RETRIEVING PATTERNS OF GENETIC DIVERSITY IN A GLOBAL SET OF CHICKEN BREEDS

Dissertation submitted

to obtain the Doctor of Philosophy (Ph.D) degree

at the Faculty of Agricultural Sciences,

Georg-August-University Göttingen, Germany

Presented by

Dorcus Kholofelo Malomane

Born in Bushbackridge, South Africa

Göttingen, September 2019

D7

Reference 1: Prof. Dr. Henner Simianer

Reference 2: Prof. Dr. Steffen Weigend

Reference 3: Prof. Dr. Armin Otto Schmitt

Day of disputation: 08 November 2019

TABLE OF CONTENTS

SUMMARY

The chicken was first domesticated about 6000 B.C in Asia. Today the species is widely spread across the globe providing a good source of quality protein. There have been concerns about the loss of genetic diversity within the species due to the rapid spread and domination of the highly intensive commercial lines which utilizes small number of breeds with limited genetic diversity. In addition, a strong phenotypic selection for fancy breeds which has become very popular in the 19th century has affected the genetic diversity of many populations. Genetic diversity in a population or a species is very important for its fitness e.g. adaptations to changing environments and resistance to diseases. Therefore, it is important to preserve the genetic resources in the chicken for its sustainability and to be able to respond to unforeseen circumstances. Genetic diversity studies are crucial in order to make informed decisions for the conservation and effective management of farm animal genetic resources.

Our study was focused on investigating the genetic diversity in a global set of chicken breeds based on the SYNBREED chicken diversity panel (SCDP). The SCDP consists of a total of 174 chicken populations from Asia, Europe, Africa and South America, which were genotyped with the 600K Affymetrix® Axiom™ HD Genotyping Array for chicken comprising of 580,961 SNPs. The panel includes two wild populations (*Gallus gallus gallus* and *Gallus gallus spadiceus*), 12 commercial lines (4 brown egg layers, 4 white egg layers and 4 broilers), 81 local breeds, and 79 fancy breeds of European and Asian backgrounds. Given the sensitivity of SNP data to ascertainment bias, we first investigated how we can mitigate the effects of ascertainment in our data when studying the genetic diversity.

In **chapter 2** we used 42 of the 174 populations from our data which had both individual genotype data as well as pool whole genome resequencing (WGS) data. We estimated various genetic diversity measures i.e. expected heterozygosity (H_e), fixation index (F_{ST}) values, phylogenetic analysis and principal components analysis (PCA), using the SNP array and WGS data, and compared the results. The array data overestimated the H_e, underestimated the pairwise F_{ST} values between breeds which had low F_{ST} values (<0.25) in the WGS data, and overestimated F_{ST} values (>0.25) for high WGS F_{ST}. The PCA and phylogenetic analysis were less affected by ascertainment bias. Subsequently, we applied different SNP filtering options such as SNPs polymorphic in the *Gallus gallus* (founder populations), linkage disequilibrium (LD) based pruning and minor allele frequency (MAF) filtering and the combinations thereof to the array data. Then we assessed the option/s that could account for the ascertainment bias effects and were therefore viable to improve the accuracy of subsequent studies. Generally, the LD based pruning of SNPs produced better results which were comparable to the WGS. The overestimation of H_e was slightly reduced and F_{ST} values were a little lower than in the WGS data, but in a systematic manner.

We performed the LD based pruning of SNPs to the array data for further genetic diversity analyses in **chapters 3** and **4**. In **chapter 3** we studied the overall genetic diversity within and between the chicken populations. PCA and admixture analysis showed a continuous separation of Asian breeds at one end and European breeds on the other. The African and South American breeds clustered mostly in between but slightly towards either the Asian or the European cluster, supporting their Asian and European backgrounds of origin. The commercial white layers clustered towards European breeds and the brown layers and broilers clustered with the Asian breeds, reflecting their parental backgrounds. Furthermore, the fancy breeds covered a wide spectrum of the genetic diversity and clustered with other breeds of similar origin. However, the fancy breeds and the

highly selected commercial layer lines showed low genetic diversity within population with average observed heterozygosity lower than 0.205 across breeds' categories. The wild and less selected African, South American and some local Asian and European breeds showed high within population genetic diversity with the average observed heterozygosity greater than 0.225.

In **chapter 4** we further investigated if the observed overall genetic diversity within the populations in **chapter 3** is a consequence of their genetic expansion from the wild populations. We studied this following the single founder migration model which asserts that the genetic diversity within populations decreases with the increase in geographic distances from their founders. Additionally, as a consequence of the geographic expansion, genetic differentiation is expected to increase between the populations and the founder population. In our study we didn't have geographic distances and the geographical location of the sampling in chicken often does not coincide with the geographical location of the breed development. Therefore, we used the Reynolds' genetic distance of the sampled breeds to the wild ancestors as a proxy for geographic distance, and verified, whether the reduction of diversity can also be found with increasing genetic distance to the domestication center. We found that 87.5% of the variation in the overall genetic diversity within the domestic populations can be explained by the Reynolds' genetic distances to the wild populations. In comparison to the other SNP classes, the non-synonymous class was the most deviating from to the overall pattern. The changes in the genetic diversity due to the distance to founder populations was found to be fastest within genes that are associated with transmembrane transport, protein transport and protein metabolic processes, and lipid metabolic processes. In general, such genes are flexible to be manipulated according to the population's needs. On the other hand, genes with major functions e.g. brain development were more static and hence changes may have detrimental effects on the chickens.

Overall, the genetic diversity in the chicken has been influenced by management and breeding practices. The study has shown that local breeds have more genetic diversity due to less artificial genetic manipulation compared to fancy and commercial breeds. The study shed insights into the global genetic diversity and provides a good future reference in the global chicken diversity studies.

ZUSAMMENFASSUNG

Das Huhn wurde erstmals um 6000 v. Chr. in Asien domestiziert. Heute ist diese Art weltweit verbreitet und stellt eine bedeutsame Quelle für hochwertiges Protein in der Ernährung dar. Mit der raschen Ausbreitung und Vorherrschaft der hochintensiven Produktionslinien, die eine geringe Anzahl von Rassen mit begrenzter genetischer Vielfalt nutzen, gehen Bedenken über einen globalen Verlust der genetischen Vielfalt innerhalb dieser Art einher. Darüber hinaus hat die im 19. Jahrhundert aufkommende Rassegeflügelzucht, welche sich vornehmlich starker Phänotypselektion bediente, die genetische Vielfalt vieler Populationen beeinflusst. Die genetische Vielfalt in einer Population oder einer Art hat direkten Einfluss auf deren Fitness, z.B. bei der Anpassung an veränderte Umweltbedingungen und die Resistenz gegen Krankheiten. Daher ist es wichtig, diese genetische Vielfalt des Huhns als Ressource zu erhalten um auf unvorhergesehene Umstände reagieren zu können. Studien zur genetischen Diversität sind daher von entscheidender Bedeutung, um fundierte Entscheidungen für die Erhaltung und das effektive Management solcher genetischen Ressourcen zu treffen.

Diese Studie konzentrierte sich auf die Untersuchung der genetischen Vielfalt in einer globalen Stichprobe von Hühnerrassen, welche in dem SYNBREED chicken diversity panel (SCDP) repräsentiert sind. Das SCDP besteht aus insgesamt 174 Hühnerpopulationen aus Asien, Europa, Afrika und Südamerika, die mit dem 600K Affymetrix® Axiom™ HD Genotyping Array für das Huhn, welcher 580.961 SNPs enthält, genotypisiert wurden. Das Panel umfasst zwei Wildpopulationen (Gallus gallus gallus und Gallus gallus spadiceus), 12 kommerzielle Linien (4 braune Legelinien, 4 weiße Legelinien und 4 Broiler), 81 lokale Rassen und 79 Schaurassen mit europäischem und asiatischem Hintergrund. Angesichts der bekannten Sensitivität von SNP-Daten gegenüber dem sogenannten „SNP Ascertainment bias“, welcher auf einer mangelnden

Repräsention der genetischen Variabilität einer bestimmten Rasse durch die auf dem SNP-Chip enthaltenen Varianten basiert, haben wir zunächst untersucht, wie wir dessen Auswirkungen auf unsere Analysen möglichst minimieren können.

In Kapitel 2 verwendeten wir die 42 der 174 Populationen, für die genetische Information sowohl in Form individueller Genotypdaten als auch als Vollgenomsequenzpoolsvorlag (WGS). Wir analysierten die Diversität mit verschiedenen Methoden, darunter die erwartete Heterozygotie (H_e), der Fixation Index (F_{ST}), phylogenetische Bäume und Hauptkomponentenanalyse (PCA), unter Verwendung der SNP-Array und WGS-Daten, und verglichen die Ergebnisse. Die Array-Daten überschätzten die H_e und unterschätzten die paarweisen F_{ST}-Werte zwischen Rassen, welche niedrige F_{ST}-Werte (<0,25) in den WGS-Daten hatten, und überschätzten für hohe F_{ST}-Werte (>0,25). Die PCA- und phylogenetische Analyse waren von der Verzerrung der Ermittlungen weniger betroffen. Anschließend wandten wir verschiedene SNP-Filteroptionen auf die Arraydaten an: Wir behielten SNPs welche polymorph im *Gallus gallus* (Gründerpopulationen) waren, Filterten zur Verminderung des Kopplungsungleichgewichtes (LD), Filterten für eine bestimmte Minor Allel Frequenz (MAF), als auch mit Kombinationen aus diesen Filtern. Dann bewerteten wir die Option(en) hinsichtlich ihrer Fähigkeit den Ascertainment bias zu vermindern. Im Allgemeinen lieferte der LD-basierte Filterung von SNPs Ergebnisse, die die besser mit denen auf den WGS Daten geschätzten vergleichbar waren. Die Überschätzung von H_e wurde leicht reduziert und die F_{ST}-Werte waren etwas, aber systematisch, niedriger als in den WGS-Daten.

Wir nutzten die LD-basierte Filterung der Array-Daten für weitere genetische Diversitätsanalysen in den Kapiteln 3 und 4. In Kapitel 3 haben wir die gesamte genetische Vielfalt innerhalb und zwischen den Hühnerpopulationen untersucht. Die PCA- und Admixtureanalyse zeigte eine

eindeutige Trennung der asiatischen Rassen an einem Ende und der europäischen Rassen am anderen Ende. Die afrikanischen und südamerikanischen Rassen konzentrierten sich hauptsächlich zwischen, jedoch leicht in Richtung entweder des asiatischen oder europäischen Clusters, was durch ihre Rassenhistorie erklärt werden kann. Die kommerziellen weißen Legelinien waren zwischen den europäische Rassen und die braunen Legelinien und Masthühner zwischen den asiatischen Rassen zu finden, was ihren elterlichen Hintergrund widerspiegelt. Darüber hinaus deckten die Schaurassen ein breites Spektrum der genetischen Vielfalt ab und gruppierten sich mit anderen Rassen ähnlicher Herkunft. Diese und die hochselektierten kommerziellen Legelinien zeigten jedoch eine geringe genetische Vielfalt innerhalb der Population mit einer durchschnittlichen beobachteten Heterozygotie von weniger als 0,205 über alle Kategorien von Rassen. Die wilden und weniger selektierten afrikanischen, südamerikanischen und lokalen asiatischen und europäischen Rassen zeigten eine hohe genetische Vielfalt innerhalb der Populationen mit einer durchschnittlichen beobachteten Heterozygotie von mehr als 0,225.

In Kapitel 4 haben wir weiterhin untersucht, ob die beobachtete gesamte genetische Vielfalt innerhalb der Populationen in Kapitel 3 eine Folge ihrer genetischen Expansion aus den Wildpopulationen ist. Wir untersuchten dies nach dem Single-Gründer-Migrationsmodell, das darauf basiert, dass die genetische Vielfalt innerhalb der Populationen mit zunehmender geographischer Entfernung von ihren Stammvätern abnimmt. Darüber hinaus wird erwartet, dass als Folge der geografischen Expansion die genetische Differenzierung zwischen den Populationen und der Gründerpopulation zunehmen wird. In unserer Studie konnten wir nicht auf geografische Informationen zurückgreifen und die geografische Lage der Stichprobe beim Huhn stimmt oft nicht mit der geografischen Lage tatsächlichen Rassenentwicklung überein. Daher haben wir die genetische Entfernung der Rassen von der Ursprungspopulation mit Hilfe der Reynoldsdistanz

geschätzt und anstelle der geographischen Entfernung verwendet, um zu überprüfen, ob die Reduktion der Vielfalt auch mit zunehmender genetischer Entfernung zum Domestizierungszentrum bestätigt werden kann. Wir fanden heraus, dass 87,5% der Variation der gesamten genetischen Vielfalt innerhalb der einheimischen Populationen durch den genetischen Abstand zu den Wildpopulationen erklärt werden kann. Im Vergleich zu den anderen SNP-Klassen war die Klasse der nicht-synonymen Substitutionen diejenige, die vom Gesamtmuster am meisten abweicht. Die Veränderungen in der genetischen Vielfalt durch die Entfernung zu den Gründerpopulationen wurden am schnellsten innerhalb von Genen gefunden, die mit Transmembrantechnologie, Proteintransport und Proteinstoffwechselprozessen sowie Lipidstoffwechselprozessen in Verbindung gebracht werden konnten. Im Allgemeinen sind dieses Gene welche Veränderungen erfahren, wenn die Zucht der zugrunde liegenden Rasse auf die Bedürfnisse der Bevölkerung ausgerichtet wird. Andererseits waren Gene mit Hauptfunktionen, wie z.B. der Gehirnentwicklung, statischer, da Veränderungen hier nachteilige Auswirkungen auf die Hühner haben würden.

Insgesamt wurde die genetische Vielfalt beim Huhn durch Management- und Zuchtpraktiken beeinflusst. Diese Studie hat gezeigt, dass lokale Rassen mehr genetische Vielfalt haben, da sie im Vergleich zu Schau- und kommerziellen Rassen weniger starke züchterische Manipulationen aufweisen. Die Studie vermittelt Einblicke in die globale genetische Vielfalt und bietet eine gute Referenz für in Zukunft am Huhn durchgeführten Diversitätsstudien.

CHAPTER 1

General introduction

The origin of chicken

Chickens are of the family of *Phasianidae* and the genus *Gallus*. Four types or species of wild chickens are reported in the modern studies of birds, the red (*Gallus gallus*), grey (*Gallus sonnerati*), green (*Gallus varius*), and Ceylon (*Gallus lafayettii*) jungle fowls [1]. The origin of all these wild ancestors is distributed across South and Southeast Asia as well as Southwest China as demonstrated in Figure 1.1 [2]. The *Gallus gallus* presumably originates from South and Southeast Asia as well as from Southwest China (Yunnan province), the *Gallus sonnerati* from India, the *Gallus varius* from Java islands, and the *Gallus lafayettii* from Sri Lanka. The red jungle fowl (RJF) consists of five subspecies, *G. gallus gallus, gallus spadiceus, gallus murghi, gallus bankiva* and *gallus jabouillei.* Although it is not clear to what extent the other three wild type species contribute to the modern chickens, it has been established that the RJF is the main progenitor of the widely distributed chicken of today, the *Gallus gallus domesticus* [1–4]. Although it has been suggested that earliest domestication of chickens may have taken place around 6000 B.C. or earlier in China [2], research studies point out that the main precursor of widely spread today's chicken diversity is from the domestication events that took place in the Indus Valley during 2500-2100 B.C.[1].

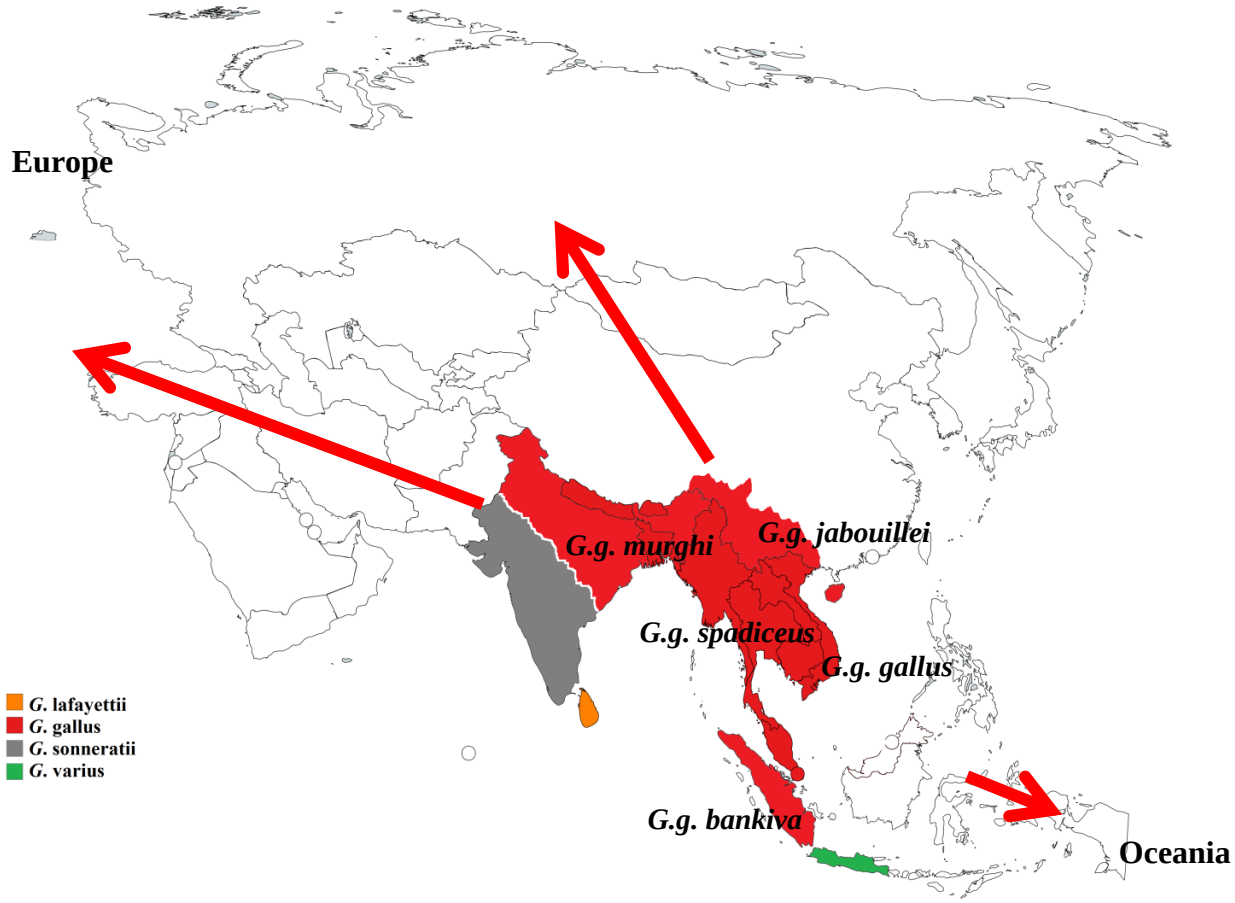


Figure 1.1: The distribution of chicken wild species. Red arrows show the westwards and eastwards directions of dispersion after domestication.

From centers of domestication in Asia to the world: a brief history of chicken dispersion

The dispersion of chicken across the world has mainly been facilitated by human migration and trading. After domestication, chickens were taken westward to Europe and further eastwards to Island Southeast Asia and the Oceania/the Pacific [5]. Domesticated livestock including chickens mainly arrived in the Oceania regions when people colonized the islands [5, 6]. It is suggested that chickens were brought to the pacific islands in varying times which might date to as early as around 4500 B.P. [5].

Europe. The main period of domestic chickens' dispersion throughout Europe was around 450 to 1100 B.C. [2]. However, reports suggest that there has been chickens in Europe from as early as at least 4000 to 3000 B.C. [2, 4]. Chickens are believed to have been brought on two main routes into Europe, one through the south via Persia and Greece to the Mediterranean region and another one through the north via China and Russia to Northern Europe [1].

Africa. Information on dispersion of domestic chickens into Africa is very sketchy. Reports suggest that chickens existed in pictorial records in Egypt before 1400 B.C. [1]. The general belief is that chickens entered Africa from Asia through the Indian ocean coastline and from Europe through the north (Horn of Africa) [7, 8]. However, there is much speculations and lack of clarity on when exactly chickens reached the African continent and by which route or entry point. Studies based on analyzing mitochondrial DNA suggested that the most common and possibly the earliest haplogroup in Africa may be originating from South Asia (Indian subcontinent) and could have entered the eastern part of Africa through three possible routes: through the Middle East into Egypt, through the Horn of Africa or directly to Coastal East Africa [8].

The Americas. Chickens presumably arrived in the Americas quite late after the time of domestication and from several sources. Chickens were introduced to South America from Polynesia [9] and from Europe in the 15th century [10], additionally it is believed that Europeans also brought forth Asian breeds to South America as well as chickens from Africa through slave trading [5].

Main categories of breeds forming the global chicken diversity

From the wild species, many breeds and lines have been developed which are currently spread across the globe. While some breeds may have evolved naturally, many other breeds were also created by cross breeding and high selection programs to enhance or produce new phenotypes for different purposes. Below the main categories of chickens (local, fancy and commercial type breeds) are described. The classification is mainly based on the utilization of different breeding and management practices that have resulted in the current status of global chicken diversity.

Local breeds refer to the native and village or traditional chicken breeds, often indigenous, and well-adapted to the country or region. These breeds are usually raised in low input production systems [11]. In many developing countries, the chickens are kept under extensive production systems, they are usually kept in backyards, sleep on trees, house corners, scavenging for food or fed on left overs with limited or no supplementary food. There is also no/limited routine health check or vaccination against diseases, no structured breeding programs nor selection programs in place [12–14], and intercrossing occurs between nearby villages or populations [15–17]. In rural villages where many of these chickens are kept, the local chickens are mostly used for household consumption, to a less extent for sales, gifts and traditional rituals [12]. Local breeds are often associated with low productivity which poses challenges for their existence as local consumers become attracted to the productivity of commercial lines [14, 17].

Fancy breeds. Fancy breeding is characterized by breeding for physical appearance in accordance with the poultry standards e.g. [18, 19]. One of the oldest poultry standards were established by the British and got published in 1865 [19]. About a decade later, North America published their first standards for fancy breeding named the ‘American Standard of Perfection’ administered by the American Poultry Association (http://www.amerpoultryassn.com/). Many poultry breed

standards were established in the 19th century by poultry breeds' associations and clubs in order to give complete description and guidelines of how a specific breed should look like. Therefore, participants aim to produce such an ideally 'perfect' description. The description could be for a particular physical appearance e.g. feather color, miniature, skin color, or for behavioral characteristics, e.g. fighters. In Europe, participants following the European Poultry Standards [18] explored many phenotypes from breeds which have been maintained in Europe for decades and new breeds which were brought to Asia during the 19th century, either keeping them as purebreds or crossing them to produce new phenotypes. German fancy breeds present an important asset of genetic diversity which covers a wide variety of breeds. However, due to the strict requirements to meet these standards or the perfect phenotypes, breeders in these associations practice very refined selection for their breeding stock and even mate very closely related individuals in order to achieve the perfect phenotypes. Such practices may be detrimental as they enhance the loss of genetic diversity, propagate negative consequences of inbreeding and could endanger the survival of the breed.

Commercial breeds refer to breeds which are raised primarily for profit. Commercial chicken breeding companies are specialized in meat (broilers) and egg production types whereby breeds with good productive characteristics for either meat or eggs have been developed and subsequently selected for such traits. Currently, the commercial egg layers consist of two main types, the white and brown egg layers. The commercial chicken breeding industry now provides most chickens and is spread all over the world. Commercial breeding companies are characterized by sophisticated breeding programs including well defined breeding goals and highly intense production systems with strict health regime and elaborate housing and feed administration.

These three main types of breeds present a wide range of phenotypic features and make up the current global chicken diversity. The diversity ranges from different aspects of production, reproduction, growth and behavior (e.g., broodiness, fighters and recognition) as well as physical aspects (e.g., comb type and color, plumage color, shank length, egg thickness and color) [20–22]. Furthermore, there are differences at the genomic level which are underlying these phenotypic differences. The conservation of genetic variation is important for the preservation of the species especially in view of the increasingly alarming changes on the planet earth, e.g. global warming. Genetic diversity studies are crucial to understand important genetic variants for different situations and conditions to effectively manage the chicken genetic resources as well as making informed decisions for current utilization and preservation of important genetic resources for the future [23].

Acquisition and use of genomic data for genetic diversity studies

Many studies have commonly used genetic markers especially microsatellites in chicken genetic diversity studies e.g. [7, 24]. Currently, the most popular types of genomic data used are the whole genome sequencing (WGS) and single nucleotide polymorphism (SNP) genotype data. Besides the WGS data, the SNP data is preferred among other genetic marker data because they are the most abundant form of genomic variation containing more in-depth information. Although the whole genome sequencing data is still the most effective way of studying the genome wide diversity, acquiring such data remains very expensive and requires additional infrastructure (e.g. good reference genomes) and therefore poses more limitations to conduct large studies. Therefore, SNP genotyping is commonly used as an affordable alternative with less infrastructural requirement, less effort and time. SNP genotyping data is acquired by using already developed SNP panels such as genotyping arrays or chips. The single nucleotide polymorphisms in the panels

are acquired by sequencing a set of individuals which are selected from a limited number of populations. Specified criteria (e.g., minimum allele frequency, even distribution of SNPs across the genome) are then applied to select the SNP panel to be used. The selected SNP panels are used for the genotyping of many individuals from different populations. Because of the selection procedure of the SNPs, the SNP panels may suffer from ascertainment bias. Ascertainment bias is a systematic deviation in statistical population genetic measures from true values due biased marker discovery protocols [25].

Some of the problems or shortfalls that come with these methods of SNP discovery include the fact that since the discovery of SNPs is dependent on the allele frequency, there is less chance of discovering rare SNPs in a small sample set compared to the common SNPs and therefore, many rare SNPs are missing [26]. Consequently, the estimates of genetic diversity that depend on the allele frequencies will also be affected [27]. There may be overestimation of genetic diversity in some populations, especially in those that are included in the discovery panel [28]. Overall the classical population genetics methods are designed for whole genome data or randomly sampled SNPs across many loci and do not account for the ascertained genotype data. Therefore, when these methods are applied to SNP genotypes without accounting for ascertainment bias, they will produce inaccurate results [27, 29].

The high density SNP genotyping array for chicken

The latest and highest density array for the chicken was released in 2003, the 600K Affymetrix® Axiom® HD genotyping array for chicken [30]. The array was designed using twenty-four chicken lines, which consisted of fifteen commercial lines (4 broilers, 6 white egg layers and 5 brown layers), eight experimental inbred layer lines and one unselected layer line. It contains 580,954 SNPs of which 21,534 are coding variants, providing the opportunity to explore the genetic

diversity of chickens [30]. The work presented in **chapters 2 to 4** is based on chicken data genotyped with the Affymetrix 600K genotyping array.

SNP annotation and functional classification

The first draft of the chicken reference genome, based on the red jungle fowl (*Gallus gallus*), was released in 2004 [31]. The development of the Affymetrix 600K genotyping array was based on the fourth version of the reference genome, the Gallus_gallus-4.0, which was released in 2011 containing 28 of the 38 autosomes, both sex chromosomes and two linkage groups. Currently, on the Affymetrix official webpage (now acquired by Thermo Fisher Scientific, https://www.thermofisher.com/de/de/home.html), the Affymetrix 600K genotyping array is accompanied by the annotation map based on the fifth version of the genome assembly, the Gallus_gallus-5.0 which contains three additional autosomes (GGA30, 31, and 33) [32]. However, a new version of the reference genome (Gallus_gallus-6.0) is already available. The Affymetrix Gallus_gallus-5.0 annotation map contains information of genes associated with the SNPs as well as the type of consequences of the SNPs. The SNP consequences were classified following the Ensembl variant predictor [33]. SNPs can be classified into two major groups, genic and intergenic (non-genic). Figure 1.2 shows the different SNP variants within and between the genes (for descriptions of SNP variants see Table 1.1).

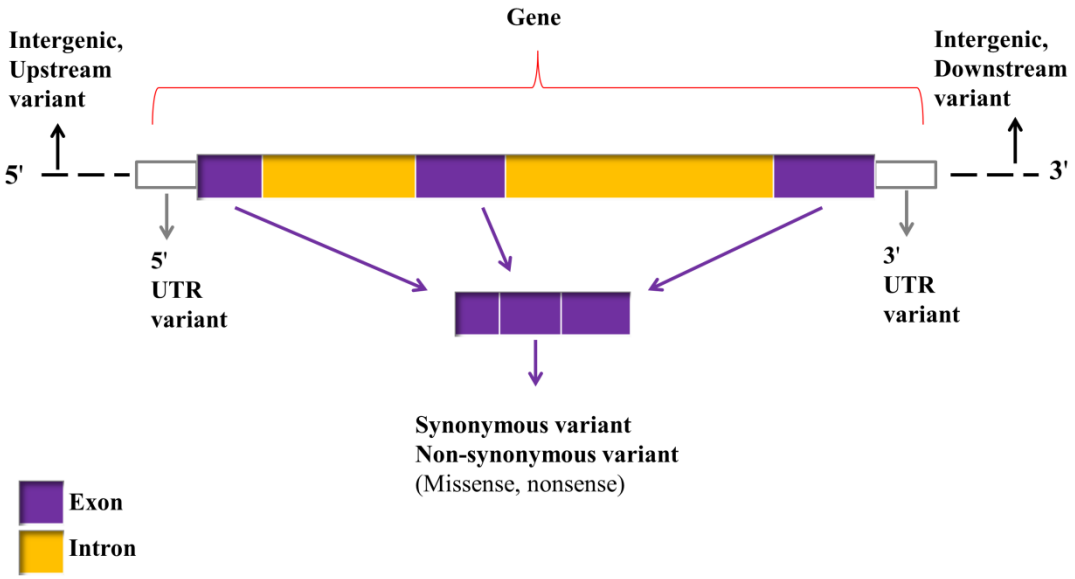


Figure 1.2: Diagram showing the location of different SNP variants (modified from McLaren et al. 2010)

Table 1.1: Description of SNP variants

Type of SNPs	Description
Exonic SNPs	Variants within the coding region of the gene, may or may not result in the alteration of the amino acids as follows: • the ***synonymous*** type refers to the sequence variant that does not change the amino acid, • the ***non-synonymous*** nucleotide substitution changes one or more bases, resulting in a different amino acid sequence.
Intronic SNPs	Variants within the noncoding region of a gene (introns) which is not translated into the protein.
5' UTR SNPs	The transcribed SNPs located at the 5' end of the gene but are not translated into the protein.
3' UTR SNPs	The transcribed SNPs located at the 3' end of the gene but are not translated into the protein.
Upstream SNPs	Variants which are located adjacent to the 5' UTR region.
Downstream SNPs	Variants which are located adjacent to the 3' UTR region.
Intergenic SNPs	SNPs which are not part of the gene but are located in the upstream or downstream region.

Model for the cause of genetic differentiation between populations

The theory of '*isolation by distance*' can be used to easily understand the cause of genetic differentiation between populations. The theory of '*isolation by distance*' refers to the decrease in genetic relatedness with the increase in geographic distance [34, 35]. The theory was first introduced by Sewall Wright to articulate the patterns of differentiation under dispersal [35]. In a large, random mating population genetic differentiation may be brought forth by local patterns of selection or random mutations, as well as limited possibility of mating between those individuals that are not in close proximity. Such differentiation may result in the creation of certain population structures. However, the differentiation is limited to some extent by lack of isolation [36]. Ishida [37] describes Wright's concept as '*ecological isolation by distance*' as it concerns the local interaction of individuals. However, where genetic associations are restricted by geographic separations due to population migration, population genetic patterns reflect the differentiation among the subpopulations. This is termed the theory of '*genetic isolation by distance'* according to Malécot [34]. Such differentiation under long geographic distances occurs because the consequences of genetic drift act more rapidly than the potential or chance of an individuals' interaction under dispersal [38]. Therefore, in general, genetic relatedness of individuals is defined by the local (ecological) aspects and large distances of geographic separations.

A single founder migration model of genetic diversity

When a small number of individuals migrate from a single large population to form a new population in a new territory, they carry only a subset of the genetic diversity of that present in the large population. This phenomenon of the change in genetic diversity is called the 'founder effect' [39]. The smaller the migrated population, the more vulnerable it is to the effects of genetic drift, e.g. population bottlenecks. Not only does the migrated population lose genetic diversity due to

genetic drift, but it may also result in population differentiation from the founder population as described by the theory of '*genetic isolation by distance'*. If there is/are subsequent migration/s from the newly formed populations to farer distances, the genetic diversity further decreases, an event termed 'serial founder effect'. Furthermore, genetic differentiation further increases between the subsequent migrants and the original founder population. Using the so called 'Out of Africa' theory, which assigns Africa as the origin of modern humans, studies have applied population genetics methods to establish the expansion of populations from a single founder as the best explanatory factor of the geographical patterns of genetic diversity within a species [40].

Measures of genetic diversity and population structure

There are many measures of genetic diversity between and within populations and in the following parts we briefly introduce some of those which have been often applied in this thesis.

Reynolds' genetic distance

Reynolds' genetic distance is a measure of population divergence by genetic drift [41]. This measure of distance is based on the coancestry coefficient and it is estimated as:

$$D_R = \frac{1}{2}\sum_i \frac{\sum_i (p_{1i} - p_{2i})^2}{1 - \sum_i (p_{1i} p_{2i})}$$

where i is the i^{th} allele at bi-allelic loci and p_{1i} is the frequency of the i^{th} allele at bi-allelic loci locus in population 1.

Heterozygosity

Heterozygosity is the state of having two different alleles at one locus. It is used as a measure of genetic variability within a population. The expected heterozygosity is calculated as:

$$H_e = 2p\,(1-p)$$

where p represents the allele frequency of one allele [42]. High fixation of alleles (e.g. by selection) results in the decrease in heterozygosity and therefore loss of genetic variability.

Wright's fixation index (F_{ST})

Wright's fixation index is a popular measure of population differentiation and was introduced by Wright [43]. It measures the proportion of genetic variance between populations based on the allele frequencies with values ranging from 0 to 1, where 0 indicates that there is no genetic differentiation between populations. It is calculated as

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}\,(1-\bar{p})}$$

where σ_S^2 is the variance of allele frequency between subpopulations and σ_T^2 is the variance of allele frequency in the total populations. In the 'Methods' section of **chapter 2** we show how different sample sizes are accounted for following the recommendations of Weir and Cockerham [44].

Principal component analysis (PCA)

Principal component analysis is a statistical technique that transforms a high number of observed variables which are possibly correlated into low dimensional data of artificial, uncorrelated variables. These key variables, called principal components, account for most of the variation of the observations [45–47]. PCA makes it easy to explore data and to visualize the relatedness of individuals or populations in a simpler form.

The aim and objectives

Studying and understanding the diversity of a species is crucial for making informed decisions for the conservation and effective management of farm animal genetic resources, as well as for understanding different evolutionary dynamics of the species. The main aim of the thesis was to investigate the genetic diversity in global chicken populations starting from the centers of chicken domestication in Asia. We had access to high density SNP genotype data and WGS data. The studies of genetic diversity based on SNP genotype data may produce misleading results due to ascertainment bias. Therefore, our first objective was to investigate different strategies to mitigate the effects of ascertainment bias when using SNP genotype data. The second objective was to investigate the overall genetic diversity between and within the globally collected chicken populations. The third objective was to investigate to what extent the observed overall genetic diversity in the chicken populations is a result of their genetic expansion from their wild founders in Asia.

References

[1] Crawford RD. Origin and history of poultry species. In: Crawford RD, editor. Poultry Breeding and Genetics. Amsterdam-Oxford-Newyork-Tokyo: Elsevier; 1990. p. 1–42.

[2] West B, Zhou BX. Did chickens go North? New evidence for domestication. J Archaeol Sci. 1988; 15: 515–33.

[3] Liu YP, Wu GS, Yao YG, Miao YW, Luikart G, Baig M, et al. Multiple maternal origins of chickens: Out of the Asian jungles. Mol Phylogenet Evol. 2006; 38: 12–9.

[4] Tixier-Boichard M, Bed'Hom B, Rognon X. Chicken domestication: From archeology to genomics. C R Biol. 2011; 334: 197–204.

[5] Storey AA, Athens JS, Bryant D, Carson M, Emery K, DeFrance S, et al. Investigating the global dispersal of chickens in prehistory using ancient mitochondrial dna signatures. PLoS One. [2012; 7: e39171.

[6] Storey AA, Spriggs M, Bedford S, Hawkins SC, Robins JH, Huynen L, et al. Mitochondrial DNA from 3000-year old chickens at the Teouma Site , Vanuatu. J Archaeol Sci. 2010; 37: 2459–68.

[7] Lyimo CM, Weigend A, Msoffe PL, Eding H, Simianer H, Weigend S. Global diversity and genetic contributions of chicken populations from African, Asian and European regions. Anim Genet. 2014; 45: 836–48.

[8] Mwacharo JM, Bjørnstad G, Mobegi V, Nomura K, Hanada H, Amano T, et al. Mitochondrial DNA reveals multiple introductions of domestic chicken in East Africa. Mol Phylogenet Evol. 2011; 58: 374–82.

[9] Storey AA, Ramírez JM, Quiroz D, Burley D V, Addison DJ, Walter R, et al. Radiocarbon and DNA evidence for a pre- Columbian introduction of Polynesian chickens to Chile Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. Proc Natl Acad Sci. 2007; 104: 10335–10339.

[10] Gongora J, Rawlence NJ, Mobegi VA, Jianlin H, Alcalde JA, Matus JT, et al. Indo-European and Asian origins for Chilean and Pacific chickens revealed by mtDNA. Proc Natl Acad Sci. 2008; 105: 10308–13.

[11] Besbes B, Tixier-Boichard M, Hoffmann I, Jain G. Future trends for poultry genetic resources. In: Proceedings of the International Conference of Poultry in the 21st Century: Avian Influenza

and Beyond. Bangkok; 2007.

[12] Ekue FN, Poné KD, Mafeni MJ, Nfi AN, Njoya J. Survey of the traditional poultry production system in the Bemenda area, Cameroon. FAO. 2002; January 2002: 15–25.

[13] Fathi M, El-Zarei M, Al-Homidan I, Abou-Emera O. Genetic diversity of Saudi native chicken breeds segregating for naked neck and frizzle genes using microsatellite markers. Asian-Australasian J Anim Sci. 2018; 31: 1871–80.

[14] Fulton JE, Berres ME, Kantanen J, Honkatukia M. MHC-B variability within the Finnish Landrace chicken conservation program. Poult Sci. 2017; 96: 3026–30.

[15] Adebambo AO, Mobegi VA, Mwacharo JM, Oladejo BM, Adewale RA, Iloro LO, et al. Lack of Phylogeographic Structure in Nigerian Village Chickens Revealed by Mitochondrial DNA D-loop Sequence Analysis Lack of Phylogeographic Structure in Nigerian Village Chickens Revealed by Mitochondrial DNA D-loop Sequence Analysis. Int J Poult Sci. 2010; 9: 503–7.

[16] Muchadeyi FC, Eding H, Wollny CBA, Groeneveld E, Makuza SM, Shamseldin R. Absence of population substructuring in Zimbabwe chicken ecotypes inferred using microsatellite analysis. Anim Genet. 2007; 38: 332–9.

[17] Qu L, Li X, Xu G, Chen K, Yang H, Zhang L, et al. Evaluation of genetic diversity in Chinese indigenous chicken breeds using microsatellite markers. Sci China, Ser C Life Sci. 2006; 49: 332–41.

[18] Rassegeflügel-Standard für Europa in Farbe. Bund Deutscher Rassegeflügelzüchter (ed.), Howa Druck & Satz GmbH, Fürth. ISBN 987–3–9806597-1-0.

[19] Tegetmeier WB. The Standard of Excellence in Exhibition Poultry, authorized by the Poultry Club. London: Groombridge and Sons; 1865.

[20] Dahloum L, Moula N, Halbouche M, Mignon-Grasteau S. Phenotypic characterization of the indigenous chickens (Gallus gallus) in the Northwest of Algeria. Arch Anim Breed. 2016; 59: 79–90.

[21] Faruque S, Siddiquee N, Afroz M, Islam M. Phenotypic characterization of Native Chicken reared under intensive management system. J Bangladesh Agric Univ. 2010; 8: 79–82.

[22] Fathi MM, Al-homidan IH. Characterisation of Saudi native chicken breeds: a case study of morphological and productive traits. Worlds Poult Sci J. 2017; 73: 916–27.

[23] Tixier-Boichard M, Leenstra F, Flock DK, Hocking PM, Weigend S. A century of poultry genetics. Worlds Poult Sci J. 2012; 68: 307–21.

[24] Hillel J, Groenen MAM, Tixier-Boichard M, Korol AB, David L, Kirzhner VM, et al. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genet Sel Evol. 2003; 35: 533–57.

[25] Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME. Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. PLoS One. 2013; 8: e74612.

[26] Clark AG, Hubisz MJ, Bustamante CD, Williamson SH. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15: 1496–502.

[27] Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is

important, and how to correct it. BioEssays. 2013; 35: 780–6.

[28] Herrero-Medrano JM, Megens H-J, Groenen MA, Bosse M, Pérez-Enciso M, Crooijmans RP. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. BMC Genomics. 2014; 15: 601.

[29] Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004; 1: 218–24.

[30] Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013; 14: 59.

[31] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004; 432: 695–716.

[32] Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3: Genes, Genomes, Genetics. 2017; 7: 109–17.

[33] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010; 26: 2069–70.

[34] Malécot G. The mathematics of heredity. Translated by Yermanos DM. San Francisco, CA USA: Freeman; 1969.

[35] Wright S. Isolation by Distance. Genetics. 1943; 28: 114–38.

[36] Kimura M, Weiss GH. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics. 1964; 49: 561–76.

[37] Ishida Y. Sewall Wright and Gustave Malécot on Isolation by Distance. Philos Sci. 2009; 76: 784–96.

[38] Aguillon SM, Fitzpatrick JW, Bowman R, Schoech SJ, Clark AG, Coop G, et al. Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. PLoS Genet. 2017; 13: e1006911.

[39] Provine WB. Ernst Mayr: Genetics and speciation. Genetics. 2004; 167: 1041–6.

[40] Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005; 15: 159–60.

[41] Reynolds J, Weir BS, Cockerham CC. Estimation o f the coancestry coefficient: Basis for a short-term genetic distance. Genetics. 1983; 105: 767–79.

[42] Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th edition. Essex, UK: Longmans Green, Harlow; 1996.

[43] Wright S. The Genetical Structure of Populations. Ann Eugen. 1951; 15: 322–54.

[44] Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984; 38: 1358–70.

[45] Raychaudhuri S, Stuart JM, Altman RB. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. In: Pacific Symposium on Biocomputing. 2000. p. 455–66.

[46] Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303–4.

[47] Malomane DK, Norris D, Banga CB, Ngambi JW. Use of factor scores for predicting body weight from linear body measurements in three South African indigenous chicken breeds. 2014; 46: 331-335.

CHAPTER 2

Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies

Dorcus Kholofelo Malomane[1,2], Christian Reimer[1,2], Steffen Weigend[2,3], Annett Weigend[3], Ahmad Reza Sharifi[1,2], Henner Simianer[1,2]

[1]Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

[2]Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

[3]Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, Neustadt, Germany

Published in BMC Genomics

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4416-9

Abstract

Background: Single nucleotide polymorphism (SNP) panels have been widely used to study genomic variations within and between populations. Methods of SNP discovery have been a matter of debate for their potential of introducing ascertainment bias, and genetic diversity results obtained from the SNP genotype data can be misleading. We used a total of 42 chicken populations where both individual genotyped array data and pool whole genome resequencing (WGS) data were available. We compared allele frequency distributions and genetic diversity measures (expected heterozygosity (H_e), fixation index (F_{ST}) values, genetic distances and principal components analysis (PCA)) between the two data types. With the array data, we applied different filtering options (SNPs polymorphic in samples of two *Gallus gallus* wild populations, linkage disequilibrium (LD) based pruning and minor allele frequency (MAF) filtering, and combinations thereof) to assess their potential to mitigate the ascertainment bias.

Results: Rare SNPs were underrepresented in the array data. Array data consistently overestimated H_e compared to WGS data, however, with a similar ranking of the breeds, as demonstrated by Spearman's rank correlations ranging between 0.956 and 0.985. LD based pruning resulted in a reduced overestimation of H_e compared to the other filters and slightly improved the relationship with the WGS results. The raw array data and those with polymorphic SNPs in the wild samples underestimated pairwise F_{ST} values between breeds which had low F_{ST} (<0.15) in the WGS, and overestimated this parameter for high WGS F_{ST} (>0.15). LD based pruned data underestimated F_{ST} in a consistent manner. The genetic distance matrix from LD pruned data was more closely related to that of WGS than the other array versions. PCA was rather robust in all array versions, since the population structure on the PCA plot was generally well captured in comparison to the WGS data.

Conclusions: Among the tested filtering strategies, LD based pruning was found to account for the effects of ascertainment bias in the relatively best way, producing results which are most comparable to those obtained from WGS data and therefore is recommended for practical use.

Background

Following the process of animal domestication, evolutionary forces such as selection and genetic drift have played a critical role in animal diversification. Such forces led to genomic alterations such as fixation of favorable alleles within a breed or species and differentiation from the ancestral state due to successful selection programs or adaptation. This concept of domestication and its subsequent impact on diversity of animal species, breeds or strains was well explored by Darwin [1, 2]. So, phylogenetic studies aim to assess these variations.

The wild, unselected native and village chicken populations retain a reservoir of and exhibit more genetic variability [3–5]. Commercial breeds are known for being intensely selected for economic purposes, i.e. meat and egg type production. Successful egg type selection programs within the commercial layers have resulted in a reduced genetic variability within these lines. In Europe, an organized and systematic breeding in chickens was developed during the 19th century. Selection programs in this case were based on producing attractive features (for entertainment) in line with the breed standards; because of this, many fancy breeds were heavily selected for their attractiveness. To date such heavily selected breeds exhibit reduced genetic diversity and high average genetic distances to other breeds [3–5]. Major components for the reduced variability within both the commercial and the fancy breeds are due to the fact that the selection was certainly based on small number of founders, small effective population size and/or high degree of inbreeding.

Using whole genome resequencing (WGS) data is considered as the best way of doing association or diversity studies [6, 7]. It provides a high resolution of the genome information capturing most (and even the finer) details underlying genomic variations. However, the cost of whole genome sequencing still is high for application in larger sample sets. Additionally, limitations such as infrastructure (e.g. WGS requires good reference genomes), work effort and time poses further constraints. So, generating WGS data for the required sample size in such studies is challenging [6].

Genotyping tools have been developed to overcome these constraints and have made genotype data available in sufficient numbers. Single nucleotide polymorphism (SNP) panels have been widely used in studies of genomic variation within species [8, 9]. For the construction of such SNP sets, a limited number of individuals selected from populations of interest (the so-called ascertainment group) are used as discovery panels. These individuals are sequenced and provide the basis to select polymorphic loci targeted for further genotyping in a larger set of individuals [9, 10]. SNPs are often selected based on quality, with predefined spacing (e.g. equally spaced) and desired frequency distribution [10], among other criteria.

These methods of SNP discovery may introduce ascertainment bias, hindering classical population genetic methods to provide correct results when applied with SNP genotype data [11, 12]. Ascertainment bias is a systematic deviation of population genetic statistics from a theoretical ‘true’ value, which arises from a non-random selection of set of individuals or biased marker discovery protocols [6, 13].

If the level of ascertainment bias is high, results of population genetic studies could be widely misinterpreted [14]. Thus, exploring the potential systematic effects that the ascertained genotype

data can have on the results of diversity studies and finding a way to minimize these effects is crucial.

Differences in the allele frequency distribution between SNP genotype data and WGS data have been commonly used to assess ascertainment bias [6, 11, 15]. An easily verifiable indicator of a potential ascertainment bias is a complete absence of SNPs or an underrepresentation of rare SNPs. Discovery of SNPs is driven by the allele frequency, and with an often small size of the discovery panel, discovering rare SNPs is mostly limited [14]. With the missing rare SNPs, the SNP data may not be an adequate representation of the WGS data. Gorlov et al. [16] argue that missing rare SNPs can lead to loss of valuable information and lessen the ability to detect those rare SNPs in association studies, which may be critical e.g. in the context of rare causal SNPs for rare diseases.

Effects of ascertainment bias on genetic diversity analysis within and between populations have been reported in several studies [9, 13, 17]. One of the assertions is that selection of subpopulations for discovery panels tends to over-represent variability of that ascertainment group. Consequently, effects of ascertainment bias on heterozygosity estimates [18, 19], fixation index (F_{ST}) values and phylogenetic relationships [9] have been reported. Herrero-Medrano et al. [18] and Albrechtsen et al. [15] observed that ascertainment bias affected some populations more than others when studying their genetic diversity with SNP chip data. McTavish & Hillis [9] concluded that both the F_{ST} and principal components analysis (PCA) estimated from SNP chip data were distorted when ascertainment bias was not accounted for. Principal components analysis is a statistical technique that captures patterns of high dimensional data and projects them into a lower dimensional space, allowing to determine key variables that explain the observations [20, 21]. PCA has been used in many studies to capture genetic structures of populations [22–26]. In contrary to McTavish & Hillis [9], McVean [27] reported that the PCA is less affected by ascertainment bias. He claims

that effects of ascertainment bias on PCA are easy to predict and only have little impact on the structuring of populations unless the bias is very severe.

Despite the available proposed schemes and several suggestions made on how to address the issue of ascertainment bias in population genetic analysis [6, 12, 15], there are still challenges on the definite measures to deal with this issue [17]. Clark et al. [14] concluded that it is not always easy to correct for ascertainment bias, success is not guaranteed, and mostly the suggested corrections are not applicable to every study [15]. Most of the suggestions were also tested using simulated data, which may miss out some of the complexities encountered when using real data.

In this study, we tried to assess the impact of ascertainment bias and the efficiency of various strategies to account for it in a chicken diversity panel, which is based on a diverse set of chicken populations for which both pooled WGS data and individual SNP genotype data obtained with a high density SNP array were available. For most of the studied populations, there is no sufficient documentation on the breed history and/or background and we are skeptic that the material used allows to identify the mechanisms causing ascertainment bias. Therefore, we based our primary focus on identifying strategies to mitigate ascertainment bias rather than to do a full analytical (or empirical) study to understand the causes of ascertainment bias. With the SNP genotyping array [10] that was used, the SNP panel was established by selecting a few populations (for details please see the “Methods” section) which are not representative for all the other populations used in our study. In addition, the SNP selection criterion included discarding low minor allele frequency (MAF) SNPs which potentially causes an underrepresentation of SNPs under selection [28]. Criteria used in our study to assess the impact of ascertainment bias and the various strategies to mitigate its effects were similarity of allele frequency spectra, expected heterozygosity, F_{ST}, PCA, distance measures and topologies of phylogenetic trees. In general, the results obtained from the

WGS data were considered as the 'reference standard' and strategies to correct for ascertainment bias were considered based on how good the WGS-based results were met.

Methods

Animals

A total of 42 chicken populations were used in this study. For each of the populations, both whole genome resequencing data based on pooled samples and individual genotype data obtained with a 600K SNP Affymetrix® Axiom® High Density Chicken Genotyping Array were available. A list of the 42 populations with their abbreviations and population sizes as used in the study is provided in Table 2.1. Samples used in this study were collected under the umbrella of the SYNBREED project (www.synbreed.tum.de) from chicken fancy breeds in Germany between 2010 and 2012. The collection was completed by samples of two Red Jungle fowl populations, *Gallus gallus gallus* (GGg) and *Gallus gallus spadiceus* (GGsc) taken from previous EU project AVIANDIV (see [29]).

For the WGS pooled data, equal amounts of DNA of the individuals of each population were pooled using *PicoGreen*® quantitation assay except for the WL_A. In the case of WL_A, 10 birds were sequenced individually and virtual pooling was performed. Thirty-nine of the 42 populations in the WGS consisted of 385 individuals of which 383 were also genotyped individually. The other 3 populations (WL_A, BL_A and BL_D) were commercial lines (see Table 2.1) and consisted of different individuals in the two data sets. In the array data set, in addition to the 383 individuals, 461 more individuals were added and their distribution is also shown in Table 2.1. So, when comparisons were made between array and WGS data with commercial breeds included, the 383 plus 461 individuals' version of array data was used. For the commercial breeds, each breed contained 20 individuals in the array data. In the WGS data, each breed contained 9-10 individuals

for the non-commercials and 10-15 individuals for the commercial breeds. The commercial breeds were among the breeds used in the discovery panel for the development of the 600K Affymetrix genotyping array.

Collection of blood samples for this study was performed in accordance with the German Animal Protection Law and was submitted to and 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).

Table 2.1: List of breeds, their abbreviations and sample sizes as used in the study

Breed and abbreviation	**Array data (n)**	**WGS data (n)**
Commercial breeds :		
WL_A – White Leghorn line A	20*	10*
BL_A – Rhode Island Red line A	20*	15*
BL_D – White Rock line D	20*	15*
Wild populations :		
GGg – Gallus Gallus Gallus	10(10)	10
GGsc – Gallus Gallus Spadiceus	9(10)	9
European populations:		
ABwa – Barbue d'Anvers quail	10(10)	10
ARsch – Rumpless Araucana black	9(11)	9
BAsch – Rosecomb Bantam black	10(10)	10
BKschg – Bergische Crower	10(22)	10
DZgh – German Bantam gold partridge	10(10)	10
FZgpo – Booted Bantam millefleur	10(10)	10
HOxx – Dutch White Crested	10(7)	10
ITrh – Leghorn brown	10(10)	10
KAsch – Castilians black	9(11)	9
KRsch – Creeper black	10(20)	10
KRw – Creeper white	10(20)	10
LER11- White Leghorn line R11	9(13)	9(1)

OMsschg - East Friesian Gulls silver penciled	10(10)	10
PAxx - Poland any colour	11(12)	11
SBsschs - Sebright Bantam silver	10(10)	10
WTs - Westphalian Chicken silver	10(10)	10
Asian populations:		
ASrb – Aseel red mottled	10(10)	10
BHrg – Brahma gold	10(10)	10
CHgesch – Japanese Bantam black tailed buff	10(12)	10
CHschw – Japanese Bantam black mottled	10(19)	10
COsch – Cochin black	10(11)	10
DLIa – German Faverolles salmon	10(10)	10
KSgw – Ko Shamo black-red	9(13)	9
MAxx – Malay black red	10(21)	10
MRschk – Marans copper black	10(10)	10
NHL68 – New Hampshire line 68	9(14)	9(1)
OFrbx – Orloff red spangled	10(15)	10
OHsh - Ohiki silver duckwing	10(10)	10
ORge - Orpington buff	10(10)	10
SAsch - Sumatra black	9(11)	9
SEw - Silkies white	10(10)	10
SHsch - Shamo black	9(11)	9
SNwsch - Sundheimer light	10(10)	10
TOgh - Toutenkou black breasted red	10(11)	10
WYw - Wyandotte white	10(9)	10
YOwr - Yokohama red saddled white	10(10)	10
ZCw - Pekin Bantam white	10(10)	10

n is number, in brackets () are additional individuals added to the population (not present in the other data type), *completely different individuals in the two data sets

WGS data and preparation

Pools of the 42 populations comprising in total 425 individuals were resequenced with 20X target coverage. The sequence reads were aligned to the chicken reference genome (galGal4) [30] using Burrows-wheeler alignment algorithm implemented in BWA [31] and sorted using Samtools [32]. Picard tools were used to mark duplicates and GATK was used for calling the SNPs [33, 34]. For more details on the preparation pipeline see Reimer et al. [35].

Genotype (array) data and filtering

The initial array data set contained 918 animals and 580, 588 SNPs. SNPs misplaced at wrong chromosomes were removed. The data was then filtered for SNP call rates of >99% and animal call rate of >95% using the SNP & Variation Suite Version (SVS) 8.1[36] which retained 904 animals and 450, 082 SNPs. From this point, the following SNP filtering pipeline was applied, with number of SNPs left at each step shown in brackets:

1. SNPs with missing positions were discarded (445,428).
2. SNPs that shared the same position on the same chromosome were discarded (e.g. if there were two SNPs sharing the same position, both of them were discarded (445,388).
3. SNPs had to be present in both array and WGS data (21,759 of array SNPs were not found in the WGS data) and only SNPs from chromosome 1-28 were considered (401,420).
4. SNPs were discarded if the reference (and/or alternative) allele of genotype (array) data didn't match the reference (and/or alternative) allele from the sequence data (401,125).

After the above filtering, a total of 401,125 SNPs remained for further analysis. This set of data was used in assessing allele frequency calling in the pooled sequence data, comparing allele frequencies between the array and WGS data, and assessing how this uncorrected ascertained data

affect genetic diversity analysis by being compared to results analyzed from WGS. The array data SNP was converted so that allele A resembled the respective reference allele.

Different filtering schemes were applied to the array dataset (Array_all in Table 2.2) to be tested for their potential to account for ascertainment bias. More specifically, we applied three different basic filtering principles:

1. LD based SNP pruning, which has been described to partially account for the effects of ascertainment bias. In our study, SNP pruning for LD was done in PLINK v1.9 [37, 38]. The parameters: *indep 50 5 2* were used, whereby *50* is the window size in SNPs, *5* is the window size step (in SNPs) after LD calculation (after LD has been calculated from the 50 SNPs window, and SNPs which exceed the VIF threshold are removed, the window is shifted 5 SNPs forward and the procedure is repeated), and *2* is the variance inflation factor VIF = $1/(1-r^2)$ [39].
2. A second filter applied was to restrict the analysis to SNPs that were found to be polymorphic in the wild chicken populations, which were represented in our study with two populations (GGsc and GGg subspecies).
3. A third filter excluded SNPs with less than 5% MAF. This MAF filtering was done in PLINK v1.9 [37, 38] using the command –maf 0.05.

These filters were applied alone and in combination, the corresponding filters and resulting data sets are presented in Table 2.2.

Table 2.2: Array data set versions with different filtering strategies applied

Given name for data set	Filter/s applied	No of SNPs
Array_all		401, 125
Array_MAF5	Filtered out SNPs with less than 5% MAF	379, 342
GG	Retained only SNPs that are polymorphic in the two *Gallus gallus* wild populations (GGg and GGsc)	289, 390
GG_MAF5	**GG** and filtered out SNPs with MAF less than 5%	284, 748
Pruned	SNPs were pruned based on LD	122, 006
Pruned_MAF5	**Pruned** and filtered out SNPs with MAF less than 5%	107, 604
Pruned_GG	**Pruned** and **GG**	86, 404
Pruned_GG_MAF5	**Pruned_GG** and filtered out SNPs with MAF less than 5%	82, 975

Allele frequency calling in the pooled sequence data

To investigate the reliability of allele frequency calling in our WGS pooled data, we estimated and compared allele frequencies between array (using all 401,125 SNPs) and WGS pooled data for corresponding loci. To avoid issues relating to sample size [40], only 39 of the 42 populations (with 383 individuals for array data and 385 individuals for WGS data) were used for this comparison, the 3 commercial populations which contained different individuals in the two data sets were excluded. Then we also compared the allele frequencies for each breed between the two sets, this time including also the 3 commercial breeds. We used Pearson's correlations between estimated allele frequencies of WGS and array data to assess the accuracy of allele frequency calling in the pool WGS data. All allele frequency calculations were based on the alternative allele

at each locus. Allele frequencies for the pooled sequences were calculated as the proportion of reads' counts for the alternative allele at each locus.

Assessing ascertainment bias in the array data

We randomly sampled 401,125 SNPs in 100 repetitions from the WGS data, computed the average allele frequency spectrum (AFS) and compared it with the AFS of the 401,125 SNPs in the array data.

Genic SNPs of *Gallus gallus* were annotated with Ensembl genes 85 [41] and the proportions of SNPs in genic and non-genic regions were calculated and compared between the two sets. The genic region was defined according to the Ensembl gene definition, comprised of any spliced transcripts with overlapping coding sequence [42]. It was further determined if there are differences in MAF distributions from the genic and non-genic regions in the two data types.

Assessing the potential effects of ascertainment bias in genetic variation analysis

Within breeds diversity analyses, population differentiation and phylogenetic structure analyses were performed and compared between the WGS data and different versions of the array data. For within breed variation, the expected heterozygosity (H_e) was estimated as: $H_e = 2p(1-p)$, where p represented the allele frequency of the alternative allele [43]. We could not use the observed heterozygosity for comparison since this one was not available for the pooled sequence data.

As a measure of population differentiation, the pairwise fixation index (F_{ST}) between breeds for each locus was estimated as: $F_{ST} = \frac{s^2}{\bar{p}\,(1-\bar{p})}$ [44]. For the same sample sizes s^2 was calculated as $\sum_i (\widetilde{p_i} - \bar{p})^2/r$ where $\widetilde{p_i}$ is the allele frequency of the i^{th} population, $\bar{p}$ is the average allele frequency across populations and r is the number of populations the F_{ST} is calculated for. For

different sample sizes the s^2 was calculated as $\sum_i n_i(\tilde{p}_i - \bar{p})^2/r\bar{n}$ and $\bar{p}$ calculated as $\sum_i n_i\tilde{p}_i/r\bar{n}$ where n_i is the sample size of the i^{th} population and $\bar{n}$ is the mean sample size. The F_{ST} values were averaged across loci.

Phylogenetic variation between populations in the different data sets was evaluated by means of phylogenetic trees and principal components analysis (PCA). Pairwise genetic distances were estimated using Nei's standard genetic distance [45]. The pairwise genetic distance matrices of the different array data versions were compared with that of WGS using Frobenius (*F*) distances, which was calculated as $F_{A,B} = \sqrt{trace((A-B)*(A-B)')}$ [46], where A and B are the two distance matrices to be compared. Since it couldn't be ruled out that there is a scale effect of the number of SNPs used in the construction of the distance matrix, we sampled 100 replicates from the WGS data with the same number of SNPs as was used in the construction of the array-based matrix in the respective comparison. We then calculated the genetic distances and compared the respective array-based matrix to the 100 replicates of the WGS-based matrices.

The phylogenetic trees were derived from the pairwise distance matrices between the breeds. The 'Ape' package in R v3.2.2 was used to compute and construct neighbor joining (NJ) trees [47, 48]. The NJ trees were then compared using their topological distances obtained from two methods:

1. Penny & Hendy [49] consider the topological distance as twice the number of internal branches defining different bipartitions of the tips. Comparisons here are made by counting the number of different partitions resulting from cutting the interior branches of the two trees. Differences in partitions are determined by having one or more different objects (in our case different populations) when the trees are cut at a branch. The topological difference is then calculated by how many partitions need to be changed in order for the two trees to be similar. This method determines how similar objects are grouped together in the two trees based on the partitions.

A value of 0 means that cutting the trees at any similar branch point results in similar objects on the partitions of the two trees; therefore, the two trees are considered to have a similar topology. The lower the value, the more similar the two trees are.

2. Billera et al. [50] consider the topological distance as the sum of the branch lengths that need to be erased to have two similar trees calculated as $d = \sqrt{\sum(X_i - Y_i)^2}$, where X and Y are two NJ trees, and i is the i^{th} population in X and Y. X_i and Y_i are the branch lengths of the i^{th} population in trees X and Y respectively. The branch length is described as the amount of evolutionary change [51], and the distance between two populations in one tree is the sum of the branch lengths connecting them. Therefore, if population *i=1* in tree X and Y has the same branch lengths but population *i=2* in tree X and Y has different length, the distance between population 1 and 2 in the two trees will be different. This method estimates the difference between the two trees for the i^{th} population and sums all the differences for every population. A value of 0 means that all pairs of populations have the same branch lengths connecting them in the two trees.

Again these comparisons were made between the different versions of the array data set and the randomly sampled 100 replicates of WGS data and with the same number of SNPs, respectively.

The "ade4" and "stats" packages in R were used to compute the PCA and the packages "factoextra" and "scatterplot3d" for visualizing the results in two dimensional (2D) and three dimensional (3D) respectively [48, 52, 53].

Results

Assessing allele frequency calling in pool whole resequencing data

We compared the estimated allele frequency for all SNPs in the 'Array_all' data set with the estimates from the pool WGS data at each corresponding locus. The allele frequency spectra of

the two data sets were found to be mostly identical (Figure 2.1). The proportion of SNPs in the frequency bin 0.025-0.125 was slightly higher in the WGS than the array data while the proportions of SNPs in the bins 0.150-0.3 were slightly higher in the array than the WGS data. A high correlation was obtained between the allele frequencies of the two sets (r = 0.983), as well as within the different breeds (ranging from r = 0.94 to 0.99).

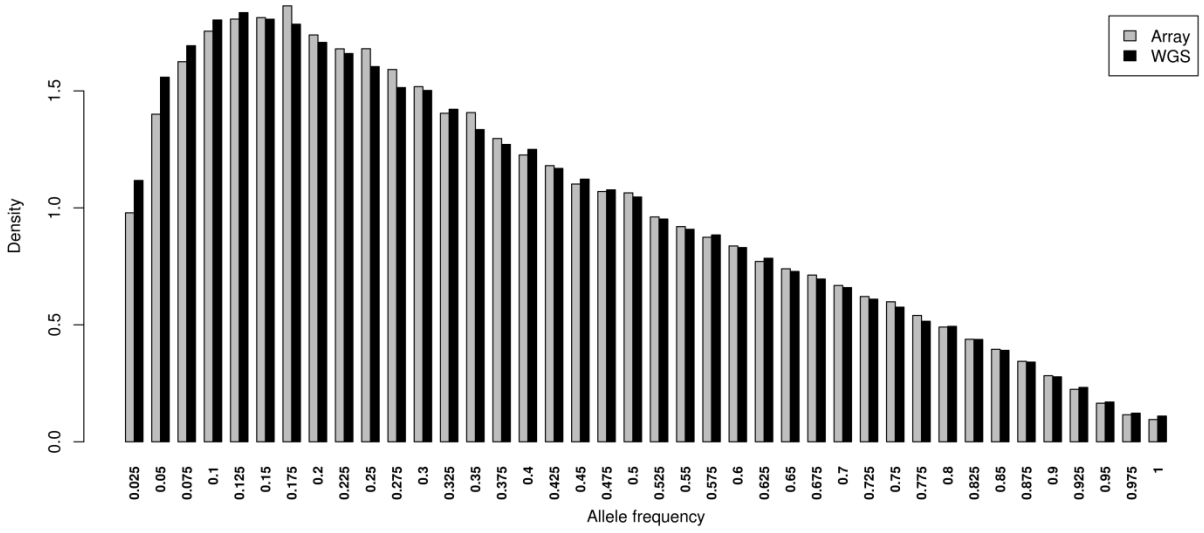


Figure 2.1: Allele frequency spectrum of array data and corresponding WGS loci for 39 populations.

Assessing the potential of ascertainment bias in the array genotype data

The allele frequency spectra showed remarkable differences for the two data types (Figure 2.2). The array data had very low but increasing numbers of SNPs at allele frequencies between 0 and 0.175 while the WGS had a very high number of rare variants between 0 and 0.025 and SNP numbers decreased with increasing frequencies, with the exception of the last window (which includes the fixation of the derived allele) which was found to be slightly over-represented.

For the individual populations (refer to Additional file 2.1), the most affected in terms of missing rare SNPs were the Marans copper black (MRschk), Araucana black (ARsch) and the wild GGsc; and the least affected were the European fancy bantam (SBsschs, BAsch, FZgpo and ABwa) breeds, the White Leghorn line R11 (LER11), the Asian long tailed (TOgh and OHsh) breeds and the commercial white layers (WL_A). In the latter, these results have shown to be related to the genetic diversity within these breeds (see H_e estimates below and the discussion thereof).

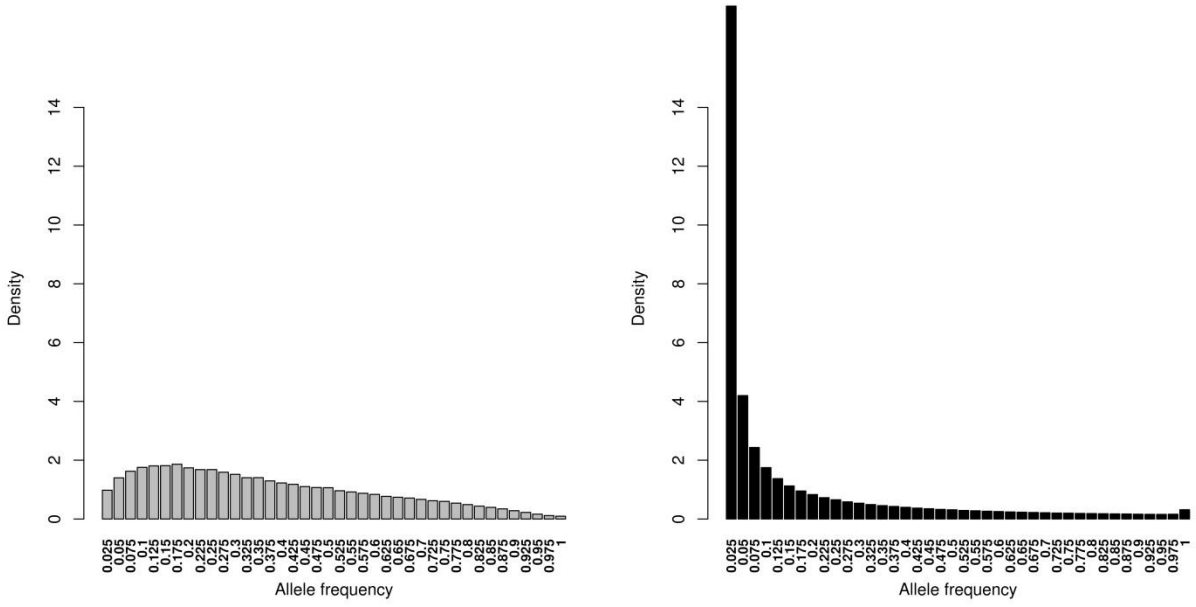


Figure 2.2: Allele frequency spectrum of array data (left) and WGS data (right) for 39 populations.

The proportion of SNPs was 39.6% and 39.9% in genes and was 0.044% and 0.012% in exons for array and WGS data, respectively (see Table S2.1 in Additional file 2.2). Differences in (minor) allele frequencies (in genic and non-genic regions) followed a similar pattern to that observed in Figure 2.2 whereby rare variants were underrepresented in the array data. The correlations between MAF proportions in genic and non-genic regions were 0.956 and 0.999 in the array and WGS data, respectively. The minor allele frequency of SNPs differed very little between the genic and non-

genic regions with the array and sequence data (Figure S2.1). From this we concluded that the selection of SNPs in the array was not biased based on their positions in genic or non-genic region, although, differences between the two sets were found to be in the exonic regions whereby the array set had an overrepresentation of SNPs.

Within breed variation was assessed by comparing the expected heterozygosity estimates between the two sets, and the results for the WGS vs. Array_all, GG and Pruned versions of array data are shown in Figure 2.3. The versions with MAF filtering barely showed any difference and are therefore not shown. In Figure 2.3, we ranked the breeds in ascending order of the estimated H_e in WGS and fitted (for each same breed) the array estimated H_e to observe if it also appears in the same ranking order as the WGS data. The red jungle fowls, which are believed to be the ancestors of domestic chickens are expected to carry more genetic information than found in most of the other populations. When using the WGS data, the highest genetic diversity was observed in the two red jungle fowls (wild: GGsc and GGg) which was not the case with the Array_all data. There was also considerable random fluctuations in the ranking of the breeds in the Array_all data. Tying up these H_e back to the allele frequency spectra of each population, the highly affected breeds in terms of AFS were also more affected in terms of the H_e ranking (estimated with Array_all) and vice versa for the less affected once. The H_e ranking of MRschk and ARsch in the array data was very high compared to the other breeds. Given the allele frequency and H_e estimates, we observed that the breeds which were least affected by ascertainment bias are mainly those with less genetic variability. After filtering the data for SNPs being polymophic in the wild populations (GG) or pruning the SNPs based on LD (Pruned), the maximum diversity in the wild populations was captured and less fluctuations appeared in the ranking order.

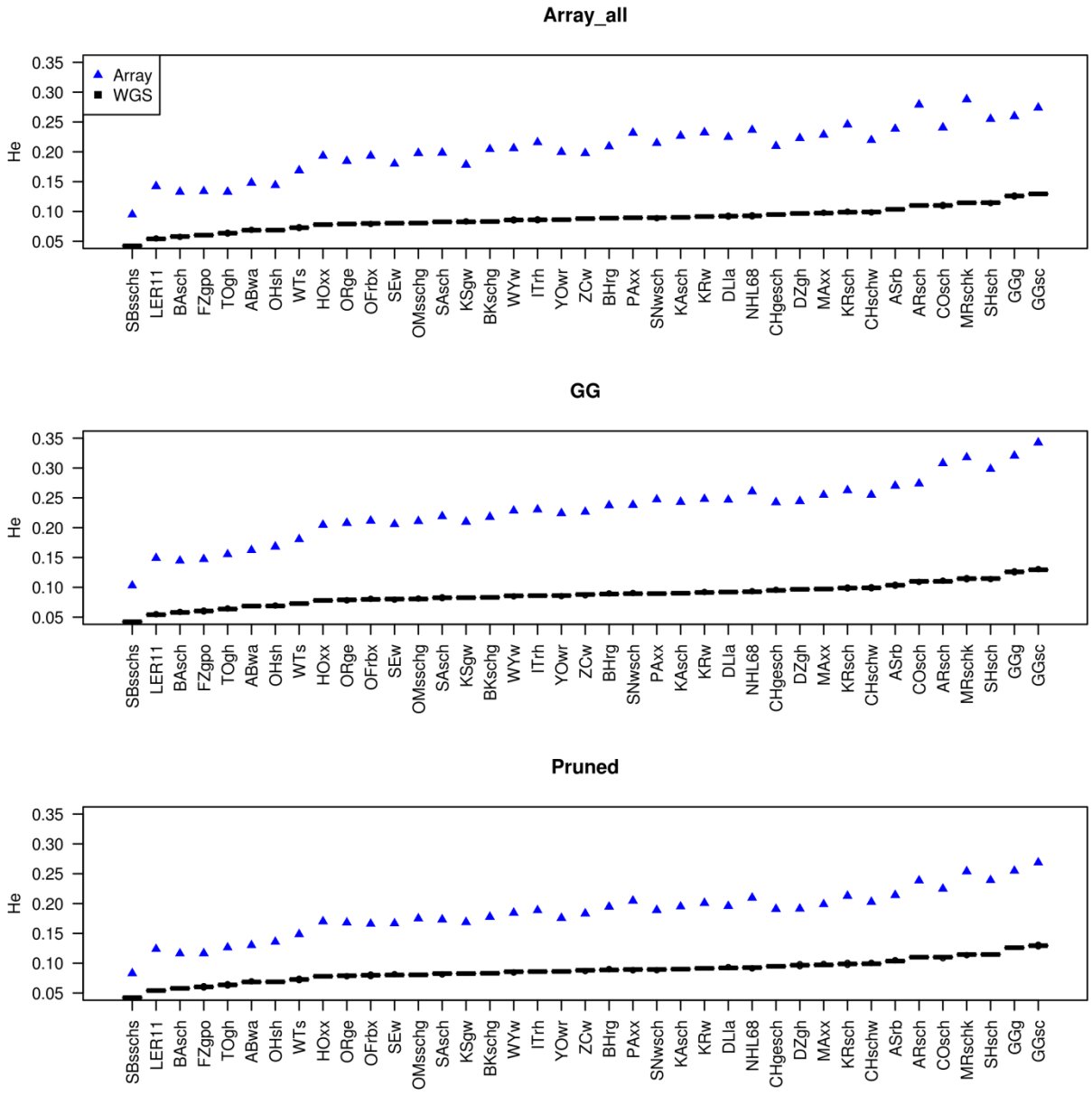


Figure 2.3: Comparisons of expected heterozygosity (H_e) estimates between WGS (boxplot of 100 replicates) and array (Array_all, GG and Pruned) data.

In agreement with e.g.[3], (based on microsatellite data) both the commercial brown (BL_A and BL_D) and white (WL_A) layers displayed reduced genetic diversity within the breed (Additional file 2.3, Figure S2.2, estimated using the data with 42 populations). The commercial white egg layers, which emerged from a single parental origin, the White Leghorn breed [5, 54], had very

low genetic diversity. The brown layers (BL_A and BL_D) with multi-parental origins of Asian and European background had more genetic diversity compared to white layers. Noting that these commercial breeds were part of the discovery panel, we investigated whether the H_e results behaved differently than in other populations when using array data. Unlike the two brown layer lines with elevated H_e ranking when using any of the array data, the white layers didn't deviate from the WGS H_e ranking when using the array data (Figure S2.2). So this makes it difficult to tie the effects of ascertainment bias on H_e estimation to the relatedness of the breeds to the discovery panel breeds. Furthermore, the fact that the commercial lines' individuals used in the array data are different to those used in the WGS could also be of impact in this context.

When fitting a linear regression of the WGS-based H_e values on array-based H_e values the slope is >2 with all considered data sets (smallest with 2.150 for the LD pruned data, see Table 2.3 and Figure S2.3 in Additional file 2.3) reflecting not only a systematic overestimation of expected heterozygosity from array data, but also a scale effect resulting in an even more severe overestimation for highly heterozygous breeds. While the underrepresentation of low MAF SNPs in the array data compared to WGS data (cf. Figure 2.2) provides a good explanation for the observed difference in the average H_e, the reason for the scale effect remains to be understood.

Table 2.3: Relationship between the H_e estimates between WGS and the array data sets

	r_s	Slope
Array_all	0.956	2.233
Array_MAF5	0.957	2.321
GG	**0.985***	2.770
GG_MAF5	0.984	2.790
Pruned	0.973	**2.150***
Pruned_MAF5	0.974	2.340
Pruned_GG	0.983	2.675
Pruned_GG_MAF5	0.983	2.717

r_s – Spearman's rank correlation. Slope – the slope of regression line when the H_e estimates of array data are regressed against those of WGS data. *Numbers in bold face represent the best value in the column. These results are based on 39 populations.

A comparison between the estimated pairwise F_{ST} values of WGS and the different filtered versions of the array data is shown in Figure 2.4. The black regression line shows the expected linear relationship between the F_{ST} of WGS and array where the pairwise F_{ST} values estimated from the two sets are equal. The Array_all, Array_MAF5 and the versions filtered for being polymorphic in the *Gallus gallus* populations (GG and GG_MAF5) underestimated the F_{ST} where WGS F_{ST} was low (0.09 to <0.15) and overestimated the F_{ST} where WGS F_{ST} was high (>0.15). The LD pruned versions (Pruned and Pruned_MAF5) and the LD pruned plus polymorphic to *Gallus gallus* populations' (Pruned_GG and Pruned_GG_MAF5) data sets consistently underestimated the pairwise F_{ST} values. The regression lines for comparing WGS F_{ST} and F_{ST} estimated from the LD pruned versions didn't cross through the expected regression line, while for versions without LD pruning the regression lines cross each other.

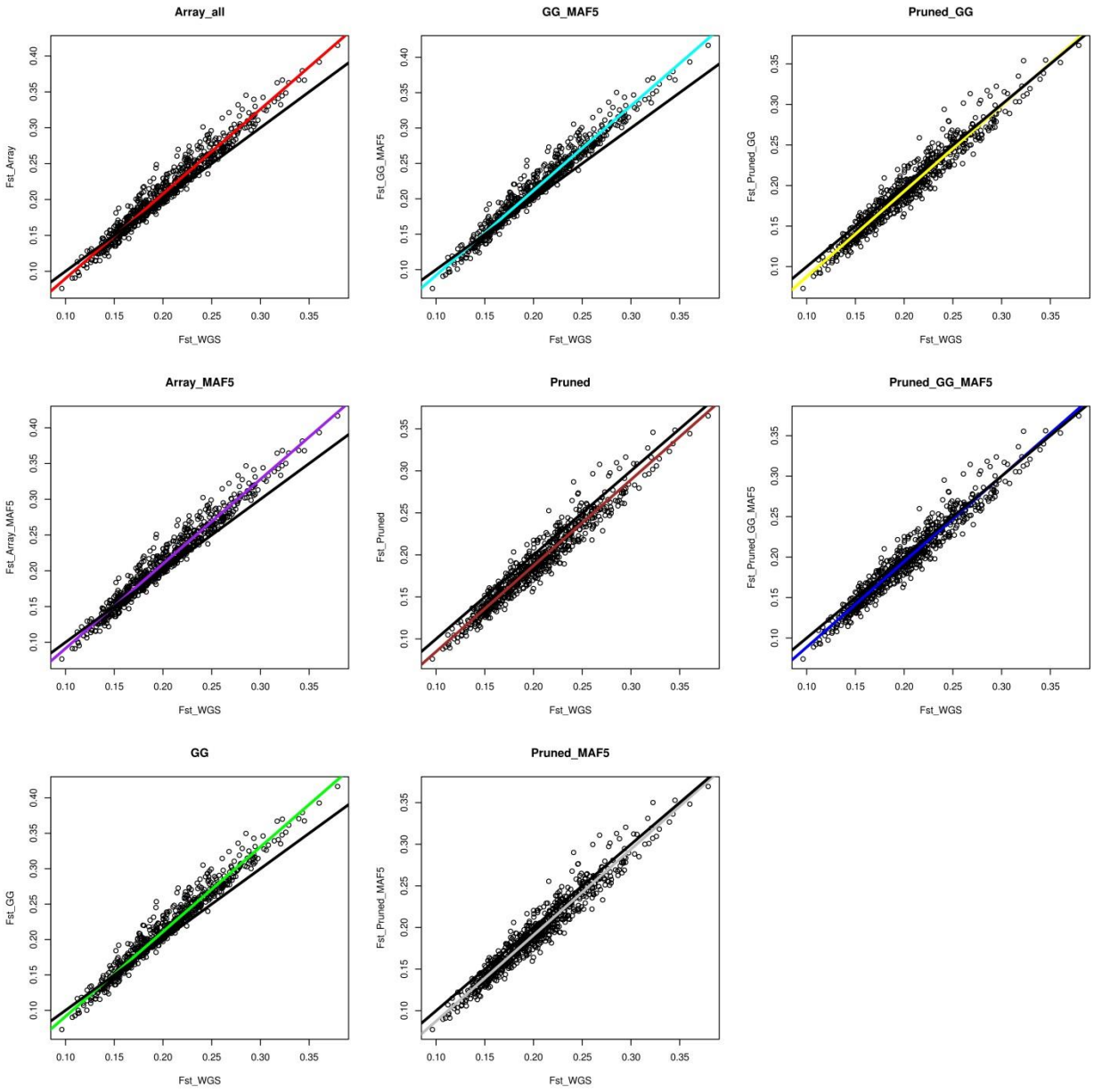


Figure 2.4: Regressions through the pairwise F_{ST} values between WGS and array data. Black lines represent the expected identity relationship between the two data sets (with a slope of 1).

The slopes and regression coefficients (R^2) of these linear relationships are presented in Table 2.4. The WGS vs. Pruned data had the lowest R^2 (0.937), however, with a slope (1.023) closer to 1 compared to the rest of the other array sets. The WGS vs. GG and GG_MAF5 had the highest R^2 (0.959 for both of them) and yet the highest slope too (1.197), whereas in this case a better slope

(close to 1) is preferred (it justifies the significance of the linear relationship between the pairwise F_{ST} values estimated from WGS and array data). A combination of filtering SNPs based on LD and retaining SNPs that are polymorphic in the wild populations (GG) improved the R^2 but compromised the slope.

Table 2.4: The relationship between the F_{ST} estimates of the WGS and array data

	WGS				
	Slope	**R^2**	**Regression constant**	**Standard error (SE)**	**Residual variance**
Array_all	1.179	0.954	-0.028	0.009	0.0001
Array_MAF5	1.183	0.954	-0.027	0.010	0.0001
GG	1.197	**0.959***	-0.028	0.009	0.0001
GG_MAF5	1.197	**0.959***	-0.028	0.009	0.0001
Pruned	**1.023***	0.937	-0.017	0.010	0.0001
Pruned_MAF5	1.033	0.939	-0.016	0.010	0.0001
Pruned_GG	1.055	0.940	-0.018	0.010	0.0001
Pruned _GG_MAF5	1.057	0.941	-0.017	0.010	0.0001

*Numbers in bold face represent the best value in the column. R^2 – regression coefficient. These results are based on 39 populations.

Table 2.5 shows the Frobenius (*F*) distances between the distance matrices of WGS and array (on the diagonal), and the different array sets among themselves. The mean *F* distance between WGS and Pruned data was the lowest (3.152) and highest between WGS and GG_MAF5 data (6.700). A lower *F* distance means two compared distant matrices are more similar. Therefore the pairwise distance matrix of Pruned data is more related to the WGS than the rest of the sets. Among the array versions, the most distant matrices were found between the Pruned version and the GG and

GG_MAF5 versions (these GG and GG_MAF5 versions had the highest distances to the matrix of WGS data).

Table 2.5: Frobenius (*F*) distances between distance matrices of WGS and array data

	Array_all	**Array_MAF5**	**GG**	**GG_MAF5**	**Pruned**	**Pruned_MAF5**	**Pruned_GG**	**Pruned_GG_MAF5**
Array_all	5.312±0.001							
Array_MAF5	0.591	5.889±0.001						
GG	1.239	0.685	6.501±0.001					
GG_MAF5	1.434	0.868	0.200	6.700±0.001				
Pruned	2.230	2.810	3.397	3.596	**3.152***±0.001			
Pruned_MAF5	1.332	1.886	2.447	2.644	0.971	4.115±0.001		
Pruned_GG	1.034	1.530	2.038	2.232	1.417	0.462	4.548±0.002	
Pruned_GG_MAF5	0.811	1.216	1.676	1.867	1.800	0.836	0.329	4.931±0.002

The diagonal is a mean of the *F* distance between the array data set and 100 WGS replicates with the standard errors (SE). *Number in bold face represents the best value on the diagonal.

The neighbor joining trees of the WGS, Array_all, Pruned and GG data sets are shown in Figure 2.5. Four clusters were identified and circled with different colors and Table S2.2 in Additional file 2.2 shows the breeds and their cluster affiliations. Three breeds were outside the clusters and are noted in Table S2.2 with an n (not assigned). All the array data sets were able to capture the same clusters as the WGS data in exception of the MAxx population which was not assigned to

any cluster when using the GG set while assigned to cluster 2 when using the other sets. Cluster 1 and 2 represent breeds from Asian origin, with cluster 1 grouping the normal sized breeds together and cluster 2 showing a cluster of dwarf birds. Similarly cluster 3 and 4 represents breeds from European origin with normal sized and dwarf birds' clusters, respectively. From visual inspection, the trees shown displayed many similarities, especially the way breeds were clustered together.

Figure 2.5: Neighbour joining trees of WGS, Array_all, GG and Pruned data sets.

To quantify the similarities statistically, we used two different methods [49, 50] to access the topological distances (Figure 2.6 and 2.7) between trees of the WGS and array data sets. Based on the Billera method, the topological distance between the WGS and the Pruned data was the lowest (with distance of 0.027) while it was highest with the GG_MAF5 data (with a distance of 0.052) (detailed in Table S2.3 in Additional file 2.2). For the WGS and GG data, the distance was 0.050 and for WGS and Array_all data it was 0.044. All the mean topological distances between WGS and the various array sets didn't fall within the same ranges as the distances between the 100 replicates of WGS (see Figure 2.6 and Table S2.3). Nonetheless the results show that there is a better relationship between the trees of WGS and the Pruned data than of WGS with any of the other array versions.

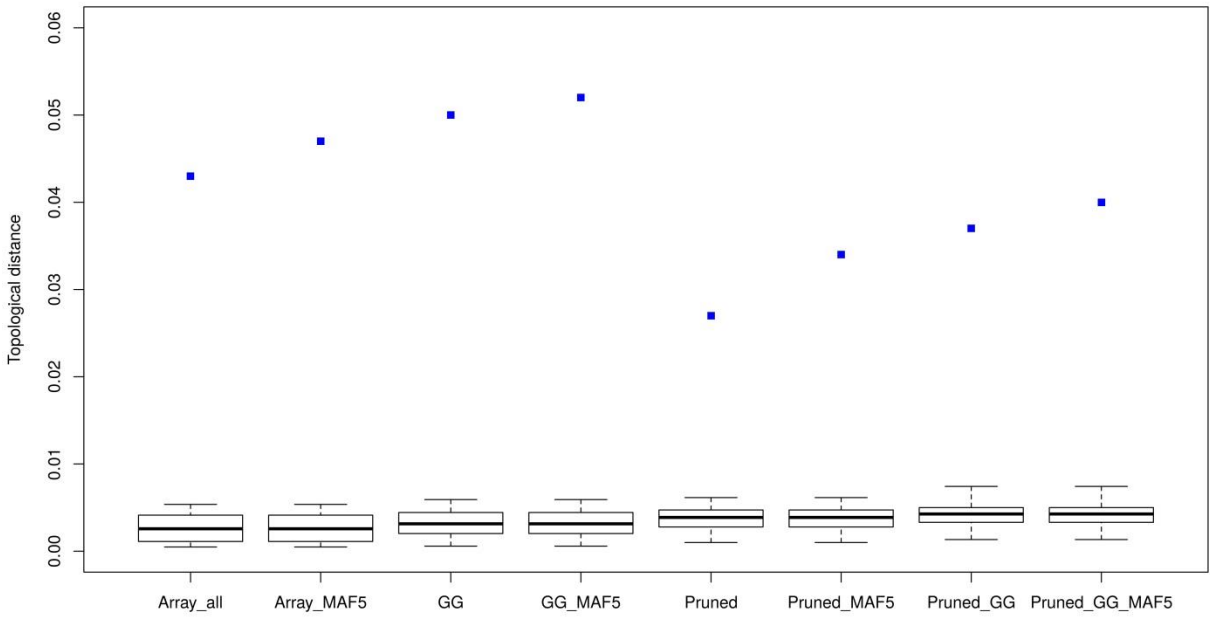


Figure 2.6: Topological distances between the NJ trees of array and 100 replicates of WGS data based on the Billera method. The boxplots reflect distances between the 100 replicates of WGS and the blue dots are mean distance between the array set and the 100 WGS replicates.

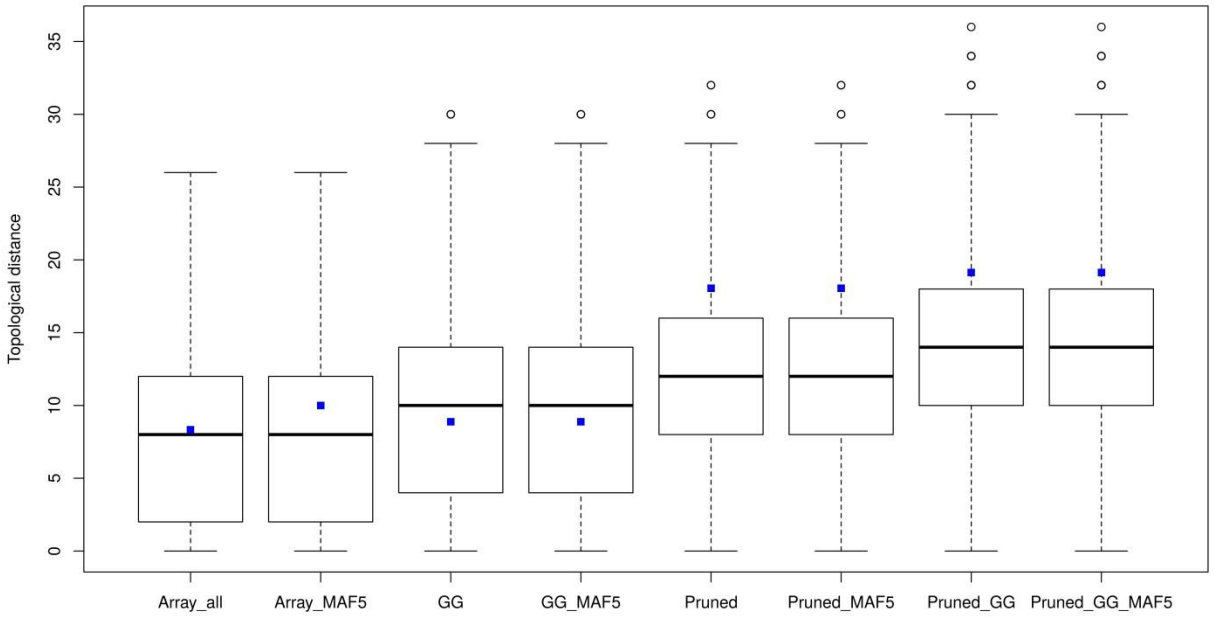


Figure 2.7: Topological distances between NJ trees of array and 100 replicates of WGS data based on the Penny and Hendy method. The boxplots reflect distances between the 100 replicates of WGS and the blue dots are mean distance between the array set and the 100 WGS replicates.

Using the Penny and Hendy method, the mean distances between WGS and all the array sets fell within the distance ranges between 100 WGS replicates (see Figure 2.7 and Table S2.4 in Additional file 2.2). However, the standard errors for the mean distances for all sets' comparisons were also high. The distances between the WGS and GG, Array_all and their MAF filtered versions were much closer to the median of the 100 replicates. These comparisons of the array and WGS trees based on trees' partitions using the Penny and Hendy method yielded closer

relationships between the two data types. These distances confirmed the visual observation whereby the trees show a relative similar clustering of breeds (Figure 2.5). Comparisons across the different array versions showed that Array_all is more related to the GG and both of them are distant to the Pruned data (Table S2.5 in Additional file 2.2).

We computed the PCA to see how population structures are captured by the array data compared to the WGS, and visualize the results in 2D and 3D plots. The 2 dimensional PCA plots showed only a very little and hardly noticeable difference between the array sets and the WGS data. Overall all the array versions were able to capture almost similar structures as that of WGS in the two-dimensional PCA. Figure 2.8 shows the PCA plots of WGS, Array_all, GG and Pruned data sets. In general, the first PC discriminates Asian (left) from European (right) breed types. The first two PCs accounted for 16.9%, 20.2%, 19.5% and 14% variation in the WGS, Array_all, GG and Pruned data respectively. So, the amount of variation explained by the first two PCs was overestimated with Array_all and GG data, and underestimated with Pruned data. The 3rd PC in these sets still seemed to capture a reasonable amount of variation very close to the same amount captured by the 2nd PC (see Figure S2.4 in Additional file 2.3). Visually, the 3D plots showed at least some noticeable, but still small differences in the population structuring compared to the 2D plots.

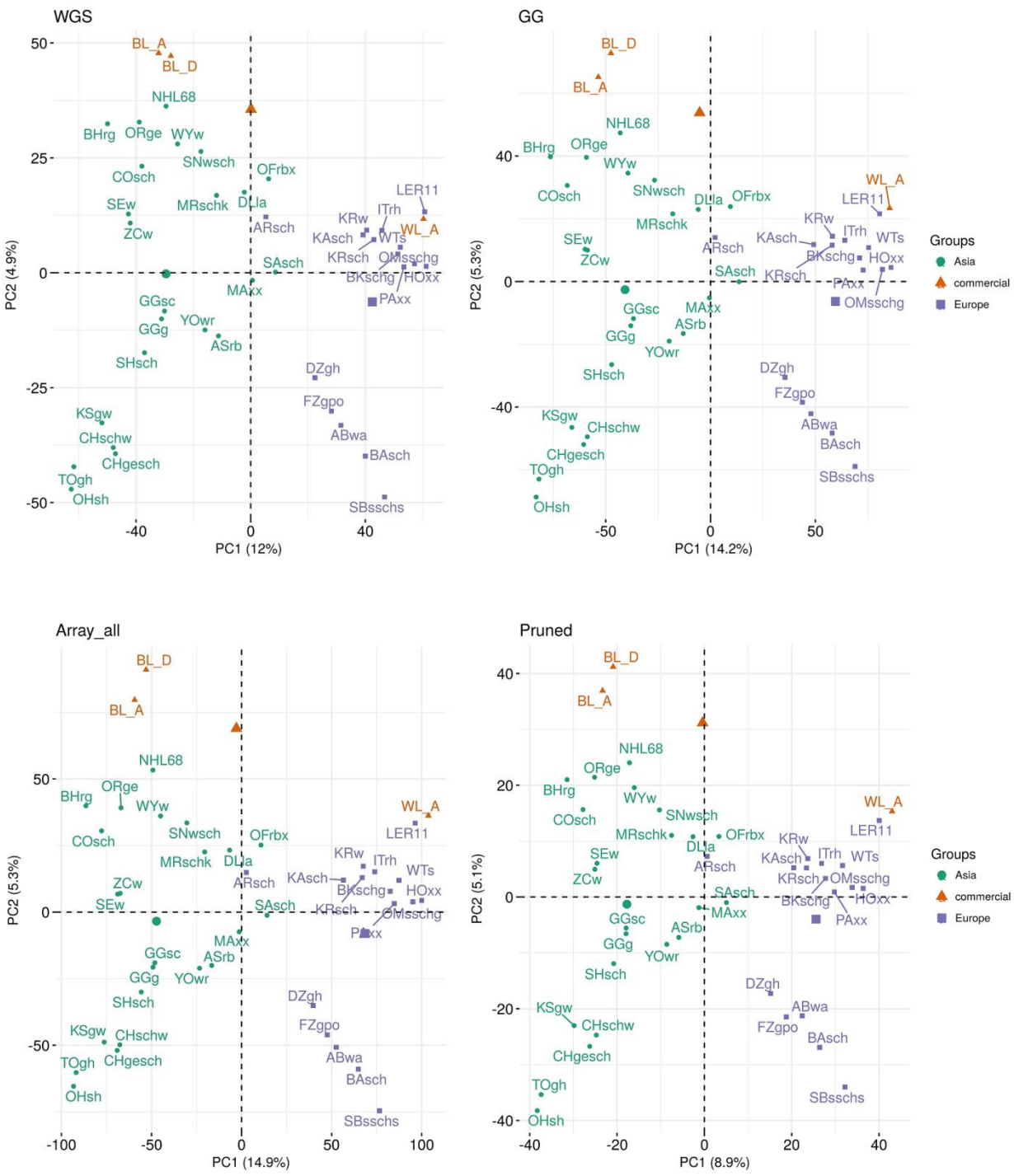


Figure 2.8: Two dimensional PCA plots of WGS and array (Array_all, GG and Pruned) data.

Discussion

When assessing allele frequency calling in the pooled WGS data, high correlations were obtained between the allele frequencies estimated with the Array_all' data set and pool WGS data set at each corresponding locus and very slight differences between the allele frequency spectra, we conclude that the estimation of allele frequencies from pooled sequences is sufficiently reliable. When comparing the AFS from the two datasets (not based on corresponding loci), the array dataset severely underrepresented the rare SNPs (Figure 2). This confirmed the already known findings of other studies on ascertained SNP data e.g. [9, 14, 15] and therefore suggests a risk for an ascertainment bias in array-based analysis of the chicken biodiversity panel.

To investigate the effects of ascertainment bias and strategies to mitigate its effects, we performed further genetic diversity analyses using the different filtered (LD based pruned, SNPs polymorphic to the GGsc and GGg populations and MAF filtering) versions of the array data and the results were compared with that obtained from the WGS data. LD based pruning of SNPs has been used in several studies presumed to produce reasonable genetic diversity comparisons between breeds [25, 55, 56]. The basic idea of LD based pruning is to remove markers which are highly correlated with other markers within a given window, leaving markers in the set with low LD to each other. This is efficient to remove the multicollinearity effects, which may result in overestimation of effects of SNPs due to highly correlated SNPs. For example, pairwise relatedness can be overestimated if the SNPs are highly correlated. LD based pruning is believed to be very effective when estimating differentiation measures between populations e.g. genetic distances, inbreeding coefficient, kinships and PCA [57].

Filtering of SNPs based on being polymorphic in wild populations not used in the SNP discovering process was discussed as a possibility to reduce ascertainment bias effects in the European Union

project (supported from the European Commission) “GLOBALDIV” (http://www.globaldiv.eu/) (not published). The idea was to use most original population within the same species or even a closely related species for selecting markers to be used in diversity studies in order to reduce the possible overestimation of diversity in the discovery panel populations.

Filtering of SNPs with less than 5% MAF is a common practice in quality control of SNP data because of concerns about lower genotyping rates, accuracy of genotype calling or perception about statistical conclusions that comes from analyzing such SNPs [58]. This filtering however will have consequences, there might be significant information behind these rare SNPs and removing them might hinder the chance of discovering such information [16].

Herrero-Medrano et al. [18] found that SNP chip data underestimated heterozygosity (both observed and expected) compared to next generation sequencing data. While Clark et al. [14] obtained higher heterozygosity estimates with ascertained HapMap data, the heterozygosity estimates were lowered after correcting for the bias. In our study, using the array data led to a systematic overestimation of the expected heterozygosity compared to WGS data. However, array data provided a very similar ranking of the breeds, as demonstrated by Spearman’s rank correlations between 0.956 (for Array_all) and 0.985 (for GG, see Table 2.3). Pruning SNPs based on LD resulted in a reduced overestimation of H_e compared to the other filters and improved the relationship with the WGS results slightly.

Estimating F_{ST} from the raw array data or with filtering for SNPs found in the wild populations resulted in inconsistence (i.e. underestimation of F_{ST} where WGS F_{ST} was low and overestimation the F_{ST} where WGS F_{ST} was high) estimates. These inconsistencies may cause misinformed conclusions on the actual differentiation among the populations. In a related study, ascertainment bias has shown to result in higher F_{ST} values from ascertained SNP data when compared with WGS

data [6]. Albrechtsen et al. [15] observed only a small difference in F_{ST} estimates between SNP chip and resequencing data. But when populations were less related to the ascertained panel, the F_{ST} estimates increased due to ascertainment bias. They therefore concluded that the bias is dependent on how the investigated populations are related to the ascertainment sample. The array used in our study was developed using several experimental and commercial broiler and layer lines [10]. Due to the multi-breed background of this discovery panel, it is challenging to relate each population to all of these discovery panel populations (including the ones that we didn't use in this study) in order to come up with a conclusion of whether the relatedness of these populations to the discovery populations affect their F_{ST} estimates. Additionally, similar to what we have observed with the H_e comparisons, the two commercial layers which we used in our study, were also affected differently (results not shown). This suggests that the effects of ascertainment bias on F_{ST} estimation in these data sets were very similar independent of whether the populations are more or less related to the discovery panel populations. The LD based pruned SNP data underestimated pairwise F_{ST} values between breeds, however in a consistent manner and thus should still be preferred over the other filtering strategies.

The clustering of populations by using both PCA and NJ trees is less affected by ascertainment bias. Even thou quantifiable measures such as Frobenius distances (for comparing the distance matrices of the two data types) and topological distances (for comparing the NJ trees) showed that the LD pruned data versions had a better relationship with the WGS data, the NJ trees computed from all array sets displayed similar clusters to the one computed with the WGS data. Ascertainment bias is expected to have limited and predictable effects on PCA. This is according to the in-depth explanation of the underlying processes, including migration, geographical isolation, and admixture in interpreting PCA projections explained by Mcvean [27]. Projections of

PCA from SNP genotype data are expected to be similar to PCA projections from WGS data unless the SNP discovery panel is very strongly biased [27]. This expectation was proven truthful in our study where all array data versions (even the Array_all) were found to exhibit structures which were visually very close to the ones obtained from WGS data.

In general, MAF filtering had very little or no effect in all comparisons done, and when its effect was noticeable it actually tended to worsen the results. Tabangin et al. [58] oppose discarding low MAF SNPs with the conception that it will inflate false positives results. Our results also discourage the MAF filtering to consequently study diversity.

Quite a number of studies [6, 8, 9, 11, 12, 46, among others] on ascertainment bias in genetic studies provide a very good background and insight on the topic. However, in most of these studies, the conclusions made on ascertainment bias and its effects on genetic analysis were based on simulated or limited real data. When investigating genome-wide genetic diversity in cattle breeds with SNP data, Edea et al. [60] also investigated the effects of ascertainment bias and most of our results are in agreement with their findings. Furthermore, we overcame the shortfalls that were not looked into in their study (i.e. we looked at more possible filtering options, we used WGS as a reference standard and our results discourages the MAF filtering)**.** To the best of our knowledge this paper presents so far the largest study on how different filtering strategies accounts for the effects of ascertainment bias in diversity studies, using real SNP genotype and WGS data. Some of our results (e.g. the only marginal difference between PCA from SNP genotype and WGS data) differ from what was claimed based on simulated (ascertained and non-ascertained) data (e.g. [12]).

Limitations of this study are due to the use of pooled WGS data with a limited number of individuals (9-15 per population) and with 20X coverage only. Due to this, low MAF SNPs may

still be missed and some measures, like observed heterozygosity and other inbreeding-related metrics, are not available for the WGS data. Nonetheless, the comparisons between the AFS of WGS and array data based on corresponding loci (Figure 2.1) has shown that estimated rare SNPs were a bit higher in the pooled sequence data than in array data therefore, implying a better detection of rare SNPs by sequence pooling (which are missed by the array data). Given these limitations, the pooled WGS data may not completely reflect all aspects of the true diversity of the studied breeds in a comprehensive way, but still our results allow a fair assessment of ascertainment bias and potential mitigation strategies for a number of relevant quantities.

Conclusions

Using the array genotype data as it is to study genetic diversity of different populations without any accountability measure for ascertainment bias runs the risk of getting misleading results. This study provides insights of how the effects of ascertainment bias can be minimized through appropriate SNP filtering strategies. A variety of populations were represented in our data, comprising possibly both close and distant to the populations in the discovery panel. The LD based pruning of SNPs has proven to yield consistent results which are highly comparable to those obtained from whole genome sequence data for the various populations used in this study in all the results. So, even though it doesn't fully account for ascertainment bias, the effects remain rather limited and are systematic and, by this, predictable. The other filtering strategies showed to be affected differently with some of the criteria (e.g. F_{ST} values between populations) and therefore may lead to inconsistent conclusions. Overall pruning of SNPs based on LD outperformed the other filtering strategies and is recommended for practical applications.

Additional files

Additional files are available online via the link:

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4416-9#Sec14

The files are named as follows:

Additional file 2.1 is named **Additional file 1.** Allele frequency spectrum figures of each population.

Additional file 2.2 is named **Additional file 2** and contains the following:

Table S2.1 named **Table S1.** Proportion of SNPs in genic and non-genic in WGS and array data.

Table S2.2 named **Table S2.** Population clusters.

Table S2.3 named **Table S3.** Topological distances between NJ trees of WGS and array data based on Billera method.

Table S2.4 named **Table S4.** Topological distances between NJ trees of WGS and array data based on Penny and Hendy method.

Table S2.5 named **Table S5.** Topological distances among the array versions.

Additional file 2.3 is named **Additional file 3** and contains the following:

Figure S2.1 named **Figure S1.** Comparion of MAF between genic and non-genic regions in array (left) and WGS (right) data.

Figure S2.2 named **Figure S2.** Comparisons of expected heterozygosity (H_e) estimates between WGS (boxplot of 100 replicates) and array (Array_all, GG and Pruned) data, for all 42 populations.

Figure S2.3 named **Figure S3.** Expected heterozygosity (H_e) estimated with array vs. WGS data for the 39 populations.

Figure S2.4 named **Figure S4.** Three dimensional PCA plot of A) WGS, B) GG and C) Pruned array data.

Additional file 2.4 is named **Additional file 4.** An excel file containing the ENA accession numbers of all sequence reads used in the WGS data.

List of abbreviations

AFS: Allele frequency spectrum, *F*: Frobenius, F_{ST}: Fixation index, H_e: Expected heterozygosity, LD: Linkage disequilibrium, MAF: Manor allele frequency, NJ: Neighbor joining,

PC: Principal component, PCA: Principal components analysis, SNP: Single nucleotide polymorphism, SVS: SNP Variation Suite, Synbreed: Synergistic Plant and Animal Breeding, WGS: Whole genome sequence

Ethics approval and consent to participate

We confirm that the collection of blood samples for this study was performed 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).

Availability of Data and Materials

The datasets analyzed during the current study consists of two data types: the SNP array (individual genotype) data and the WGS data. The array data is available at https://owncloud-shib.gwdg.de/index.php/s/yuRx70UYcXNrFN4 as PLINK binary files containing genotype and map information. The WGS data has been uploaded to the European Nucleotide Sequence Database (ENA) and the accession numbers are in Additional file 2.4.

Funding

This research was funded by the German Federal Ministry of Education and Research (FKZ 0315528E) through the "Synbreed - *Synergistic Plant and Animal Breeding*" project. This work is part of DKM's Doctoral programme which is supported financially by the Erasmus Mundus (through the INSPIRE project).

Authors' contributions

DKM contributed to the conception and design of the study, analyzed and interpreted the data; wrote the final draft and prepared the submitted manuscript. HS supervised the study, substantially contributed to the conception and design of the study as well as the revision and editing of the manuscript. CR contributed to the data analysis, conception and design of the study; and revision and editing of the manuscript. SW contributed to the conception and design of the study, participated in the provision of the data as well as reviewing and editing of the manuscript. AW participated in the provision and preparation/editing of the data. ARS contributed to the revision of the manuscript and performing some statistics. All authors read and approved the manuscript.

Acknowledgements

We are grateful to all the participated breeders for their assistance in sampling.

References

[1] Darwin C. The variation of animals and plants under domestication. London, UK: John Murray; 1868.

[2] Darwin C. The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London, UK: John Murray; 1859.

[3] Hillel J, Groenen MAM, Tixier-Boichard M, Korol AB, David L, Kirzhner VM, et al.

Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genet Sel Evol. 2003; 35: 533–57.

[4] Granevitze Z, Hillel J, Feldman M, Six A, Eding H, Weigend S. Genetic structure of a wide-spectrum chicken gene pool. Anim Genet. 2009; 40: 686–93.

[5] Eltanany M, Distl O. Genetic diversity and genealogical origins of domestic chicken. Worlds Poult Sci J. 2010; 66: 715–26.

[6] Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays. 2013; 35: 780–6.

[7] Qanbari S, Simianer H. Mapping signatures of positive selection in the genome of livestock. Livest Sci. 2014; 166: 133–43.

[8] Brumfield RT, Beerli P, Nickerson DA, Edwards SV. The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol. 2003; 18: 249–56.

[9] McTavish EJ, Hillis DM. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics. 2015; 16: 266.

[10] Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013; 14: 59.

[11] Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004; 168: 2373–82.

[12] Nielsen R, Signorovitch J. Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. Theor Popul Biol. 2003; 63: 245–55.

[13] Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME. Impact of Marker Ascertainment

Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. PLoS One. 2013; 8: e74612.

[14] Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15: 1496–502.

[15] Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010; 27: 2534–47.

[16] Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. Am J Hum Genet. 2008; 82: 100–12.

[17] Rosenblum EB, Novembre J. Ascertainment bias in spatially structured populations: A case study in the Eastern Fence Lizard. J Hered. 2007; 98: 331–6.

[18] Herrero-Medrano JM, Megens H-J, Groenen MA, Bosse M, Pérez-Enciso M, Crooijmans RP. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. BMC Genomics. 2014; 15: 601.

[19] Rogers AR, Jorde LB. Ascertainment Bias in Estimates of Average Heterozygosity. Am J Hum Genet. 1996; 58: 1033–41.

[20] Raychaudhuri S, Stuart JM, Altman RB. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. In: Pacific Symposium on Biocomputing. 2000. p. 455–66.

[21] Ringnér M. What is principal component analysis? Nat Biotechnol. 2008; 26: 303–4.

[22] Ma J, Amos CI. Principal Components Analysis of Population Admixture. PLoS One. 2012; 7: e40115.

[23] Yu K, Wang Z, Li Q, Wacholder S, Hunter DJ, Hoover RN, et al. Population Substructure and Control Selection in Genome-Wide Association Studies. PLoS One. 2008; 3: e2551.

[24] Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006; 2.

[25] López Herráez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, et al. Genetic variation and recent positive selection in worldwide human populations: Evidence from nearly 1 million SNPs. PLoS One. 2009; 4: e7888.

[26] Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 2010; 107: 786–791.

[27] McVean G. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet. 2009; 5: e1000686.

[28] Qanbari S, Strom TM, Haberer G, Weigend S, Gheyas AA, Turner F, et al. A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens. PLoS One. 2012; 7: e49525.

[29] Lyimo CM, Weigend A, Msoffe PL, Eding H, Simianer H, Weigend S. Global diversity and genetic contributions of chicken populations from African, Asian and European regions. Anim Genet. 2014; 45: 836–48.

[30] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004; 432: 695–716.

[31] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.

Bioinformatics. 2009; 25: 1754–60.

[32] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–9.

[33] Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297–303.

[34] Depristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next- generation DNA sequencing data. Nat Genet. 2011; 43: 491–8.

[35] Reimer C, Rubin CJ, Weigend S, Waldmann KH, Distl O, Simianer H. The Minipig Genome Harbors Regions of Selection for Growth. In: 10th World Congress on Genetics Applied to Livestock Production. Vancouver, BC, Canada; 2014.

[36] SNP & Variation Suite ™ (Version 8.1). Bozeman, MT: Golden Helix, Inc. Available at: http://goldenhelix.com/.

[37] Purcell S, Chang C. PLINK 1.9. https://www.cog-genomics.org/plink2. Accessed 12 Mar 2017.

[38] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 7.

[39] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559–75.

[40] Gautier M, Foucaud J, Gharbi K, Cézard T, Galan M, Loiseau A, et al. Estimation of

population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. Mol Ecol. 2013; 22: 3766–79.

[41] Yates A, Akanni W, Ridwan Amode M, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016; 44.

[42] Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl Gene Annotation System. Database. 2016; 1-19.

[43] Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th edition. Essex, UK: Longmans Green, Harlow; 1996.

[44] Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution (N Y). 1984; 38: 1358–70.

[45] Nei M. Genetic distance and molecular phylogeny. In: Population genetics and fishery management. Ryman N, Utter F, editors. Washington: Washington Sea Grant Program; 1987. p. 193–223.

[46] Weisstein EW. Frobenius Norm. http://mathworld.wolfram.com/FrobeniusNorm.html. Accessed 13 Mar 2017.

[47] Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinforma . 2004; 20: 289–90.

[48] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. URL https://www.R-project.org/.

[49] Penny D, Hendy M. The use of tree comparison metrics. Syst Zool. 1985; 34: 75–82.

[50] Billera LJ, Holmes SP, Vogtmann K. Geometry of the Space of Phylogenetic Trees. Adv Appl

Math. 2001; 27: 733–767.

[51] Vellend M, Cornwell WK, Magnuson-Ford K, Mooers AØ. Measuring phylogenetic biodiversity. In: Frontiers in Measuring Biological Diversity. Magurran, A.E. and McGill BJ, editor. Oxford University Press; 2010. p. 194–207.

[52] Dray S, Dufour A-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. J Stat Softw. 2007; 22.

[53] Ligges U, Martin M. Scatterplot3d - an R package for Visualizing Multivariate Data. J Stat Softw. 2003; 8: 1–20.

[54] Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen M a M, Crooijmans RPM a, et al. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc Natl Acad Sci. 2008; 105: 17312–7.

[55] Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, et al. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One. 2009; 4: e4668.

[56] Makina SO, Muchadeyi FC, van Marle-Koster E, MacNeil MD, Maiwashe A. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front Genet. 2014; 5: 1–7.

[57] Double Helix Inc. Determining the best LD Pruning options. http://blog.goldenhelix.com/jbartole/determining-best-ld-pruning-options/. Accessed 12 Mar 2017.

[58] Tabangin ME, Woo JG, Martin LJ. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc. 2009; 3.

[59] Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004; 1: 218–24.

[60] Edea Z, Bhuiyan MSA, Dessie T, Rothschild MF, Dadi H, Kim KS. Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. Animal. 2015; 9: 218–26.

CHAPTER 3

The SYNBREED chicken diversity panel: A global resource to assess chicken diversity at high genomic resolution

Dorcus Kholofelo Malomane[1,2], Henner Simianer[1,2], Annett Weigend[3], Christian Reimer[1,2], Armin Otto Schmitt[2,4], Steffen Weigend[2,3]

[1]Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, 37075 Goettingen, Germany

[2]Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, 37075 Goettingen, Germany

[3]Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, 31535 Neustadt, Germany

[4]Breeding Informatics Group, Department of Animal Sciences, University of Goettingen, 37075 Goettingen, Germany

Published in BMC Genomics

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5727-9

Abstract

Background: Since domestication, chickens did not only disperse into the different parts of the world but they have also undergone significant genomic changes in this process. Many breeds, strains or lines have been formed and those represent the diversity of the species. However, other than the natural evolutionary forces, management practices (including those that threaten the persistence of genetic diversity) following domestication have shaped the genetic make-up of and diversity between today's chicken breeds. As part of the SYNBREED project, samples from a wide variety of chicken populations have been collected across the globe and were genotyped with a high density SNP array. The panel consists of the wild types, commercial layers and broilers, indigenous village/local type and fancy chicken breeds. The SYNBREED chicken diversity panel (SCDP) is made available to serve as a public basis to study the genetic structure of chicken diversity. In the current study we analyzed the genetic diversity between and within the populations in the SCDP, which is important for making informed decisions for effective management of farm animal genetic resources.

Results: Many of the fancy breeds cover a wide spectrum and clustered with other breeds of similar supposed origin as shown by the phylogenetic tree and principal component analysis. However, the fancy breeds as well as the highly selected commercial layer lines have reduced genetic diversity within the population, with the average observed heterozygosity estimates lower than 0.205 across their breeds' categories and the average proportion of polymorphic loci lower than 0.680. We show that there is still a lot of genetic diversity preserved within the wild and less selected African, South American and some local Asian and European breeds with the average observed heterozygosity greater than 0.225 and the average proportion of polymorphic loci larger than 0.720 within their breeds' categories.

Conclusions: It is important that such highly diverse breeds are maintained for the sustainability and flexibility of future chicken breeding. This diversity panel provides opportunities for exploitation for further chicken molecular genetic studies. With the possibility to further expand, it constitutes a very useful community resource for chicken genetic diversity research.

Background

Chickens are of major and increasing importance for agricultural production as an efficient source of high quality protein. There have been concerns about loss of animal genetic resources and erosion of many genotypes due to crossbreeding or replacement by the high performing commercial hybrids resulting from highly efficient selection programs [1, 2]. Such loss of valuable genetic material will put a strain on animal production and could make it vulnerable to challenges in the future. It is therefore important to preserve genetic resources that may help to meet future demands in animal breeding [3, 4]. Studying and understanding the diversity between and within populations clearly is crucial for effective management of farm animal genetic resources [5].

Domestication history of chickens is still a matter of scientific debate, and has enjoyed the interest of researchers and scholars, from tracing the centers of domestication to exploring the archeology and dispersion of the chickens across different parts of the world [6–10]. One widely accepted hypothesis is that the main source of today's chickens which are diffused across the world comes from domestication events that have taken place in the Indus Valley during 2500-2100 B.C. [6, 11]. Since domestication, chickens have been widely dispersed from Asia to the different parts of the world. Several routes from the centers of domestication to Europe, Africa and South America have been reported [9, 10, 12–14]. From Asia, chickens are believed to have reached Europe through the Mediterranean region and through the north via China and Russia to Northern Europe [6]. It is supposed that chickens in Africa have descended from both European and Asian chicken

stocks [6, 8]. Despite the debate on whether the South American chickens originated from Polynesian or European breeds [9, 13, 15, 16], it is clear that both European and Asian flocks have contributed to the South American chicken breeds. Several local Asian and European breeds have formed the founder stocks to develop commercial egg laying and broiler chickens. Subsequently, the commercial lines have been highly selected for production purposes (e.g. meat, egg production and feed conversion efficiency) [5, 12, 17].

In Europe, many local type breeds were developed mostly by intense selection and crossbreeding for desired phenotypic traits. In the 19th century, with an increasing popularity local strains maintained for centuries in Europe have been developed into standardized chicken breeds. At the same time, Asian breeds such as Cochin and Langshan were imported to Europe. In addition to keeping them as purebred populations, many new breeds evolved from crossing European breeds and newly imported Asian breeds following the European Poultry Standards [18]. Fancy chicken breeding in Europe is characterized by limited exchange of mating individuals resulting in population fragmentation, which promotes inbreeding when population sizes are small. In Asian, African and South American countries, however, local chicken breeds are often raised by villagers under extensive farming systems and with little to no selection, and exchange of breeding stocks across close villages [1, 19–23]. Due to the often low productivity of local, unselected breeds in many developing countries, the production of local breeds has been threatened by the commercial breeding and the introduction of crossbreeding to improve productivity [8, 20, 24].

The history of the origin of chickens together with management practices following domestication provides an important backbone to assess the genetic make-up and diversity between today's chicken breeds. Low resolution studies of chicken biodiversity using microsatellites have shown that genetic diversity has been greatly affected by management practices. Highly selected layer

lines, in particular white layers, showed reduced genetic variability while the wild type and less improved indigenous village chickens retained high genetic diversity [25, 26]. In this study we used single nucleotide polymorphism (SNP) genotype data to study the biodiversity of a wide range of globally sampled chicken populations at a high genomic resolution. This data was acquired under the umbrella of the SYNBREED (www.synbreed.tum.de) project. The SYNBREED chicken diversity panel used here consists of 174 chicken populations, representing four continents (Asia, Europe, South America and Africa). The SCDP also includes broiler and layer purebred lines, as well as two wild populations (*Gallus gallus gallus* and *Gallus gallus spadiceus*). We have included some commercial lines in our analyses as representatives of the most favored stocks in breeding programs whose end products are distributed globally. They are not at risk for extinction, but may threaten local breeds by crossbreeding. We show their share of genetic diversity with a much wider spectrum of chicken breeds in SCDP set which these commercial lines do not cover. We have analyzed the genetic diversity within and between the populations and report here the current status of global genetic diversity based on this panel.

Methods

Data acquisition

Animals. Deoxyribonucleic acid (DNA) samples were collected from a wide range of chicken populations across the globe under the umbrella of the SYNBREED project (project lifetime 2009 – 2014). First, samples were collected from 80 fancy chicken breeds in Germany between 2010 and 2012. Fancy breeds are chickens which have been developed following hobbyists' selection programs to create phenotypes which meet the requirements of the poultry standards (i.e. the European Poultry Standards). The German Association of Poultry Breeders (Bund Deutscher Rassgeflügelzüchter e.V., BDRG) maintains a wide spectrum of traditional and fancy poultry

breeds. They reflect various types of breeds of very different origins and breed histories according to the European Poultry Standards. Additional samples were collected from chicken breeds kept by farmers organized in “The Society for the Conservation of Old and Endangered Livestock Breeds (Gesellschaft zur Erhaltung alter und gefährdeter Haustierrassen e.V., GEH)”. Blood samples were collected from the wing vein using EDTA as anticoagulant. Sampling was carried out in strict accordance to the German Animal Welfare regulations, and notice was given to the authorities of Lower Saxony according to § 8 of the German Animal Welfare Act (33.9-42502-05-10A064) and with the written consent of the animal owners. The collection was completed by samples of two Red Jungle Fowl populations, *Gallus gallus gallus* and *Gallus gallus spadiceus*, as well as samples of nine local breeds and four broiler lines taken from the previous EU project AVIANDIV (https://aviandiv.tzv.fal.de/, see also [25]). In addition, four commercial purebred white layer lines and four commercial purebred brown layer lines were added from other subprojects of the SYNBREED project.

After 2012, the panel was complemented with DNA samples of 71 populations from 22 countries provided by partners (see Table 3.1) or taken from previous collaborations. The total data used in this study consisted of 3,235 individuals from 162 populations (from 32 countries, representing the Africa, South America, Asia, and Europe) and 12 commercial purebred lines (4 white egg layers, 4 brown egg layers and 4 broilers). The breeds’ information (i.e. names, acronyms, samples sizes and other information) is presented in Table S3.1 in Additional file 3.1.

Table 3.1: The SYNBREED Chicken Diversity Consortium

Contact	Sampling region	Institution
Olivier Hanotte	Albania	School of Life Sciences, University of Nottingham, United Kingdom
Miika Tapio/Mervi Honkatukia	Finland	Luke Natural Resources Institute, Finland
Steffen Weigend	Germany	Friedrich-Loeffler-Institut, Germany
Henner Simianer	Germany	Georg-August-Universität, Germany
András Hidas	Hungary	Institute for Small Animal Research, Hungary
Amadeu Francesch	Spain	IRTA-Centre Mas de Bover, Spain
Christine Flury	Switzerland	School of Agricultural Forest and Food Sciences, Bern University of Applied Sciences, Switzerland
Asmaa Abushady	Egypt	Genetics Department, Faculty of Agriculture, Ain Shams University, Cairo, Egypt.
Olivier Hanotte/Takele Desta	Ethiopia	School of Life Sciences, University of Nottingham, United Kingdom
Ahmad Ali	Pakistan	Department of Bioscience COMSATS, University Islamabad, Pakistan
Mohyeldein Berima	Sudan	Department of Animal Production, Faculty of Agriculture, University of Zalingei, Sudan
Charles Lyimo	Tanzania	Sokoine University of Agriculture, Tanzania
Farai Muchadeyi	Zimbabwe	Agricultural Research Council-Biotechnology Platform, South Africa
Raed M. Al-Atiyat/Riyadh S. Aljumaah	Saudi Arabia	King Saud University, Kingdom of Saudi Arabia
Mohammad Shamsul Alam Bhuiyan	Bangladesh	Department of Animal Breeding and Genetics, Bangladesh Agricultural University, Bangladesh
Guohong Chen	China	Yangzhou University, Jiangsu Province, People's Republic of China
Mehmet Ali Yildiz	Turkey	Animal Science, Faculty of Agriculture, Ankara University, Turkey

Cuc, Ngo Thi Kim	Vietnam	National Institute of Animal Science, Vietnam
Jeremy Austin / Michael Herrera	Pacific/Philippines	School of Biological Sciences, University of Adelaide, Australia
Maria Rosa Lanari	Argentina	National Institute of Agricultural Technology, Argentina
Fernando Mujica	Chile	Universidad Austral de Chile, Chile
Carl Schmidt	Rwanda/Uganda	University of Delaware, Delaware, USA

Samples from Iceland,Norway, Poland, Russia, Ukraine, France, Italy, Israel, Thailand were taken from the AVIANDIV project (https://aviandiv.tzv.fal.de/, EC project BIO4CT980342)

The populations labeled (in the 'Label' column) with an acronym ending with 'xx' include individuals that belong to different color varieties or that were sampled from different regions, even though they belonged to the same breed. They either were kept by different breeders with unknown exchange of genetic material or were sampled in different regions within a country. Therefore, the definition of a population in our study refers to the sampling population rather than a breeding population because this does not apply to some of the populations. For all fancy breeds sampled in Germany, breed names follow the European Poultry Standards [18]. The breed named "Italiener" (Italian), with different color varieties, is a Leghorn type breed for which a separate breed standard exists in Germany.

The populations were classified into twelve categories based on their continent of origin and/or type as shown in Table 3.2 and Table S3.1 in Additional file 3.1. In the case of populations of Asian and European origin collected in Germany, the sampling location was also included in the category name (as "DE"). The category Asia_local included native chicken breeds sampled in Asia. Likewise, the category Europe_local comprises breeds of European background sampled in different parts of Europe. The DE_Europe_Ban and DE_Asia_Ban categories consist of bantam

type chickens from European and Asian origin, respectively, which were both sampled in Germany. Some of the breeds have already been characterized in other studies (references provided in the last column of Table S3.1 in Additional file 3.1) mainly using microsatellites.

Genotyping. DNA samples were genotyped with the Affymetrix® Axiom™ Genome-Wide Chicken Genotyping Array encompassing over 580K SNPs [27]. Genotyping was performed at the Technische Universität München (Prof. R. Fries). In a few cases, which are marked with an asterisk in Table S3.1, genotype data was provided by partners.

Table 3.2: Categories of chicken breeds

Category	Full name	Number of breeds	Number of individuals
Wild	Wild type chicken	2	38
Com_WL	Commercial white layers	4	80
Com_BL	Commercial brown layers	4	80
Com_BRO	Commercial broilers	4	73
DE_Europe_Ban	European bantams sampled in Germany	8	156
DE_Europe	European breeds sampled in Germany	35	660
DE_Asia_Ban	Asian bantams sampled in Germany	8	177
DE_Asia	Asian breeds sampled in Germany	28	531
Europe_local	European local breeds sampled across Europe	25	443
Asia_local	Asian local breeds sampled across Asia	30	509
South_America	South American breeds	4	78
Africa	African breeds	22	410
Overall		**174**	**3,235**

Data editing and filtering

In total, genotype information for 580,961 SNPs was obtained from the array. 579,621 of the SNPs were annotated using the genome assembly Gallus_gallus-5.0 [28]. We deleted 134 duplicated SNPs (both SNPs deleted). We only considered SNPs from the 28 autosomal chromosomes and removed 26,839 SNPs from the two sex chromosomes. Furthermore, we deleted 499 SNPs with ambiguous chromosome annotation. We filtered the data for an animal call rate of ≥95% and SNP call rate of ≥99% (leaving 436,581 SNPs) using the SNP & Variation Suite (SVS) version 8.1 [29]. We then performed LD based pruning which has been found to effectively reduce the effects of ascertainment bias in diversity analysis when using SNP data [30]. LD based pruning of SNPs was performed using SVS with the parameters "50 5 0.2", which represent window size, window shift and r^2 (pruning of markers with a pairwise r^2 of greater than 0.2), respectively, leaving 123,273 SNPs for further analysis. Furthermore, imputation was performed on the remaining SNPs using Beagle 3.3 [31] to recover missing genotypes.

Data analysis

Genetic diversity between populations and assessing the population structure

Genetic distances and phylogenetic tree: We estimated Reynolds' genetic distances [32] between the sampled populations. These distances were used to construct an unweighted neighbor joining (NJ) tree using SplitsTree software (version 4.14.4) [33]. Based on the tree we identified possible clusters and labeled them accordingly.

Principal component analysis: We performed a principal component analysis (PCA) using SVS. Because of the large number of 3,235 individuals, we calculated the average principal component (PC) scores for each population to make their positions in the PCA plot presentable. We then

plotted the average scores of each population for PC 1 & 2 with different colors to highlight the breeds' categories.

Admixture analysis: We evaluated the relatedness of the populations through admixture analysis using ADMIXTURE 1.3 software [34]. ADMIXTURE determines population relatedness and assigns populations to ancestral clusters. It includes a cross-validation procedure that allows the identification of a number of populations K which fits best the model based upon cross-validation (CV) error. We analyzed our data set up to a value of K = 80, however without reaching a minimum of the CV error (data not shown). We display results for K = 2 to 11 according to the number of clusters identified with the NJ tree, to illustrate population relatedness and assignment of populations to clusters with proportions to ancestral populations.

Genetic diversity within populations.

Genetic variability measures such as proportion of polymorphic SNP loci, levels of observed (H_o) and expected (H_e) heterozygosity were used to evaluate the genetic diversity within populations. The observed heterozygosity was calculated directly, while the expected heterozygosity was estimated as: $H_e = 2q(1-q)$ where q was the frequency of one of the alleles [35]. The H_e and H_o estimates for all individuals within each population were averaged over all SNPs. Because of low sample sizes and the fact that a number of the populations did not form mating groups, the calculated expected heterozygosity values should be treated with caution. Consequently, for many of the breeds we avoided making conclusions based on the Hardy–Weinberg principles.

Results

Genetic diversity between populations and the population structure

Neighbor joining tree and cluster assignment. The Reynolds' genetic distances between populations were used to construct a neighbor joining tree which is presented in Figure 3.1. We labeled observed clusters on the tree. It should be noted that these clusters were identified manually according to our visual interpretation. Below we provide a general description of the clustering results. More detailed information about the clusters and breeds within each cluster is presented in Document S3.1 in Additional file 3.2.

In cluster 1, the White Leghorn lines of both commercial and fancy breeds are grouped together. Cluster 2 consists of breeds of European background (green). Cluster 3 encompasses mainly breeds from the Middle East and geographically nearby areas, sampled in Saudi Arabia, Egypt, Pakistan, Israel, Sudan, Ethiopia, Turkey, and Italy. The close relationship of the breeds in this cluster is likely due to their neighboring geographic distribution and distribution routes of chickens in these regions. The NJ tree further shows a very small cluster (cluster 4) which consists of two populations of Vorwerkhuhn (VWco and VWcoE) and Lakenfelder (LAco). Vorwerkhuhn was recognized as a standardized breed in Germany in 1919 and one of the founder breeds was the Lakenfelder breed. Cluster 5 consists of European bantam breeds as well as some Asian game birds which were sampled in Germany. Cluster 6 consists exclusively of chickens sampled in Finland. Following this group, several populations of European background were arranged in the middle of the tree, but were not forming a visually distinct cluster. They were found between cluster 6 and 7. Among these breeds there were three populations of the Araucanas. Cluster 7 branches into two sub-clusters with African populations on the one side and South American Mapuche populations on the other. Among the African populations, there were two Tanzanian

ecotypes (MOxx and CWxx). Though five ecotypes from Tanzania were included in this study, they did not cluster together in the NJ tree. The remaining three breeds from Tanzania clustered with populations in cluster 9. The second sub-cluster including the South American sub-cluster also contained some populations from Eastern Europe (Russia, Ukraine and Albania).

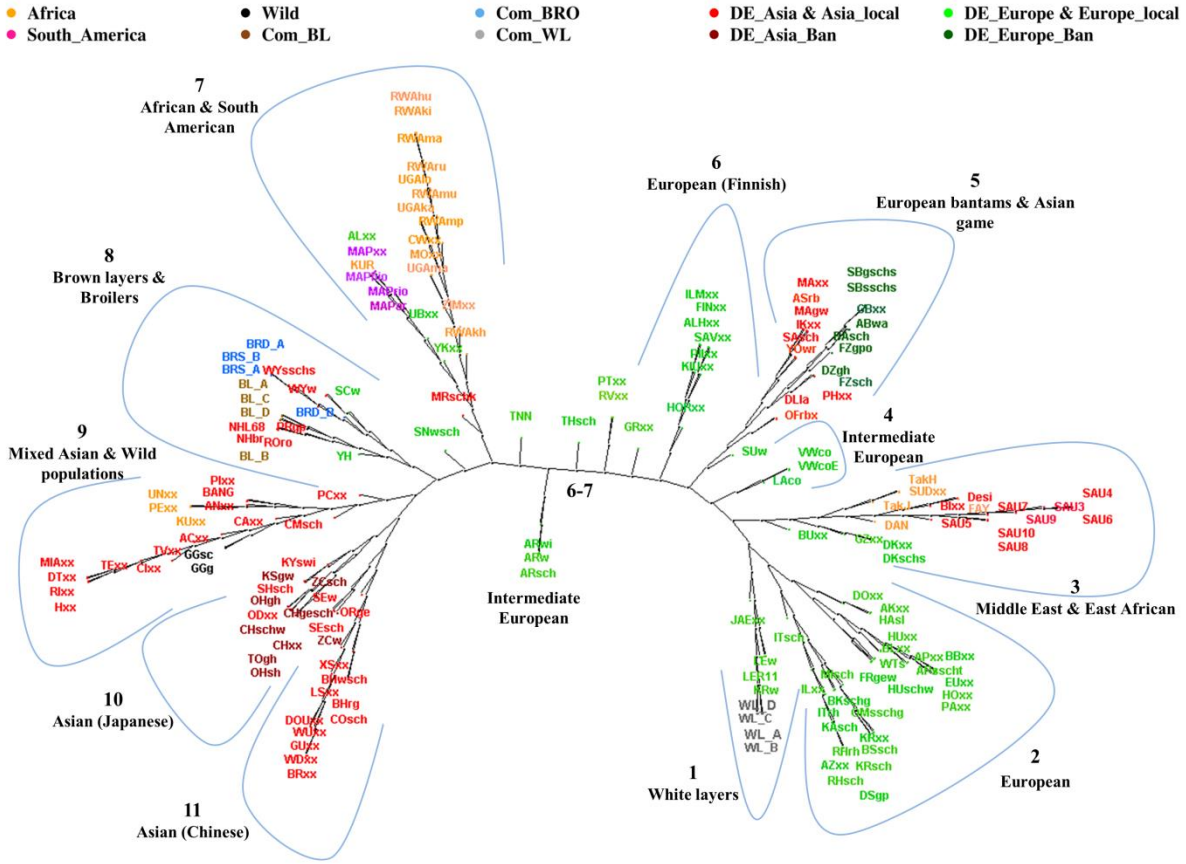


Figure 3.1: Neighbor Joining tree of 174 chicken populations based on Reynolds' genetic distances calculated from SNP genotypes. Clusters 1 to 11 are described in the main text and in detail in Document S3.1 in Additional file 3.2.

In cluster 8 commercial brown layers and broilers are found. Close to the four commercial purebred brown layer lines (BL_A-D), there were also two lines of New Hampshire (NHL68 and NHbr) and the fancy breed of Rhode Island Red (ROro). Two of the brown layer lines (BL_A and BL_B)

originated from the breed Rhode Island Red while the other two lines (BL_C and BL_D) are based on White Plymouth Rock. New Hampshires may have formed a part of the dam lines used in the development of brown layer lines [17]. The Plymouth Rocks (PRgp) sampled from fancy breeders clustered close to the purebred broiler lines (BRS_A, BRS_B, BRD_A and BRD_B). Plymouth Rock was part of the female lines for the development of broiler chickens [12]. Even though modern broiler lines became very different from these main founders, it is interesting to see that they clustered together with the fancy breed of Plymouth Rock. Cluster 9 is dominated by breeds of Asian background, mainly from Vietnam. They clustered with three of the Tanzanian ecotypes. Notably, in this cluster the two wild populations (GGg and GGsc) sampled in Thailand were also found. Both clusters 10 and 11 consist of breeds of exclusively Asian background. The breeds in cluster 10 are mainly Japanese and were sampled in Germany. Cluster 11 is dominated by Chinese breeds sampled in both Europe (Germany) and Asia. All the Asian bantam breeds which were sampled in Germany were also found in clusters 10 and 11.

Principal component analysis. Average scores of each population for PC1 versus PC2 are shown in Figure 3.2. Populations sampled in Germany are denoted by triangle symbols. The commercial breeds and the two wild populations are displayed as squares, while the rest of the populations are marked as dots. The first PC shows a gradually increasing separation of the European type breeds (green) on the one side from the Asian breeds (red) on the other side, with the African (orange) and South American (pink) breeds in the middle. The Asian breeds sampled in Germany clustered with chicken breeds sampled in Asia. The Mapuche chickens sampled in South America clustered mostly towards the Asians side of the PCA plot, while the African types were separated, with some of them clustering towards the Asian and others towards the European breeds. The Asian breeds (populations also seen in NJ cluster 3) from the Middle East and nearby regions (i.e. Saudi

Arabians, Bedouin from Israel, and Desi from Pakistan), Indian game breed (IKxx), Sumatra black (SAsch) and Orloff (OFrbx) clustered with the European breeds and some of the African breeds. A few breeds of (mostly eastern) European origin found in cluster 7 and 8 of the NJ tree included breeds such as the Hungarian Yellows (YH), the Albanian Crowers (ALxx), the Ukrainian bearded (UBxx), the Yurlov crower (YKxx) from Russia, as well as ALH (ALHxx) from Finland and Swiss chicken (SCw). They clustered in the Asian side of the PCA plot with broilers, South American and some of the African breeds. PC1 also shows a wide separation between the two layer line types, the commercial brown layers (in brown colour) on the one end and the white (gray) egg layers on the other. Commercial broilers (light blue) are between them, but much closer to the brown layers.

It is noteworthy that the second PC could be related to the breed's body size. This is because the PC2 shows a transitioning of mainly the small sized (mostly bantams) birds at the lower part of the PCA plot and the normal sized birds towards the upper part. However, the separation is much clearer for the European type breeds than for the Asian types.

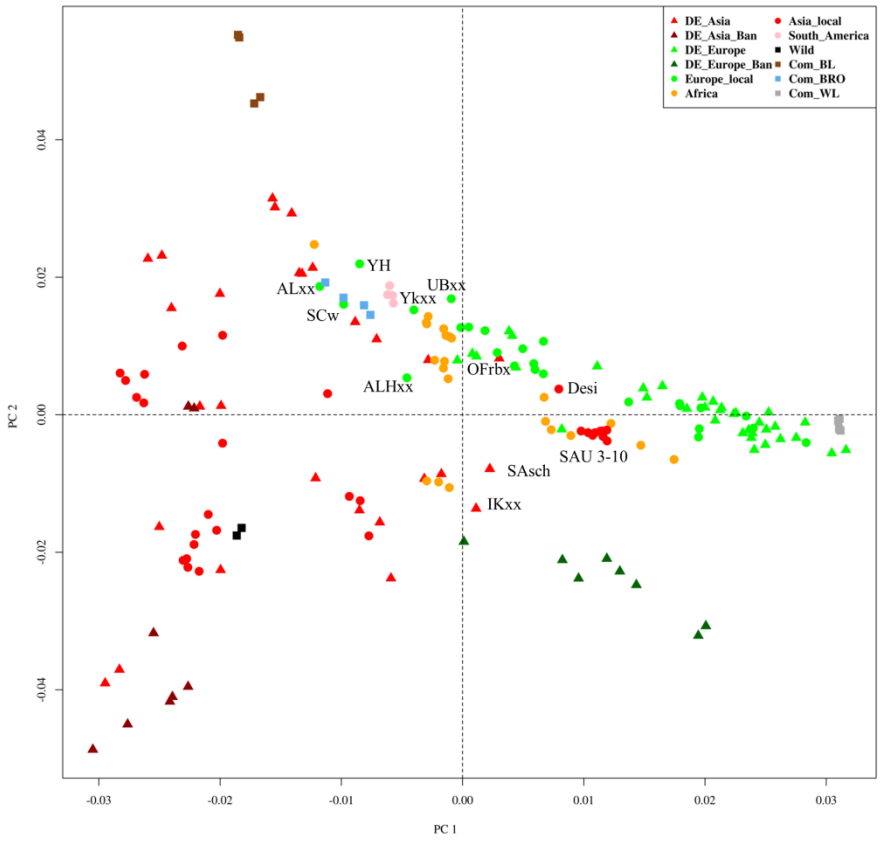


Figure 3.2: Principal component analysis with components averaged across populations. Breeds which are labelled, their names are mentioned in the main text.

Admixture analysis. Admixture analysis results for K = 2 and K = 5 are displayed in Figure 3.3; results for the other K-values up to K = 11 are shown in Figure S3.1 in Additional file 3.3. We only included in the main text the results for K = 2 to show the overall structure of the studied populations, K = 5 to show the extent of admixture in these populations because K = 5 was visually clear and less dense than K values greater than 5. We transformed the NJ tree from Figure 3.1 into a cladogram in order to relate the tree to the admixture plots. Since the Araucana populations are found in the center of the NJ tree in Figure 3.1, we used one of them, the Araucana black (ARsch), as the first breed in the cladogram. We then adopted the order of the breeds obtained from the

cladogram (clusters 1 to 11 as in Figure 3.1) as the order of the breeds in the admixture plots (Figure 3.3).

In agreement with the PCA results, the admixture exhibited a gradually increasing separation of breeds from European (green) background from breeds of Asian (red) background with K = 2, with the African and South American breeds situated in the middle of the spectrum. The commercial white layers were completely homogeneous in the European gene pool (green) at K = 2 while the brown layers and broilers were admixed, however, with more proportion of the Asian ancestry cluster. In the NJ tree, cluster 3 is made up of populations from Asia (Saudi Arabia, Pakistan, and Israel) and Africa (Sudan, Egypt and Ethiopia) clustering in the middle of European clusters. On the admixture plot (Figure 3.3) these populations of cluster 3 display a larger genome share with Europeans (green). Regarding the African populations, the populations found in NJ clusters 7 and 9 had more affiliation to the Asian gene pool except for the Tanzanian ecotypes Morogoro and Ching'wekwe in NJ cluster 7. The admixture analysis shows that the Morogoro and Ching'wekwe share more European lineage similarly to the African breeds of cluster 3 rather than those of cluster 7. The assignment of these two breeds to cluster 7 on the NJ tree could be due to the one-dimensional nature on the phylogenetic tree with limited capability to resolve the membership when breeds are more related to several other breeds. The South American Mapuche chickens were more affiliated with the Asians at K = 2 as in the PCA plot.

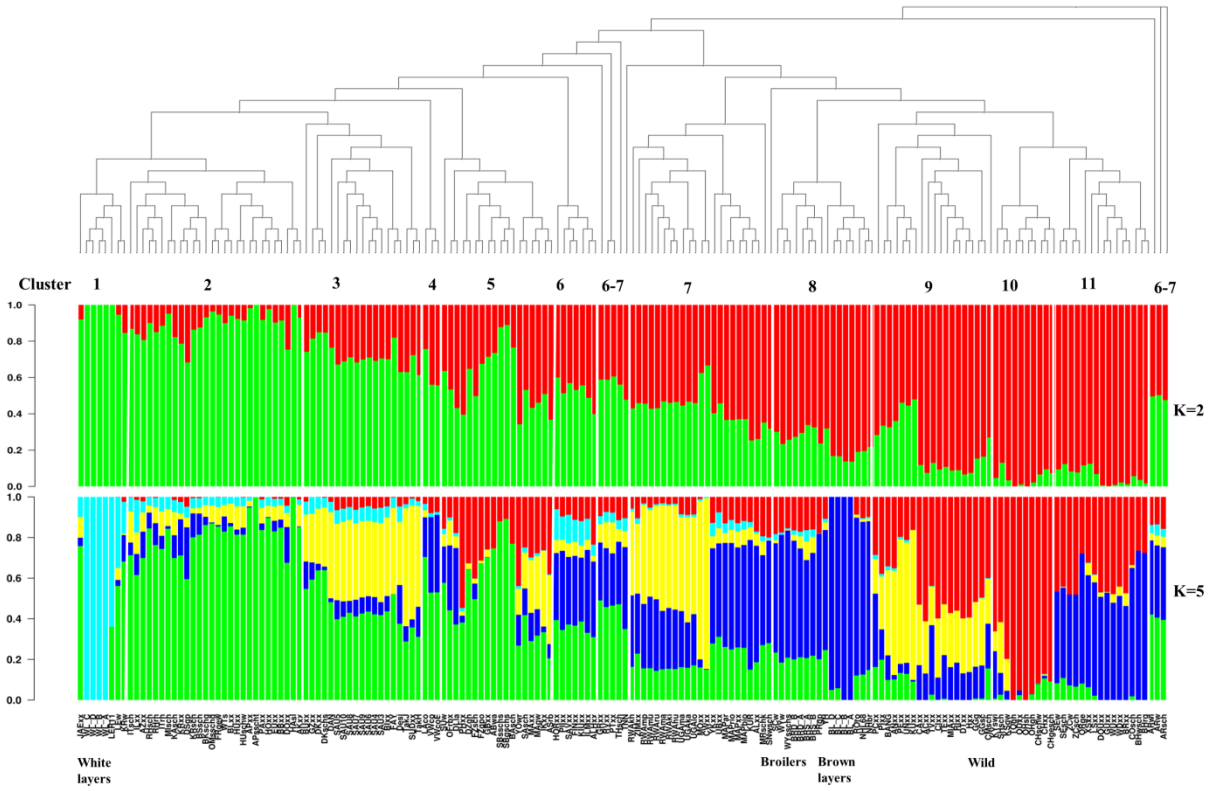


Figure 3.3: Neighbor Joining tree and admixture analysis of the 174 chicken populations. At the bottom of the NJ tree the cluster numbers are given. Different clusters are separated by white vertical lines in the admixture plots. On the right side of the plots, the assumed numbers of ancestors (K values) used in the admixture analysis are given.

At K = 5, the white layers displayed their own homogenous cluster (light blue) which is not shared among many breeds. Thereby, the White Leghorn line (LER11) was more affiliated to this white layers' cluster, while the other populations which clustered with the commercial white layer populations in NJ cluster 1 (Figure 3.1) were admixed, with more contribution from the European gene pool (green). Two of the brown layer lines (BL_A and BL_B), the two purebred lines based on Rhode Island Red, were also homogeneous (in the blue gene pool) while the other two brown

layer lines showed very little admixture with the European gene pool (green). The blue gene pool also dominated in the broiler lines, the South American (Mapuche) and Chinese breeds (NJ cluster 11). It should be noted that the breeds which showed high affiliation to this blue gene pool were located on the upper left box of the PCA plot (see Figure 3.2), which is another illustration of their relationship. The yellow gene pool is shared among all the NJ clusters that contained African breeds (clusters 3, 7 and 9). In those NJ clusters one also finds the Middle Eastern populations, a few European and Asian breeds including the two wild populations which also have reasonable affiliation to this gene pool. The Asian breeds in NJ cluster 10 were very little admixed. They were all sampled in Germany and probably have been kept in small flocks with inbreeding taking place. Among these breeds (NJ cluster 10) the Ohiki and Totenko (OHsh and TOgh) were completely homogeneous.

Overall, the populations with Asian background had a higher degree of admixture (with exception of those sampled in Germany) than those of European background which constitute a large proportion of the European (green) cluster. This suggests a higher diversity within the Asian breeds than within the European breeds.

Genetic diversity within populations

In Figure 3.4 we show the proportion of polymorphic loci and mean observed heterozygosity of populations within each category. The proportion of polymorphic loci (p), observed (H_o) and expected (H_e) heterozygosity for each population are shown in Table S3.1.

The proportion of polymorphic loci was lowest in commercial white layers with $\bar{p}$ (average p within category) = 0.394 (Figure 3.4A), followed by the European bantam breeds sampled in Germany (DE_Europe_Ban) with $\bar{p}$ = 0.511 and commercial brown layers with $\bar{p}$ = 0.570. However, the proportion of polymorphic loci varied considerably within the European bantams.

Among the commercial lines, the broilers had the highest degree of SNP polymorphism with $\bar{p}$ = 0.794. However, one broiler line (BRD_B) showed a rather low polymorphism, p = 0.685, compared to the other three broiler lines with p > 0.800. The European breeds sampled in Germany had on average a lower proportion of polymorphic SNP loci ($\bar{p} = 0.511$) than those from other parts of Europe ($\bar{p} = 0.724$), which are labelled as "Europe_local". Within these two European categories, there were three extreme outliers with a very low average proportion (p < 0.410) of polymorphic loci, i.e. the Leghorn line (LER11), and the Hamburger silver spangled (HAsl) and the Jaerhoens (JAExx) breeds.

Asian breeds sampled in Germany also had lower proportions of polymorphic loci ($\bar{p}$ = 0.662 and 0.679 for DE_Asia_Ban and DE_Asia, respectively) than those sampled in Asia ($\bar{p} = 0.863$). Among the Asian bantams sampled in Germany, the Ohiki silver Duckwing (OHsh) breed had an extremely low mean proportion of polymophic loci ($p = 0.483$, Table S3.1) while the remaining populations of this group displayed average values above 0.600. The Totenko black breasted red (TOgh) and Koeyoshi (KYswi), both breeds of Japanese origin, were outliers among the Asian breeds sampled in Germany with a very low proportion of polymorphic loci (p = 0.438 and 0.453 respectively, Table S3.1). They formed a homogeneous cluster in the admixture analysis plot (part of NJ cluster 10, Figure 3.3) which may be due to reduced diversity. The breeds sampled in Asia (Asian_local), the South American and the African breeds showed high variability of SNPs, which was even higher than that of the wild populations on average. The wild populations had a $\bar{p}$ of 0.849 while the South Americans and the Africans had 0.923 and 0.912, respectively.

The $\bar{H}_o$ over all populations combined was 0.232. Similar to the variation in SNP polymorphism, the level of heterozygosity was very low and deviated more from the overall mean (≤ 0.150) in white layers, some European breeds and European bantam breeds sampled in Germany.

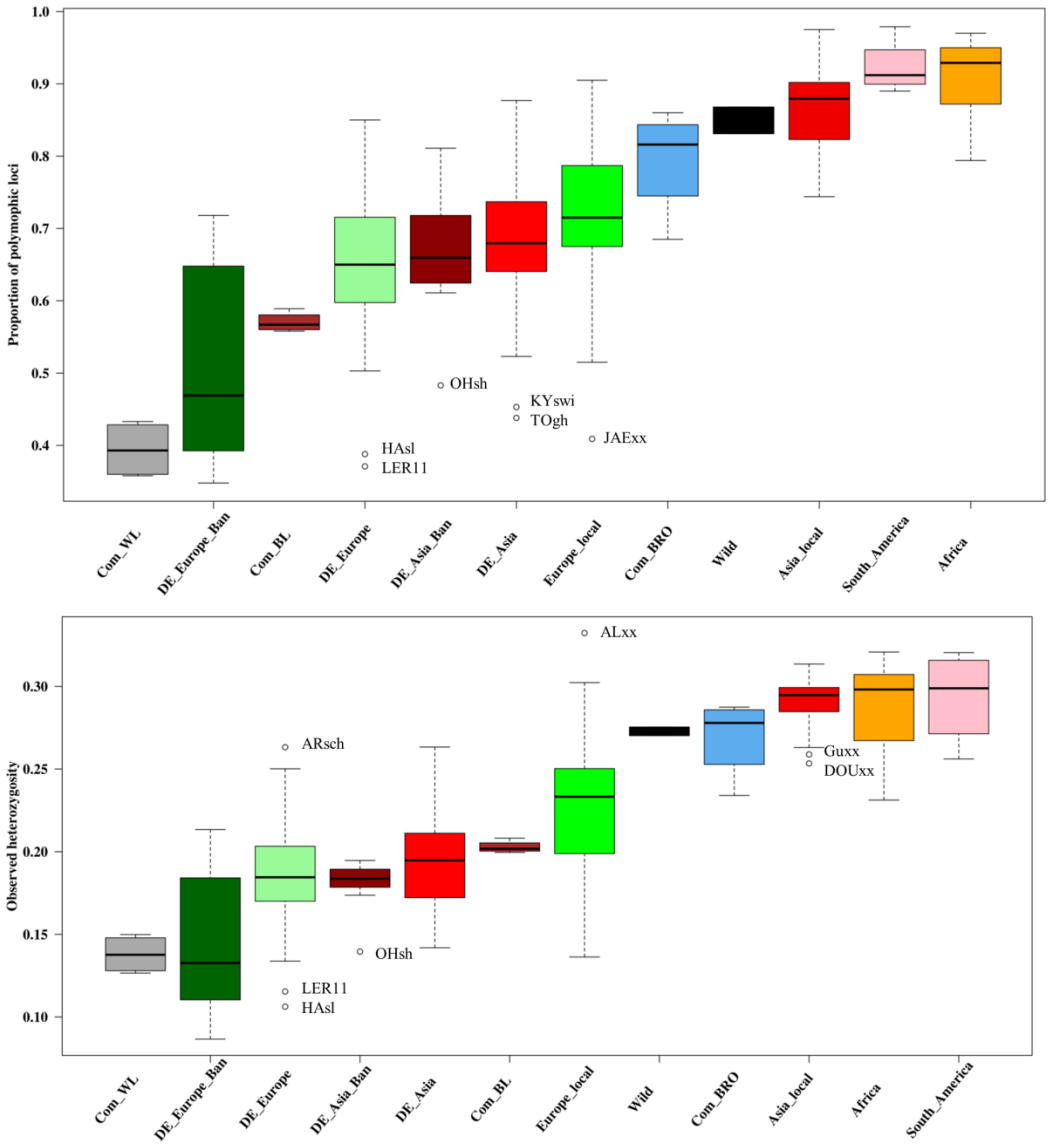


Figure 3.4: Proportion of polymorphic loci (A) and observed heterozygosity (B) within the populations grouped by chicken category. ALxx - Albanian Crowers, ARsch - Rumpless Araucana black, DOxx - Dou (Henan game), GUxx - Gushi chicken, HAsl - Hamburgh silver spangled, JAExx - Jaerhoens, KYswi - Koeyoshi Longcrower, LER11 - White Leghorn,OHsh - Ohiki bantam, silver duckwing, TOgh - Toutenko black breasted red.

The white layers had an $\overline{H}_o$ of 0.138. The commercial brown layers had moderate levels of heterozygosity (ranging from 0.200 to 0.208), while the broilers were the most heterozygous among the commercial lines with estimates ranging from 0.234 to 0.287.

The European breeds which were sampled in various parts of Europe other than Germany (Europe_local) had a higher proportion of heterozygous SNPs (with $\overline{H}_o$ = 0.228, which is very close to the overall mean heterozygosity of all the studied populations) than the European breeds sampled in Germany (with $\overline{H}_o$ = 0.185). Two of the European bantams, the gold and silver Sebright (SBgschs and SBsschs), had the lowest level of heterozygosity among all the breeds. The Sebrights are reported to be highly inbred with small population sizes according to the Central Documentation on Animal Genetic Resources in Germany [36], which goes along with the high degree of homozygosity found.

The Asian breeds sampled in Germany exhibited lower heterozygosity ($\overline{H}_o$ = 0.196) than those sampled in Asia ($\overline{H}_o$ = 0.289). The lowest proportion of heterozygous SNPs among the Asian populations was observed for Ohiki silver Duckwing (bantam), Totenko black breasted red and Koeyoshi (which were sampled in Germany), which are also low in the proportion of polymorphic SNPs. Both the African and the Mapuche populations from South America had very high levels of heterozygosity, with an $\overline{H}_o$ of 0.288 and 0.294, respectively, while for the two wild populations (GGsc and GGg) the proportion of heterozygous SNPs was slightly lower ($\overline{H}_o$ = 0.273).

Discussion

The SYNBREED chicken diversity panel encompasses a global set of chicken breeds. This extensive collection of genetic variability, combined with a high-resolution characterisation of the genome allows deep insights into the diversity within the species, and makes the panel a valuable

resource for research. In this study, we focused our analyses on the assessment of genetic relationships between populations to evaluate the distribution of diversity at a global scale, as far as this is represented by the present collection. In addition, we studied the degree of variability within population and compared it between the various categories of breeds. We compared the results of various analyses of the diversity spectrum with our expectations, which were based on sampling sites, historical records, known results from earlier studies and personal knowledge of the breeds' history.

Genetic clustering of populations

The various approaches used to assess genetic relationship between chicken populations of the spectrum consistently identified a gradual separation of genomic diversity from Asian to European breeds, with populations from Africa and South America located towards the center. This becomes evident in the Admixture analysis results, in particular at a resolution level of K = 2 clusters, as well as in the plot of the first two PCs, but also in assessing the origin of cluster members in the phylogenetic tree. The majority of Asian breeds sampled either in Asia (China, Vietnam, Pakistan, Bangladesh, Southeast Asia) or sampled in Germany from fancy breeders grouped together in the NJ tree (clusters 9, 10, and 11) as well as in the PCA plot, but separated from the majority of the European breeds which segregate in NJ clusters 1 and 2. The wild populations fitted well into the Asian cluster. They display high levels of genetic diversity as shown by the levels of heterozygosity, SNP polymorphism and their high admixture. This finding is in agreement with the widely accepted opinion that chickens were first domesticated in Asia, predominantly from the Red Jungle Fowl (*Gallus gallus*), with some contribution from *Gallus sonneratii* in Southwest India [37] and probably *Gallus lafayettii* in Sri Lanka (reviewed by [3, 10]), and then spread to various continents. Another general observation confirming earlier studies based upon

microsatellites is that commercial white layer and brown layer breeds clustered separately at opposite ends of the diversity spectrum [25, 38]. Together with broiler lines they cover only a limited part of the spectrum and a wide diversity exists complementary to the commercial lines.

Chicken of Asian origin

The Asian breeds covered a huge spectrum of genetic diversity. Despite some of the breeds being sampled in Germany (see 'Methods' section and Table S3.1), they blended very well on the PCA plot, NJ tree and admixture analysis with those sampled in Asia. This was also observed in a previous study based on microsatellite markers [19]. It shows that the breeds of Asian background that are kept by fancy breeders in Germany, even though some of them have been kept for over 150 years, they still belong to the Asian gene pool. They are mostly kept as purebreds to maintain their specific phenotypic features which are of interest to fancy breeders. For example, the Cemani (CMsch) breed which was sampled in Germany has its roots from Indonesia. A typical phenotypic trait of the breed is dermal hyperpigmentation (fibromelanosis), a mutation which makes the chicken entirely black [39]. The Indonesian local type of this breed is closely related to the Green and Red Jungle Fowls due to continuous interbreeding of the breed with the Jungle Fowls and other domestic chickens [40]. Likewise, in our study it is clustered closely to the Red Jungle Fowls in NJ cluster 9 so they didn't lose such relatedness. On the other note, the fanciers chose to keep the Asian ornamental breeds for their miniature features (i.e. Ohiki, Chabo and Ko Shamo) and long crowing and/or fighting features (i.e. Totenkou, Koeyoshi, Shamo and Onaga dori) and their ornamental long tail traits as well [46]. So these breeds remained closely related to the local Asian breeds. Another notable observation is that in cluster 3 of the NJ tree, some of the Asian breeds sampled in the Middle East clustered with African and European breeds. This is also supported by the PCA plot as well as the admixture analysis plot. The close relationship of these breeds could

be supported by their geographic distribution, though it is not clear whether this resulted from migration of chickens from Asia to Africa along the Indian Ocean, and from Europe and the Arabian Peninsula via the Mediterranean and the Red Sea [46, 47], or from a continuous exchange of the Mediterranean region with that part of Asia.

Chickens of European origin

The European breeds sampled in Germany clustered very well with the rest of the European breeds. Consistent with that, breeds of European origin are represented close together in the PCA plot, distinct from the breeds of Asian origin (PC1). The second PC distinguishes bantam breeds from large chicken breeds in the European gene pool. The Iron Age is assumed to be the main period for dispersion of chickens through Europe. Our results suggest that the majority of breeds categorized as typical European breeds according to the European Poultry Standards (those categorized as DE_Europe and DE_Europe_Ban) have been little or not exposed to crossbreeding with Asian breeds as they do not overlap with Asian breeds. However, there are some exceptions for local breeds sampled in Europe. In the PCA plot (Figure 3.2), a few breeds, mainly from Eastern Europe (Russia, Hungary, Albania, and Ukraine), but also from Switzerland and Finland are found away from other European breeds and clustered more towards the Asian breeds. As mentioned above, one of the routes for chickens from Asia to Europe was through Russia and Eastern Europe. Given the history of separation between the East and the West of Europe, the affiliation of the Eastern European breeds (found in clusters 6-8) to the Asian breeds might suggest that they have been bred rather isolated from other European (Western and Northwestern) breeds, and therefore have not yet lost their relatedness to breeds of their origin even after being in Europe for a long time. Subsequently breeds such as the Finnish lines (in cluster 6), Hungarian and Polish Green legged Partridge (GRxx) chickens have been kept in conservation flocks after the

reunification of the East and the West [38, 43, 44]. Finland has been part of the East under the Russian Empire until 1917. Further information on these Finnish, Hungarian and Polish chickens can be found in Additonal file 2. Alternatively, some of the European breeds clustering in the neigbourhood of Asian breeds might have been exposed to crossbreeding with Asian type breeds as is documented for the Swiss chicken (Schweizer Huhn) (http://www.fao.org/dad-is/browse-by-country-and-species/en/). Indeed the Swiss chicken, Transylvanian Naked Neck and Hungarian Yellow do show slightly higher levels of observed heterozygosity than expected which may suggest possible crossbreeding.

Chickens of African and South American origins

Chickens in Africa and South America originated from both Asian and European chickens [6, 9]. None of the African and South American populations appeared at any extreme points, neither in the NJ tree nor in the PCA plot but were in the middle either slightly towards an Asian or a European affiliation. However, South American populations were underrepresented in this panel which is not representing a complete picture of South American chicken diversity, while African populations were better represented and therefore can potentially cover a reasonable spectrum of the African diversity.

African. The split of African breeds between both Asian and European clusters supports the reports on their origin from both an Asian and a European origin [8]. Mitochondrial DNA studies have also shown that the common haplogroup in the African chickens is shared with some Asian and European chickens [3, 55] while other haplogroups observed in Africa (less common and possibly of more recent arrival) included those also observed in commercial layers and broilers as well as in Northwest Europe [21]. Consistant with that, some of the African breeds are clustered not far from the commercial broilers in the PCA plot, while on the admixture plot (K=5) they have a good

share of the gene pool (blue) which segregate in the commercial brown layers, broilers and Chinese populations. We believe this relationship is possibly due to the fact that they share some similar ancestries tracing back from Asia. The studied African populations were sampled in the North, East and South of Africa. The North (from Egypt and Sudan) and the East (from Ethiopia, Horn of Africa) African breeds were grouped closely together with the Saudi Arabian breeds and share a high proportion of the European gene pool (Figure 3.3). The relationship between these breeds is explained above. The breeds from Uganda, Rwanda, Tanzania (partly), and Zimbabwean ecotypes were clustered together. The distribution of the African breeds suggests that there might be some gene flow between them as they were sampled in geographically close countries. The splitting of the Tanzanian breeds into two groups (clusters 7 and 9) supports the two maternal origins reported previously [45]. The Ugandan chickens were clustered together in the same sub-cluster of NJ cluster 7. The Kuroiler breed, however, also sampled in Uganda, did not cluster very close with the other Ugandan breeds. It is reported that Kuroiler chickens were derived by crossbreeding either colored broiler males with Rhode Island Red females, or White Leghorn males with female Rhode Island Reds [46]. They have recently been brought from India to Uganda through a project which aimed at improving sustainability and productivity (meat and eggs) of chickens in Africa. In this line, African populations also shared very little of the genome with Kuroilers as displayed in the admixture plot (Figure 3.3; the yellow part prevailing in African populations is almost missing in Kuroilers). Instead, there was a high degree of overlap of the Kuroiler genome with breeds in cluster 8 where brown layers and broilers dominated, as well as a shared ancestry with cluster 11 of Chinese breeds.

South American. South American breeds were exclusively represented by the Mapuche chickens in this study. These populations showed a good share of affiliation with the Chinese populations

at K=5 (Figure 3.3) of the admixture analysis, but also with some membership into the European (green) lineage. In the PCA plot they seemed slightly more related to Asian breeds (Figure 3.2), while they can be found in a rather central position between European and Asian breeds in the NJ tree (Cluster 7). Even though a previous study of the eggshell coloration in Chilean breeds suggested a possible Chinese origin [47], Wang et al. [48] and Wragg et al. [49] later reported that the blue egg shell trait in the Chilean and Chinese breeds has a different genetic basis in the two origins. It was reported that many phenotypic features of the Mapuche chickens resemble those of breeds of Asian origin rather than of European origin [50], but these populations showed a level of admixture with both the Asian and European gene pool in our analysis. Our current results do not really solve the debate regarding the origin of the Mapuche chickens. Another point of interest is that the Mapuche did not cluster with the European Araucanas, but still the admixture plot shows a lot of overlap between them. In fact, all the gene pools segregating in Mapuche chicken also segregate in the Araucanas for all the K values; the only difference is the proportion of affiliations to the gene pools. For example, the Araucanas at K= 9 - 11show higher proportions of the lineages segregating in European populations (green and gray) than the Mapuche. Therefore, it is possible that the European Araucanas might have been mixed with the European breeds, or some of their genomes are getting fixed rapidly as they do show lower levels of observed heterozygosity than expected and their genetic diversity is highly reduced compared to that of the Mapuche.

The distinction of within breed diversity between local and fancy breeds

The highest genetic diversity was observed within populations sampled in Africa, South America and Asia, some of which exhibited even higher diversity than the wild populations. Generally, the local type breeds from the four continents exhibit more genetic diversity than those from German fancy breeders and commercial layer lines (Figure 3.4). The local chickens are often kept by

villagers under extensive management systems and without controlled breeding programs, but in some cases they are also kept in conservation facilities with the purpose to preserve their genetic architecture [20, 44, 51, 52]. Therefore, the high genetic diversity persists due to the fact that the pool for mating individual is generally larger, and hence a lower rate of inbreeding, and there is some exchange of genetic material by intercrossing of breeds, and little artificial selection is practiced [21, 40, 53]. Although the chicken breeds kept by German fancy breeders cover a wide spectrum of diversity overlapping with local breeds, the management followed by the fancy breeders only preserved the genetic relatedness of these breeds to their ancestral genetic background which, however, caused a drastic reduction in the level of genetic diversity within the breeds. This is likely due to several reasons: Firslty, for the fancy breeds that were imported from Asia to Europe, the number of animals was limited. Therefore, this founder effect contributed to a reduced level of diversity compared to the original populations. Secondly, fancy breeds in Germany (of both Asian and European background) are generally kept in small flock sizes with little or no exchange of mating stocks between breeders. Taking the example of Cemani breed again, the local Cemani breed in Indonesia has shown a similar level of nucleotide diversity as the Jungle Fowls [40]. However, in our study, the Cemani from fancy breeders, even though it was brought only recently to Germany, already has reduced genetic diversity compared to the *Gallus gallus* species and also had the lowest genetic diversity among the local Asian populations of its respective NJ cluster 9. This is probably the result of a limited number of breeding individuals and the absence of a continuous gene flow from other breeds while trying to keep the breed pure (as it is for many fancy breeds in Germany) for its interesting fibromelanosis trait. Another examples are the Leghorn line (LER11) and Hamburger breeds which have the lowest genetic diversity within the European categories. The LER11 is a White Leghorn line which has been kept as a

closed population at least since 1965 when it came to the Institute for Small Animal Breeding in Celle (Germany) and was most likely based on a narrow gene pool (as other commercial White Leghorn lines). The Hamburger breed is a fancy breed with a small effective population size [54]. Additionally, the selection practices to meet the European breed standards may also have had a huge impact on the reduction of genetic diversity within the fancy breeds. These standards are very strict and breeders aim at an almost "perfect" phenotype. To achieve this, they practice even matings of very close relatives. This is very evident as almost all the Asian and most of the European breeds, which were sampled from fancy breeders, exhibited lower observed heterozygosity within the population than expected (Table S3.1). A systematic management of diversity in small populations is almost completely missing, and hence all these breeds and color variants display a low level of within breed diversity.

Genetic diversity of the commercial lines

With respect to commercial purebred lines, white layers are clearly distinct from brown egg layers, while broiler lines cluster more in the center (Figure 3.2) and closer to Asian than to European breeds. This, in turn, fits well with the history of these chicken lines. Commercial white egg layer lines originated from an Italian breed located in Livorno (Tuscany, central Italy), the single comb White Leghorn. Consequently, they clustered with other European breeds, especially the fancy White Leghorn lines (LER11 and LEW). The genetic basis of brown egg layers is broader than that of white layers, utilizing Rhode Island Red, Plymouth Rock, Australorps, and New Hampshire among others and the broilers were mainly based on Cornish (Indian Game) and Plymouth Rock [12]. The latter might also explain the closer relationship of broilers to brown layers than to white layers as they share some parental background in the White Plymoth Rock. There is a reported loss of ancestral genetic diversity by 50% in commercial lines [17]. In our study, the single-parented

white layers have shown much reduced genetic diversity and displayed a homogeneous cluster in the admixture analysis for all K-values that were analyzed. The brown layers had low to moderate genetic diversity. Compared to the layer lines, commercial broilers were more diverse, which is almost at the same level as that of wild populations. This might be related to a broader genetic basis of founder populations and a larger effective population sizes in selection programs.

Measures are needed to preserve genetic diversity, between and within breeds. There is a large spectrum of between breed diversity preserved in the local breeds from different origins and the fancy breeds from Germany. Additionally, the local chickens have proven to be great reservoirs of within breed genetic diversity. However, the low genetic diversity within the fancy breeds, and the non-structured, non-monitored breeding programs of local breeds raised concerns about their vulnerability to go extinct [20, 44, 55]. So measures for preserving and maintaining genetic diversity should include new utilization possibilities of local breeds, but from a genetic point, such breeds should be included in conservation programs. These programs will include both cryopreservation in gene banks and in-situ conservation flocks managed properly to minimize the rate of inbreeding. Flocks should be kept in high numbers to avoid non-random mating and vulnerability to genetic drift effects otherwise smaller number of birds in conversation flocks would results in reduction of genetic diversity over time as it has been observed in some of the already established facilities e.g. [20, 52].

Conclusions

In this study we assessed genetic diversity between and within breeds from chickens collected across the world, from various backgrounds. It is very evident that the origin, geographic expansion, selection and different management practices have had a major impact on the global pattern of chicken diversity. Overall, the commercial white layers had the lowest variation among

the commercial lines, the bantams displayed lower genetic diversity than the normal sized breeds in the respective category of origin (e.g. Asian bantams vs Asian locals or Asians sampled in Germany); and breeds that were sampled in Germany (both European and Asian breeds) had lower genetic diversity than those sampled in various places in the respective continent of origin. At the current state, the commercial breeding lines seem to have not yet reached selection plateau in the current breeding programs and are still responding to the breeders' objectives. However, the limited diversity they cover (as shown by PCA and NJ tree), and the very low within breed diversity, in particular within the white layer purebred lines, might limit the flexibility to respond to unforeseen future needs. There is still more genetic diversity within the less selected African, South American and some local Asian and European breeds. Therefore, it is required that genetic diversity in these chickens be maintained in order to have the opportunity to respond to future challenges.

As conservation measures are costly, it was stressed by [56] that "conservation decisions must be based on a global inventory of the species diversity". The data of SCDP can be seen as a step towards establishing such a reference collection in chicken. In this way, it is supporting international initiatives as at the European level with the EU project IMAGE (http://www.imageh2020.eu/index.php), with collaborative effort to characterize and manage genetic diversity in livestock and poultry species. Not only is this panel the biggest gene pool of chicken data by far; it also has the potential to expand as new breeds and other sources of genetic materials will be added from other parts of the world. The SCDP data set presents ample opportunities for exploitation for further chicken molecular genetic studies and is made available for public access (see section "Availability of data and materials" for details).

Additional files

Additional files are available online:

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5727-9#Sec24

The file names are as follows:

Additional file 3.1: Table S3.1 is named **Additional file 1: Table S1.** Population information, population diversity measures and origin of DNA samples.

Additional file 3.2: Document S3.1 is named **Additional file 2: Document S1.** Description of clusters.

Additional file 3.3: Figure S3.1 is named **Additional file 3: Figure S1.** Admixture Analysis.

List of abbreviations

B.C.: Before Christ, BDRG: Bund Deutscher Rassgeflügelzüchter, CV: Cross Validation, DNA: Deoxyribonucleic acid, EDTA: Ethylenediamine Tetraacetic acid, EU: European Union, GEH: The Society for the Conservation of Old and Endangered Livestock Breeds, H_e: Expected heterozygosity, H_o: Observed heterozygosity, LD: Linkage disequilibrium, mtDNA: Mitochondrial DNA, NJ: Neighbor joining, PC: Principal component, PCA: Principal components analysis, SCDP: SYNBREED chicken diversity panel, SNP: Single nucleotide polymorphism, SVS: SNP Variation Suite, SYNBREED: Synergistic Plant and Animal Breeding.

Ethics approval and consent to participate

We confirm that the collection of blood samples for this study was performed 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).

Availability of Data and Materials

The dataset used in the current study is deposited in the Figshare repository and can be accessed through this link: 10.6084/m9.figshare.8003909.

Funding

The "SYNBREED - Synergistic Plant and Animal Breeding" project was funded by the German Federal Ministry of Education and Research (FKZ 0315528E) and we are grateful for the financial support. This work is part of DKM's Doctoral programme which is supported financially by the Erasmus Mundus programme of the European Union (through the INSPIRE project) and by the IMAGE project. The funding bodies were not involved in the design of the study, collection, analysis as well as the interpretation of data and writing of the manuscript.

Authors' contributions

DKM, SW and HS were involved in designing the study. DKM and SW analyzed the data and drafted the manuscript. SW was involved in sampling and coordinating data collection. HS, CR and AOS contributed to the critical revisions and edits of the manuscript. AW contributed to the provision, preparation and editing of the data. All authors read and approved the manuscript.

Acknowledgements

The members of the SCDP (Box 1) as well as the breeders of the "Bund Deutscher Rassegeflügelzüchter e.V." and the "Gesellschaft zur Erhaltung alter und gefährdeter Haustierrassen e.V" in Germany significantly contributed to this study by providing samples or SNP data, or gave access to their animals for sampling. These contributions formed the backbone of this study, and we are very grateful to all the participating colleagues and breeders. We thank Stephan M. Funk, Nature Heritage, and Janet Fulton, Hyline International, for the constructive

comments on the manuscript. We also thank the Technische Universität München (Prof. Fries) for genotyping the samples. Sampling and genome sequencing of the Saudi chickens was supported through a grant (12-AGR2555-02) from the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia.

References

[1] Kitalyi AJ. Village chicken production systems in rural Africa: household food security and gender issues. Rome Italy: Food and Agriculture Organization of the United Nations; 1998.

[2] Edea Z, Bhuiyan MSA, Dessie T, Rothschild MF, Dadi H, Kim KS. Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. Animal. 2015; 9: 218–26.

[3] Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, et al. Genetic diversity in farm animals - A review. Anim Genet. 2010; 41 SUPPL. 1: 6–31.

[4] FAO. Global plan of action for animal genetic resources and the Interlaken Declaration. In: International Technical Conference on Animal Genetic Resources for Food and Agriculture. Rome, Italy; 2007.

[5] Tixier-Boichard M, Leenstra F, Flock DK, Hocking PM, Weigend S. A century of poultry genetics. Worlds Poult Sci J. 2012; 68: 307–21.

[6] Crawford RD. Origin and history of poultry species. In: Crawford RD, editor. Poultry Breeding and Genetics. Amsterdam-Oxford-Newyork-Tokyo: Elsevier; 1990. p. 1–42.

[7] Liu YP, Wu GS, Yao YG, Miao YW, Luikart G, Baig M, et al. Multiple maternal origins of chickens: Out of the Asian jungles. Mol Phylogenet Evol. 2006; 38: 12–9.

[8] Mwacharo JM, Bjørnstad G, Mobegi V, Nomura K, Hanada H, Amano T, et al. Mitochondrial DNA reveals multiple introductions of domestic chicken in East Africa. Mol Phylogenet Evol. 2011; 58: 374–82.

[9] Storey AA, Ramírez JM, Quiroz D, Burley DV, Addison DJ, Walter R, et al. Radiocarbon and DNA evidence for a pre- Columbian introduction of Polynesian chickens to Chile Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. Proc Natl Acad Sci. 2007; 104: 10335–10339.

[10] Tixier-Boichard M, Bed'Hom B, Rognon X. Chicken domestication: From archeology to genomics. C R Biol. 2011; 334: 197–204.

[11] West B, Zhou BX. Did chickens go North? New evidence for domestication. J Archaeol Sci. 1988; 15: 515–33.

[12] Crawford RD. Poultry genetic resources: evolution, diversity, and conservation. In: Poultry Breeding and Genetics. Amsterdam-Oxford-Newyork-Tokyo: Elsevier; 1990. p. 43–60.

[13] Gongora J, Rawlence NJ, Mobegi VA, Jianlin H, Alcalde JA, Matus JT, et al. Indo-European and Asian origins for Chilean and Pacific chickens revealed by mtDNA. Proc Natl Acad Sci. 2008; 105: 10308–13.

[14] Mwacharo JM, Bjørnstad G, Han JL, Hanotte O. The History of African Village Chickens : an Archaeological and Molecular Perspective. African Archaeol Rev. 2013; 30: 97–114.

[15] Gongora J, Rawlence NJ, Mobegi VA, Jianlin H, Alcalde JA, Matus JT, et al. Reply to Storey et al.: More DNA and dating studies needed for ancient El Arenal-1 chickens. Proc Natl Acad Sci. 2008; 105: E100–E100.

[16] Storey AA, Quiroz D, Ramirez JM, Beavan-Athfield N, Addison DJ, Walter R, et al. Pre-Columbian chickens, dates, isotopes, and mtDNA. Proc Natl Acad Sci. 2008; 105: E99–E99.

[17] Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen MM, Crooijmans RPM, et al. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc Natl Acad Sci. 2008; 105: 17312–7.

[18] Rassegeflügel-Standard für Europa in Farbe. Bund Deutscher Rassegeflügelzüchter (ed.), Howa Druck & Satz GmbH, Fürth. ISBN 987-3-9806597-1-0.

[19] Wimmers K, Ponsuksili S, Hardge T, Mathur PK, Horst P. Genetic distinctness of African, Asian and South American local chickens. Anim Genet. 2000; 31: 159–65.

[20] Qu L, Li X, Xu G, Chen K, Yang H, Longchao Z, et al. Evaluation of genetic diversity in Chinese indigenous chicken breeds using microsatellite markers. Sci China Ser C Life Sci. 2006; 49: 332–41.

[21] Muchadeyi FC, Eding H, Wollny CBA, Groeneveld E, Makuza SM, Shamseldin R. Absence of population substructuring in Zimbabwe chicken ecotypes inferred using microsatellite analysis. Anim Genet. 2007; 38: 332–9.

[22] Adebambo AO, Mobegi VA, Mwacharo JM, Oladejo BM, Adewale RA, Iloro LO, et al. Lack of Phylogeographic Structure in Nigerian Village Chickens Revealed by Mitochondrial DNA D-loop Sequence Analysis Lack of Phylogeographic Structure in Nigerian Village Chickens Revealed by Mitochondrial DNA D-loop Sequence Analysis. Int J Poult Sci. 2010; 9: 503–7.

[23] Muchadeyi FC, Eding H, Simianer H, Wollny CBA, Groeneveld E, Weigend S. Mitochondrial DNA D-loop sequences suggest a Southeast Asian and Indian origin of Zimbabwean village

chickens. Anim Genet. 2008; 39: 615–22.

[24] Leroy G, Kayang BB, Youssao IAK, Yapi-Gnaoré CV, Osei-Amponsah R, Loukou NE, et al. Gene diversity, agroecological structure and introgression patterns among village chicken populations across North, West and Central Africa. BMC Genet. 2012; 13: 34.

[25] Lyimo CM, Weigend A, Msoffe PL, Eding H, Simianer H, Weigend S. Global diversity and genetic contributions of chicken populations from African, Asian and European regions. Anim Genet. 2014; 45: 836–48.

[26] Hillel J, Groenen MAM, Tixier-Boichard M, Korol AB, David L, Kirzhner VM, et al. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genet Sel Evol. 2003; 35: 533–57.

[27] Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013; 14: 59.

[28] Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 Genes, Genomes, Genet. 2017; 7: 109–17.

[29] SNP & Variation Suite ™ (Version 8.1). Bozeman, MT: Golden Helix, Inc. http://goldenhelix.com/.

[30] Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics. 2018; 19: 22.

[31] Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase

inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2008; 84: 210–23.

[32] Reynolds J, Weir BS, Cockerham CC. Estimation o f the coancestry coefficient: Basis for a short-term genetic distance. Genetics. 1983; 105: 767–79.

[33] Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 2006; 23: 254–67.

[34] Alexander DH, Novembre J, Lange K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. Genome Res. 2009; 19: 1655–64.

[35] Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th edition. Essex, UK: Longmans Green, Harlow; 1996.

[36] TGRDEU. Zentrale Dokumentation Tiergenetischer Ressourcen in Deutschland. https://tgrdeu.genres.de/default/index/index/?lang=en. Accessed 11 May 2018.

[37] Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al. Identification of the Yellow skin gene reveals a hybrid origin of the domestic chicken. PLoS Genet. 2008; 4.

[38] Bodzsar N, Eding H, Revay T, Hidas A, Weigend S. Genetic diversity of Hungarian indigenous chicken breeds based on microsatellite markers. Anim Genet. 2009; 40: 516–23.

[39] Dorshorst B, Molin AM, Rubin CJ, Johansson AM, Strömstedt L, Pham MH, et al. A complex genomic rearrangement involving the Endothelin 3 locus causes dermal hyperpigmentation in the chicken. PLoS Genet. 2011; 7: e1002412.

[40] Dharmayanthi AB, Terai Y, Sulandari S, Zein MSA, Akiyama T, Satta Y. The origin and evolution of fibromelanosis in domesticated chickens: Genomic comparison of Indonesian Cemani and Chinese Silkie breeds. PLoS One. 2017; 12: 1–24.

[41] Macdonald KC, Edwards DN. Chickens in Africa: the importance of Qasr Ibrim. Antiquity. 1993; 67: 584–90.

[42] Masonen P. Trans-Saharan Trade and the West African Discovery of the Mediterranean World. In: The third Nordic conference on Middle Eastern Studies: Ethnic encounter and culture change. Joensuu, Finland; 1995.

[43] Siwek M, Wragg D, Sławińska A, Malek M, Hanotte O, Mwacharo JM. Insights into the genetic history of Green-legged Partridgelike fowl: MtDNA and genome-wide SNP analysis. Anim Genet. 2013; 44: 522–32.

[44] Fulton JE, Berres ME, Kantanen J, Honkatukia M. MHC-B variability within the Finnish Landrace chicken conservation program. Poult Sci. 2017; 96: 3026–30.

[45] Lyimo CM, Weigend A, Janßen-Tapken U, Msoffe PL, Simianer H, Weigend S. Assessing the genetic diversity of five Tanzanian chicken ecotypes using molecular tools. S Afr J Anim Sci. 2013; 43.

[46] Kuroiler Poultry Guide. https://www.africa-uganda-business-travel-guide.com/Kuroiler.html#BackgroundofKroiler. Accessed 6 Jul 2018.

[47] Zhao R, Xu GY, Liu ZZ, Li JY, Yang N. A study on eggshell pigmentation: biliverdin in blue-shelled chickens. Poult Sci. 2006; 85: 546–9.

[48] Wang Z, Qu L, Yao J, Yang X, Li G, Zhang Y, et al. An EAV-HP Insertion in 5′ Flanking

Region of SLCO1B3 Causes Blue Eggshell in the Chicken. PLoS Genet. 2013; 9: e1003183.

[49] Wragg D, Mwacharo JM, Alcalde JA, Wang C, Han JL, Gongora J, et al. Endogenous retrovirus EAV-HP linked to blue egg phenotype in Mapuche fowl. PLoS One. 2013; 8: e71393.

[50] Alcalde A. Ethno-ornithology and history of the Mapuche fowl. Rev Chil Ornitol. 2016; 22: 126–32.

[51] Cuc NTK, Simianer H, Eding H, Tieu HV, Cuong VC, Wollny CBA, et al. Assessing genetic diversity of Vietnamese local chicken breeds using microsatellites. Anim Genet. 2010; 41: 545–7.

[52] Mtileni BJ, Muchadeyi FC, Weigend S, Maiwashe A, Groeneveld E. A comparison of genetic diversity between South African conserved and field chicken populations using microsatellite markers. 2010; 40: 462–6.

[53] Mtileni B, Dzama K, Nephawe K, Rhode C. Estimates of effective population size and inbreeding in South African indigenous chicken populations: implications for the conservation of unique genetic resources. Trop Anim Health Prod. 2016; 48: 943–50.

[54] TGRDEU. Zentrale Dokumentation Tiergenetischer Ressourcen in Deutschland. https://tgrdeu.genres.de/default/hausundnutztiere/detailansicht/detail/63E5D466-BB56-FD58-E040-A8C0286E751D. Accessed 24 Feb 2019.

[55] Chen G, Bao W, Shu J, Ji C, Wang M, Eding H, et al. Assessment of population structure and genetic diversity of 15 Chinese indigenous chicken breeds using microsatellite markers. Asian-Australasian J Anim Sci. 2008: 21: 331-339.

[56] Simianer H. Decision making in livestock conservation. Ecol Econ. 2005; 53: 559–72.

CHAPTER 4

Genetic diversity in global chicken breeds as a function of genetic distance to the wild populations

Dorcus Kholofelo Malomane[1,2], Steffen Weigend[2,3], Armin Otto Schmitt[2,4], Annett Weigend[3], Christian Reimer[1,2], Henner Simianer[1,2]

[1]Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

[2]Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

[3]Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, Neustadt, Germany

[4]Breeding Informatics Group, Department of Animal Sciences, University of Goettingen, Germany

Submitted to Plos Genetics Journal

Abstract

Migration of populations from their founder population is expected to cause a reduction in genetic diversity and facilitates population differentiation between the populations and their founder population as predicted by the theory of genetic isolation by distance. Consistent with that, a model of expansion from a single founder predicts that patterns of genetic diversity in populations can be well explained by their geographic expansion from the founders, which is correlated to the genetic differentiation. To investigate this in the chicken, we have estimated the relationship between the genetic diversity in 172 domesticated chicken populations and their genetic distances to wild populations. We have found a strong inverse relationship whereby 87.5% of the variation in the overall genetic diversity of domesticated chicken can be explained by the genetic distance to the wild populations. We also investigated if different types of SNPs and genes present similar patterns of genetic diversity as the overall genome. Among different SNP classes, the non-synonymous ones were the most deviating from the overall genome. However, the genetic distances to wild populations still explained more variation in domesticated chicken diversity in all SNP classes ranging from 81.7 to 88.7%. The genetic diversity seemed to change at a faster rate within the chicken in genes that are associated with transmembrane transport, protein transport and protein metabolic processes, and lipid metabolic processes. In general, such genes are flexible to be manipulated according to the population needs. On the other hand, genes which the genetic diversity hardly changes despite the genetic distance to the wild populations are associated with major functions e.g. brain development. Therefore, changes in the genes may be detrimental to the chickens. These results contribute to the knowledge of different evolutionary patterns of different functional genomic regions in the chicken.

Introduction

Domesticated chickens (*Gallus gallus domesticus*) are one of the most widely distributed domestic animal species in the world. Some of the reasons are due to their portability and flexibility of transportation through human migration, stock trading, and expansion in the agricultural practices [1, 2], in addition their use for nutrition is not suffering from any religious or cultural reservations. It is commonly accepted that the world-spread chickens of today originate predominantly from domestication of the red jungle fowl (*Gallus gallus* species) in Asia (reviewed by Tixier-Boichard et al [3]). From the centers of domestication, chickens have dispersed into different parts of the world. There has been formation of new breeds or lines as populations moved outward from ancestral territories and settled in new colonies. One of the expectations from such expansion processes is the increase of genetic distances (increased differentiation) of the outward populations to the original ancestors, and the loss of genetic diversity within such populations due to genetic drift and subsequent serial founder effects [4–6]. In Malomane et al [7] we studied the overall genetic diversity between and within the chicken breeds. In the current study we aimed at investigating if the observed genetic diversity in the chicken breeds is a result of their genetic expansion from the chicken wild populations following the concepts behind the theory of genetic isolation by distance [8–10] and the model of expansion from a single location such as the 'Out of Africa' migration model [4]. The theory of genetic isolation by distance refers to the population genetic patterns whereby genetic differentiation increases with the increase in geographic distance between populations. This is because the exchange of genetic material between the populations (i.e. mating opportunities) is confined by the distance [8, 11]. Likewise, movements of individuals further apart from their founders would be expected to increase genetic differentiation. This has been established with the 'Out of Africa' theory which asserts that modern humans originate from

Africa [12] and human populations worldwide resulted in a reduction in genetic diversity with the increasing geographic distance from east Africa (Ethiopia) [4, 5, 13, 14]. Similar studies in cattle also reported a decreasing genetic diversity with increasing geographic distance to the cattle domestication center in Southwest Asia [15, 16].

The loss of genetic diversity within the migrated populations, which can be explained by the geographic distance from their founders, is believed to be a good measure of neutral genetic diversity as a consequence of genetic drift. However, the overall genetic diversity is also a result of population specific events such as mutations, natural selection to favor adaptation in the current environments and/or artificial selection (e.g. in livestock production practices) as well as population specific drift [5]. Consequences of selection are often measured by non-neutral genetic variation as it is assumed that non-neutral regions with functional fitness effects in the genome evolve differently to the neutral genome. In this study we used the global collection of chicken breeds [7] to investigate the pattern of the overall genetic diversity moving outwards the centers of chicken domestication, given all events taking place in the genome. Furthermore, we investigate if different functional regions of the genome present similar patterns as the overall genome. We hypothesized that changes in genetic diversity may be faster in some genes or functional categories depending on their functions and changes may also be different in different breeds or breed groups due to different adaptive or artificial selection targets. Therefore, the pattern of relationship between genetic diversity and genetic distance may behave differently, less complying with the overall genome and more dynamic than the non-genic regions due to differences in selection patterns in addition to other population specific events.

Studying the theory of genetic isolation by distance and/or the concept of migration from a single location with chickens poses some challenges because the physical locations do not always

represent their geographic origin (following migration from founders). For many chicken breeds the time point when they have migrated to their current locations is unknown. We also believe that geographic distances may not be the best predictor of the genetic diversity in the chicken. This is because unlike in humans where genetic evolution is mostly driven by natural circumstances, rapid migration, crossbreeding forced by man, refined breeding programs and artificial selection for desired traits have largely shaped the evolution of domesticated chickens. The changes in genetic diversity and evolutionary rates are often rapid in domesticated livestock and the genetic architecture of chickens around the same geographic location may also differ greatly depending on different breeding practices or selection targets. Therefore, in our study we used Reynolds' genetic distances [17] instead of geographic distances but following similar concepts as the genetic isolation by distance and model of expansion from a single founder [5, 8, 9]. Reynolds' distances estimate differences under the assumptions that genetic differentiation occurs by genetic drift.

Materials and Methods

Ethics statement

The data used was derived from a previous study [7], sourced from the SYNBREED (http://www.synbreed.tum.de/) project which was funded by the German Federal Ministry of Education and Research (FKZ 0315528E). Sampling of chickens followed the German Animal Welfare regulations, the authorities of Lower Saxony were notified according to §8 of the German Animal Welfare Act (33.9-42502-05-10A064) and with the written consent of the animal owners.

Data description and quality control

Data consisted of 3,235 chicken individuals from 174 chicken populations collected in Asia, Africa, South America and Europe. The populations were classified into twelve breed categories which were based on their continent of origin and/or type as described in Table S4.1. The chickens

were genotyped with the 600K Affymetrix® Axiom™ Genome-Wide Chicken Genotyping Array [18]. We used only the SNPs from the 28 autosomal chromosomes and removed 499 SNPs with ambiguous chromosome annotation. The data was filtered for an animal call rate of ≥95% and SNP call rate of ≥99% using the SNP & Variation Suite (SVS) version 8.1 [19]. We performed LD based pruning to account for ascertainment bias [20] using the PLINK software v1.9 [21, 22] with the parameters *indep 50 5 2*. After the filtering steps 156,753 SNPs were left for further analysis and imputation was performed to recover missing genotypes using Beagle 3.3 [23]. A further description of the data can be found in Malomane et al. [7].

Classification of the SNPs

We classified SNPs according to their functional consequences and assigned them to their associated genes using the Affymetrix Galgal5 annotation map [24]. SNPs were classified into the following categories: non-synonymous which is made of the missense and nonsense (only eight in total) variants, synonymous, exonic (a combination of the non-synonymous and synonymous SNPs as well as other coding and non-coding exonic SNPs which were not assigned as non-synonymous or synonymous), intronic, 5' untranslated region (5' UTR), 3' untranslated region (3' UTR), upstream, downstream and intergenic classes. SNPs assignments were prioritized in the order as they appear on Table 4.1. For example, if one SNP is associated with two genes but has different functional consequences for the two genes (e.g. non-synonymous for one gene and synonymous for the other gene) then a non-synonymous functional consequence was considered first instead of the other consequences, followed by synonymous and so forth. As for the up- and downstream variants, a SNP was assigned to the upstream class if it was located within 5 kb upstream of the gene and in analogy for the downstream SNPs. The distribution of SNPs into their functional classes is shown in column 1 of Table 4.1.

For assigning SNPs to indiviadual genes, the 156K SNPs were mapped to a total of 10,456 associated genes [24].

Estimation of genetic diversity outward from wild populations

Two subspecies of the wild populations (*Gallus gallus*), the *G. gallus spadiceus* and *G. gallus gallus*, sampled about 20 years ago were used as reference for original founders, and reflect genetic diversity in centers of domestication.

We estimated the pairwise Reynolds' genetic distances [17] between the two wild type populations (*G. gallus* ssp.) and the domesticated populations, and then calculated the mean genetic distance of each domesticated population to the two wild populations. Furthermore, observed heterozygosity was estimated within each population. Then, we estimated the linear relationship between the overall genetic diversity within the domesticated populations and their mean genetic distances to the two wild type populations. The amount of variation in genetic diversity within the populations which can be explained by the genetic distance was measured by the R^2 value. To investigate if different SNP classes and genes show similar patterns as the overall genome pattern (when using all SNPs), we also estimated the genetic diversity in the different SNP classes and in genes and subsequently estimated the linear relationship with the genetic distances to the wild populations. We used the likelihood ratio test implemented in the R lmtest package (v0.9-36) [25] which uses the χ^2 test to compare the linear regression coefficients of the overall pattern to the patterns of the different SNP classes.

For the individual genes, because some of the genes were annotated with only one or very few associated SNPs while others were annotated with more, we only considered genes with at least ten associated SNPs (resulting in 6,303 in total) for making comparisons with the overall pattern.

We evaluated the rate of change in the genetic diversity within the genes due to the change in genetic distances of populations to the wild populations using the regression coefficients of the linear relationship between the two parameters.

Functional annotation of genes

Genes within the lowest and highest 5% ranges of regression coefficients in the relationship between genetic diversity within populations and genetic distances to the wild populations were grouped into functional terms using the ClueGO (v2.5.1) [26] ontology enrichment package in Cytoscape (v3.6.1) [27]. Additionally, individual gene functions were annotated using the DAVID functional annotation tool (v6.8) [28].

Results and discussion

The relationship between the overall genetic diversity and the genetic distance to wild populations

The relationship between the observed heterozygosity within domestic chicken (*Gallus gallus domesticus*) populations and the genetic distance to the wild populations (*Gallus gallus*) is shown in Figure 4.1. The different breed categories as described in Table S4.1 are represented by symbols of different colours and shapes. There is a strong inverse relationship between the genetic diversity within populations and their genetic distances to the wild populations. This relationship is similar even when using just neutral markers (intergenic SNPs, Figure 4.2). Across these chicken populations, 87.5% (Table 4.1) of the total variation in the heterozygosity can be explained by the genetic distance to the wild populations. This figure is slightly higher than those obtained in several human studies when using geographic distances. Geographic distances of humans out of Africa explained 76.3% of microsatellite heterozygosity and 78.4% of fixation index F_{ST} variation in [5] and 85% of microsatellite heterozygosity in [14]. They had a correlation of -0.910 with SNP

haplotype heterozygosity and -0.870 with microsatellite heterozygosity in the same study in [29]. Furthermore, studies in humans have shown that there is a high correlation (e.g. 0.765 to 0.885 [5]) between the genetic distances (using different genetic distance measures) and geographic distance. However the correlations were not as high in domesticated cattle studies compared to humans. For example, a correlation of 0.624 was reported by [30] and while [15] reported a correlation of 0.750 for ancient cattle samples, the correlation was 0.540 in modern cattle samples. The weakening relationship between geographic and genetic distances in modern domesticated cattle was suggested to be due to the human manipulation of genetic diversity among other reasons, as it is with many domesticated livestock [15].

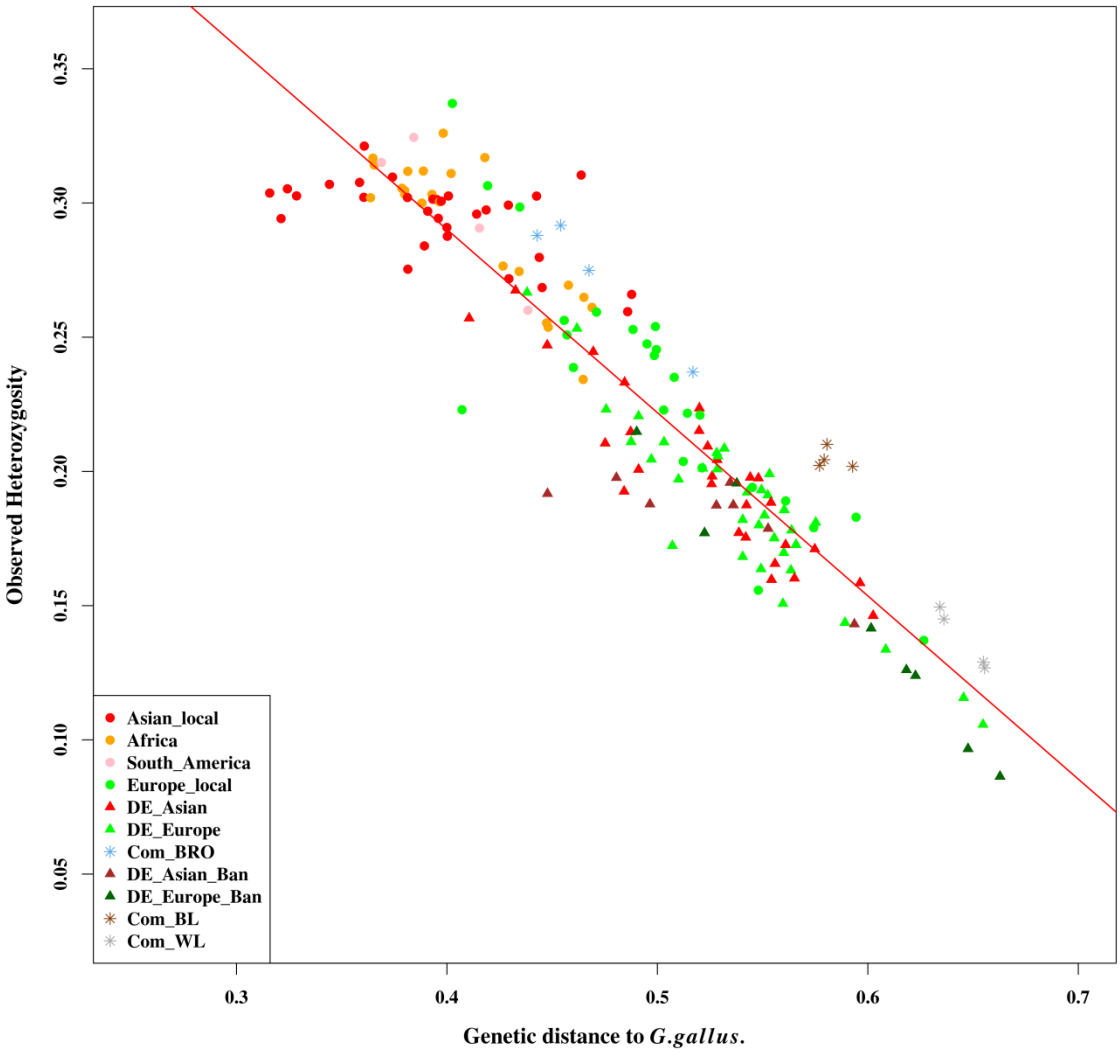


Figure 4.1: The relationship between the overall genetic diversity within populations and their genetic distance to *Gallus gallus*. The full names of the categories and description can be found in Table S4.1. The fitted regression line to the data with the equation heterozygosity = 0.563 - 0.683 x (genetic distance to *G. gallus*) is drawn in red. The R^2 for the linear regression is 0.875 ($p < 0.001$).

Since we had different population sizes whereby some population samples consisted of less than 15 individuals, we checked if this affects the estimates. We estimated the genetic diversity when only populations with 15 or more individuals were considered and found that the population sizes did not affect the estimates. We also sampled 1000 SNPs in 100 replicates to validate that the relationship between heterozygosity and genetic distance does not happen by chance. The percentages of variation explained in the 100 replicates ranged from 85.1 to 88.3% with a mean of 86.7%. Figure S4.1 shows the regression plots of the 100 replicates with their 95% confidence intervals. Furthermore, we permuted the SNPs to investigate whether the decreasing heterozygosity is not generally an artefact of the Reynolds' distances. We found that the relationship between the observed heterozygosity and the genetic distance based on permutated SNPs was almost non-existing with an R^2 value of 0.01. We also used the fixation index (F_{ST}) as an alternative measure of differentiation, and found the Mantel correlation coefficient (r_m) of pairwise F_{ST} values with the corresponding Reynolds' distances to be 0.976. Reynolds' genetic distances to the wild populations (*G .gallus*) and the F_{ST} values were highly correlated with a Pearson's r = 0.990 and their relationship is shown in S4.2 with an R^2 value of 0.990. When using F_{ST}, the genetic differentiation of the breeds from the wild populations (*G. gallus*) explained 86.2% of the variation in genetic diversity (Figure S4.2).

Given our results we can conclude that the variance in genetic diversity within the domesticated chicken populations can be well explained by the genetic distance to the *Gallus gallus*. Although our current study may not directly prove this due to lack of geographic sampling coordinates, given the whole data set it is evident that the geographic distance alone may not well predict the observed genetic variations in the chickens because:

i) breeds of the same geographic origin are found scattered across the genetic diversity spectrum. This is the case for Asian (red symbols) and European (green symbols) type breeds. As it is shown in Figure 4.1 and as well highlighted in [7], the Asian and European chickens sampled from the German fancy breeders (denoted with prefix DE_) have highly reduced genetic diversity as well as higher genetic distance to the wild chickens (*G .gallus*) than their respective local breeds. However, when considering the sampling areas, the genetic diversity may correlates to the geographic distances to the *G. gallus* within the Asian breed categories but not in the European breeds. Many of the fancy breeds presumably originate from a small number of breeding birds imported from Asia to Europe. Following that, they have been subjected to strong phenotypic selection, with small effective population sizes, population bottlenecks, and intended inbreeding to keep the desired traits. Therefore, such practices are responsible for most of the variations in the genetic diversity of the fancy Asian and European type breeds vs. the respective local types.

ii) the concept of isolation by distance assumes that individuals from nearby locations are likely to be related due to mating possibilities. This is often the case in traditional breeding systems but it is not the case with the fancy and commercial breeding and management practices. Individuals within a commercial breeding herd are more related to each other than to other lines despite the geographic distances. In fancy breeds, there may be gene flow between small stocks based on personal contacts or personal relationships of breeders, but not related to geographic distance forming a substructure within the breed. Actually such gene flow between fancy breeds is also very limited. Furthermore, if geographic distance was a better predictor for the loss of genetic diversity and increased differentiation of breeds to the wild populations, then the African and South American breeds might be expected to have highly reduced genetic

diversity due to geographic distances. They also would be expected to have high genetic distances to the wild populations as well as to the rest of the Asian populations; in fact, both expectations are not fulfilled, and some of the African populations were found to be clustered with the wild type breeds [7].

Therefore, the observed variations in genetic diversity may not well be predicted only by geographic expansion but rather by a combination with other aspects or subsequent events e.g. effective population sizes, types of breeding practices, and possibly subsequent series of founder events following the geographic expansion, as previously suggested [5, 6]. Such events which have taken place after geographic expansion have definitely contributed to the variations in allele frequencies and thus the genetic distances of domestic chickens to the wild populations. In addition, equilibrium between genetic drift, migration and mutation has probably not been reached in all studied populations, which would be compatible with the theory of genetic isolation by distance [5, 8, 9]. The theoretical expansion models are also based on 'natural' expansion through migration, while chickens and other livestock were actively transported by humans (e.g. with ships) to distant places.

Comparisons of the patterns of genetic diversity between the overall genome (all SNPs) and different functional SNP classes.

We compared the patterns of the relationship between the genetic diversity and genetic distances to the *Gallus gallus* species when using the overall SNPs to that from different SNP classes as shown in Figure 4.2 and Table 4.1. The rate of change in genetic diversity due to the genetic distance to the wild populations is represented by the slope in column 4. Compared to other SNP classes, the non-synonymous class showed a relevant deviation from the overall pattern whereby the observed heterozygosity across the breeds was lower than that of the overall genome. The non-

synonymous class also had the most deviating slope among the classes (-0.624 compared to -0.683 for all SNPs). To investigate if the different pattern in the non-synonymous class is not due to the sample size, we resampled the same number of SNPs as in the non-synonymous class (1,082 SNPs) from the overall set (156K SNPs) 100 times. We estimated the heterozygosity and plotted the 100 samples to compare with the non-synonymous set. It is shown in Figure S4.3 that the difference in pattern of the non-synonymous class to the overall genome pattern is not due to the sample size.

Furthermore, the intergenic and intronic classes had the highest proportion of SNPs than the other SNP classes (Table 4.1). In order to validate that the similarity of these two classes to the overall is not an artefact of the sample sizes, we sampled 1,000 SNPs a 100 times from the intergenic and intronic classes (separately). Then we estimated the heterozygosity and compared the results to the overall SNPs, showing that the similarities are not due to the larger sample sizes (Figures S4.4 and S4.5). In comparing the regression models using the likelihood ratio test, the exonic (including both the synonymous and non-synonymous separately) and 5' UTR SNP classes showed highly significant differences to the overall SNPs ($p < 0.001$, Table 4.1 last coloumn). Nonetheless, all SNP classes show a reduction in genetic diversity across populations with the increase in genetic distance to the wild types, with the R^2 values ranging from 81.7% to 88.7%.

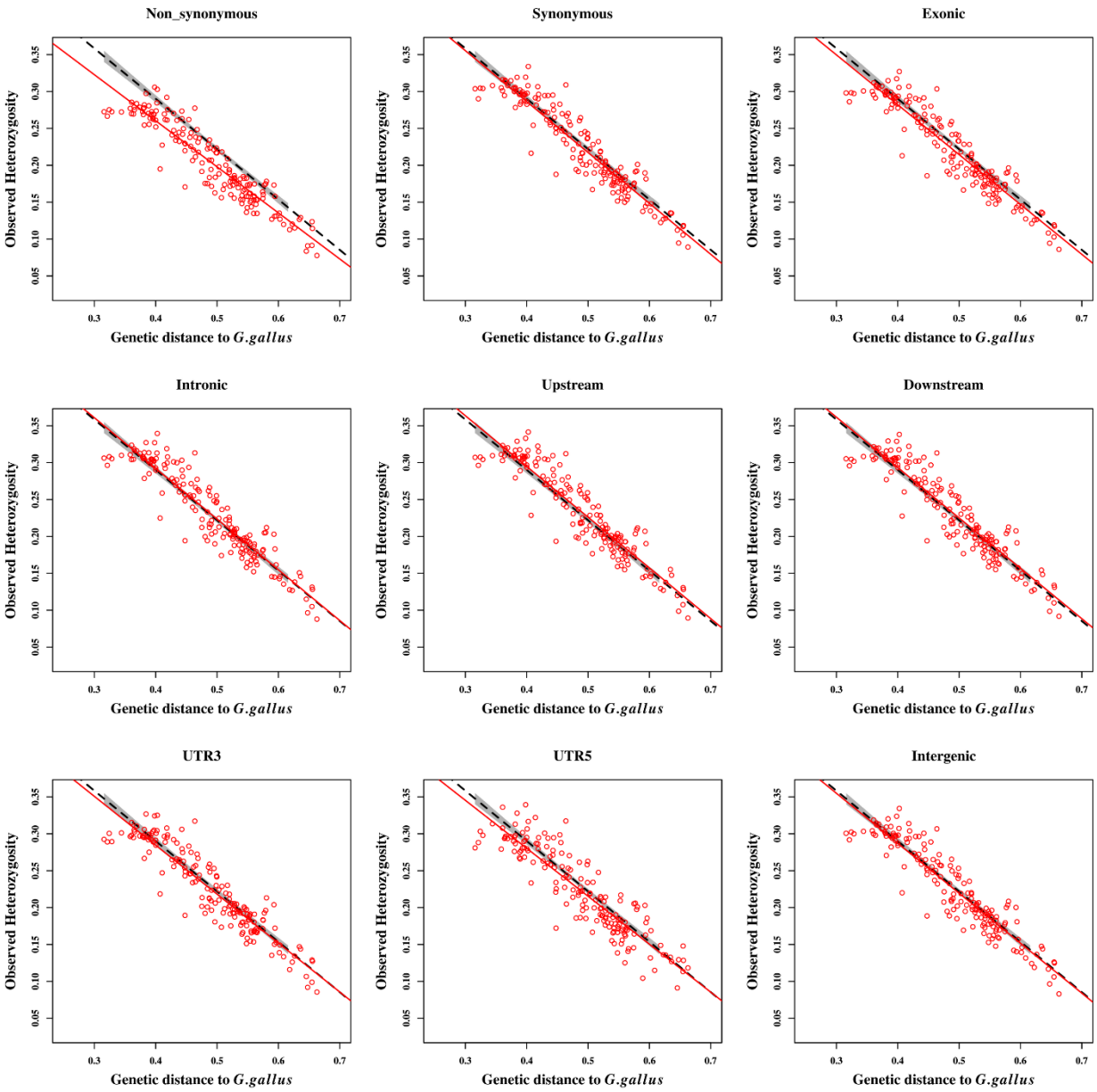


Figure 4.2: Genetic diversity within populations estimated from different SNP classes vs. their Reynolds' genetic distance to *Gallus gallus* ssp. The red circles represent the 172 domesticated populations for the corresponding SNP class. Dashed black lines represent the regression lines for the relationship between observed heterozygosity and the genetic distance to *G. gallus* for the overall pattern and the red lines are for the SNP classes. The areas shaded in gray

represent a 95% confidence interval. The R^2 values and slopes of the linear relationships are shown in Table 4.1. UTR5 and UTR3 refer to the 5' and 3' UTR classes, respectively.

Table 4.1: Comparisons of the linear relationship between genetic diversity and genetic distances of populations to *Gallus gallus* ssp. for different SNP classes.

SNP class	Number of SNPs	R^2	Slope	SE of slope	Likelihood ratio χ^2 test
All SNPs	156,753	0.875	-0.683	0.020	
Non-synonymous	1,082	0.871	-0.624	0.018	p< 0.001
Synonymous	3,891	0.887	-0.690	0.019	p< 0.001
Exonic	5,959	0.885	-0.676	0.019	p< 0.001
Intronic	71,175	0.876	-0.687	0.020	p> 0.050
5' UTR	118	0.817	-0.650	0.020	p< 0.001
3' UTR	1,383	0.864	-0.663	0.020	p> 0.050
Upstream	11,559	0.871	-0.688	0.020	p< 0.050
Downstream	8,777	0.871	-0.683	0.024	p > 0.050
Intergenic	57,782	0.872	-0.677	0.020	p > 0.050

The number of exonic SNPs is the sum of non-synonymous and synonymous SNPs plus the coding and non-coding exonic SNPs which were not assigned to neither the non-synonymous nor synonymous classes. All R^2 values are significant, $p < 0.001$. SE – standard error.

The results show that for the synonymous SNPs, 88.7% of the variation in the heterozygosity across populations can be explained by their genetic distance to *G. gallus* while in the non-synonymous sites it explains 87.1%, and the lowest percentage was observed for 5' UTR (81.7%). However, it is important to note that the 5' UTR class had only 118 SNPs and hence the differences could be an effect of the sample size. To test this, we have randomly sampled 118 SNPs in 100 replicates from the overall set and estimated the relationship as we have done with the non-

synonymous SNPs, and the R^2 of the replicates ranged from 77.9% to 86.5% with a mean of 81.7%, suggesting that this result is most likely an artefact caused by small sample size.

Figure 4.3 shows the mean observed heterozygosity in the different SNP classes. Generally, the observed heterozygosity was lower in genic than in non-genic SNP classes. Within the genic class, lower heterozygosity was observed in exonic than in intronic SNPs. Consistent with Figure 4.2, the non-synonymous SNPs presented the lowest genetic diversity among all the SNP classes. This could be expected since non-synonymous changes can present favourable or disadvantagous consequences. The theoretical assumption is that selection acts rapidly towards fixation of the favourable alleles and purging of the non-favourable ones, thus leading to more homozygosity in these protein altering variants. The exonic and 5' UTR classes followed the non-synonymous class with lowest mean heterozygosity. UTR variants can play a role in the regulation of gene expression and translation. For example, 3' UTR could interfere with microRNA to facilitate the translation of critical disease genes (e.g. cancer genes in humans) [31, 32]. It is also claimed that positive selection for the adaptation of humans in different habitats has been achived with high differentiation in the 5' UTR gene variants [33]. Such examples highlight the importance of UTR variants as possible targets for selection.

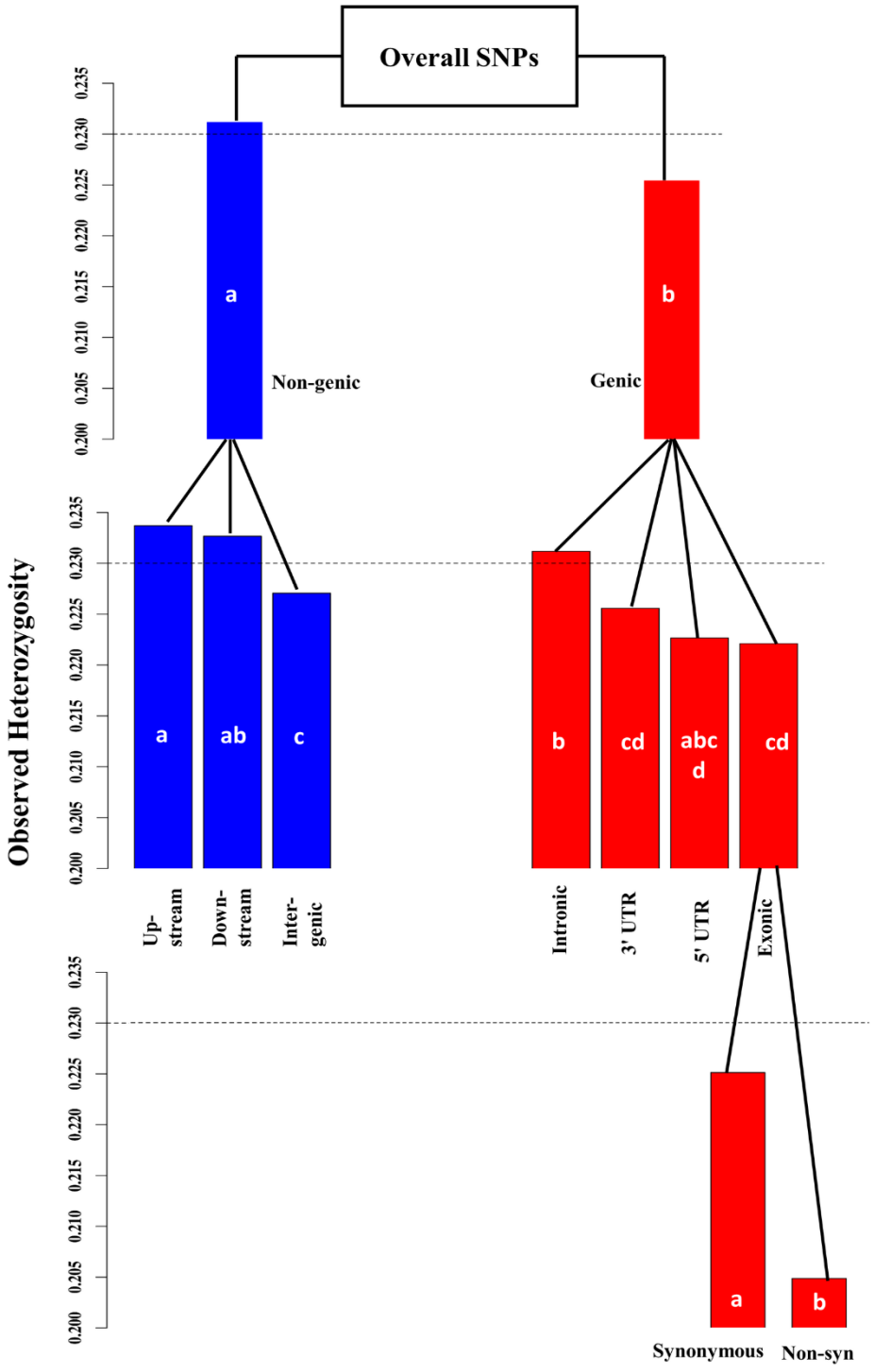


Figure 4.3: The mean observed heterozygosity in the different SNP classes. The gray dotted lines represent the overall mean observed heterozygosity when all SNPs are considered. Non-syn – Non-synonymous. The mean heterozygosities of the SNP classes were significantly different to

the overall mean (Welch two sample t-test $p < 0.05$) except for the 3' UTR and 5' UTR classes. The standard errors (SEs) of the means were lower than 0.005 in all the SNP classes and the overall except for the 5' UTR with SE = 0.009. Different letters in the bars means that there is significant difference in the mean heterozygosity within the same level, e.g. difference between 'Non-genic' and 'Genic' classes on the first level or difference between 'Non-synonymous' and 'Synonymous' classes on the third level.

Patterns of genetic diversity in different genes

We have investigated the patterns of genetic diversity in those 6,303 chicken genes, for which at least 10 SNPs were mapped to the gene, in comparison to the overall genetic diversity pattern. In particular, we wanted to find out if the decrease in genetic diversity is faster or slower in certain genes. Reliabilities (R^2) of the linear regression of the genetic distance from the wild ancestor on heterozygosity for the genes ranged from 0.036 to 0.701 with a mean R^2 of 0.450 and the slopes ranged from -0.110 to -1.099. However, the R^2 values were correlated to the number of SNPs within the genes with r = 0.562. The slopes were independent of the SNP numbers within genes with r = 0.026. The correlation between the slopes and R^2 values was -0.556. We evaluated the regression coefficients (slopes) of the relationship between the heterozygosity and genetic distance for the genes in the top and lowest 5% ranges, which were in total 32 genes at each end. Based on these slope classifications, functional annotations of the genes were done for the combination of molecular function, biological and immune system processes as well as KEGG pathways using the ClueGo package. Based on the ClueGo results, none of the genes in the top 5% range formed any functional clusters while 4 of the genes (namely: EGFR, PAFAH1B1, PTPRS and RTN4) in the lowest 5% were associated with brain development.

Genes in the lowest 5% had slopes ranging from -0.110 to -0.319 while the top 5% ranged from -0.960 to -1.099 (Table S4.2). The genes in the top 5% indicate rapid changes in genetic diversity due to the genetic distance of the chicken breeds to *G. gallus* while those in the lowest 5% indicate genetic diversity changes at a very slow rate in relation to the genetic distance. We obtained the individual gene functions for these genes in the lowest and top ranges from DAVID annotation platform (Table S4.2). The figures showing the relationship between genetic diversity and genetic distance in these genes are shown in Figure S4.6 and S4.7 for the top and lowest 5% ranges, respectively.

The genes in the top 5% slope range were associated with transmembrane transport (SLC25A6, SLC22A15, SLC4A3), protein transport and protein metabolic processes (SLMO1, ERO1L, UCHL5, KCNB1, CSE1L), and lipid metabolic processes (PLCXD1, MIR33, HADHA) among other functions. The transmembrane transport refers to the transportation of solute/s across the protein embedded lipid bilayer. The lipid bilayer facilitates the distribution of molecules such as ions and proteins between different membrane compartments by allowing them to cross to different areas only when it is necessary [34]. Proteins are responsible to perform a wide range of important biochemical functions including those relating to adaptation, survival and performance. Proteins and lipids are also core biological molecules of living organisms and key molecules for energy generation. The energy and nutrient requirements differ for different types of breeds or strains and are as well influenced by other factors such as breeding goals and management systems [35, 36]. Hence the high flexibility of these genes to change may also be associated with such factors in addition to the change in genetic diversity which was initially due to the populations' physical expansion from the *G. gallus*. In general, these genes are flexible to change without necessarily causing harm to the individuals but probably to complement the evolution of the populations. The

genes in this range had R^2 ranging from 0.419 to 0.628 indicating the good association of the genetic diversity and the genetic distance to the wild populations.

Most of the genes in the lowest 5% slope range have consistently lower genetic diversity across the breeds despite the genetic distance to the *Gallus gallus* (see Figure S4.7) and they are mainly related to critical functions which may be absolutely necessary for normal functioning of the individuals. Among all the genes, the slopes were the lowest and much closer to zero for the DPYSL2 (-0.112) and GRB2 (-0.110) genes which also had the lowest R^2 values of 0.036 and 0.038, respectively among all the genes. The GRB2 gene, which is involved in many pathways and functional processes, is assumed to be highly conserved in chicken as well as in humans and was reported to be under very strong evolutionary constraint [37]. Other than some of the genes, which are mentioned above for being related to the development of the brain, genes in the lowest 5% range were also found to be associated with other important developmental processes, functions and pathways. Such include positive regulation of cell proliferation (NTF3, ESRP2, EGFR, FGFR1), positive regulation of reactive oxygen species metabolic process (GRB2, STK17A), regulation of cell death, cell and structure morphogenesis (GRB2, NTF3, DOCK5, EGFR, STK17A), positive regulation of reproduction (GNRH1), development of spinal cord (PTPRS), salivary gland morphogenesis (FGFR1, ESRP2, EGFR), lung morphogenesis (FGFR1, ESRP2), brain morphogenesis and development (FGFR1, PAFAH1B1, DPYSL2), axon development (NEFM, RAB8A, RTN4, DPYSL2) among others functions. ADAM28 belongs to the family of ADAMs genes, being a family of transmembrane proteins involved in several processes including embryonic morphogenesis and tissue development, neurogenesis, cell adhesion, cell migration, axon outgrowth and guidance, cell proliferation and cell differentiation during development [38]. In humans, the ADAMs are said to be involved in the regulation of

growth factor activities, promoting cell growth and invasion. They may alter cell communication or signaling in cancer cells causing an increase in cancer cell proliferation and progression [39]. The allele frequency in our study (results not shown) showed a very rapid fixation of the alternative allele in the ADAM28 in all breed categories supporting the assumption that the mutations might be of importance. In general, the consistent lower genetic diversity in the lowest 5% slope range and limited/lack of response to the changes in genetic distance to *G. gallus* can be due to several reasons such as i) some genes may be under evolutionary constraints such that changes of the genes may be generally critical for normal development or functioning of the animal and changes in the genes may have detrimental effects. ii) Purifying selection may be acting to remove the non-favorable alleles and is, therefore, leading to rapid fixation of the other allele. iii) On the other hand, genetic diversity might have been already reduced from the founders i.e. selection and fixation of the preferred variants took place prior to domestication; hence no or less feasibility for further reduction in genetic diversity is being possible. In this line, we investigated the genes which have the lowest estimated heterozygosity within the *Gallus gallus* populations. We found out that 27 of the 32 genes in the lowest 5% slope range were among genes with the lowest 1% of estimated heterozygosity within *Gallus gallus*. Furthermore, seven of those 27 genes also were among the genes with the lowest 5% of estimated heterozygosity within all breed categories.

We have analyzed the patterns of genetic diversity within a wide range of chicken breeds as a function of genetic distances from the chicken wild types. Given all forces taking place in the genome, we can conclude that the overall genetic diversity in the chicken can be well explained by the genetic distance to the wild populations. However, different functional genomic regions, genes and pathways have shown different evolutionary dynamics across the breeds resulting in different patterns of the genetic diversity compared to the overall genome and the neutral loci. The

non-synonymous sites in particular have shown to be the most deviating from the overall pattern of genetic diversity compared to other genomic sites. Furthermore, we have found that genetic diversity changed at a faster rate in genes which are flexible to be manipulated according to the population needs e.g. genes involved in energy metabolism. On the other hand, genes which show resistance to change are associated with critical vital functions e.g. brain development, crucial for normal functioning of the individuals. Such genes presumably have maintained similar low levels of genetic diversity across all populations by selection or by evolutionary constraints, and the variations or the lack thereof in the genomic diversity between the breeds (within these genes) does not reflect the genetic distances to the wild type populations. This study presents insights and contributes to the knowledge of evolutionary dynamics of different functional genomic regions in the chicken.

Acknowledgements

We acknowledge the members of the SCDP consortium as well as the breeders of the "Bund Deutscher Rassegeflügelzüchter e.V." and the "Gesellschaft zur Erhaltung alter und gefährdeter Haustierrassen e.V" in Germany for providing samples or SNP data, or for giving access to their animals for sampling. We also thank the German Federal Ministry of Education and Research (FKZ 0315528E) for funding the "Synbreed - Synergistic Plant and Animal Breeding" project. DKM also acknowledges the financial support from the IMAGE project which enabled her to conduct the study as part of her PhD work.

References

[1] Wragg D, Mwacharo JM, Alcalde JA, Hocking PM, Hanotte O. Analysis of genome-wide structure, diversity and fine mapping of Mendelian traits in traditional and village chickens. Heredity. 2012; 109: 6–18.

[2] Mwacharo JM, Bjørnstad G, Mobegi V, Nomura K, Hanada H, Amano T, et al. Mitochondrial DNA reveals multiple introductions of domestic chicken in East Africa. Mol Phylogenet Evol. 2011; 58: 374–82.

[3] Tixier-Boichard M, Bed'Hom B, Rognon X. Chicken domestication: From archeology to genomics. C R Biol. 2011; 334: 197–204.

[4] Deshpande O, Batzoglou S, Feldman MW, Luca Cavalli-Sforza L. A serial founder effect model for human settlement out of Africa. Proc R Soc B Biol Sci. 2009; 276: 291–300.

[5] Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci. 2005; 102: 15942–7.

[6] Hunley KL, Healy ME, Long JC. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: Implications for biological race. Am J Phys Anthropol. 2009; 139: 35–46.

[7] Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The SYNBREED chicken diversity panel : A global resource to assess chicken diversity at high genomic resolution. BMC Genomics. 2019; 20: 345.

[8] Malécot G. The mathematics of heredity. Translated by Yermanos DM. San Francisco, CA USA: Freeman; 1969.

[9] Wright S. Isolation by Distance. Genetics. 1943; 28: 114–38.

[10] Ishida Y. Sewall Wright and Gustave Malécot on Isolation by Distance. Philos Sci. 2009; 76: 784–96.

[11] Cavalli-Sforza LL, Barrai I, Edwards AWF. Analysis of human evolution under random genetic drift. Cold Spring Harb Symp Quant Biol. 1964; 29: 9–20.

[12] Stringer C, Andrews P. Genetic and fossil evidence for the origin of modern humans. Science. 1988; 239: 1263–8.

[13] Pemberton TJ, DeGiorgio M, Rosenberg NA. Population Structure in a Comprehensive Genomic Data Set on Human Microsatellite Variation. G3 Genes, Genomes, Genet. 2013; 3: 891–907.

[14] Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005; 15: R159–R160.

[15] Scheu A, Powell A, Bollongino R, Vigne JD, Tresset A, Çakirlar C, et al. The genetic prehistory of domesticated cattle from their origin to the spread across Europe. BMC Genet. 2015; 16: 54.

[16] Utsunomiya Y, Bomba L, Lucente G, Colli L, Negrini R, Lenstra J, et al. Revisiting AFLP fingerprinting for an unbiased assessment of genetic structure and differentiation of taurine and zebu cattle. BMC Genet. 2014; 15: 47.

[17] Reynolds J, Weir BS, Cockerham CC. Estimation o f the coancestry coefficient: Basis for a short-term genetic distance. Genetics. 1983; 105: 767–79.

[18] Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013; 14: 59.

[19] SNP & Variation Suite ™ (Version 8.1). Bozeman, MT: Golden Helix, Inc. http://goldenhelix.com/.

[20] Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics. 2018; 19: 22.

[21] Purcell S, Chang C. PLINK 1.9. https://www.cog-genomics.org/plink2. Accessed 12 Mar 2017.

[22] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 7.

[23] Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2008; 84: 210–23.

[24] Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 Genes, Genomes, Genet. 2017; 7: 109–17.

[25] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. URL https://www.R-project.org/.

[26] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009; 25: 1091–3.

[27] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003; 13: 2498–504.

[28] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane H, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4: R60.

[29] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319: 1100–4.

[30] Wang Y. Genetic and Geographic Diversity of Gyr (Bos Indicus) Cattle in Brazil. University of Natural Resources and Life Sciences, Vienna; 2015.

[31] Nicoloso MS, Sun H, Riccardo Spizzo, Kim H, Wickramasinghe P, Shimizu M, et al. Single nucleotide polymorphisms inside microrna target sites influence tumor susceptibility. Cancer Res. 2010; 70: 2789–2798.

[32] Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, et al. Genome-wide identification of SNPs in MicroRNA genes and the SNP effects on MicroRNA target binding and biogenesis. Hum Mutat. 2012; 33: 254–63.

[33] Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008; 40: 340–5.

[34] Andersen OS, Koeppe RE. Bilayer thickness and membrane protein function: An energetic perspective. Annu Rev Biophys Biomol Struct. 2007; 36: 107–30.

[35] Buzała M, Janicki B, Czarnecki R. Consequences of different growth rates in broiler breeder and layer hens on embryogenesis, metabolism and metabolic rate: A review. Poult Sci. 2015; 94: 728–33.

[36] Boschiero C, Costa G, Moreira M, Gheyas AA, Godoy TF, Gasparin G, et al. Genome-wide

characterization of genetic variants and putative regions under selection in meat and egg-type chicken lines. BMC Genomics. 2018; 19: 83.

[37] Wasenius VM, Meriläinen J, Lehto VP. Sequence of a chicken cDNA encoding a GRB2 protein. Gene. 1993; 134: 299–300.

[38] Lin J, Yan X, Markus A, Redies C, Rolfs A, Luo J. Expression of seven members of the ADAM family in developing chicken spinal cord. Dev Dyn. 2010; 239: 1246–54.

[39] Mochizuki S, Okada Y. ADAMs in cancer cell proliferation and progression. Cancer Sci. 2007; 98: 621–8.

[40] Shao F, Wang X, Yu J, Jiang H, Zhu B, Gu Z. Expression of miR-33 from an SREBF2 intron targets the FTO gene in the chicken. PLoS One. 2014; 9: e91236.

Supplementary materials

Table S4.1: Categories of the chicken breeds

Category	Full name	Number of breeds	Number of individuals
Wild	Wild type chicken	2	38
Com_WL	Commercial white layers	4	80
Com_BL	Commercial brown layers	4	80
Com_BRO	Commercial broilers	4	73
DE_Europe_Ban	European bantams sampled in Germany	8	156
DE_Europe	European breeds sampled in Germany	35	660
DE_Asia_Ban	Asian bantams sampled in Germany	8	177
DE_Asia	Asian breeds sampled in Germany	28	531
Europe_local	European local breeds sampled across Europe	25	443
Asia_local	Asian local breeds sampled across Asia	30	509
South_America	South American breeds	4	78
Africa	African breeds	22	410
Overall		**174**	**3,235**

Source: Malomane et al. [7]

Table S4.2: List and functions of the genes in the top and lowest 5% slope ranges

Gene	Slope	R^2	SNP no	Function
Top 5% genes				
SLC25A6	-1.099	0.573	14	Transmembrane transport, mitochondrial inner membrane, integral component of membrane. Calcium signaling pathway.
PLCXD1	-1.066	0.529	10	Lipid metabolic process.
CDK8	-1.062	0.584	11	Mediator complex.
SLC22A15	-1.031	0.532	12	Substrate-specific transmembrane transporter activity, integral component of membrane.
SCGN	-1.023	0.527	10	Regulation of cytosolic calcium ion concentration.
FAM46D	-1.012	0.509	11	Domain of unknown function DUF1693.
IPCEF1	-1.007	0.450	10	Pleckstrin homology domain, Pleckstrin homology-like domain.
SLMO1	-1.007	0.582	29	Positive regulation of protein targeting to mitochondrion.
CCM2	-1.002	0.505	10	Vasculogenesis, endothelial cell development, multicellular organism growth, cell-cell junction organization, inner ear development, venous blood vessel morphogenesis, pericardium development, blood vessel endothelial cell differentiation, endothelial tube morphogenesis.
PSTPIP1	-1.000	0.592	18	Endocytosis, cell migration.
GTF3C6	-0.997	0.510	13	Transcription from RNA polymerase III promoter.
MIR33	-0.997	0.553	21	Lipid metabolism and energy homeostasis [40].
ERO1L	-0.995	0.489	15	Cellular protein modification process, 4-hydroxyproline metabolic process, protein maturation by protein folding, extracellular matrix organization, endoplasmic reticulum unfolded protein response, protein folding in endoplasmic reticulum, cell redox homeostasis, brown fat cell differentiation, chaperone mediated protein folding requiring cofactor, release of sequestered calcium ion into cytosol, intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress.
C1GALT1	-0.995	0.499	17	Angiogenesis, kidney development, O-glycan processing, core 1, intestinal epithelial cell development.

SLC4A3	-0.992	0.509	10	Regulation of intracellular pH, chloride transmembrane transport.
TSPAN3	-0.992	0.624	32	Cell surface receptor signaling pathway.
TRAF3IP2	-0.989	0.464	14	B cell apoptotic process, positive regulation of defense response to virus by host, humoral immune response, positive regulation of I-kappaB kinase/NF-kappaB signaling, immunoglobulin secretion.
UCHL5	-0.989	0.570	17	Ubiquitin-dependent protein catabolic process, protein deubiquitination, Ino80 complex.
KCNB1	-0.988	0.628	24	Action potential, negative regulation of insulin secretion, protein homooligomerization, cellular response to glucose stimulus, positive regulation of protein targeting to membrane, regulation of action potential, positive regulation of long term synaptic depression.
TUBGCP2	-0.987	0.517	11	Meiotic nuclear division, cytoplasmic microtubule organization, centrosome duplication, interphase microtubule nucleation by interphase microtubule organizing center, mitotic spindle assembly.
CSE1L	-0.984	0.511	10	Protein import into nucleus, protein export from nucleus.
LOC101748919	-0.982	0.533	14	Function not well known.
WASF1	-0.981	0.455	10	Rac protein signal transduction, actin cytoskeleton organization, positive regulation of Arp2/3 complex-mediated actin nucleation.
GIPC2	-0.978	0.538	12	Extracellular exosome.
BIK	-0.977	0.487	14	Apoptotic process, Bcl-2-interacting killer.
HADHA	-0.969	0.466	12	Fatty acid beta-oxidation, response to insulin. Fatty acid elongation, fatty acid degradation, valine, leucine and isoleucine degradation, lysine degradation, tryptophan metabolism, beta-alanine metabolism, propanoate metabolism, butanoate metabolism, biosynthesis of unsaturated fatty acids, metabolic pathways, biosynthesis of antibiotics, carbon metabolism and fatty acid metabolism pathways.
PPP2R5A	-0.967	0.419	11	Signal transduction, positive regulation of protein dephosphorylation, negative regulation of establishment

				of protein localization to plasma membrane, negative regulation of lipid kinase activity.
LOC427547	-0.967	0.433	10	Homophilic cell adhesion via plasma membrane adhesion molecules.
MIR1685	-0.965	0.499	31	Not well known.
WWC2	-0.964	0.567	17	Negative regulation of transcription from RNA polymerase II promoter, negative regulation of hippo signaling, negative regulation of organ growth.
LSM6	-0.963	0.520	11	mRNA splicing, via spliceosome, maturation of SSU-rRNA. RNA degradation and spliceosome pathways.
LOC420419	-0.960	0.559	28	MAPK signaling pathway, GnRH signaling pathway.
Lowest 5% genes				
TRIM2	-0.319	0.202	33	Cytoplasm, Zinc ion binding.
OPTC	-0.311	0.178	15	Leucine-rich repeat-containing N-terminal, Leucine-rich repeat, Leucine-rich repeat, typical subtype.
CKMT1A	-0.311	0.246	15	Phosphorylation.
SPTBN1	-0.306	0.183	15	Mitotic cytokinesis, common-partner SMAD protein phosphorylation, SMAD protein import into nucleus, Golgi to plasma membrane protein transport, membrane assembly, protein targeting to plasma membrane, positive regulation of interleukin-2 secretion, positive regulation of protein localization to plasma membrane.
KIF1B	-0.305	0.125	16	Microtubule-based movement, neuron-neuron synaptic transmission, neuromuscular synaptic transmission, anterograde axonal transport, cytoskeleton-dependent intracellular transport, mitochondrion transport along microtubule.
PTPRS	-0.304	0.283	44	Spinal cord development, cerebellum development, hippocampus development, cerebral cortex development, corpus callosum development, extracellular matrix organization, establishment of endothelial intestinal barrier.
FGFR1	-0.297	0.253	28	Negative regulation of transcription from RNA polymerase II promoter, angiogenesis, ureteric bud development, organ induction, positive regulation of mesenchymal cell proliferation, chondrocyte

				differentiation, epicardial cell to mesenchymal cell transition, vacuolar phosphate transport, sensory perception of sound, positive regulation of cell proliferation, fibroblast growth factor receptor signaling pathway, positive regulation of phospholipase C activity, positive regulation of neuron projection development, peptidyl-tyrosine phosphorylation, ventricular zone neuroblast division, embryonic limb morphogenesis, midbrain development, fibroblast growth factor receptor signaling pathway involved in orbitofrontal cortex development, inner ear morphogenesis, outer ear morphogenesis, middle ear morphogenesis, positive regulation of MAP kinase activity, positive regulation of cell cycle, positive regulation of transcription, DNA-templated, protein autophosphorylation, paraxial mesoderm development, regulation of lateral mesodermal cell fate specification, cell maturation, mesenchymal cell differentiation, positive regulation of cardiac muscle cell proliferation, auditory receptor cell development, branching involved in salivary gland morphogenesis, lung-associated mesenchyme development, regulation of branching involved in salivary gland morphogenesis by mesenchymal-epithelial signaling, vitamin D3 metabolic process, positive regulation of MAPKKK cascade by fibroblast growth factor receptor signaling pathway, negative regulation of fibroblast growth factor production, positive regulation of endothelial cell chemotaxis to fibroblast growth factor, positive regulation of parathyroid hormone secretion, regulation of extrinsic apoptotic signaling pathway in absence of ligand. MAPK signaling pathway, adherens junction, and regulation of actin cytoskeleton pathways.
STXBP6	-0.293	0.219	25	Exocytosis, Golgi to plasma membrane transport, regulation of SNARE complex assembly, exocyst localization.

NDUFA9	-0.292	0.150	14	Cell envelope biogenesis, outer membrane, carbohydrate transport and metabolism. Oxidative phosphorylation, Metabolic pathways.
PAFAH1B1	-0.284	0.262	12	Establishment of mitotic spindle orientation, ameboidal-type cell migration, acrosome assembly, neuron migration, positive regulation of cytokine-mediated signaling pathway, mitotic nuclear division, nuclear migration, chemical synaptic transmission, germ cell development, neuroblast proliferation, learning or memory, retrograde axonal transport, adult locomotory behavior, protein secretion, transmission of nerve impulse, corpus callosum morphogenesis, hippocampus development, layer formation in cerebral cortex, neurogenesis, actin cytoskeleton organization, microtubule organizing center organization, osteoclast development, positive regulation of embryonic development, establishment of planar polarity of embryonic epithelium, regulation of GTPase activity, cortical microtubule organization, negative regulation of JNK cascade, vesicle transport along microtubule, brain morphogenesis, neuromuscular process controlling balance, nuclear envelope disassembly, cell division, maintenance of centrosome location, auditory receptor cell development, positive regulation of dendritic spine morphogenesis, regulation of microtubule cytoskeleton organization, cochlea development, microtubule cytoskeleton organization involved in establishment of planar polarity, regulation of microtubule motor activity.
LOC416354	-0.283	0.177	10	Energy production and conversion / Coenzyme metabolism, negative regulation of transcription from RNA polymerase II promoter. Wnt signaling and Notch signaling pathway.
C20H20ORF4	-0.282	0.157	15	Not well known.
RTN4	-0.280	0.154	13	Axonogenesis, axonal fasciculation, endoplasmic reticulum membrane, endomembrane system, integral component of membrane.

NEFM	-0.276	0.279	33	Axon development, mitophagy in response to mitochondrial depolarization.
PRMT7	-0.271	0.167	18	Spliceosomal snRNP assembly, regulation of gene expression, DNA-templated, regulation of transcription, cell differentiation, DNA methylation involved in gamete generation, histone H4-R3 methylation.
BRIP1	-0.270	0.227	18	DNA repair, regulation of transcription from RNA polymerase II promoter, DNA duplex unwinding.
MYL9	-0.267	0.105	12	Signal transduction mechanisms, cytoskeleton, cell division and chromosome partitioning, calcium ion binding, myosin heavy chain binding, platelet aggregation.
TMEM39B	-0.267	0.122	10	Not well known
EGFR	-0.264	0.115	14	Cell, digestive tract and salivary gland morphogenesis, transmembrane receptor protein tyrosine kinase and epidermal growth factor receptor signaling pathways, learning or memory, cell proliferation, epidermis development, single organismal cell-cell adhesion, cerebral cortex cell migration, positive regulation of cell growth, positive regulation of cell migration, positive regulation of cyclin-dependent protein serine/threonine kinase activity involved in G1/S transition of mitotic cell cycle, positive regulation of catenin import into nucleus, negative regulation of protein catabolic process, negative regulation of apoptotic process, positive regulation of MAP kinase activity, positive regulation of nitric oxide biosynthetic process, positive regulation of DNA repair and replication, positive regulation of transcription, protein autophosphorylation, positive regulation of fibroblastand epithelial cell proliferation, regulation of peptidyl-tyrosine phosphorylation, regulation of nitric-oxide synthase activity, positive regulation of protein kinase B signaling, eyelid development in camera-type eye, response to UV-A, positive regulation of ERK1 and ERK2 cascade, cellular response to amino acid stimulus, epidermal growth factor stimulus and estradiol stimulus, positive regulation of production of miRNAs involved in gene silencing by miRNA. MAPK signaling pathway,

				ErbB signaling pathway, Calcium signaling pathway, FoxO signaling pathway, Endocytosis, Dorso-ventral axis formation, Focal adhesion, Adherens junction, Gap junction, Regulation of actin cytoskeleton, GnRH signaling pathway.
RAB8A	-0.264	0.239	33	Vesicle docking involved in exocytosis, small GTPase mediated signal transduction, axonogenesis, protein secretion, cilium assembly, Golgi vesicle fusion to target membrane. Endocytosis pathway.
KCTD9	-0.257	0.243	11	Protein homooligomerization.
ADAM28	-0.252	0.169	13	Integral component of membrane, metalloendopeptidase activity, epidermal growth factor-like domain, peptidase M12B, ADAM/reprolysin, blood coagulation inhibitor EGF-like conserved site, disintegrin conserved site.
WHSC1L1	-0.246	0.143	11	Chromatin structure and dynamics, zinc ion binding, histone-lysine N-methyltransferase activity. Lysine degradation.
ESRP2	-0.245	0.135	15	mRNA processing, regulation of RNA splicing, positive regulation of epithelial cell proliferation, epithelial tube branching involved in lung morphogenesis, branching involved in salivary gland morphogenesis.
LOC415713	-0.240	0.129	11	Not well known.
EBF2	-0.231	0.270	31	Cell fate determination, transcription, DNA-templated, regulation of transcription, DNA-templated, multicellular organism development, positive regulation of chromatin binding, positive regulation of transcription from RNA polymerase II promoter, brown fat cell differentiation, adipose tissue development.
DOCK5	-0.213	0.113	12	Small GTPase mediated signal transduction, negative regulation of vascular smooth muscle contraction, positive regulation of vascular associated smooth muscle cell migration.
NTF3	-0.199	0.076	14	Activation of MAPK activity, positive regulation of receptor internalization, transmembrane receptor protein tyrosine kinase signaling pathway, cell-cell signaling, positive regulation of cell proliferation, positive regulation

				of cell migration, activation of protein kinase B activity, positive regulation of peptidyl-serine phosphorylation, regulation of apoptotic process, negative regulation of neuron apoptotic process, regulation of neuron differentiation and morphogenesis, regulation of peptidyl-tyrosine phosphorylation, induction of positive chemotaxis, activation of GTPase activity, positive regulation of actin cytoskeleton reorganization. MAPK pathway.
GNRH1	-0.198	0.098	14	Reproduction, multicellular organism development, regulation of gene expression, positive regulation of luteinizing hormone secretion, response to alkaloid, response to ethanol, regulation of hormone biosynthetic process, response to steroid hormone, response to serotonin, regulation of ovarian follicle development, regulation of testosterone secretion, negative regulation of neuron migration. GnRH signaling pathway.
STK17A	-0.189	0.062	13	Intracellular signal transduction, positive regulation of fibroblast apoptotic process, regulation of reactive oxygen species metabolic process.
DPYSL2	-0.112	0.036	10	Endocytosis, cytoskeleton organization, axon guidance, brain development, regulation of axon extension.
GRB2	-0.110	0.038	16	Insulin receptor signaling pathway, positive regulation of signal transduction, cell differentiation, positive regulation of actin filament polymerization, receptor internalization, signal transduction in response to DNA damage, regulation of MAPK cascade, anatomical structure formation involved in morphogenesis, cellular response to ionizing radiation, positive regulation of reactive oxygen species metabolic process. MAPK signaling pathway, ErbB signaling pathway, FoxO signaling pathway, Dorso-ventral axis formation, Focal adhesion, Gap junction, Jak-STAT signaling pathway, Insulin signaling pathway, GnRH signaling pathway.

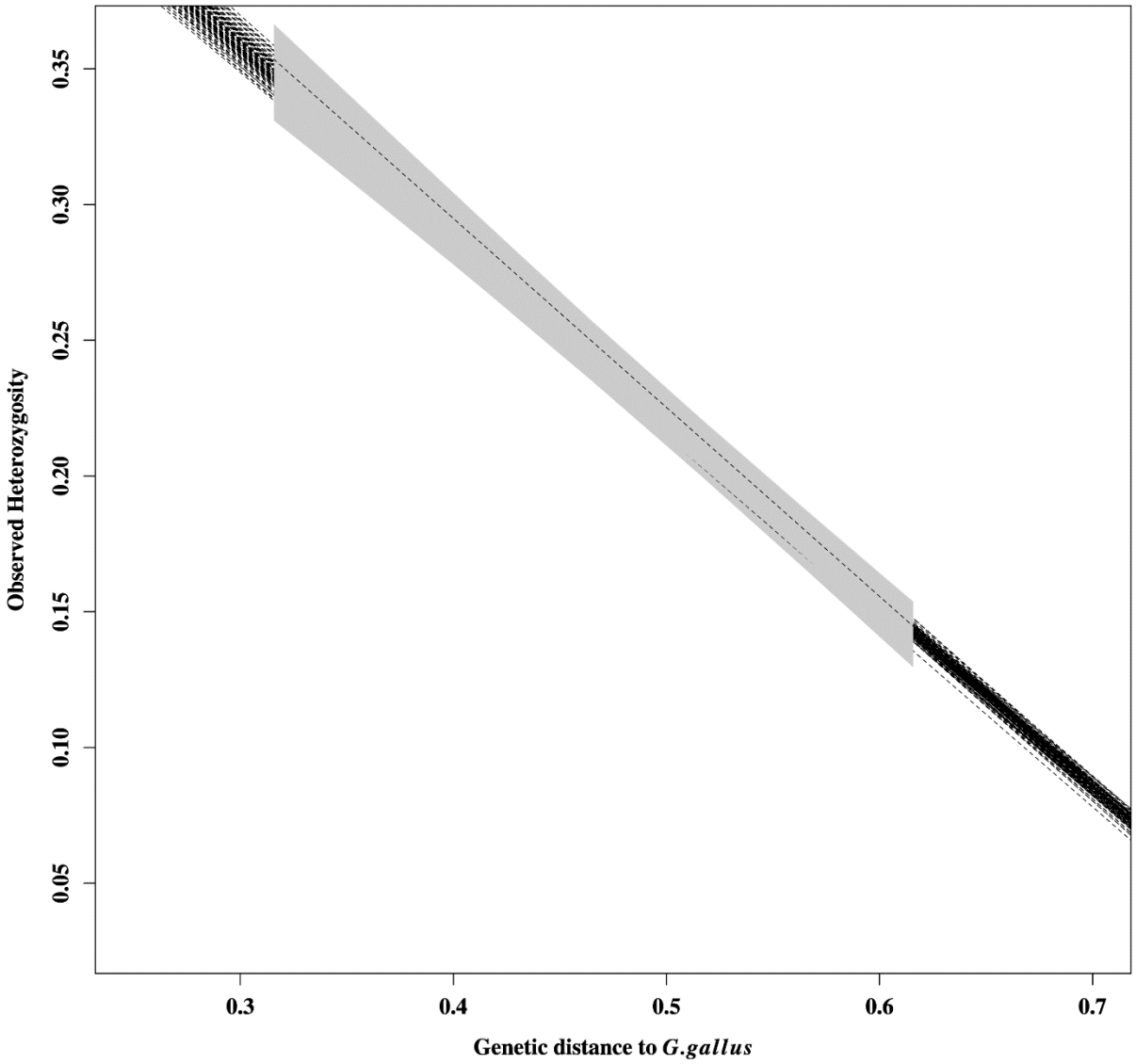


Figure S4.1: Genetic diversity vs. Reynolds' genetic distance to the *Gallus gallus* estimated from 1000 SNP samples in 100 replicates. The dashed lines represent the 100 sample sets and the gray area shows a 95% confidence interval.

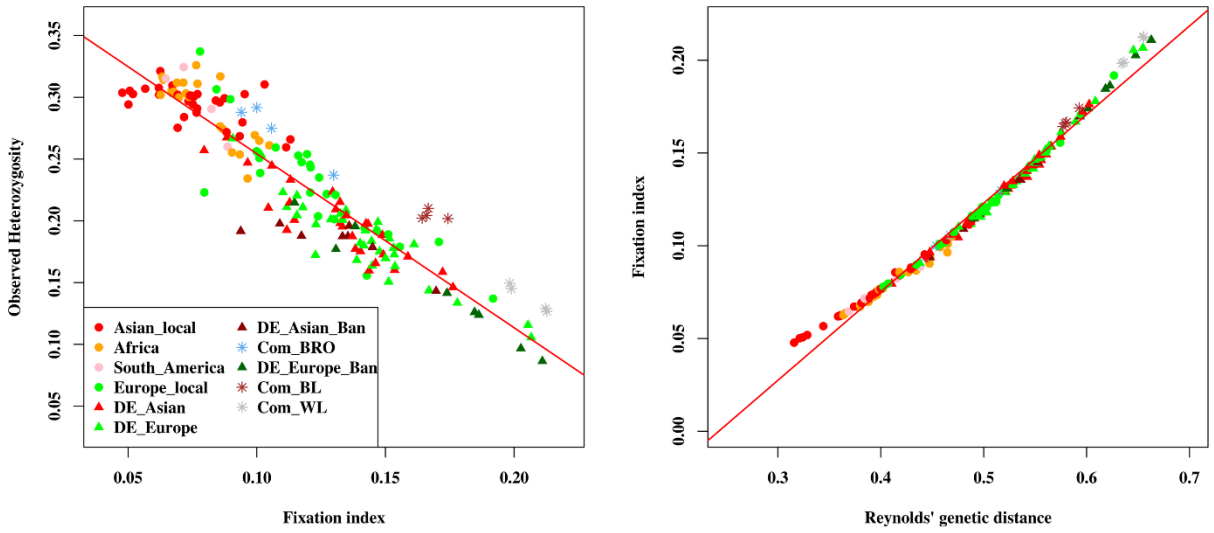


Figure S4.2: The relationship between the observed heterozygosity and genetic differentiation (F_{ST}) from *G. gallus* (left), and the relationship between F_{ST} and Reynolds' genetic distances to *G. gallus* (right). The regression lines of the relationships are drawn in red. The R^2 of the left figure is 0.862 and 0.990 for the right figure. Different breed categories are denoted in different colors and/or shapes.

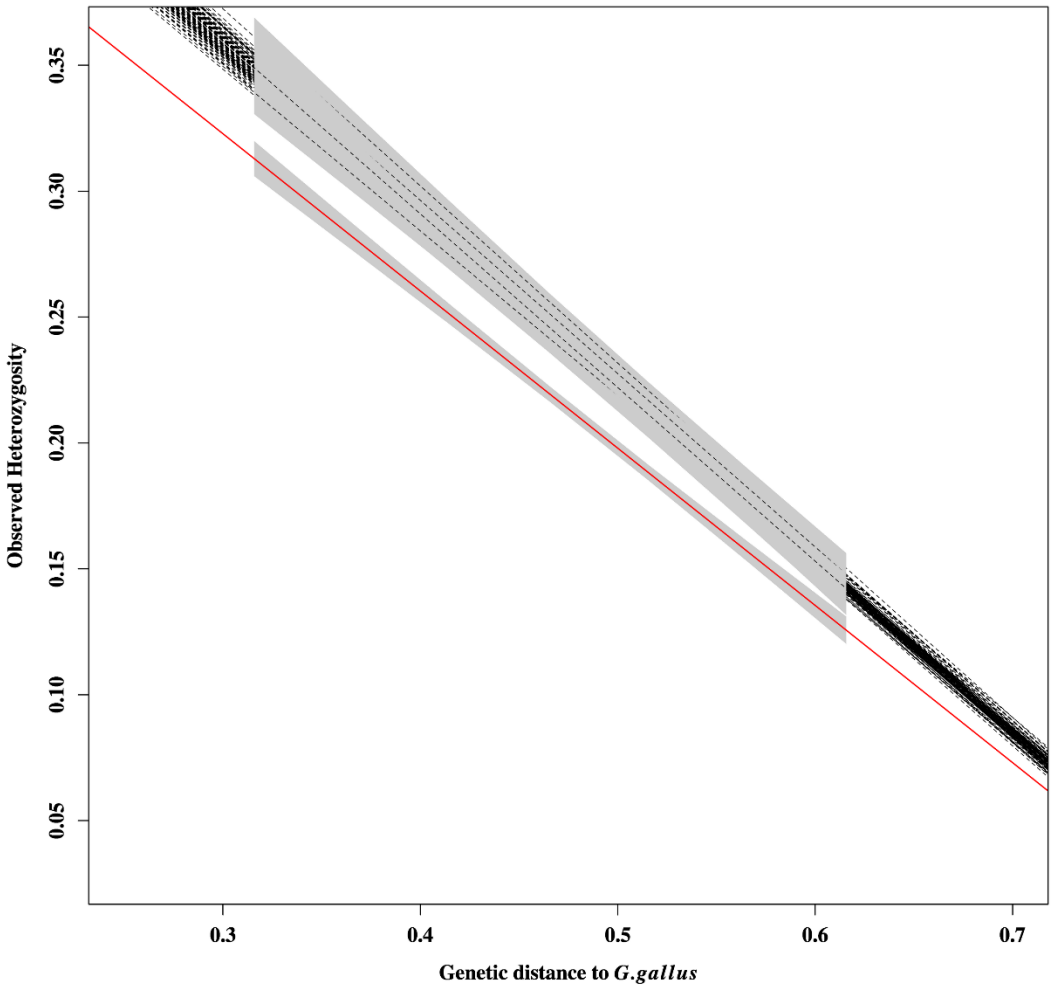


Figure S4.3: Comparison of the relationship between the genetic distances to *G. g*allus and the genetic diversity estimated from the non-synonymous class vs. 100 random samples of the same number of SNPs as the non-synonymous class from the overall SNPs. The black dotted lines represent estimations with the overall SNPs, the red solid line represents the non-synonymous SNPs. The shaded areas represent the 95% confidence intervals of the regression lines. The mean R^2 of the 100 samples is 0.869 and the mean slope is -0.684.

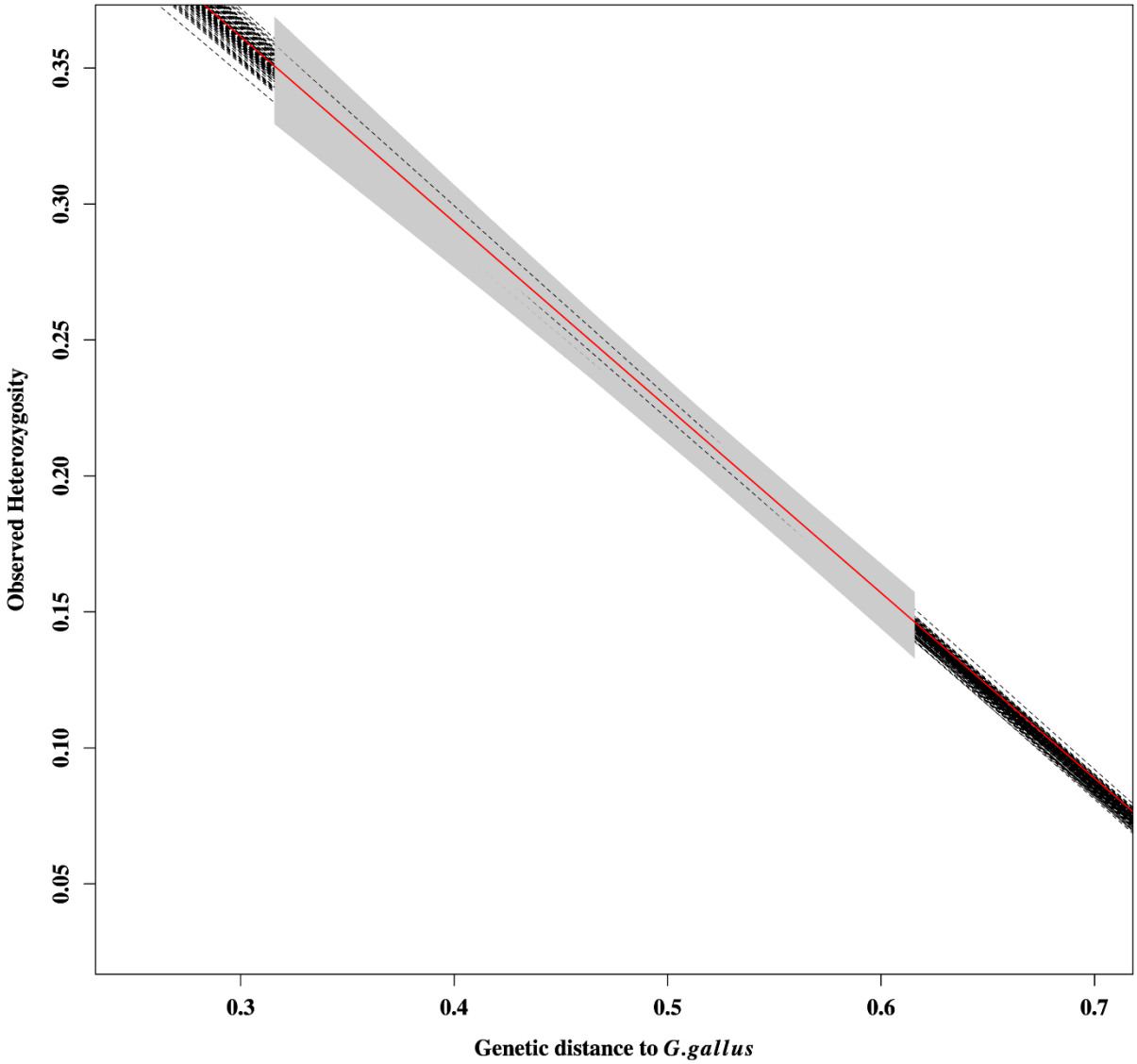


Figure S4.4: Comparison of the relationship between the genetic distances to *G. gallus* and the observed heterozygosity estimated from intronic SNPs vs. the overall set. The black dashed lines represent estimations with the 100 replicates from randomly sampling 1000 SNPs from the intronic SNPs and the red solid line represents overall SNPs. The 95% confidence intervals are shaded in gray. The mean R^2 and slope of the 100 samples are 0.869 and -0.686, respectively.

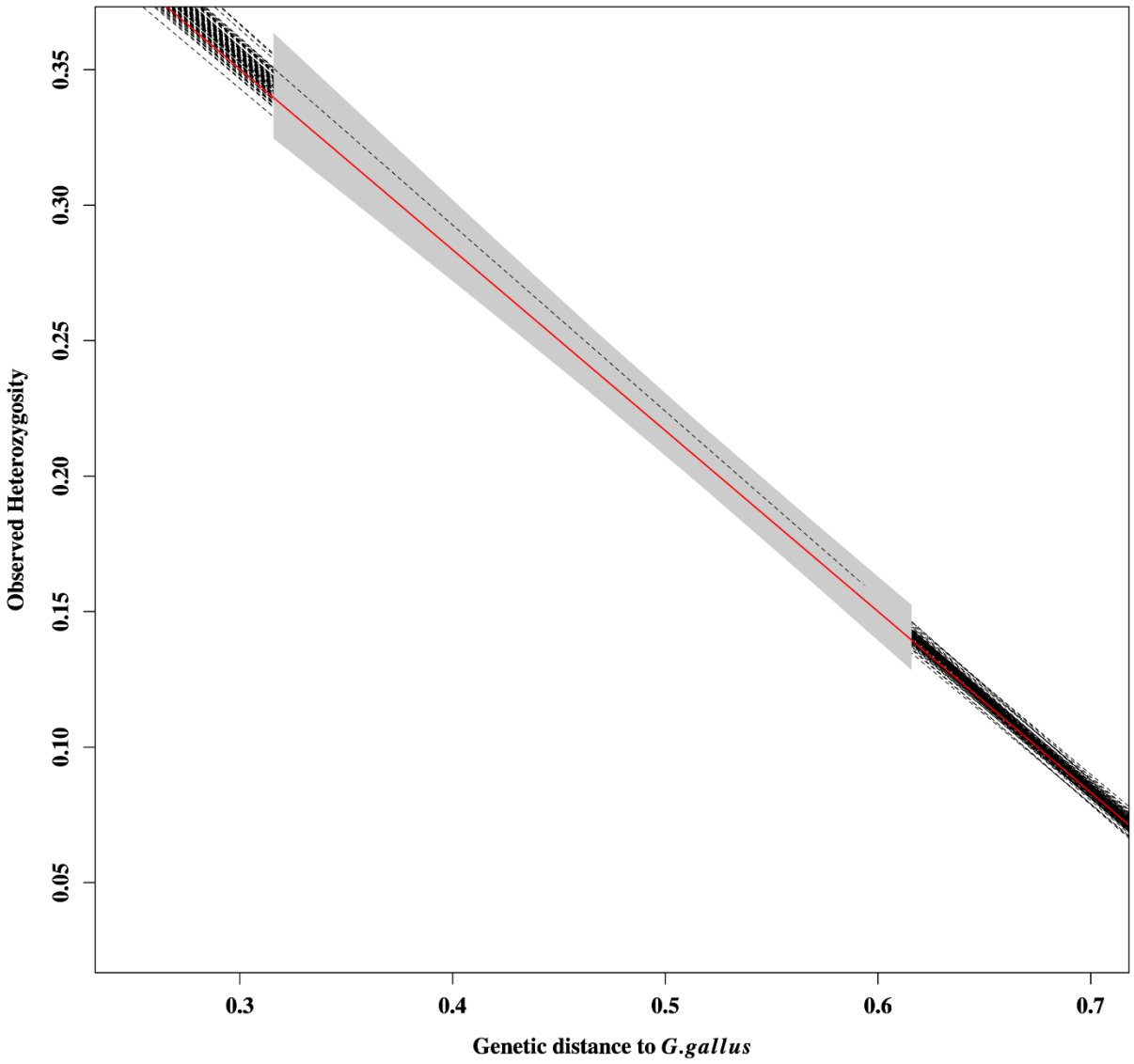
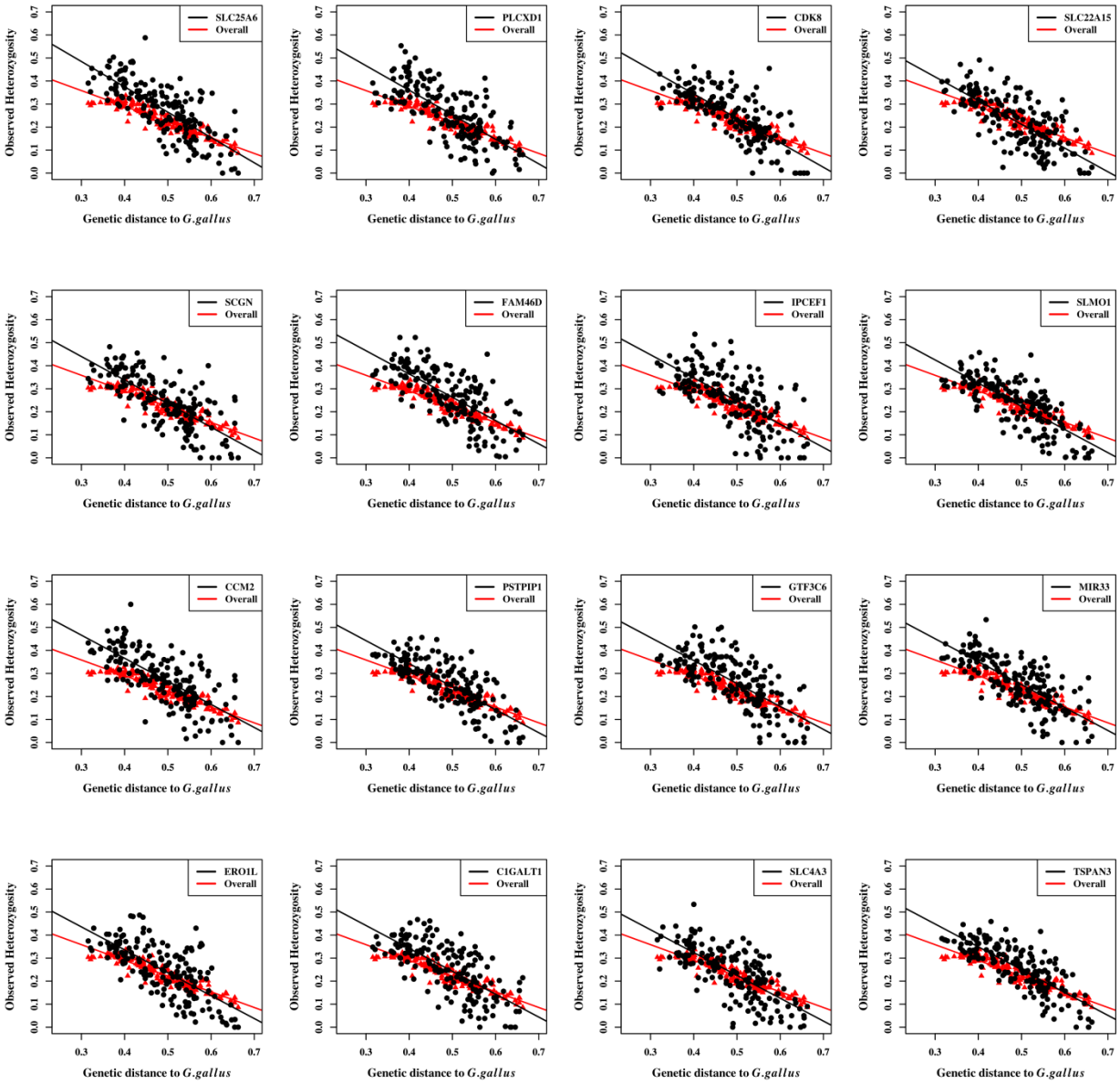


Figure S4.5: Comparison of the relationship between the genetic distances to *G. gallus* and the observed heterozygosity estimated from intergenic SNPs vs. the overall set. The black dashed lines represent estimations with the 100 replicates from randomly sampling 1000 SNPs from the intergenic SNPs and the red solid line represents overall SNPs. The 95% confidence intervals are shaded in gray. The mean R^2 and slope of the 100 samples are 0.865 and -0.678, respectively.


SLC25A6
Overall
PLCXD1
Overall
CDK8
Overall
SLC22A15
Overall
SCGN
Overall
FAM46D
Overall
IPCEF1
Overall
SLMO1
Overall
CCM2
Overall
PSTPIP1
Overall
GTF3C6
Overall
MIR33
Overall
ERO1L
Overall
C1GALT1
Overall
SLC4A3
Overall
TSPAN3
Overall
Observed Heterozygosity
Genetic distance to G.gallus

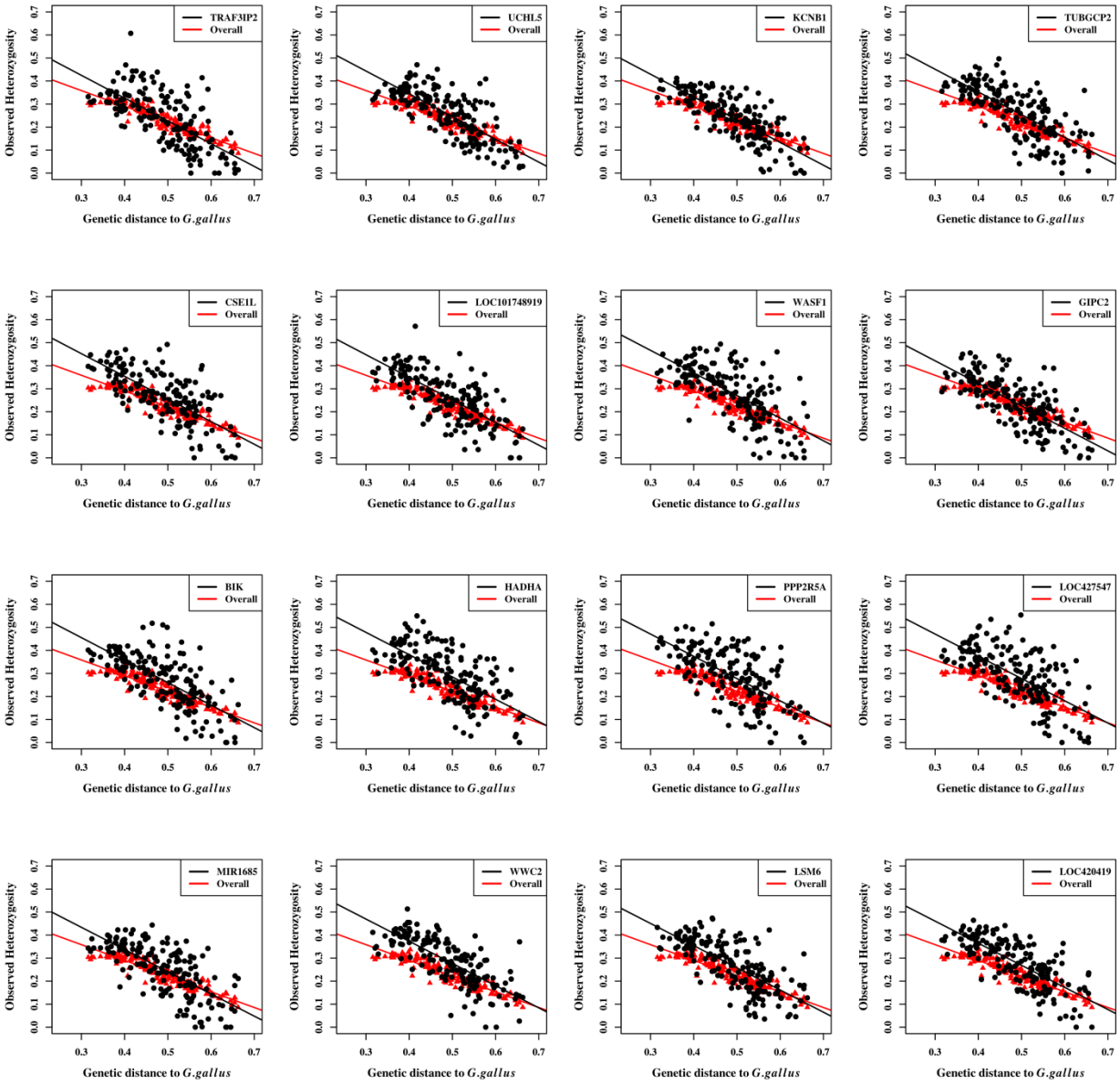
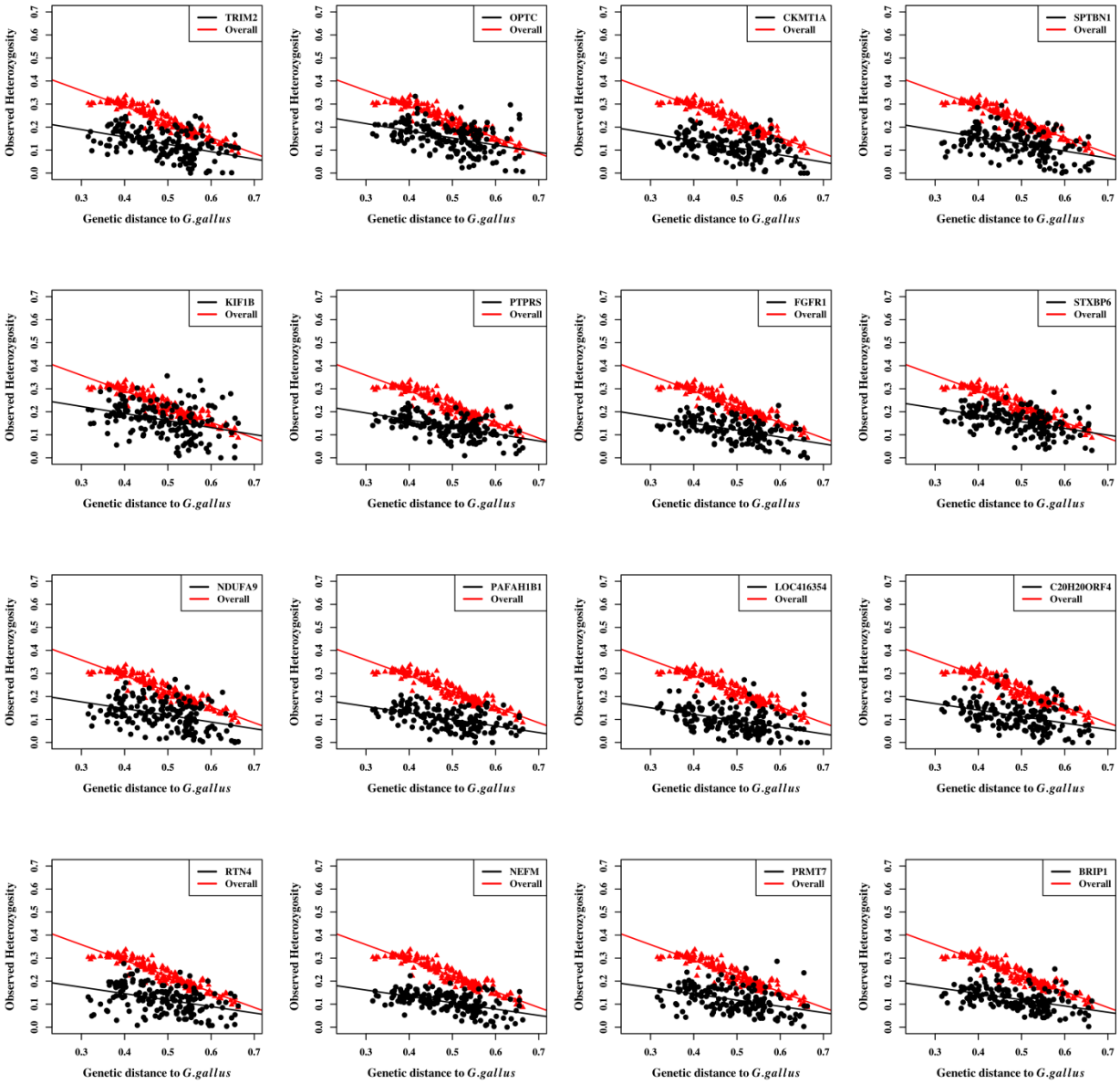


Figure S4.6: The relationship between genetic diversity and genetic distance to *G.gallus* for genes in the top 5% slope range


TRIM2
OPTC
CKMT1A
SPTBN1
KIF1B
PTPRS
FGFR1
STXBP6
NDUFA9
PAFAH1B1
LOC416354
C20H20ORF4
RTN4
NEFM
PRMT7
BRIP1
Overall
Observed Heterozygosity
Genetic distance to G.gallus

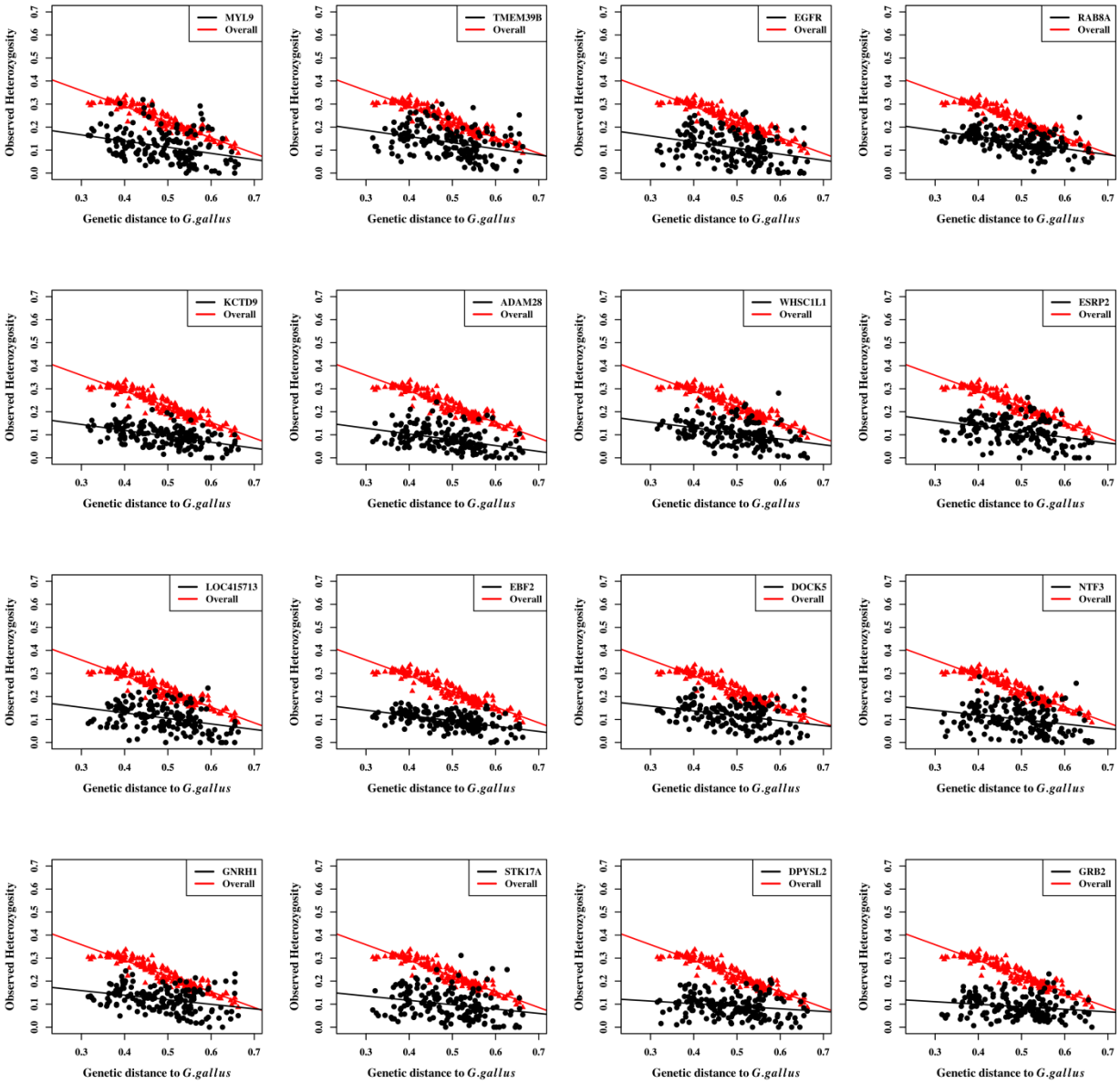


Figure S4.7: The relationship between genetic diversity and genetic distance to *G.gallus* for genes in the lowest 5% slope range

CHAPTER 5

General discussion

The effects of ascertainment bias on genetic diversity estimates

The best way to study the genetic diversity is through the use of whole genome sequencing (WGS) data which not only presents high resolution of genetic information but also may escape the consequences of SNP ascertainment compared to genotyping platforms [1]. However, due to the cost and infrastructure requirements for the WGS, it is very unrealistic to sequence the often desired large quantities of individuals or populations for every genetic diversity study. Therefore, the SNP genotyping platforms are alternative popular options to get genetic information to study the diversity especially in large sets of individuals and populations at lower costs and infrastructure. However, due to the SNP ascertainment bias of genotyping platforms, they may not well represent the 'true' genetic diversity and interpretation of such results may be misleading [2]. Correcting for ascertainment needs an understanding of what are actually the consequences or effects of it. Therefore, in the following we discuss some of the consequences of SNP ascertainment bias in estimating genetic diversity from genotyping platforms which are based on our findings and the literature.

The most common and easily verifiable consequence of ascertainment bias in SNP genotyping data is the underrepresentation of rare SNPs [2, 3]. In **chapter 2** we estimated the allele frequency (AF) from a total of 42 chicken populations where both individual genotyped array data and pool whole genome resequencing (WGS) data were available. Our results confirmed what was already reported in the literature [3, 4] and show a severe underrepresentation of rare SNPs in the array data which were almost missing. The effects of ascertainment bias on population based allele frequency are complex. We observed that the rare SNPs were much underrepresented in the highly diverse breeds compared to the low diversity populations (see example in Figure 5.1). In addition, the low diversity populations also had slightly higher number of SNPs segregating at the AF bin

of 1 in the array data. The ranking of within population genetic diversity was based on the heterozygosity estimated with the WGS data in **chapter 2**. Population genetics analyses often rely on the allele frequencies. Therefore, these missing rare SNPs and the inconsistence effects of ascertainment bias on AF among populations are expected to affect such analyses [1]. Some of the population genetics parameters such as the heterozygosity and F_{ST} may be more affected than others [2, 5].

The general consequence of ascertainment bias on heterozygosity estimation is that, due to the missing rare SNPs, the heterozygosity is overestimated especially in the breeds that were included in the SNP discovery panel. Herrero-Medrano et al [6] found that, among populations of pigs, the expected heterozygosity estimated from the 60K SNP data was extremely higher than in the Next Generation Sequencing data. The bias was not quite systematic across the studied breeds such that when using the 60K SNP data, due to their relationship to the porcine 60K discovery populations, local European and commercial pig breeds showed much higher heterozygosity than the local Asian and wild pigs. Systematic overestimation of heterozygosity due to the relationship of populations to the discovery panel was also observed for European human populations [7]. In **chapter 2** we also observed overestimation of expected heterozygosity among the 42 chicken populations when using SNP array data compared to the WGS data. Since the commercial white and brown layers were part of the discovery panel of the 600K chicken genotyping array, we investigated whether the H_e results behaved differently than in other populations when using array data. We ranked the heterozygosity estimates in the WGS data and followed the same ranking for the array data. The white layers didn't show deviation from the WGS H_e ranking when using the array data. However, the brown layer lines did. There were some populations which are related to some of the breeds in the discovery panel and they deviated from the WGS ranking. However,

there were other populations which are also related to some of the breeds in the discovery panel but they didn't show deviation from the WGS ranking. These results show that it can be challenging to correctly predict effects of ascertainment on population based heterozygosity.

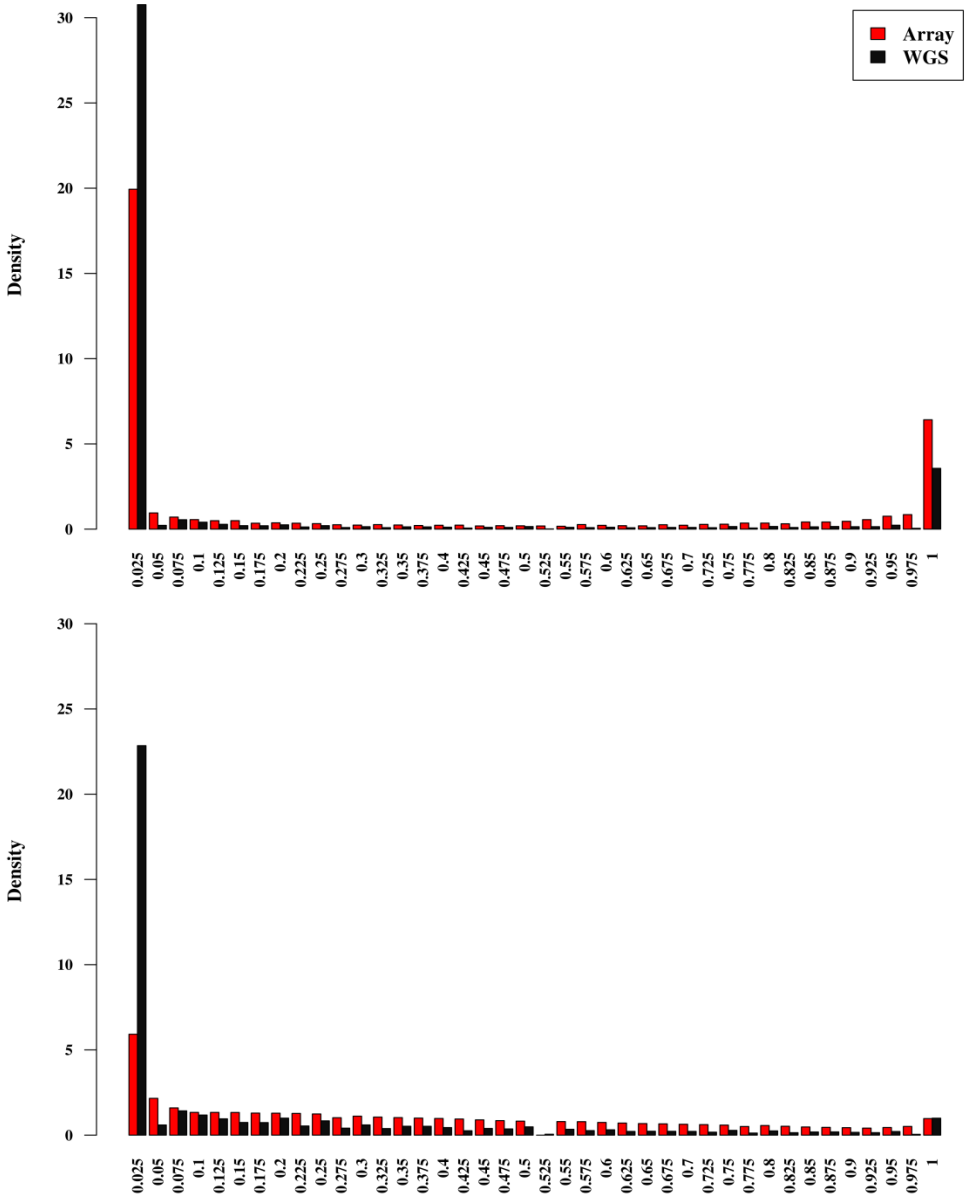


Figure 5.1: Allele frequency spectra of the lowest diversity Sebright population (top) and the highest diversity *Gallus gallus spadiceus* (GGsc) population (below).

Different effects of ascertainment bias on F_{ST} estmates were reported by several scholars. Some reported that ascertainment bias have low effects on F_{ST} estimates [7] while higher estimation of F_{ST} was reported in [1]. Albrechtsen et al [3] additonally reported small differences in F_{ST} estimates from ascertained SNP data vs. WGS data, however, when populations were less related to the ascertained panel, the F_{ST} estimates increased due to ascertainment bias. Our results in **chapter 2** show that the estimation of F_{ST} was highly affected by ascertainment bias. Pairwise F_{ST} values were underestimated in the array data between populations where WGS F_{ST} was low. On the other hand they were overestimated for populations where WGS F_{ST} was high.

The analysis of principal components based on SNP data was reported to be less affected by ascertainment bias [7, 8]. Eller [7] used empirical data and McVean [8] used mathematical methods to come to this conclusion. However, studies based on simulated data claimed that PCA estimated from SNP chip data were distorted when ascertainment bias was not accounted for [3, 9]. Consistent with the results obtained with empirical data [7], our PCA with the array data was less affected by ascertainment and was able to capture a similar population structure as the WGS data. Therefore, simulated studies may be over-emphasizing ascertainment bias in the simulated data which may differ from reality.

Correcting for ascertainment bias

Several ways for correcting for ascertainment bias have been suggested e.g. [1, 3, 4, 10, 11]. However, such suggestions are hardly adopted by scholars due to their complexity and/or requirements of additional data to make the corrections. For example, [10] used complex mathematical formulas to model the ascertainment schemes and incorporating such information into maximum likelihood estimators to correct for the ascertainment bias. In addition, most of the suggestions were also tested using simulated data, which may miss out some of the complexities

of real data. Some of the scholars also admit that these suggested methods are not applicable to every study [3] and it is not always easy to correct for the ascertainment bias [2]. As those studies based on the simulated data often over-emphasize the level of ascertainment bias compared to the true level in real data, one could also argue that some of those methods to correct for ascertainment bias in simulated data may not apply for real cases. Therefore, in **chapter 2** we focused on investigating alternative methods which are tested on real data and are easy to adopt as well. Seeing how misleading the genetic diversity estimates from the ascertained array data could be, we investigated how different SNP filtering options based on minor allele frequency (MAF), polymorphism of SNPs in the wild populations and linkage disequilibrium (LD) pruning could effectively mitigate the effects of ascertainment bias. Many studies of genetic diversity use <5% MAF filtering of SNPs as a quality control when using SNP data because of concerns about lower genotyping rates or accuracy of genotype calling. Given that rare SNPs are already missing in the SNP data, this might cause further implications. There also have been claims about LD pruning to aid in minimizing ascertainment bias in SNP data [12, 13]. The filtering out of SNPs that are not polymorphic in the wild populations was suggested as an ascertainment bias potential accountability measure in the "GLOBALDIV" (http:// http://www.globaldiv.eu/, unpublished) project. Using these filtering options and combinations thereof, a substantial improvement in the results was obtained with array data when LD pruning was applied compared to the other filtering options. The improvement included reduction in the overestimation of H_e and improved rank correlation from 0.957 in the raw array data to 0.973 in the LD pruned data. A huge improvement was also on the estimation of F_{ST} values. Although the estimated pairwise F_{ST} between the populations were a little lower after LD pruning of SNPs, the effect was systematic, comparable to the WGS and therefore, predictable. Hence we recommend this option for mitigating

ascertainment bias when using SNP data for genetic diversity studies. The MAF filtering didn't improve the estimates and in some cases even worsened the estimates. The option of LD based pruning removes markers which are highly correlated with other markers within a given window, leaving markers in the set with low LD to each other. This option is efficient to remove the multicollinearity effects, which may result in overestimation of effects of SNPs due to highly correlated SNPs. In **chapter 3** and **4**, we applied the SNP LD pruning to account for ascertainment bias in our array data.

The SYNBREED chicken diversity panel (SCDP) and data availability

A brief overview of SCDP collection

The SYNBREED chicken diversity panel is an extensive collection of chicken breeds across the globe. The collection consists of 174 chicken populations which were collected from 32 countries (see map in Figure 5.2). It consists of wild populations, commercial lines, local breeds (from Asia, Europe, Africa and South America), and fancy breeds of European and Asian background but kept in Germany. The chickens were genotyped with the 600K Affymetrix® Axiom™ Genome-Wide Chicken Genotyping Array which consists of 580,961 SNPs. The data is made available to the public in the Figshare repository (https://doi.org/10.6084/m9.figshare. 8003909).

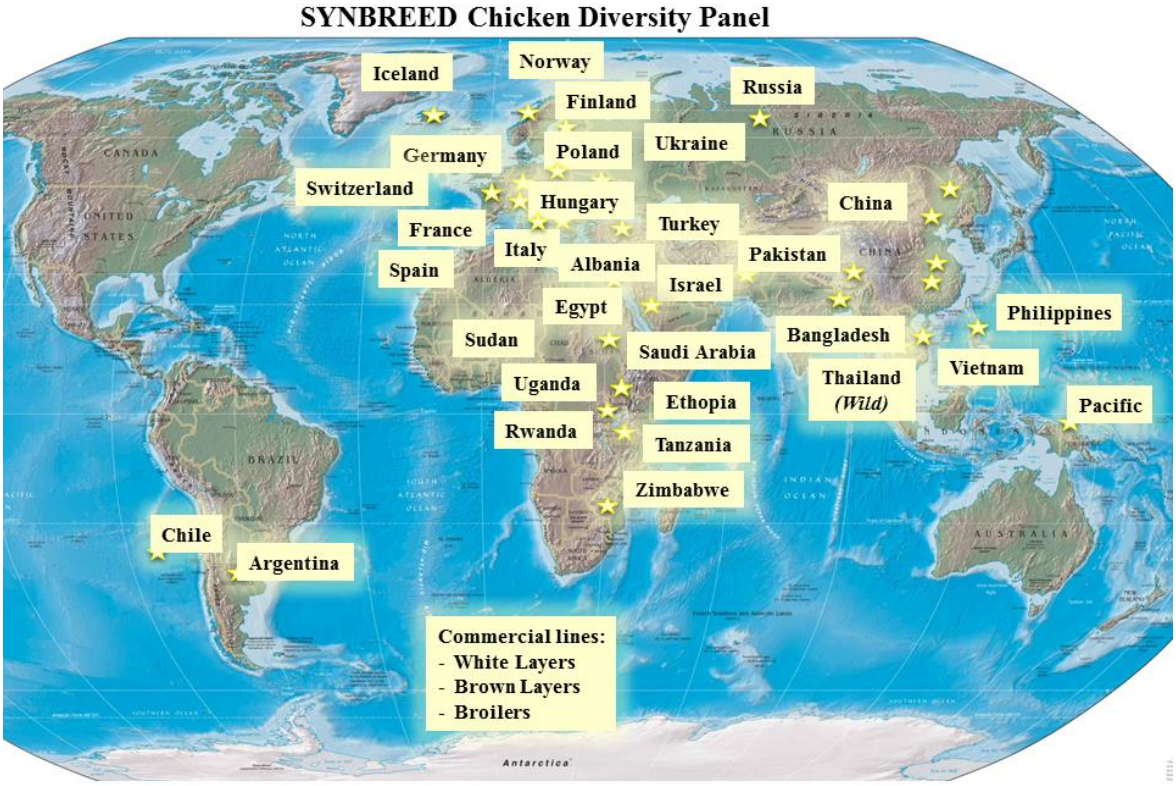


Figure 5.2: Sampling map of the SYNBREED chicken diversity panel. Map edited from http://www.emapas.com/mapa/Mapamundi/217.html

The general findings on the SCDP genetic diversity

In **chapter 3** we analyzed the genetic diversity in the SCDP. The panel covered a wider range of the global chicken genetic diversity than what was previously reported, especially based on microsatellites. For example, [14] analyzed the genetic diversity in 113 chicken populations from Asia, Africa and Europe while [15] studied genetic diversity in 64 Asian and European chicken populations and one African population using microsatellites. In addition, [16] studied the diversity in 52 chicken breeds from Europe and Asia using microsatellite markers as well. In the three studies the wild type and commercial lines were included. Otherwise a vast number of studies have based their analysis on regional, countrywide or continental diversity (e.g., [17–19]). The

SCDP combines a very diverse set of many populations with high resolution genomic data, obtained from the high density genotyping array. Thus, it is a very useful and attractive resource for exploring the global chicken genetic diversity. The findings on the genetic diversity analysis of the SCDP is consistent with the general findings drawn from the previous studies such that the wild type, local African, South American and local Asian populations possess more genetic diversity than European, fancy and commercial breeds. In the SCDP analysis we concluded that, the genetic diversity within global chicken has been driven largely by management and breeding practices. The increased genetic diversity in local breeds comes from the utilization of extensive management practices, with often intercrossing between nearby populations, and no or little artificial selection practices. The intensely managed and highly selected commercial egg layers have low genetic diversity. Interestingly, the commercial broilers in the SCDP, with presumably a management system comparable to the commercial egg layers, have high genetic diversity. We believe this may be related to the age of the samples, the broader genetic basis of founder populations and possibly the larger effective population sizes. Finally, lower genetic diversity found in fancy breeds is due to many generations of phenotypic selection, low effective population sizes and subsequent inbreeding. In the latter we discuss how these practices in the global chicken production have influenced the genetic differentiation of the populations from the founder populations (wild types).

The SCDP have also shown that the between breeds' genetic diversity is widespread and supports two major sources of the globally distributed chicken from Asian and European backgrounds. This is shown by the cluster and structure analyses using the PCA, phylogenetic tree and admixture analysis. The Asian and European breeds are mainly segregated occupying both ends of the diversity spectrum while African and Asian breeds are clustered mostly in between the two, but

slightly towards either the Asian or the European cluster. In addition, commercial white layers clustered with European breeds agreeing with their development from the Italian white leghorn breed, and the brown layers and broilers cluster with the Asian breeds closer to the breeds of their parental background. The clear continuous separation of the European breeds from Asian breeds may be due to the fact that, following domestication of chickens in Asia, Europe was one of the places where chickens were brought to [20, 21]. The chickens arrived in Europe long ago and hence experienced a lot of evolutionary changes. On the other hand, chickens arrived in Africa and South America a little later brought from Asia and Europe.

The need to conserve chicken genetic resources

The Food and Agriculture Organization (FAO) of the United Unions has raised the concerns about the marginalization of local livestock which utilize traditional production systems [22]. The existence of the local breeds is threatened by the crossbreeding or replacement by the high performing commercial hybrids from the intensive commercial livestock market which utilizes a very narrow range of breeds [17, 23]. A good example of this is the Finnish Landrace chicken [24]. When the large scale commercial breeds were introduced to Finland, they almost replaced the entire native Finnish chickens. Luckily, a few populations which still existed in isolated areas were brought together and formed the today's Finnish Landrace, of which conservation programs are now in place to keep the breed existing. The commercial hybrids are highly spread across the world, but they have reduced genetic diversity and therefore, may not be able to sustain the chicken industry in the rise of unforeseen challenges in the future. Therefore, it is very necessary that the genetic diversity in the chicken be preserved. Following the mandate of the FAO [22] to conserve the animal genetic resources, the SYNBREED chicken diversity panel was established. As conservation measures are costly, Simianer [25] emphasized that conservation decisions should be

made upon the global diversity of the species. Therefore, the SCDP data can be seen as a step towards establishing such a reference of global diversity of the species, with the potential to expand as new breeds and other sources of genetic materials will be added from other parts of the world.

The applicability or the limitations of the 'single founder migration model' in domesticated chicken

The migration model from a single founder is based upon geographic patterns of genetic diversity as a result of genetic drift. According to the expectations of genetic isolation by distance [26], genetic differentiation between subpopulations and their founder population increases with the increase in geographic distance. This concept measures the variance in genetic diversity within populations which can be explained by the geographic distance to the founder population due to genetic drift. But can the geographic distances best explain patterns of genetic drift in domestic livestock?

In **chapter 4** we wanted to explore if the genetic diversity of global domesticated chickens can be predicted by their expansion from the wild populations which represent the founders of the domesticated chickens in order to understand some aspects of the species evolution. Previous studies have shown high correlation between the geographic and genetic distances in humans [27]. Such correlations were a little lower in domestic cattle [28, 29]. Consequently, the geographic distance could better explain the variations in genetic diversity in human [27, 30–32] than in cattle. In **chapter 4** we used Reynolds' genetic distances, which assume differences between populations due to genetic drift, instead of using geographic distances. We reported in **chapter 4** that 87.5% of the variations in the overall genetic diversity within the domestic populations can be explained by the Reynolds' genetic distances to the wild populations.

In the following some limitations of geographic distances in explaining the genetic diversity in chicken populations due to genetic drift are highlighted, as well as some of the factors facilitating genetic differentiations and the severity of genetic drift in the chicken. In this way the genetic distances is justified as a better fit than the geographic distances in predicting the genetic diversity within domesticated chickens.

The migration process. Migration processes in chickens are mostly facilitated by humans. Chickens can be transported over long distances at one go without leaving footprints of genetic exchange with the regions in-between. Also, other populations may be transported in short migration steps to reach the same final distance, thus, losing genetic diversity along the way due to the serial founder effect. Therefore, the flexibility of chicken transportation over different distances challenges the use of geographic or physical distance to predict genetic diversity within the chickens.

Breeding and management practices, and the exchange of genetic material. The concept of genetic isolation by distance is based upon the natural populations' mating aspects including possibilities of mating between nearby populations [26, 33, 34]. Often in populations of local breeds under traditional management systems there are chances for the exchange of mating individuals around same geographic areas and hence the genetic diversity between such populations becomes less varied [17, 18, 35, 36]. The similarity of the genetic diversity will also correspond to their neighboring geographic distances and to the geographic distances to the founder populations.

For fancy breeds and commercial breeds, the exchange of mating stock between neighbors is affected by different breeding goals. For example, there are different goals in production of game

vs. bantam chicken breeds or brown vs. white egg laying chicken breeds. The commercial lines are related to each other by type but do not exchange mating individuals with the different populations around the same geographic regions. In fancy breeds, there may be gene flow between small stocks based on personal contacts or personal relationships of breeders but not related to geographic distances. Actually such gene flow between fancy breeds is also very limited and undocumented. Even in cases where fancy breeders may be breeding for the same goals (e.g. two game breed farmers around the same region), they may be competitors and will not exchange breeding animals.

Generally, the genetic diversity within European fancy breeds, with limited or no exchange of mating chickens, differs from that within local European breeds in spite of them located around the same geographic area. The genetic diversity of the fancy breed becomes lower than that of the local type breeds because of the lower number of mating individuals (the effective population size). Many fancy breeds which were brought from Asia to Europe in small numbers, have been kept as purebreds, practicing inbreeding due to low effective population sizes. Therefore, the effects of genetic drift are severe in the fancy populations.

Main conclusions

Based on the studies in **chapters 2** to **4** the main conclusions of the thesis can be narrowed down as follows: When using SNP genotype data to study genetic diversity, LD based pruning of the SNPs can effectively mitigate the effects of ascertainment bias. Although this filtering doesn't account for all the ascertainment bias in the array data, it produces results which are better related to WGS data and hence the interpretation is not distorted.

The SCDP shows that the current global genetic diversity in chicken is absolutely not in a hopeless state. In addition to the wide spread of genetic diversity found between the SCDP populations, the local populations also showed high genetic diversity within the population. Therefore, measures for preserving and maintaining genetic diversity should include new utilization possibilities of local breeds. From a genetic point of view, such breeds should be included in conservation programs.

Lastly, the increase in genetic distances of the domesticated populations from the wild type populations is highly correlated to the reduction in the genetic diversity within the domesticated populations. This has not only been driven by the initial expansion of subpopulations from the founder populations but also by other subsequent events including breeding practices and effective population sizes.

References

[1] Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays. 2013; 35: 780–6.

[2] Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15: 1496–502.

[3] Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010; 27: 2534–47.

[4] Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004; 168: 2373–82.

[5] Qanbari S, Simianer H. Mapping signatures of positive selection in the genome of livestock.

Livest Sci. 2014; 166: 133–43.

[6] Herrero-Medrano JM, Megens H-J, Groenen MA, Bosse M, Pérez-Enciso M, Crooijmans RP. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. BMC Genomics. 2014; 15: 601.

[7] Eller E. Effects of ascertainment bias on recovering human demographic history. Hum Biol. 2001; 73: 411–27.

[8] McVean G. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet. 2009; 5: e1000686.

[9] McTavish EJ, Hillis DM. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics. 2015; 16: 266.

[10] Nielsen R, Signorovitch J. Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. Theor Popul Biol. 2003; 63: 245–55.

[11] Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004; 1: 218–24.

[12] Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 2010; 107: 786–791.

[13] Makina SO, Muchadeyi FC, van Marle-Köster E, MacNeil MD, Maiwashe A. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front Genet. 2014; 5: 1-7.

[14] Lyimo CM, Weigend A, Msoffe PL, Eding H, Simianer H, Weigend S. Global diversity and genetic contributions of chicken populations from African, Asian and European regions. Anim Genet. 2014; 45: 836–48.

[15] Granevitze Z, Hillel J, Feldman M, Six A, Eding H, Weigend S. Genetic structure of a wide-spectrum chicken gene pool. Anim Genet. 2009; 40: 686–93.

[16] Hillel J, Groenen MAM, Tixier-Boichard M, Korol AB, David L, Kirzhner VM, et al. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genet Sel Evol. 2003; 35: 533–57.

[17] Qu L, Li X, Xu G, Chen K, Yang H, Zhang L, et al. Evaluation of genetic diversity in Chinese indigenous chicken breeds using microsatellite markers. Sci China, Ser C Life Sci. 2006; 49: 332–41.

[18] Muchadeyi FC, Eding H, Wollny CBA, Groeneveld E, Makuza SM, Shamseldin R. Absence of population substructuring in Zimbabwe chicken ecotypes inferred using microsatellite analysis. Anim Genet. 2007; 38: 332–9.

[19] Cuc NTK, Simianer H, Eding H, Tieu HV., Cuong VC, Wollny CBA, et al. Assessing genetic diversity of Vietnamese local chicken breeds using microsatellites. Anim Genet. 2010; 41: 545–7.

[20] Storey AA, Athens JS, Bryant D, Carson M, Emery K, DeFrance S, et al. Investigating the global dispersal of chickens in prehistory using ancient mitochondrial dna signatures. PLoS One. 2012; 7: e39171.

[21] West B, Zhou BX. Did chickens go North? New evidence for domestication. J Archaeol Sci. 1988; 15: 515–33.

[22] FAO. Global plan of action for animal genetic resources and the Interlaken Declaration. In: International Technical Conference on Animal Genetic Resources for Food and Agriculture. Rome, Italy; 2007.

[23] Leroy G, Kayang BB, Youssao IAK, Yapi-Gnaoré CV, Osei-Amponsah R, Loukou NE, et al. Gene diversity, agroecological structure and introgression patterns among village chicken populations across North, West and Central Africa. BMC Genet. 2012; 13: 34.

[24] Fulton JE, Berres ME, Kantanen J, Honkatukia M. MHC-B variability within the Finnish Landrace chicken conservation program. Poult Sci. 2017; 96: 3026–30.

[25] Simianer H. Decision making in livestock conservation. Ecol Econ. 2005; 53: 559–72.

[26. Malécot G. The mathematics of heredity. San Francisco, CA USA: Freeman; 1969.

[27] Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci. 2005; 102: 15942–7.

[28] Scheu A, Powell A, Bollongino R, Vigne JD, Tresset A, Çakirlar C, et al. The genetic prehistory of domesticated cattle from their origin to the spread across Europe. BMC Genet. 2015; 16: 54.

[29] Wang Y. Genetic and Geographic Diversity of Gyr (Bos Indicus) Cattle in Brazil. University of Natural Resources and Life Sciences, Vienna; 2015.

[30] Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005; 15: R159–R160.

[31] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319: 1100–4.

[32] Deshpande O, Batzoglou S, Feldman MW, Luca Cavalli-Sforza L. A serial founder effect model for human settlement out of Africa. Proc R Soc B Biol Sci. 2009; 276: 291–300.

[33] Wright S. Isolation by Distance. Genetics. 1943; 28: 114–38.

[34] Kimura M, Weiss GH. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics. 1964; 49: 561–76.

[35] Adebambo AO, Mobegi VA, Mwacharo JM, Oladejo BM, Adewale RA, Iiri LO, et al. Lack of phylogeographic structure in Nigerian village chickens revealed by mitochondrial DNA D-loop sequence analysis. Int J Poult Sci. 2010; 9: 503–7.

[36] Dharmayanthi AB, Terai Y, Sulandari S, Zein MSA, Akiyama T, Satta Y. The origin and evolution of fibromelanosis in domesticated chickens: Genomic comparison of Indonesian Cemani and Chinese Silkie breeds. PLoS One. 2017; 12: e0173147.

APPENDIX

Acknowledgements

In all I thank **God** for giving me the strength and wisdom to carry me through this PhD thesis.

I would like to express my sincere and great appreciation to my main supervisor **Prof. Dr. Henner Simianer**. I thank you for accepting me in your group and have given me the opportunity to work on this interesting project. You have groomed and guided me without losing patience. You have given me many opportunities to grow in this field in form of outsourced courses and academic conferences. I am very grateful.

To my second supervisor **Prof. Dr. Steffen Weigend**, thank you for giving me the opportunity to be part of the Synbreed work. Thank you for your guidance and support throughout this project. You and your wife **Mrs. Annett Weigend** have been very gracious to me during my visits to Mariensee.

Prof. Dr. Armin Otto Schmitt, thank you for accepting to be my third supervisor. I appreciate all your valuable contributions to my thesis work.

I thank **all my colleagues** from the Animal Breeding and Genetics Group for the academic and personal support but as well as the good times we have shared during my PhD period. Special thanks to **Dr. Christian Reimer** for taking me through the 'baby steps' of my PhD and continuous support. I thank you **Dr. Birgit Jutta Zumbach** for the last minute proof reading of the English grammar on my thesis.

Ms. Döring you have helped me a lot since the day I arrived in Goettingen, not just with work related stuff but with many personal issues and you remained very supportive throughout. Thank you very much. **Mrs. Tanja Nolte**, thank you for being an awesome office mate.

I really appreciate all the financial support that I have received during the PhD period from the **Erasmus Mundus** through the INSPIRE project, the **Agricultural Research Council** through the professional development program and support from the **IMAGE** project.

To my family, **my mother and three sisters**, you have been my pillar of strength. Your love and support throughout my life, including this PhD period have been and still is endearing. **Charity**

Ngoatle, during my weakest moments you had a way to pick me up, not just being a sister but being a best friend as well. The ‘Majas’ I thank you.

Thank you to all **my friends** who stood by me and supported me during this PhD journey, through all the good and the bad times.

CURRICULUM VITAE

Dorcus Kholofelo Malomane

Date of birth:	20 March 1989
Sex:	Female
Place of Birth:	Bushbackridge, South Africa

Education

10/2015 - to date **PhD Student in Agricultural Sciences at the University of Göttingen, Göttingen, Germany**
Thesis title: Retrieving patterns of genetic diversity in global set of chicken breeds.
Thesis supervisor: Prof. Dr. Henner Simianer

09/2014 - 10/2015 **Master of Science (Business) at the University College Dublin, Ireland**
Major: Project Management
Thesis title: Assessment of risks leading to project failure of non-profit organizations in South Africa and Uganda.
Thesis supervisor: Mr Joe Houghton

01/2011 – 06/2013 **MSc. Agriculture (Animal Production) at the University of Limpopo, South Africa**
Major: Animal breeding and genetics
Thesis title: Application of principal component analysis to body weight and morphological characteristics of three indigenous chicken breeds found in the Limpopo province.
Thesis supervisor: Prof. Dr. David Norris

01/2007 – 12/2010 **BSc. Agriculture at the University of Limpopo, South Africa**
Major: Animal Production

Peer-reviewed Publications

2019 **Malomane, D. K**., H. Simianer, A. Weigend, C. Reimer, A. O. Schmitt & S. Weigend, 2019. The SYNBREED chicken diversity panel: a global resource to assess chicken diversity at high genomic resolution. *BMC Genomics*, 20: 345.

2018 **Malomane, D. K**., C. Reimer, S. Weigend, A. Weigend, A. R. Sharifi, & H. Simianer, 2018. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics*, 19: 22.

2014 **Malomane, D. K**., D. Norris, C.B. Banga, J. W. N'gambi, 2014. The Use of factor scores for predicting body weight from linear body measurements in three South African indigenous chicken breeds. *Tropical Animal Health and Production journal*, 46: 331–335.

Latest Scientific Participations

09/2018 XVth European Poultry Conference (EPC2018), Dubrovnik, Croatia. Oral presentation.

05/2018 Population, Evolutionary and Quantitative Genetics Conference (PEQG), Madison, USA. Poster presentation.

02/2018 World Congress on Genetics Applied to Livestock Production (WCGALP), Auckland, New Zealand. Poster presentation.

06/2017 Xth European Symposium on Poultry Genetics (ESPG), ST. Malo, France. Oral presentation.

09/2016	Deutsche Gesellschaft für Züchtungskunde eV (DGfZ), Hanover, Germany. Oral presentation.

Academic Fellowships And Awards

2018	World Congress on Genetics Applied to Livestock Production (WCGALP) 2018 scholarship
2017	European Symposium of Poultry Genetics (ESPG) 2017 PhD award
09/2015 – 07/2018	Erasmus Mundus INSPIRE PhD scholarship (for PhD studies)
2013	Agricultural Biotechnology International Conference (ABIC 2013) bursary award
2011-2012	National Research Foundation (NRF) scholarship (for MSc. studies)
2011	Animal Production degree 2010 best student award
2007-2010	Limpopo Department of Agriculture bursary award (for Bachelor's studies)