergence in miniature breeds of japanese

native chickens compared to red junglefowl, as revealed by microsatellite analysis. *Animal Genetics*, 39(1):71–78, 2008.

[281] J. K. Takeuchi, K. Koshiba-Takeuchi, K. Matsumoto, A. Vogel-Höpker, M. Naitoh-Matsuo, K. Ogura, N. Takahashi, K. Yasuda, and T. Ogura. Tbx5 and tbx4 genes determine the wing/leg identity of limb buds. *Nature*, 398(6730):810–814, 1999.

[282] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

[283] R. C. Team et al. R: A language and environment for statistical computing. 2013.

[284] J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303, 2017.

[285] E. Theron, K. Hawkins, E. Bermingham, R. Ricklefs, and N. Mundy. The molecular basis of an avian plumage polymorphism in the wild: a point mutation in the mc1r gene is perfectly associated with the melanic plumage morph in the bananaquit (coereba flaveola). *Current Biology*, 11:550–557, 2001.

[286] M. Tixier-Boichard, B. Bed'hom, and X. Rognon. Chicken domestication: from archeology to genomics. *Comptes rendus biologies*, 334(3):197–204, 2011.

[287] M. A. Toro, J. Fernández, and A. Caballero. Molecular characterization of breeds and its use in conservation. *Livestock Science*, 120(3):174–195, 2009.

[288] C. Truong, L. Oudre, and N. Vayatis. ruptures: change point detection in python. *arXiv preprint arXiv:1801.00826*, 2018.

[289] S. D. Turner. qqman: an r package for visualizing gwas results using qq and manhattan plots. *bioRxiv*, page 005165, 2014.

[290] A. E. van Breukelen, H. P. Doekes, J. J. Windig, and K. Oldenbroek. Characterization of genetic diversity conserved in the gene bank for dutch cattle breeds. *Diversity*, 11(12):229, 2019.

[291] T. van der Valk, M. de Manuel, T. Marques-Bonet, and K. Guschanski. Estimates of genetic load in small populations suggest extensive purging of deleterious alleles. *bioRxiv*, page 696831, 2019.

[292] T. van der Valk, D. Díez-del Molino, T. Marques-Bonet, K. Guschanski, and L. Dalén. Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. *Current Biology*, 29(1):165–170, 2019.

[293] K. J. van der Velde, E. N. de Boer, C. C. van Diemen, B. Sikkema-Raddatz, K. M. Abbott, A. Knopperts, L. Franke, R. H. Sijmons, T. J. de Koning, C. Wijmenga, et al. Gavin: Gene-

aware variant interpretation for medical sequencing. *Genome Biology*, 18(1):6, 2017.

[294] K. J. van der Velde, J. Kuiper, B. A. Thompson, J.-P. Plazzer, G. van Valkenhoef, M. de Haan, J. D. Jongbloed, C. Wijmenga, T. J. de Koning, K. M. Abbott, et al. Evaluation of cadd scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization. *Human mutation*, 36(7):712–719, 2015.

[295] E. Van Marle-Köster, C. Hefer, L. Nel, and M. Groenen. Genetic diversity and population structure of locally adapted south african chicken lines: Implications for conservation. *South African Journal of Animal Science*, 38(4):271–281, 2008.

[296] E. Van Marle-Koster and L. Nel. Genetic characterization of native southern african chicken populations: evaluation and selection of polymorphic microsatellite markers. *South African Journal of Animal Science*, 30(1):1–6, 2000.

[297] C. van Oosterhout. Mutation load is the spectre of species conservation. *Nature Ecology & Evolution*, pages 1–3, 2020.

[298] M. Van Son, M. S. Lopes, H. J. Martell, M. F. Derks, L. E. Gangsei, J. Kongsro, M. N. Wass, E. H. Grindflek, and B. Harlizius. A qtl for number of teats shows breed specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics. *Frontiers in Genetics*, 10:272, 2019.

[299] P. VanRaden, K. Olson, D. Null, and J. Hutchison. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *Journal of dairy science*, 94(12):6153–6161, 2011.

[300] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[301] J. Wang, E. Santiago, and A. Caballero. Prediction and estimation of effective population size. *Heredity*, 117(4):193–206, 2016.

[302] M.-S. Wang, M. Thakur, M.-S. Peng, Y. Jiang, L. A. F. Frantz, M. Li, J.-J. Zhang, S. Wang, J. Peters, N. O. Otecko, et al. 863 genomes reveal the origin and domestication of chicken. *Cell Research*, pages 1–9, 2020.

[303] S. Wang, C. Haynes, F. Barany, and J. Ott. Genome-wide autozygosity mapping in human populations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(2):172–180, 2009.

[304] Y. Wang, Y. Gao, F. Imsland, X. Gu, C. Feng, R. Liu, C. Song, M. Tixier-Boichard, D. Gourichon, Q. Li, et al. The crest phenotype in chicken is associated with ectopic expression

of hoxc8 in cranial skin. *PloS One*, 7(4), 2012.

[305] R. K. Waples, W. A. Larson, and R. S. Waples. Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, 117(4):233–240, 2016.

[306] R. S. Waples and P. R. England. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*, 189(2):633–644, 2011.

[307] W. C. Warren, D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Künstner, S. Searle, S. White, A. J. Vilella, S. Fairley, et al. The genome of a songbird. *Nature*, 464(7289):757–762, 2010.

[308] W. C. Warren, L. W. Hillier, C. Tomlinson, P. Minx, M. Kremitzki, T. Graves, C. Markovic, N. Bouk, K. D. Pruitt, F. Thibaud-Nissen, et al. A new chicken genome assembly provides insight into avian genome structure. *G3: Genes, Genomes, Genetics*, 7(1):109–117, 2017.

[309] K. Watanabe, E. Taskesen, A. Van Bochoven, and D. Posthuma. Functional mapping and annotation of genetic associations with fuma. *Nature communications*, 8(1):1–11, 2017.

[310] G. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975.

[311] M. V. Westbury, S. Hartmann, A. Barlow, I. Wiesel, V. Leo, R. Welch, D. M. Parker, F. Sicks, A. Ludwig, L. Dalén, et al. Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Molecular Biology and Evolution*, 35(5):1225–1237, 2018.

[312] S. Wilkinson, P. Wiener, D. Teverson, C. Haley, and P. Hocking. Characterization of the genetic diversity, structure and admixture of british chicken breeds. *Animal Genetics*, 43(5):552–563, 2012.

[313] R. J. Williamson, E. B. Josephs, A. E. Platts, K. M. Hazzouri, A. Haudry, M. Blanchette, and S. I. Wright. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of capsella grandiflora. *PLoS Genetics*, 10(9), 2014.

[314] H. Woelders, C. Zuidberg, and S. Hiemstra. Animal genetic resources conservation in the netherlands and europe: poultry perspective. *Poultry Science*, 85(2):216–222, 2006.

[315] P. Wu, J. Yan, Y.-C. Lai, C. S. Ng, A. Li, X. Jiang, R. M. Elsey, R. Widelitz, R. Bajpai, W.-H. Li, et al. Multiple regulatory modules are required for scale-to-feather conversion. *Molecular Biology and Evolution*, 35(2):417–430, 2018.

[316] Z. Wu, M. F. Derks, B. Dibbits, H.-J. Megens, M. A. Groenen, and R. P. Crooijmans. A novel loss-of-function variant in transmembrane protein 263 (tmem263) of autosomal dwarfism in chicken. *Frontiers in Genetics*, 9:193, 2018.

[317] V. Wucher, F. Legeai, B. Hedan, G. Rizk, L. Lagoutte, T. Leeb, V. Jagannathan, E. Cadieu, A. David, H. Lohi, et al. Feelnc: a tool for long non-coding rna annotation and its application to the dog transcriptome. *Nucleic Acids Research*, 45(8):e57–e57, 2017.

[318] Y. Xue, J. Prado-Martinez, P. H. Sudmant, V. Narasimhan, Q. Ayub, M. Szpak, P. Frandsen, Y. Chen, B. Yngvadottir, D. N. Cooper, et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 348(6231):242–245, 2015.

[319] N. Yu, M. I. Jensen-Seaman, L. Chemnick, O. Ryder, and W.-H. Li. Nucleotide diversity in gorillas. *Genetics*, 166(3):1375–1383, 2004.

[320] E. Zanetti, M. De Marchi, C. Dalvit, and M. Cassandro. Genetic characterization of local italian breeds of chickens undergoing in situ conservation. *Poultry Science*, 89(3):420–427, 2010.

[321] G. Zhang. Bird sequencing project takes off. *Nature*, 522(7554):34–34, 2015.

[322] G. Zhang. The bird's-eye view on chromosome evolution. *Genome Biology*, 19(1):1–3, 2018.

[323] G. Zhang, C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, 346(6215):1311–1320, 2014.

[324] M. Zhang, L. Zhou, R. Bawa, H. Suren, and J. A. Holliday. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Molecular Biology and Evolution*, 33(11):2899–2910, 2016.

[325] Q. Zhang, B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics*, 16(1):542, 2015.

[326] H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, and L. Wang. Crossmap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, 2014.

[327] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, 28(24):3326–3328, 2012.

[328] T. Zhou, L. Yang, Y. Lu, et al. Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale. *Nucleic Acids Research*, 41:56–62, 2013.

Summary

The genetic diversity harboured in a genome can have major phenotypic effects, both beneficial (e.g. adaptation) or detrimental (inbreeding depression). The work described in this thesis contributes to our understanding of the complex interplay between demography and selection in the chicken genome. I analysed genotype and sequence data from several local chicken breeds to answer questions about the underlying mechanisms that shape genomic and deleterious variation. Moreover, by means of population genomics, comparative genomics, and functional genomics I was able to unravel the genetic basis of ptilopody and develop a tool to rank variants based on their likelihood of being functional. With this thesis I provide a comprehensive overview on the importance of demographic and functional characterisation studies aiming to conserve diversity in a species genome.

In **Chapter 2** I characterise the genetic diversity, demographic history, and level of inbreeding of 37 traditional Dutch chicken breeds to guide conservation efforts and management strategies. I show that large fowls and true bantams were the source populations of recently created neo-bantam breeds, with which they share a high proportion of their alleles. I observe that neo-bantams display genetic signatures of back-crossing, often pursued for phenotype selection. By contrasting the genetic diversity and level of inbreeding of traditional breeds with that of commercial white egg layers, I highlight the importance of using markers to inform breeding programmes on potentially harmful homozygosity to prevent loss of genetic diversity.

In **Chapter 3** I examine the genomic consequences of the severe domestication bottleneck and the recent breed formation bottleneck. I show that, despite the rather similar genome-wide heterozygosity, recently bottlenecked populations have a higher proportion of deleterious variants relative to populations that have been small for longer time. In fact, in recently bottlenecked populations, genetic drift and recent inbreeding are mostly responsible for the observed genome-wide homozygosity. I also observe that deleterious variants tend to be found in long tracts of homozygous genotypes (ROHs), possibly suggesting a link with inbreeding depression.

In **Chapter 4** I use temporal sequencing data to quantify temporal genomic erosion in small populations under a recently established conservation programme. I show that, because of the relatively small number of founding individuals, a reduction in genetic diversity ($\Delta\pi$) and increase in inbreeding (ΔF_{ROH}) are expected at the start of the conservation programme. I also demonstrate that management can control genetic drift, allowing purging of deleterious alleles. In this chapter I reinforce the imperative to establish and incorporate genomic information into management practices that aim to keep local at-risk breeds from the brink of extinction.

In **Chapter 5** I provide evidence for a parallel genetic origin of ptilopody in chicken and pigeon. I show that ptilopody (or foot feathering) is determined in both species by two loci, in which similar mutations and regulatory pathways are involved. At one loci, I identify a 17 kb deletion affecting *PITX1* expression, a gene known to encode a transcription regulator of hindlimb identity and development. At the second loci, I find a foot-feathered 4 kb haplotype upstream *TBX5*, a gene involved in forelimb identity and a key determinant of foot feather development. The haplotype and causal non-protein-coding mutations likely affect *TBX5* ectopic expression in foot feathered birds during embryonic development.

In **Chapter 6** I examine the functional importance of variants found in conserved non-protein-coding elements (CNEs) of the chicken genome. To do that, I develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD) model, a variant effect prediction tool that can score and rank variants based on their likelihood of being functional. I show that CNEs display SNP densities and allele frequency distributions characteristic of genomic regions constrained by purifying selection. Moreover, by annotating SNPs with the chCADD score, I was able to identify specific subregions of higher functional importance. In fact, SNPs found in these subregions are associated with known disease genes in human, mouse, and rat. I anticipate the chCADD score to be of great use to the scientific community and breeding companies in future functional studies in chicken.

Acknowledgements

I would like to spend a few words to thank the people I had the privilege to meet during this long journey, as without them this achievement would not have been possible.

Martien, thank you for always keeping the doors open and listening to someone's challenges. Thank you for caring and for seeing in each of us the potential and determination that it takes to become a better scientist, but even more a better person. Hendrik-Jan, you know better than anyone else that these past four years have not always been easy. Nonetheless, thank you for always giving me the freedom to follow my interests and research ideas. Mirte, you jumped into the project half-way through and quickly caught up with it. Thank you for being more like a mother than a simple supervisor; thank you for constantly caring and listening to my (often too many) fluxes of conscience. Richard, even though you were not directly involved in my PhD thesis, you nonetheless left a trace behind. Thank you for being such a lovely, friendly, and happy person. These are qualities that not many people have all at once, so please keep treasuring them. Thanks to you I have found my antidote to panic attacks, which is: think about all the stuff that Richard is busy with to realize that nothing can be worse than that :)!.

I would like to thank two extraordinary people for my time in Jouy en Josas. Michele, I admire so much your passion, determination, critical thinking, and integrity that I can only hope to become one day like you. You are not only an incredible scientist, but most of all you are an incredible woman. I sincerely hope we will have the opportunity to keep in touch in the years to come. Gwendal, aside from Milad and Marco, you were among the first people I interacted with EVERY DAY at INRAE. There was no single day in which you didn't stop in front of my office to check up on me, my work, and my well-being. You have no idea how much I appreciated this kind gesture. Thank you so much for making the time spent in Jouy a great and memorable experience! (by the way, I still owe you a running :)). Aside from my collaborators at INRAE, I would also like to thank a few more people. Marco and Roberta, thank you so much for all the funny discussions we had, for all the lunches, and coffee breaks that always made me feel at home. A special thanks to the international community of the INRAE dormitory (I wish I had met you earlier): Giulia, Marco, Ana, Inssa, Elsa, Tristan and Tatiani.

Martijn, there are so many things I would like to say about you that this acknowledgements section would start to sound a bit weird, so let me keep it short :). Even though we had the chance to share the office for only two years (yes, you have no idea how much I was mad at Richard and Lucia for moving me to another office!), the chemistry for a long-lasting friendship was there since the beginning. You may appear shy and reserved at first, but the more you open up, the more amazing and unexpected you turn out to be. You are the classic example that, indeed, appearances can be deceiving :). Thank you for everything you have done, for listening and debating my strong opinions always with a smile, for opening up and reaching out, and above all, thank you for always coming to the office to call us out to join the coffee break: that was the best moment of the day! Henri, there is only one word to describe you: Henri. You have such a unique personality that the first time I met you I felt puzzled, surprised, and curious to know and understand you more. Even though I only scratched the surface with you, I call myself satisfied to have reached the point of playing rock-paper-scissors in Finland with you ....and you know for what!:). Thank you for the past three years, for being special and always yourself no matter what.

Zhou and Lanqging, I cannot talk about one of you without mentioning the other one, so here you go. You guys are so special and incredible that no words can describe my admiration and love for you. I became the person I am today also thanks to your, almost unconditional, friendship. Thank you so much for everything you have done and still are doing. I am sure we will meet somewhere else again, but until them がんばって！ Ibrahim, the first time we met was in 2014 during the Animal Breeding and Genetics course. Our friendship is a very special one, as we couldn't be more different: I am more like a pinball, you are just the most chill person I have ever met in my life, always telling me to slow down and relax....in other words, be less chiaristic :)!. Thank you for the beautiful friendship we have built over the years, for always caring, and listening to my highs and lows. I wish you all the best, my dear Ibrahim! Harmen, it is true that our paths have been similar, but never I would have imagined that someone would think we were dating....that comes as quite of a surprise!. Anyway, thank you for the time spent together, sometime in the same hotel room as IMAGE fellows (ok, now I can see why the dating part...), and for co-organizing the WSD (your love for details was always the cherry on top of the cake, we just loved it!). A special thanks also to all the other people I had the opportunity to work/interact with, and in particular Maria, Rayner, Jani, Raymond, Esther, Shuwen, Vinicius, Maulik, Raka, Daniel, Eli, Marieke, Yale, Francesca, Sem, and Joelle.

I would also like to thank a few people from the Bioinformatics Group, since I spent most of my post-5p.m. office time there for obvious reasons :). Raul, thank you for your kindness and help in the time of need. I wish you all the best in Germany! Vittorio, thank you for bringing some Italian vibe into the group (although I don't think I will miss you talking about Italian politics :)), Mehjani, thank you for being so unique, nerdy, and special. Carlos and Salima, thank you for always welcoming us even without appointments :), Ronald, thank you for making me smile (I heard words...) and for all the manga/anime talks. Christian, thank you for giving the chance to work with you on the chicken CADD score (I am so proud of that paper!!). A special thanks also goes to Sevgin, Barbara, Rens, Elvira, Serina, and Jelle.

Besides my direct colleagues, I was lucky enough to meet many interesting people during my time spent in the Netherlands and Denmark. I would like to here acknowledge two of them. Jonas, my best running partner ever since the time of the MSc!. You know I have to thank you for many things, so let me start from the cleaning experience of the washing machine. That was such a gross experience that, although I felt so ashamed at first, I cannot think about it now without laughing (maybe you are still not) :D. Thank you for always making me reach my limits with the running, for being there when I needed to talk to someone about what was happening in my life, and for your unexpected visits to Wageningen. Thank you for being an amazing and special friend in all these years. Marzieh, I still remember the first time we met as it was yesterday and you know that it wasn't a friendly encounter (you got angry at me because I was using the cutting board to keep the entrance door open since I locked myself out of the room in the dormitory in Denmark the first day I arrived...). Fortunately for me, you let yourself go and slowly opened up. Thank you for allowing me to know you, because, trust me, if it wasn't for your friendship I wouldn't have survived Denmark (do you remember how much we talked about this?)! Thank you for the most weird evening discussions (you know which ones... :D), for always caring, and for keeping in contact all these years despite the distance. I wish you all the best in Canada my dear friend!

Alla mia famiglia. Ben sei anni sono trascorsi da quel lontano giorno di agosto del 2014 in cui mi avete accompagnata in Olanda in una macchina talmente carica di bagagli da sembrare dei veri e propri emigranti. Nonostante di sacrifici ne siano stati fatti, questa tesi vuole essere la dimostrazione che ne é valsa effettivamente la pena. Papi, grazie per aver instillato in me il desiderio di scoprire e conoscere nuove cose. Mamma, grazie per avermi reso la donna independente e forte che sono oggi. Fratellone, grazie per essere la mia guida e la copia (a volte irrazionale) di me stessa. Zoe, per gli amici Zobetella porcella, grazie per il desiderio costante di coccole, e Messi, il gatto incazzato, grazie per essere l'ennesima prova dell'intelligenze felina. Un grazie di cuore a tutte quelle persone e familiari in giro per il mondo che nel loro piccolo hanno, e continuano ancor'ora, a segnare la mia vita. Un grazie particulare alle due fantastiche, ultra novantenni nonne, ai parenti in Francia e Belgio, a Camilla, agli amici della triennale e di una vita che mai smettono di stupirmi (Manuel, Cecilia, Gloria, Deva).

And last, but not least, to my registered partner :). ¿Amor?....¿Qué tal? Sabes que eres mi chocoLatino muy barbudo y muy holaquetal! Since the first time I met you I thought: if I was his girlfriend I would never let him go. And that's what happened (although you actually freaked me out at first during the Machine learning course) :). Even though you hypnotized, maximized me since the start, at the beginning I had quite some trouble understanding you (I remember I had to repeatedly ask you "What do you mean?" as it was all a "Shabadabadio Shabadabadio Babababa"). But ¿hey? Meeting you was the best thing that ever happened to my life. The passions, interests, and ideas we share are so unique and special that go beyond human comprehension. In the past months there has been a lot work, work, work, work, work, work, but you never stopped to be my favorite collector that collects anything he finds (it seems especially computers and phones recently ...). Thank you for supporting me, for being my best-friend, my love, my husband (ops, registered partner), for caring for me, and for loving the crazy person inside of me! Thank you baby, baby, baby oooooh!

Curriculum Vitae

Chiara Bortoluzzi was born on the 20^{th} October 1992 in Belluno, Italy. The international family and environment in which she grew up played an important role in her decision later in life to move abroad. After obtaining a diploma in classical studies, she decided to enroll in the interfaculty bachelor programme of Animal Sciences and Technologies at the University of Padova. Additionally, to pursue her passion for music, she also enrolled at the Conservatory of Music ”C. Pollini” in Padova, where she spent two years as a piano student. After graduating *cum laude*, she decided to join the MSc programme in Animal Breeding and Genetics at Wageningen University. Here, she did her first major thesis in genomics under the supervision of Dr. Ole Madsen and Dr. Kyle Schachtschneider. Although the thesis made her realize that a PhD in genomics would be in line with her research interests, she nonetheless decided to do another major thesis in quantitative genetics. She therefore joined Dr. Janss and Dr. Alemu at the Center for Quantitative Genetics and Genomics in Denmark. Part of her thesis was published in 2019 in the journal *Genetics Selection Evolution*. In August 2016, she started a PhD under the supervision of Dr. Hendrik-Jan Megens and Dr. Mirte Bosse. The PhD contributed to the European Union’s funded project Innovative Management of Animal GEnetic Resources (IMAGE). Thanks to the research freedom that was given to her, she was able to further develop her research interests in population genomics and comparative genomics. During her PhD, she also received funding from the WIAS graduate school to spend 6 months at the National Institute of Agricultural Research (INRAE) in Jouy en Josas. Here, she further contributed to the IMAGE project, working under the supervision of Dr. Gwendal Restoux and Dr. Michele Tixier-Boichard. She recently submitted a Marie Skłodowska-Curie fellowship to join Prof. Guojie Zhang at the University of Copenhagen, Denmark.

WIAS Training and Supervision

The basic package (1.8 ECTS)	
WIAS Introduction Day	2016
Research Integrity & Ethics and Animal Science	2016
Disciplinary competences (15.5 ECTS)	
Writing WIAS research proposal	2016
MSc-course Machine learning (FTE-35306)	2017
Summer course in ChIP seq (wet-lab)	2019
External training period - INRAE, France	2019-2020
Professional competences (21.1 ECTS)	
Project and time management	2016
The Essentials of Scientific Writing and Presenting	2016
Scientific Writing	2017
Communication with the media and the general public	2017
Writing Scientific Proposal	2017
Societal impact of your research	2018
Career perspectives	2018
Scientific Artwork – Vector graphics and images	2018
Supervising Bsc & Msc thesis students	2019
WIAS Science day chair	2017-2018
WIAS PhD council chair	2018-2019
Organization Ecological Genomics session (NAEM)	2019

Presentation skills (maximum 4 ECTS)	
Benelux Congress of Zoology, the Netherlands - oral	2017
IMAGE meeting, Göttingen, Germany - oral	2017
BioSB, Lunteren, the Netherlands – poster	2018
IMAGE meeting, Vienna, Austria - oral	2018
SMBE, Japan - poster	2018
WIAS Science Day, Lunteren, the Netherlands - oral & poster	2019
ESEB, Finland - oral	2019
IMAGE meeting, Brescia, Italy - oral	2019
AIEM, France - oral	2019
Conference on Animal Genetic Resources, Madrid, Spain - oral	2020
FAANG, Prague, Czech Republic - oral	2020
Teaching competences (maximum 6 ECTS)	
Assisting BSc and MSc thesis ring	2017
Assisting MSc-course Population & Quantitative Genetics	2017
Assisting MSc-course Genomics (3 times)	2017-2018
Assisting PhD-course (IMAGE)	2018
Supervising MSc thesis student (2 times)	2018-2019
Supervising BSc thesis student	2019
Total	**48.4 ECTS**

Publications

Peer-reviewed publications

Groß, C.*, **Bortoluzzi, C.***, de Ridder, D., Megens, H.J., Groenen, M.A.M., Reinders, M., Bosse, M. (2020). Prioritizing sequence variants in conserved non-coding elements in the chicken genome using chCADD. *PLoS Genetics* (accepted for publication).

Bortoluzzi, C., Megens, H. J, Bosse, M., Derks, M. F. L., Dibbits, B., Laport, K., Weigend, S., Groenen, M. A., Crooijmans, R. P. (2020). Parallel evolution of foot feathering in birds is mirrored by parallel evolution of genes. *Molecular Biology and Evolution* (https://doi.org/10.1093/molbev/msaa092).

Bortoluzzi, C., Bosse, M., Derks, M. F., Crooijmans, R. P., Groenen, M. A., Megens, H. J. (2020). The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. *Evolutionary applications*, 13(2), 330-341.

Upadhyay, M., **Bortoluzzi, C**., Barbato, M., Marsan, P. A., Colli, L., Ginja, C., Sonstegard, T., Bosse, M., Groenen, M.A., Crooijmans, R. P. Deciphering the patterns of genetic admixture and diversity in southern European cattle using Genome-wide SNPs. *Evolutionary applications*, 12(5), 951-963.

Heidaritabar, M., Bijma, P., Janss, L., **Bortoluzzi, C.**, Nielsen, H. M., Madsen, P., Ask, B., Christensen, O. F. (2019). Models with indirect genetic effects depending on group sizes: a simulation study assessing the precision of the estimates of the dilution parameter. *Genetics Selection Evolution*, 51(1), 24.

Bortoluzzi, C., Crooijmans, R. P., Bosse, M., Hiemstra, S. J., Groenen, M. A., Megens, H. J. (2018). The effects of recent changes in breeding preferences on maintaining traditional Dutch chicken genomic diversity. *Heredity*, 121(6), 564.

In preparation and under review

Wu, Z., **Bortoluzzi, C.**, Derks, M. F. L., Langqing, L., Bosse, M., Hiemstra, S.J., Groenen, M. A., Crooijmans, R. P. (2020). Heterogeneity of a dwarf phenotype in Dutch traditional chicken breeds revealed by genomic analyses. *Evolutionary applications* (Under review)

Bortoluzzi, C., Restoux, G., Rougé, R., Desnoues, B., Petitjean, F., Bosse, M., Tixier-Boichard, M. (2020). Quantifying temporal genomic erosion in small managed populations under a recently established conservation programme. *In preparation.*

Li, J., **Bortoluzzi, C.**, Hodge, M., Bertrand, B., Davis, B. W., Dorshorst, B. J., Siegel, P. B., Tixier-Boichard, M., Andersson, L. (2020). The mottling phenotype in chickens shows genetic heterogeneity and is caused by mutations in EDNRB2. *In preparation.*

Conference proceedings and abstracts

Bortoluzzi, C., Groß, C, Bosse, M., de Ridder, D., Reinders, M., Megens, H.J. (2020). Conserved non-coding regions in the chicken genome harbour functionally important variation, *Functional Annotation of Animal Genomes (FAANG) workshop*, Prague, Czech Republic.

Bortoluzzi, C., Megens, H. J, Bosse, M., Derks, M. F. L., Dibbits, B., Laport, K., Weigend, S., Groenen, M. A., Crooijmans, R. P. (2020). Which genes to trick to grow feathered feet, *Conference on Animal Genetic Resources*, Madrid, Spain.

Bortoluzzi, C., Restoux, G., Tixier-Boichard, M. (2019). Quantifying temporal changes in genetic diversity and mutational load in recently managed populations, *Approche Interdisciplinaire de l'Evolution Moléculaire (AIEM)*, Toulouse, France.

Bortoluzzi, C., Bosse, M., Derks, M. F, Crooijmans, R. P, Groenen, M. A., Megens, H. J. (2019) . A genetic and evolutionary perspective on foot feathering in a domestic avian species, *European Society for Evolutionary Biology (ESEB)*, Turku, Finland.

Bortoluzzi, C., Bosse, M., Derks, M. F., Crooijmans, R. P., Groenen, M. A., Megens, H. J. (2019). Patterns of deleterious variation are shaped by population history, *WIAS Science Day*, Lunteren, the Netherlands.

Heidaritabar, M., Bijma, P., Janss, L., **Bortoluzzi, C.,** Nielsen, H. M., Ask, B., Christensen, O. F. (2018). Models with indirect genetic effects depending on group sizes – A simulation study assessing the precision of the estimates of the dilution. *World Congress on Genetics Applied to Livestock Production (WCGALP)*, New Zealand.

Bortoluzzi, C., Derks, M. F., Weigend, S., Groenen, M. A., Megens, H.J. (2018). Detection and characterization of deleterious variants in traditional chicken breeds. *Society of Molecular Biology and Evolution (SMBE)*, Yokohama, Japan.

Bortoluzzi, C., Bosse., M., Groenen, M. A., Megens. H. J. (2018). The influence of demography and recombination on the homozygosity landscape in the poultry genome. *4th Dutch Bioinformatics & Systems Biology conference (BioSB)*, Lunteren, Wageningen.

Bortoluzzi, C., Crooijmans, R. P., Groenen, M. A., Megens, H. J. (2017). Unravelling the genetic basis of founder phenotypes shared among traditional chicken breeds of divergent demographic history. *Benelux Congress of Zoology*, Wageningen, the Netherlands.

Colophon

The research has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353. Additional funding to join the Génétique Animale et Biologie Intégrative (GABI) group at INRAE was provided by the Wageningen Institute of Animal Sciences (WIAS) as PhD fellowship under project number 4169000100.

Cover designed by Moreno Monti e Matteo Tranchellini
www.chic-ken.it
Instagram @chicken_ph

Printed by: DigiForce — Proefschriftmaken.nl