GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Arrays and beyond:

Evaluation of marker technologies for chicken genomics

Dissertation
to attain the doctoral degree Dr. sc. agr.
of the Faculty of Agricultural Sciences,
Georg-August-Universität Göttingen

Submitted by
Johannes Geibel
born on the 02.10.1989 in Freising

Göttingen, 14.09.2021

1st Referee:	Prof. Dr. Henner Simianer
2nd Referee:	apl. Prof. Dr. Steffen Weigend
3rd Referee:	Prof. Dr. Timothy M. Beissinger

Date of oral examination: 12.11.2021

Table of Contents

Summary

A key research question in livestock research is how livestock's phenotypic diversity is shaped by its genomic diversity. Genomic diversity is thereby assessed through genomic markers. The use and definition of genomic markers is strongly technology driven and therefore changes through time. During the last years, single nucleotide polymorphisms (SNPs) have become the main marker class. Additionally, SNP arrays have been the genotyping technology of choice during the last years due to their early availability. They are, however, currently partially displaced by whole-genome-sequencing (WGS) for SNP calling. Further, structural variants (SV) are moving more and more into the focus of researchers. In this context, the thesis aims in evaluating the value of SNP markers in various ways with its main focus on chickens as a diverse livestock species with major agricultural value.

In **Chapter 1**, the current knowledge of genomic variation, marker technologies, and their use in livestock sciences, especially in chickens, is reviewed. **Chapter 2** and **3** then address a systematic error of SNP arrays, the SNP ascertainment bias. SNP ascertainment bias is a systematic shift of the allele frequency spectrum of SNP arrays towards more common SNPs due to the pre-selection of SNPs in a limited number of individuals of few populations.

Chapter 2 aims in assessing the magnitude of the bias for a standard chicken SNP array and the steps of array design that created the bias. In the study, we therefore remodeled the design process of the chicken array based on (pooled) WGS of various chicken populations. This revealed a sequential reduction of rare alleles during the design process, which was mainly caused by the initial limitation of the discovery set and a later within-population selection of common SNPs while aiming for equidistant spacing. Increasing the discovery set had the largest impact on limiting ascertainment bias. Other steps, as e.g. validation of the SNPs in a broader set of populations did not show relevant effects.

Correction methods for ascertainment bias are by now often unfeasible in studies. **Chapter 3** therefore proposes to use imputation of the array data to WGS level as an *in silico* correction method of the allele frequency spectrum. The study revealed that imputation is able to strongly reduce the effects of ascertainment bias, even when a very sparse reference panel was used. However, it became also obvious that the reference panel then has the same effect as the discovery panel during array design. It is therefore crucial to select samples for the reference panel evenly spaced across the intended range of populations.

SVs are harder to call and genotype than SNPs. Therefore, the question arises whether effects of SV are captured by SNP-based studies due to strong linkage disequilibrium between SNPs and SVs. This is assessed in **Chapter 4** for three commercial chicken breeds, based on WGS data. The study showed that LD between deletions and SNPs was on the same level as LD between SNPs and other SNPs, indicating that deletion effects are captured by SNP marker panels as good as SNP effects. LD between SNPs and other SVs was strongly reduced. The main factor for this reduction was local differences to SNPs in terms of minor allele frequency. However, a reduction of homozygous variant calls for non-deletion SVs compared to the Hardy-Weinberg-expectation may indicate problems of the used SV genotypers.

In the last chapter (**Chapter 5**), the impact of ascertainment bias and possibilities to deal with it in chicken genomics (and also more general in livestock genomics) is discussed. Further, the potentials of including SVs into studies are evaluated. It also discusses what is necessary to combine the information of different genomic data sets to leverage the value of analyses. Finally, an outlook on what information will be additionally available in near future based on recent technological advances is given.

Zusammenfassung

Eine zentrale Forschungsfrage in der Nutztierforschung ist, wie die phänotypische Vielfalt von Nutztieren durch ihre genomische Vielfalt geprägt wird. Die genomische Vielfalt wird dabei durch genomische Marker beschrieben. Die Verwendung und Definition von genomischen Markern ist stark technologieabhängig und ändert sich daher im Laufe der Zeit. In den letzten Jahren haben sich Einzelnukleotidpolymorphismen (SNPs) zur wichtigsten Markerklasse entwickelt. Außerdem waren SNP-Arrays in den letzten Jahren aufgrund ihrer frühen Verfügbarkeit die Genotypisierungstechnologie der Wahl. Sie werden jedoch derzeit teilweise durch die Ganzgenomsequenzierung (WGS) zur SNP-Bestimmung verdrängt. Darüber hinaus rücken Strukturelle Varianten (SV) mehr und mehr in den Fokus der Forschung. In diesem Zusammenhang zielt die vorliegende Arbeit darauf ab, die Aussagekraft von SNP-Markern auf verschiedene Weise zu bewerten, wobei der Schwerpunkt auf Hühnern als einer vielfältigen Nutztierart mit großer landwirtschaftlicher Bedeutung liegt.

In **Kapitel 1** wird der aktuelle Wissensstand über genomische Variation, Markertechnologien und deren Einsatz in der Nutztierwissenschaft, insbesondere bei Hühnern, dargestellt. **Kapitel 2** und **3** befassen sich dann mit einem systematischen Fehler von SNP-Arrays, dem SNP Ascertainment Bias. Der SNP Ascertainment Bias ist eine systematische Verschiebung des Allelfrequenzspektrums von SNP-Arrays hin zu häufigeren SNPs aufgrund der Vorauswahl von SNPs in einer begrenzten Anzahl von Individuen aus wenigen Populationen.

Kapitel 2 zielt darauf ab, das Ausmaß des Bias für einen Standard-SNP-Array für Hühner und die Schritte des Array-Designs, die den Bias verursacht haben, zu bewerten. In der Studie haben wir daher den Designprozess des Hühnerarrays auf der Grundlage von (gepoolten) WGS verschiedener Hühnerpopulationen nachgestellt. Dabei zeigte sich eine sequentielle Reduktion seltener Allele während des Designprozesses, die vor allem durch die anfängliche Begrenzung des Discovery Sets und eine spätere Selektion von häufigen SNPs innerhalb der Populationen bei gleichzeitigem anstreben von äquidistanten Abständen verursacht wurde. Eine Vergrößerung des Discovery Panels hatte den größten Einfluss auf eine Begrenzung des Ascertainment Bias. Andere Schritte, wie z. B. die Validierung der SNPs in einem breiteren Set von Populationen, zeigten keine relevanten Auswirkungen.

Korrekturmethoden für den Ascertainment Bias sind in Studien bisher meist nicht durchführbar. In **Kapitel 3** wird daher vorgeschlagen, die Imputation der Array-Daten auf WGS-Niveau als *in silico* Korrekturmethode für das Allelfrequenzspektrum zu verwenden. Die Studie zeigte, dass die Imputation in der Lage ist, die Auswirkungen von Erhebungsfehlern stark zu reduzieren, selbst wenn ein sehr kleines Referenzpanel verwendet wurde. Es wurde jedoch auch deutlich, dass das Referenzpanel dann den gleichen Effekt wie das Discovery-Panel während des Array-Designs hat.

Daher ist es von entscheidender Bedeutung, dass die Proben für das Referenzpanel gleichmäßig über das Populationsspektrum verteilt ausgewählt werden.

SVs sind schwieriger zu bestimmen und zu genotypisieren als SNPs. Daher stellt sich die Frage, ob die Effekte von SV auch durch SNP-basierte Studien erfasst werden. Das wäre der Fall, wenn zwischen SNPs und SVs ein starkes Kopplungsungleichgewicht (LD) besteht. Dies wird in **Kapitel 4** für drei kommerzielle Hühnerrassen auf der Grundlage von WGS-Daten untersucht. Die Studie zeigte, dass das LD zwischen Deletionen und SNPs auf dem gleichen Niveau lag wie das LD zwischen SNPs und anderen SNPs, was darauf hindeutet, dass Effekte von Deletionen von SNP-Marker-Panels genauso gut erfasst werden wie SNP-Effekte. Das LD zwischen SNPs und anderen SVs war stark reduziert. Der Hauptfaktor für diese Verringerung waren lokale Unterschiede zu SNPs in Bezug auf die Minor-Allel-Frequenz. Eine Reduktion der homozygoten Varianten für Nicht-Deletions-SVs im Vergleich zur Erwartung unter Hardy-Weinberg-Gleichgewicht kann jedoch auf Probleme der verwendeten SV-Genotypisierer hinweisen.

Im letzten Kapitel (**Kapitel 5**) werden die Auswirkungen des Ascertainment Bias und die Möglichkeiten, damit in der Hühnergenomforschung (und auch generell in der Nutztiergenomforschung) umzugehen, diskutiert. Außerdem werden die Möglichkeiten der Einbeziehung von SV in Studien bewertet. Es wird auch erörtert, was notwendig ist, um die Informationen aus verschiedenen genomischen Datensätzen zu kombinieren damit der Aussagewert von Studien erhöht wird. Abschließend wird ein Ausblick darauf gegeben, welche Informationen aufgrund der jüngsten technologischen Fortschritte in naher Zukunft zusätzlich verfügbar sein werden.

Chapter 1

General Introduction

Source and types of genomic variation

The genomic information of eukaryotes is purely encrypted in form of deoxyribonucleic acid (DNA). DNA consists of two counter-rotating strands of nucleotides, forming a double helix (Watson and Crick 1953). A single nucleotide is the combination of a central deoxyribose, a phosphate group and one out of four nucleobases. The nucleobases are thereby either purine bases (adenine, A; guanine G) or pyrimidine bases (thymine, T; cytosine, C). The phosphate group binds to the deoxyribose of the next nucleotide via a covalent binding and thereby is responsible for establishing the backbone of the DNA strand. The nucleobases connect the opposing strands via hydrogen bounds. In this scope, A always binds to T via two hydrogen bounds, while C binds to G via three bounds (Knippers 2015). The sequence of bases allows the coding of information in form of (protein-coding) genes and according regulatory elements, available on both of the two complementary strands (Nordheim 2015). Further, the existence of the two strands is the primary basis for replicative processes (Dröge 2015a).

The nuclear DNA of animals is thereby organized in chromosomes. They can be divided into autosomes and heterosomes. While autosomes exist pair-wise, one inherited by the sire and one by the dam, heterosomes show a sex-linked pattern. In mammals, females carry two X chromosomes, while males carry an X and a shorter Y chromosome (Graves and Watson 1991). In contrast, male birds carry two Z chromosomes, and female birds have a Z and a shorter W chromosome (Stevens 1997). Genetic sex determination in fish species is due to an XY, ZW, polygenic or clonal system, often combined with environmental plasticity (Devlin and Nagahama 2002). Note that the larger heterosome regularly also carries parts of the information of the smaller heterosome in the so-called pseudo-autosomal region (Smeds *et al.* 2014; Raudsepp and Chowdhary 2015). Besides nuclear DNA, animal cells also carry mitochondrial DNA, which is organized in circular form and, besides some rare and often pathogenic cases, exclusively inherited from the dam (Hiendleder 2007).

Genomic variants are typically classified by the way they change the genome. The simplest and currently most evaluated form of polymorphisms are single nucleotide polymorphisms (SNP), which describe a single base exchange at a specified position in the genome. SNPs can thereby have up to four states in a population, even though commonly only bivariate SNPs are analyzed. Mutations generating SNPs are separated into transitions (Ti) and transversions (Tv). While Ti refers to the exchange of a purine base by the other one (A↔G) or of a pyrimidine base by the other one (C↔T), Tv describes the switch between purine and pyrimidine bases. The Ti/Tv ratio is species-dependent, but commonly larger than one, meaning that Ti are more common than Tv (Purvis and Bromham 1997).

Variants that are more than a simple base exchange are classified as structural variants (SV). However, it is common to regard short (< 50 bp) insertion-deletion (InDel) polymorphisms separately due to technical reasons. SV (> 50 bp) are generally separated into unbalanced SV, which change the overall genome size (deletions: DEL; duplications: DUP; insertions: INS) and balanced SV, which change the structural confirmation, but not the overall size of the genome (inversions: INV; translocations: TRA). Further classifications commonly tackle special cases of those variants, e.g. microsatellites, also called simple sequence repeats (SSR), as multiple repetitions of short segments with variable repetition number (Li *et al.* 2002), or have a purely technical basis, e.g. restriction site length polymorphisms (RFLP; Botstein *et al.* 1980) that are length fragments of DNA after digestion by restriction endonucleases and therefore effectively represent any possible mutation of restriction sites. Further, classifications may be mainly used in a specific technical context, e.g. copy number variants (CNV) as summary for DEL and DUP identified from sequencing read depth or array probe intensities (Wang *et al.* 2007; Abyzov *et al.* 2011).

Genomic variation is initially generated by mutation (Falconer and Mackay 1996). Mutations either change single bases, or insert, duplicate, delete, invert or translocate parts of the DNA up to the size of the complete genome, or lead to complex rearrangements. Mutational events thereby can e.g. happen due to repair mechanisms of strand breaks or due to errors in replication and crossovers (Dröge 2015b). Germline point mutations, which result in SNPs, are typically considered as being rare events with their frequency being related to the genome size (Lynch 2010). So are mutation rates per site in vertebrates estimated to be between 0.4 x 10^{-8} and 1.3 × 10^{-8} (Yoder and Tiley 2021). Mutation rates can differ throughout the chromosome. Axelsson *et al.* (2005) held the increased CpG content on micro chromosomes of chickens responsible for increased mutation rates compared to macro chromosomes. Further, Itsara *et al.* (2010) estimated the mutation rate of CNV with 1.22 x 10^{-2} mutations per generation in humans much higher than the rate of single nucleotide variants (SNV). This is in line with assumptions that the presence of segmental duplications, also called low copy repeats, can trigger SV formation mechanisms as non-homologous allelic recombination (NAHR) and thereby leads to hotspots of recurrent and non-recurrent mutation (Gu *et al.* 2008). Additionally, Carvalho *et al.* (2013) found complex genomic rearrangements to trigger further mutation in breakpoint junctions and thereby mutation rates of SNV to be increased by a factor of 10^{-4} in those regions.

Newly mutated alleles can then increase or decrease in frequency by random drift or selection (Falconer and Mackay 1996). Given the neutral theory of molecular evolution by Kimura (1968), most of the mutations are selectively neutral and thereby have a high chance of quickly getting lost by random drift, and only few get enriched in the population. This leads to a specific allele frequency spectrum, which will be handled in detail later. Further, the few mutations that come with a selective advantage leave distinct patterns in the genome (Nielsen 2005), allowing to trace them in the genome, which will be handled also later.

Mutations on the same chromosome of an individual are commonly inherited physically linked as a so-called haplotype. Changes in the frequency of the haplotypes due to drift or selection affect those linked variants therefore equally, leading to a non-random co-occurrence of alleles. This physical linkage can be broken by events of crossing over, with the chance being higher the larger the distance between the two variants is. The co-occurrence of alleles in terms of a correlation between alleles is commonly referred to as linkage disequilibrium (LD), independently from the existence of physical linkage (Qanbari 2020). The strength of LD thereby is a function of physical distance, recombination rate, and effective population size of a population (Sved 1971). LD thereby also changes over generations. The LD between variants allows to use an easy to genotype variant as a predictor (marker) for a close-by, not necessarily known, variant of interest. This means that a part of the genomic variance of interest can be predicted by a subset of the genomic variants, with the effectiveness being due to the strength of LD between markers and variants of interest (los Campos *et al.* 2020).

Advantages and limitations of genotyping and sequencing technologies

Since Watson and Crick (1953) published the basic structure of DNA, huge research effort was spent to gain a deeper understanding of the blueprint of living organisms. Accompanied by revolutionary technological breakthroughs (Sanger *et al.* 1977; Mullis *et al.* 1986), this led to the publication of the first human reference genome less than 50 years later (Lander *et al.* 2001). The growing availability of technology strongly shaped the use of genomic markers. For a long time, molecular insights were constrained to markers like RFLP (Botstein *et al.* 1980) or microsatellites (Li *et al.* 2002) that are only sparsely distributed over the genome, or to the sequencing of small genomic fragments like mitochondria (e.g. Hiendleder *et al.* 2008). Their use was quickly replaced by single nucleotide polymorphisms (SNP) with the beginning 21st century due to the development of SNP arrays and short-read sequencing technologies (LaFramboise 2009; Novembre and Ramachandran 2011; Mardis 2017). Especially the quick decrease in sequencing costs (NHGRI 2020), also known as genomic revolution, led to the discovery of millions of SNPs and InDels (**Table 1.1**). Due to problems with resolving longer SVs by short sequencing reads, recent discoveries of more than 30,000 SVs per human genome became only possible by the development of long-read sequencing technologies as PacBio and Nanopore sequencing (Ho *et al.* 2019). The following chapter will therefore explain the properties of some current state-of-the-art technologies and the bioinformatics needs to call markers. The technologies can be roughly divided into short and long-read sequencing as well as genotyping through SNP arrays.

Table 1.1: Numbers of published short variants for selected vertebrate species

Species	Reference	Assembly length [Gb]	# SNPs and InDels	Number [kb^{-1}]
Human	GRCh38.p13	3.099	700,532,304	226.05
Mouse	GRCm39	2.728	82,972,037	30.41
Chicken	GRCg6a	1.065	23,425,227	22.00
Turkey	Turkey_5.1	1.115	5,390	< 0.01
Cow	ARS-UCD1.2	2.715	97,127,239	35.77
Goat	ARS1	2.922	33,996,710	11.63
Horse	EquCab3.0	2.506	20,355,608	8.12
Pig	Sscrofa11.1	2.501	63,845,860	25.53
Sheep	Oar_v3.1	2.619	60,248,438	23.00

Numbers of published SNPs and InDels available on ENSEMBL 104 (Howe *et al.* 2021). Species were selected based on data availability and relevance for farming. Human and mouse were added for comparison.

Illumina short-read sequencing

While enhanced variants of the original Sanger sequencing approach (Sanger *et al.* 1977) are still used to re-sequence single genes, sequencing of complete vertebrate genomes is nowadays overwhelmingly performed by the use of Illumina's sequencing by synthesis. Briefly, this approach starts from fragmenting extracted DNA into parts with a specific length distribution that has typically an average (mean insert size) of several hundred bp and a specific variance. Oligonucleotide adapters, which later enable binding to the flow cells and may contain library-specific barcodes for multiplexing, are then bound to both ends of the fragments. The oligonucleotides then bind to matching oligonucleotides on the surface of the flow cell and a step called bridge amplification generates spots of multiple identical copies of the DNA fragments on the flow cell. The actual sequencing then happens by using a polymerase to bind one fluorescence-marked nucleotide to the amplicons per sequencing cycle, which emits a base-specific light signal that is captured by a camera. The process is typically repeated for 100 – 300 cycles and leads to reads with according lengths. Optionally, this is followed by a further round of bridge amplification to bind the fragments to the opposite side and repeat the same round of sequencing cycles. This then results in read pairs with opposite read directions (paired-end sequencing; Fuller *et al.* 2009; Mardis 2017).

There are some non-random error sources, appearing at different steps of the workflow, which affect Illumina short-read sequencing. Ross *et al.* (2013) identified regions with extreme GC content to be under-covered, most likely due to problems in DNA amplification by PCR. This, however, should be less problematic with modern PCR-free library preparation. They additionally showed strongly increased error rates in longer homopolymeric stretches. Nakamura *et al.* (2011) assumed inverted repeat sequences to lead to hairpin structures during sequencing and thereby to delays in nucleotide elongation and accumulated sequencing errors. Finally, Li (2014) showed that low complexity regions of the genome are highly affected by erroneous mapping of the short Illumina reads.

Calling variants from short-read data requires computationally expensive bioinformatics. It is mainly done by re-sequencing based on a reference genome instead of de-novo assembling the genome if a suitable reference genome is present. Pipelines for the discovery of short variants (SNPs and InDels) are usually based on the GATK best practices workflow (van der Auwera *et al.* 2013). This involves mapping of the reads to a reference genome (usually bwa-mem; Li 2013), marking of PCR and optical duplicates, recalibration of base quality scores, per-sample calling of variants with minimal thresholds followed by a consolidating population-wide joint calling and a final filtering step. This workflow sometimes is modified, e.g. by the choice of the variant caller (GATK haplotype caller vs. freebayes; McKenna *et al.* 2010; Garrison and Marth 2012), or whether the filtering approach relies on hard filters or a supervised machine learning algorithm (van der Auwera *et al.* 2013).

In contrast to SNPs and short InDels, SVs cannot be called directly from short reads due to their size. Instead, callers use combinations of auxiliary information as local read depth, insert size distributions and orientation of paired-end reads, split read information, and local reassembly (Ho *et al.* 2019). The strong algorithmic differences between the callers lead to different performances in regard to sensitivity and specificity for various SV- and length classes (Ho *et al.* 2019; Kosugi *et al.* 2019). To overcome those issues, ensemble approaches (e.g. parliament2; Zarate *et al.* 2020) try to combine the results of multiple callers and to balance sensitivity and specificity based on the number of supporting callers. Nevertheless, the calling of SVs from short-read sequencing is associated with a high rate of false-positive calls, requiring strict filtering strategies. It is thereby still common to include time-consuming visual scoring in those filtering procedures (Bertolotti *et al.* 2020; Bouwman *et al.* 2020). A pipeline to reduce the time needed for scoring is SV-plaudit (Belyeu *et al.* 2018). It combines the automated production of quality plots by samplot (Belyeu *et al.* 2021) with a cloud-based distribution of work across different assessors. Possibilities to speed this process up by supervised machine learning algorithms are currently evaluated (Chowdhury and Layer 2020). However, there are unsolved problems with SV calling in regions with a high share of repetitive elements and the calling of INS relative to the reference genome (Delage *et al.* 2020) due to the missing ability of short reads to accurately resolve them.

Reduction of sequencing costs may be realized by reduced library approaches such as restriction site-associated DNA sequencing (RADseq, Andrews *et al.* 2016) techniques such as genotyping by sequencing (GBS, Elshire *et al.* 2011). They, however, only give insight into special regions of the genome and results may be influenced by variations of the restriction sites that hinder cutting by the restriction enzymes (Davey *et al.* 2011; Andrews *et al.* 2016). Another method to reduce sequencing costs may be sequencing of DNA that was pooled from multiple samples (Futschik and Schlötterer 2010). Pooled sequencing, however, does not allow the observation of genotypes from single samples and comes with a series of problems regarding biased allele frequencies and technical limitations for variant calling (Futschik and Schlötterer 2010; Boitard *et al.* 2012; Chen *et al.* 2012; Gautier *et al.* 2013; Schlötterer *et al.* 2014; see also the supplementary material to **Chapter 2** and **Chapter 3**). Further, low-coverage sequencing of populations combined with imputation techniques is discussed for larger populations (Pook *et al.* 2021).

Long read sequencing technologies

Recent developments in sequencing technology are intended to overcome the short read lengths and the need for DNA amplification of Illumina sequencing by single molecule sequencing. The first technology, Single Molecule Real Time (SMRT) sequencing of Pacific Biosciences (PacBio), also uses DNA synthesis by fluorescence-marked nucleotides through a polymerase. However, in contrast to Illumina, the polymerase is attached to a zero-mode waveguide and the unamplified DNA is led to the polymerase, which allows getting rid of the amplification steps and increases read lengths to > 10 kb (Rhoads and Au 2015). An advanced protocol, called circular consensus sequencing (CCS), ligates hairpin adaptors to both ends of the DNA template to form a circular template that combines the previous forward and backward strand separated by the hairpin sequences. This is sequenced multiple times and allows for *in silico* error correction. CCS reads, also called high-fidelity (HiFi) reads, then can have accuracies of > 99.5 % which is comparable to Illumina short-reads (Wenger *et al.* 2019). Note that the error profile is especially prone to InDels in a homopolymeric context (Wenger *et al.* 2019).

In contrast, nanopore sequencing from Oxford Nanopore Technologies (ONT) comes with a completely different approach. For ONT nanopore sequencing, adapters are ligated to DNA strands. These adapters then guide the DNA through a nanopore, located at a membrane and set under a certain electric current. The passing of the DNA leads to characteristic changes of the current, which can later be used to determine the base sequence of the DNA molecules (Jain *et al.* 2016). ONT nanopore sequencing allows for huge read lengths with records of larger than one Mb, and thereby e.g. allowed for the first telomere-to-telomere assembly of a human chromosome (Miga *et al.* 2020). However, especially the translation to the base sequence, also called base calling, is computationally demanding and prone to high error rates (Wick *et al.* 2019). Recent developments try to tackle this problem by enhanced designs of the nanopore and advanced base-calling algorithms (Wick *et al.* 2019). Further, the 1D2 protocol tends to sequence forward and reverse strand to correct for sequencing errors. This, however, comes with reduced throughput of a flowcell and therefore increased costs. A special advantage of ONT nanopore sequencing is its scalability and potential usability in the field as e.g.

shown during the latest Ebola outbreaks (Quick *et al.* 2016) and also being available for the monitoring of livestock disease outbreaks (Hansen *et al.* 2018).

SNP arrays

In contrast to sequencing technologies, microarray technology is only able to genotype previously known SNPs. SNP densities are commonly either low (≤10 k SNPs; e.g. Boichard *et al.* 2012; IMAGE 2020), medium (~50 k SNPs; e.g. Matukumalli *et al.* 2009; Groenen *et al.* 2011), or high (≥500 k SNPs; e.g. Kranis *et al.* 2013; Unterseer *et al.* 2014; Illumina 2015). Two platforms are common: Affymetrix Genotyping Arrays and Illumina Bead Chips. They have both in common that they have multiple oligonucleotides, the so-called probes, for each SNP attached to the surface of the array. In the case of Affymetrix, there exist two different probes for each SNP, one for the A-allele and one for the B-allele. The DNA then binds to the probes, resulting in specific match and mismatch patterns (**Figure 1.1 a**). The combination of the signals is then translated to the AA, AB, or BB genotype (LaFramboise 2009). However, potential effects of off-target SNPs need to be taken into account (Wan *et al.* 2009). In contrast, the recent generation of Illumina Bead Chips contain multiple probes that represent only one flanking region of the SNP on the beads (**Figure 1.1 b**). The DNA then binds to the probes and the probes are extended at the SNP position by the, to the template complementary, fluorescence-marked base (Steemers *et al.* 2006). Thereby, A and T emit a red signal, and C and G a green signal. The combinations of intensity and color signals per bead result in three distinct clusters for the three genotypes (LaFramboise 2009). The restriction to two colors limits the SNPs on Illumina Bead Chips to {AT}/{CG} SNPs (Steemers *et al.* 2006). Note that earlier Illumina platforms used two different bead types per SNP, one for the A- and one for the B-allele, in combination with multi- instead of single-base extension (Gunderson *et al.* 2005).

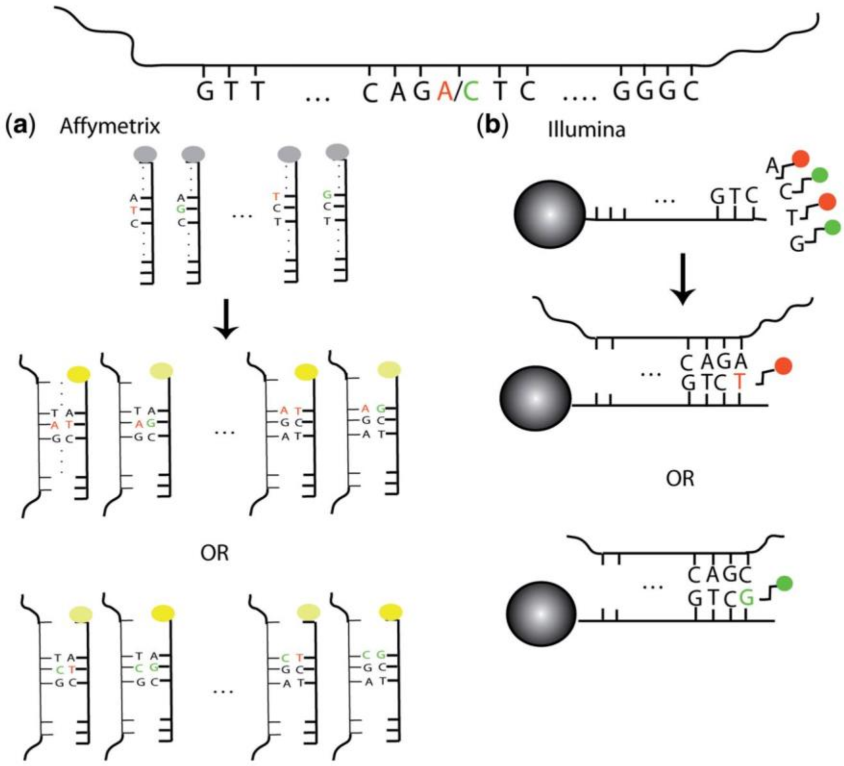

Figure 1.1: Comparison between the Affymetrix and Illumina SNP array platforms. The top represents an example DNA fragment with an A/C SNP. The Affymetrix array (a) contains match probes for both alleles with varying SNP locations. The DNA fragments then bind to the probes, resulting in perfect matches (bright yellow) or mismatches (dimed yellow). The Illumina beads (b) contain only one probe type per SNP. The DNA fragments bind to the probes, which are then extended by a single fluorescence marked base. The emitted color signal allows distinguishing the SNP allele (source: LaFramboise 2009).

The design process of arrays is based on two main selection decisions. The one with major implications on downstream analyses is the pre-selection of known candidate SNPs due to wanted characteristics, which is described in detail in **Chapter 2**. The second decision is based on technical characteristics of the platform such as invariable sites around the SNP for probe binding (Kranis *et al.* 2013).

CNV can be discovered from SNP arrays by analysis of auxiliary characteristics in populations. This may be done by screening the genotypes of a population for physically clustering mendelian errors, deviations from Hardy-Weinberg-Equilibrium, and missing genotypes as indications of DEL (Conrad *et al.* 2006; McCarroll *et al.* 2006). The current default software PennCNV (Wang *et al.* 2007) directly utilizes fluorescent intensity signals, SNP allele frequencies, and pedigree information in a Hidden Markov Model to call CNVs. Callable length classes and breakpoint resolution are thereby dependent on the SNP density of the array. Due to the bad breakpoint resolution, it is common to merge overlapping CNVs into copy number variable regions (CNVR) in array-based analysis (Lee *et al.* 2020), which implies that a CNVR may in fact consist of multiple independent CNVs.

Imputation to switch between marker maps

As shown in the previous paragraphs, marker maps can heavily differ in the type of markers and density due to the used techniques. This is especially prevailing in studies that need to combine datasets that stem from different SNP arrays and potentially include WGS data for a subset of individuals. The different maps typically show a certain share of overlap, and neighboring markers are not independent of each other due to LD. This allows estimating missing marker genotypes in the less complete data set by utilization of the information of the more complete set, known as imputation.

Imputation is typically performed by deriving information on haplotypes (Marchini *et al.* 2007; Browning and Browning 2009; Howie *et al.* 2011; Sargolzaei *et al.* 2014; Browning *et al.* 2018) or LD between markers (Money *et al.* 2015) in the denser set, the reference set. Sometimes, pedigree information is also used (Sargolzaei *et al.* 2014). Based on those information sources, the tools impute missing markers in the less dense set, the study set, with the most likely genotypes. Note that haplotype-based imputation methods always require a phasing step, whose accuracy affects the later imputation accuracy (Pook *et al.* 2019). This is commonly implemented as pre-phasing based on the study genotypes alone (Browning *et al.* 2018), increasing the speed of imputation while having only a minor impact on accuracy (Howie *et al.* 2012; Pausch *et al.* 2013).

Imputation results further strongly depend on the setup of the used reference panel. A general rule is that the genetic distance between reference panel and study set should be as small as possible (Hickey *et al.* 2012; Berry *et al.* 2014; Roshyara and Scholz 2015; Pook *et al.* 2019) and larger reference panels increase imputation accuracy (Pausch *et al.* 2013; Pook *et al.* 2019). However, as increasing the reference panel often means including more distant reference samples, the performance of multi-breed reference panels is of major interest. While e.g. IMPUTE2 (Howie *et al.* 2011) should be robust in this sense, as it limits the reference panel to k nearest haplotypes for an increase in speed, Beagle (Browning *et al.* 2018) uses the complete reference panel. This resulted in reduced accuracies for multi-breed reference panels in some studies (e.g. Berry *et al.* 2014; Korkuć *et al.* 2019; Nolte *et al.* 2020). Korkuć *et al.* (2019) especially showed the need for a strongly increased multi-breed reference panel to gain equal accuracies as for a small closely related panel. Other studies, however, could show increased imputation performance for admixed breeds and rare SNPs when using multi-breed reference panels (Brøndum *et al.* 2014; Rowan *et al.* 2019; Ye *et al.* 2019). Alleles with low frequency are further harder to impute (Hickey *et al.* 2012; Kreiner-Møller *et al.* 2015) and profit more from increased reference panel sizes (Kreiner-Møller *et al.* 2015; Rowan *et al.* 2019).

A common question is whether to impute low-density panels initially to an intermediate density and then to the targeted density, or directly to the targeted density. Studies, that had an additional intermediate reference panel (VanRaden *et al.* 2013; van Binsbergen *et al.* 2014; Kreiner-Møller *et al.*

2015) could show superior performance of the two-step procedure. However, subsetting the high-density panel to an intermediate panel for a first imputation step could not compete with direct imputation (Korkuć *et al.* 2019).

A practical issue often lies in how to measure imputation accuracy. The probably simplest solution is to calculate the mean number of imputation errors (genotype discordance), or its counterpart the genotype concordance (one minus discordance). Since this penalizes homozygote to heterozygote errors as much as homozygote to opposite homozygote errors, the allelic concordance is often used as a refined measure (Pook *et al.* 2019; Zhang *et al.* 2021). It describes one minus the mean absolute difference between the true and imputed number of alternative alleles divided by two. A problem of concordance and discordance rates is that they do not evaluate the performance of a method relative to simply imputing the most frequent genotype and thereby underestimate errors for rare alleles (Hickey *et al.* 2012). Therefore, Pearson correlations between true and imputed alternative allele counts are more appropriate (Hickey *et al.* 2012). However, if calculated per marker, and a marker becomes fixed after imputation, correlations cannot be calculated (Pook *et al.* 2019). This may require using more complex statistics as e.g. the imputation quality score (IQS; Lin *et al.* 2010).

The use of genomic markers in livestock sciences

Genomic prediction

Indicated by the Breeders Equation (Falconer and Mackay 1996), one of the main interests of a breeder is to select the best parents for the next generation as early and as accurately as possible. This interest has strongly driven the idea of not selecting based on the phenotype of an animal, or an auxiliary phenotype if the phenotype of interest cannot be observed at the time point of selection. Those are often bad estimators of the underlying genotypic background. Better estimates were initially achieved by utilizing information of relatives to predict breeding values, first by the selection index theory (Lush 1933) and later by Henderson's Best Linear Unbiased Prediction method (BLUP; development history summarized by Schaeffer 1991). In the 1990s, the idea of using associations between sparsely distributed genomic markers and phenotypes to assist traditional selection procedures, known as marker-assisted selection (MAS), was heavily evaluated (Kumar *et al.* 2011; Wakchaure and Ganguly 2015). The limitation of MAS to marker effects above a certain significance threshold, however, neglects the contribution of small effects to the total genetic value and furthermore results in a bias towards overestimated effects (Meuwissen *et al.* 2001). However, as most of the variance of relevant traits in animal breeding is based on those small and neglected effects, MAS did not establish in animal breeding (Meuwissen *et al.* 2016).

The limitations of MAS led to the groundbreaking suggestion of Meuwissen *et al.* (2001) to estimate the breeding value of an animal as the sum of all available marker effects. Estimation of those effects was based on a BLUP model that assumes the effects to come from a joint normal distribution and is known as random (or sometimes ridge) regression BLUP (rrBLUP). Even though the necessary SNP array technology to derive the needed dense set of markers was not available at that time (Koning 2016), genomic breeding programs in dairy cattle were implemented within ten years. This was possible by the proposal of a genomic dairy cattle breeding program by Schaeffer (2006), which implements the use of young bulls before progeny testing and by this approximately doubles genetic gain per time, and the availability of the first cattle array (Matukumalli *et al.* 2009). Successively, the genomic selection was also adapted in other livestock (Meuwissen *et al.* 2016) and plant breeding programs (Koning 2016).

The initial method of Meuwissen *et al.* (2001) is mainly implemented by a slightly changed method, genomic BLUP (GBLUP; VanRaden 2008). It derives similar results by using the marker genotypes to set up a genomic relationship matrix and then directly estimating genomic breeding values from a BLUP model. This is more efficient if the number of individuals is less than the number of SNPs (Koning 2016; Meuwissen *et al.* 2016). Further, a series of Bayesian nonlinear methods (also known as the Bayesian Alphabet; Gianola *et al.* 2009) tries to break with the assumption of normally distributed SNP effects by allowing a fraction of the SNPs to have zero effect (Meuwissen *et al.* 2001), or even to come from different distributions (Erbe *et al.* 2012).

Besides the size of the training set and population structure, a key factor to derive high prediction accuracies is the marker density (Erbe *et al.* 2013). This has driven the interest in whether an investment in WGS data may lead to the best results. Ober *et al.* (2012) tested this in a Drosophila dataset and showed an asymptotic trend of the accuracy when transitioning from low density to WGS. The same was shown by Perez-Enciso *et al.* (2015) through simulation. Further, the assumption of Meuwissen (2009) that Bayesian methods profit more from WGS data than GBLUP was not confirmed by Ober *et al.* (2012). As large WGS training sets are still unavailable, van Binsbergen *et al.* (2015) tested the performance of genotypes imputed to WGS, but they could not outperform high-density array data.

Mapping of quantitative trait loci

Another interest is in revealing the genetic basis of phenotypic traits to gain a better insight in the underlying biological mechanisms. Earlier linkage mapping methods relied on the decrease of marker-QTL LD over time in experimental families (Mackay and Powell 2007). Fine mapping of QTLs thereby required either large families or multi-generation breeding experiments (Mackay and Powell 2007). The availability of dense marker maps in form of genotyping arrays and later WGS data for larger phenotyped populations has strongly modified the methodology for QTL detection by switching from family-based linkage mapping, to population-based genome-wide association studies (GWAS; Visscher *et al.* 2012). The basic principle of GWAS is to statistically test each marker (mostly SNPs, but may also be other variants) of a dense marker map independently from the other markers in the panel for association with the phenotypic variance of a trait. Significantly associated markers are then assumed to be in strong LD and thereby close physical distance to a causal genomic variant, or even represent the causal variant. Besides the choice of the statistical test, a main technical issue in GWAS is to appropriately control for multiple testing and background effects due to population stratification. GWAS are thereby well suited to identify QTLs of medium to large effect size that segregate with high MAF in a population, but get problems when identifying effects of rare or fixated variants (Visscher *et al.* 2012).

Other approaches to map QTL are selection signature analyses, by Qanbari and Simianer (2014) also referred to as “genome to phenotype” approaches. The idea behind this is that artificial or natural directional selection for certain traits increases the allele frequencies of effect alleles more than what is expected from random drift. This also pulls frequencies of linked variants with it until recombination events happen, known as hitchhiking effect (Smith and Haigh 1974; Fay and Wu 2000). Selection, therefore, leaves specific patterns in the genome, e.g. increased regional differentiation between populations (Akey *et al.* 2002), local differences to the expectation under neutral molecular evolution (Tajima 1989), or the excessively high frequency of long haplotypes (Sabeti *et al.* 2002). See e.g. Nielsen (2005) and Vitti *et al.* (2013) for detailed reviews. To overcome the problem that single selection signature detection methods are specific for certain frequency- or age classes of alleles under selection, combinations of the approaches as e.g. suggested by Ma *et al.* (2015) may be helpful. Note that it is often not possible to connect selection signatures with a specific phenotype. Studies rather discuss candidate regions based on known functions of genes in those regions (e.g. Qanbari *et al.* 2019; Peripolli *et al.* 2020).

A special case is the identification of potentially lethal recessive haplotypes in livestock populations without knowledge of the actual defect. The idea behind this is that haplotypes that carry a lethal recessive allele, and are therefore used as markers for the unknown causal defect allele, do not appear homozygous in vital populations. Methods, therefore, aim at identifying those haplotypes and test whether the missing homozygosity is non-random (VanRaden *et al.* 2011). Sensitive and accurate identification of (assumed to be overwhelmingly rare) lethal haplotypes thereby depends on very large sample sizes (Hoff *et al.* 2017), by now only available through routine genotyping of major breeds. Knowledge about those lethal haplotypes allows mating regimes that specifically avoid matings of two carrier individuals and thereby ensures that no affected offspring are produced (Hoff *et al.* 2017). Besides the avoidance of direct economic loss due to reduced fertility (VanRaden *et al.* 2011; Wobbe *et al.* 2019), this is also advised by animal welfare laws, as knowingly mating two carriers may be classified as torture breeding.

Population genomics

Population genomics in livestock generally fulfill different goals. Interests are typically in inferring knowledge about domestication history (Groenen *et al.* 2012; MacHugh *et al.* 2017; Orlando 2020; Wang *et al.* 2020), describing current population structures (Bortoluzzi *et al.* 2018; Malomane *et al.* 2019; Perini 2020), monitoring small populations (Bortoluzzi *et al.* 2018; Reimer *et al.* 2020; Schäler *et al.* 2020), or the characterization and delimitation of breeds (Upadhyay *et al.* 2019; Perini 2020; Reimer *et al.* 2020).

The exploration of a population's diversity commonly relates the population of interest to a comparable ideal population given the Wright-Fisher model. This model assumes an isolated random mating population with distinct generations and constant population size as a sample of an infinitely sized base population. It further disregards mutation and selection (Falconer and Mackay 1996). Any limitation of the size of a population will necessarily result in inbreeding, meaning that parents of an individual have at least one common ancestor. The two alleles at a locus then have the chance to be identical by descent (i.b.d.). As the handling of populations often differs from the idealized conditions of the Wright-Fisher model (e.g. by overlapping generations or non-random mating), the comparison between populations in regard to their size is commonly done by the effective population size (N_e). N_e describes the size of an ideal population with the same rate of inbreeding (ΔF) as observed from the population of interest ($N_e = 1/2\Delta F$; Falconer and Mackay 1996). As N_e is often used to define the risk status of livestock populations (e.g. for German livestock breeds; BMELV 2008), monitoring of inbreeding development is a routine task. Classical pedigree-based methods are thereby gradually replaced by marker-based methods. A relative straight-forward approach to derive inbreeding coefficients of individuals (F_x) is to set up a genomic relationship matrix (e.g. VanRaden 2007) and to extract them from the diagonal elements, which are $1 + F_x$. This, however, may sometimes be problematic, as the accurate scaling of the approach by VanRaden (2007) relies on allele frequencies of an unselected founder population. Further, note that this estimate describes identical by state (i.b.s.) instead of i.b.d. probabilities. Another way to estimate inbreeding is by runs of homozygosity (ROH). Longer homozygous stretches in the genome are signs of i.b.d. haplotypes (Broman and Weber 1999). The ROH-based inbreeding coefficient (F_{ROH}) is then the proportion of the autosomal genome covered by ROH (McQuillan *et al.* 2008). ROH are, due to recombination, shorter if the common ancestor of the parents can be found more distant in the pedigree (McQuillan *et al.* 2008). This allows setting length restrictions for ROH to trace inbreeding over time (McQuillan *et al.* 2008). The identification of ROH can thereby depend strongly on the density of the marker map (Herrero-

Medrano *et al.* 2014) and parameter settings for identification algorithms (e.g. MAF filtering or LD pruning; Meyermans *et al.* 2020).

Inbreeding in a population reduces the variance at loci. Two important measures that describe the variance, and thereby the diversity, at a locus, are expected (H_E) and observed heterozygosity (H_O) (Fernández and Bennewitz 2017). H_E defines the expected proportion of heterozygote samples given a certain allele frequency (p) and Hardy-Weinberg-Equilibrium (HWE). As $H_E = 2p(1-p)$, it equals the binominal variance and describes the expected allelic variance of a diploid individual at a certain locus. H_O as the observed state may deviate on average, if the population is not in HWE. Reasons may be non-random mating schemes or selection (Falconer and Mackay 1996).

Inferring information on population substructures can be done by Wright's F statistics (Wright 1949), which relate inbreeding coefficients of a structured population to the expectation given random mating. Weir and Cockerham (1984) described it slightly differently in a variance-analytic framework that relates the genomic variance of the total population, between subpopulations, between individuals within subpopulations, and between gametes within individuals to each other in a way to extract information on inbreeding and population subdivision. Population subdivision is thereby expressed through Wright's Fixation Index (F_{ST}), which relates the between subpopulation variance to the total variance. There exist multiple F_{ST} estimators, with $\widehat{\Theta}$ by Weir and Cockerham (1984) probably being the most widely accepted one, which however requires individual-level genotype data and therefore is not always usable. When F_{ST} is estimated from two populations, it can also be understood as distance between the two populations. However, other pairwise distance measures rather try to express differences in relation to coalescent times with different underlying model assumptions. So does e.g. Nei's distance (D; Nei 1972) assume constant mutation rates and Reynolds distance (Reynolds *et al.* 1983) a pure drift-only model. Distance measures are generally based on estimates of allele frequencies.

A problem of pairwise similarity/ distance measures is that they quickly create multidimensional spaces when multiple individuals/ populations are involved. Techniques of dimension reduction as prime component analysis (PCA) are therefore extensively used. PCA extracts a series of uncorrelated vectors based on a genetic covariance matrix, the eigenvectors or prime components (PC; Patterson *et al.* 2006). This rotates the observation space in a way that the first PC explains the maximum variation that can be explained in a one-dimensional space, the second one opens the two-dimensional space that explains as much variation as possible given the first PC and so on. The most famous example of a PCA is by Novembre *et al.* (2008) who were able to show that the first two PCs of a PCA on European humans were able to reflect the geographic sampling location. A problem of PCA is that uneven sampling strongly affects the projections (McVean 2009). Further, when used for a broad set of populations, PCA may hide information due to a common reduction to only two dimensions. An alternative to PCA is multidimensional scaling (MDS). MDS has the same intention as PCA, but is based on a pairwise distance matrix (Backhaus 2003; Li and Yu 2008). By this, it is more flexible than PCA as it allows the use of different distance measures.

A common population genetic question is whether a set of observed populations shares common ancestry and how they cluster in that sense. Clustering of individuals/ populations may be performed by reconstruction of phylogenetic trees. Based on genetic distance matrixes, classical hierarchical clustering methods like the unweighted pair group method with arithmetic mean (UPGMA) iteratively collapse the distance matrix for the least distant pair of populations and calculate new distances between the collapsed group and the other remaining population(group)s. In the case of UPGMA, the new distance is simply the arithmetic mean of the old distances. This is then graphically represented by a dendrogram whose branch lengths reflect the coalescence time if mutation rates are equal along all branches (Weir 1996). An alternative approach is the construction of a neighbor-joining tree (Saitou and Nei 1987). The neighbor-joining algorithm thereby starts from a star-like phylogeny and iteratively joins pairs of populations with the goal to minimize the total tree length. This results in an unrooted phylogenetic tree. Additionally, there exist maximum parsimony methods, which simply cluster populations by the least differences without obtaining branch lengths, and maximum likelihood methods, which search for the tree that shows the maximum likelihood given a specific evolutionary model (Weir 1996). A general problem of phylogenetic trees is that they depend on models of bifurcating trees and, by this, deny the role of hybridization in evolutionary and domestic processes. The Treemix (Pickrell and Pritchard 2012) method tries to overcome this issue by representing a phylogeny as a directed network graph. Further, Patterson’s D statistic (Green *et al.* 2010; Patterson *et al.* 2012) and related estimates of admixture fractions as implemented in the Dsuite tool (Malinsky *et al.* 2021) allow testing for hybridization events.

An alternative cluster approach, which is not based on a tree-like representation, is the STRUCTURE model by Pritchard *et al.* (2000). The model assumes a set of k unknown populations, characterized by their allele frequencies. It then (partly) assigns individuals to these unknown populations through a Bayesian clustering approach while simultaneously estimating the allele frequencies of the unknown populations. This results in a vector Q for each individual that specifies which proportion of the genome belongs to which of the k populations and, by this, allows for admixed individuals. As the computational effort for the original STRUCTURE method is relatively high (Novembre and Ramachandran 2011; Novembre 2016), nowadays default implementation is the faster maximum-likelihood-based ADMIXTURE algorithm by Alexander *et al.* (2009). A still existing problem is to find the ‘right’ number of k populations with different methods coming with unstable results (Novembre 2016). This often results in studies that exploratory examine different k to interpret the results (e.g. Malomane *et al.*

2019). Nevertheless, as the STRUCTURE model has some tight assumptions (e.g. linkage equilibrium between markers, HWE, special population histories; Pritchard *et al.* 2000), Lawson *et al.* (2018) showed that over-interpretation of the results may easily happen. Note that methods like HAPMIX (Price *et al.* 2009) and RFMix (Maples *et al.* 2013), that allow local ancestry estimations of admixed chromosomes based on haplotypes from phased reference populations, may also be of interest in this context.

SNP ascertainment bias and mitigation procedures

Besides already noted potential impacts of marker density, a major drawback of using array SNPs as markers for all kinds of genomic analyses is their non-random selection. As SNPs for arrays need to be selected before array production, the first step of array design is to screen public databases and/ or a limited set of sequenced discovery samples for potential SNPs.

Following Nielsen (2004), the unfolded allele frequency spectrum describes the probability to observe a certain number of mutant alleles at a certain locus in a population of n haplotypes. Assuming neutral evolution (Kimura 1991), the expected unfolded frequency spectrum (X) is defined as $P(X = x) = x^{-1}/\sum_{i=1}^{n-1}(1/i)$ $(0 < x < 1$; Nielsen 2004). Selection (ascertainment) of all variable SNPs in a subset of the total population then biases the spectrum towards more common alleles, whereby the strength of the bias increases with decreasing number of discovery samples (**Figure 1.2**). This bias is called SNP ascertainment bias (Nielsen 2004; Clark *et al.* 2005; Albrechtsen *et al.* 2010) and is present for each SNP array with different intensities. The bias can be further increased if multiple subpopulations are present and ascertainment is only performed in one of those subpopulations. While for the discovery population the effect of ascertainment bias is as described above, the shift towards common variants is less in all non-discovery populations with extremes resulting in a shift towards rare variants (Nielsen 2004). The strength of this effect is thereby strongly affected by the distance of the population to the discovery population (Dokan *et al.* 2021; Geibel *et al.* 2021b).

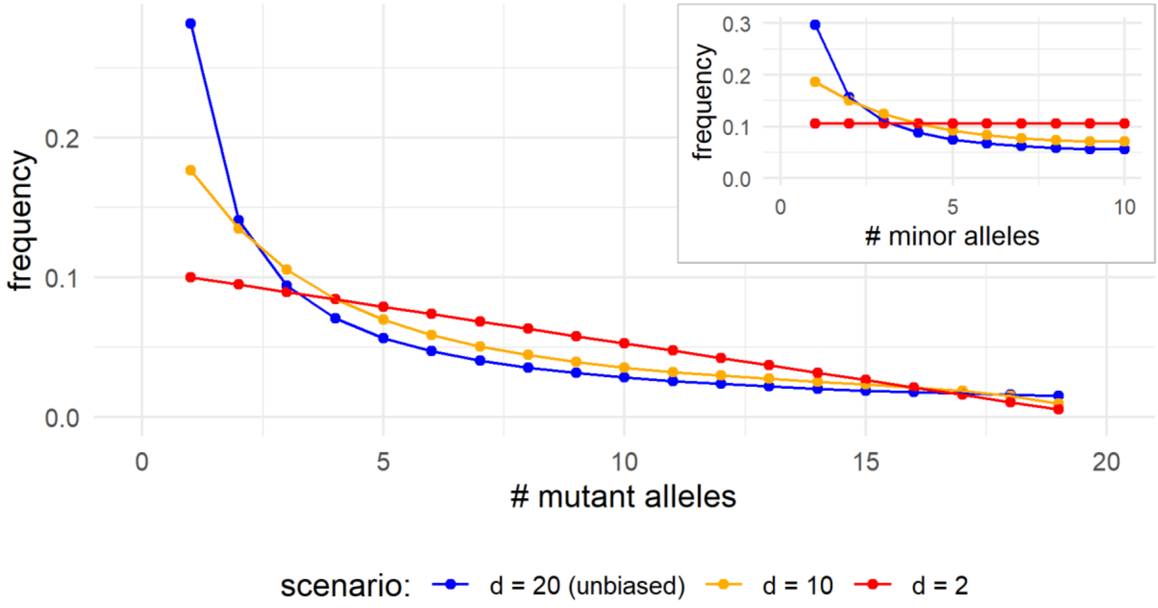

Figure 1.2: Expected allele frequency spectra under different ascertainment schemes. The spectra present the number of expected mutant (unfolded spectrum) and minor (folded spectrum; inset) alleles in a population of n = 20 haplotypes, assuming neutral molecular evolution. The three scenarios represent the unbiased case and ascertainment from discovery samples of d = 10 haplotypes vs. d = 2 haplotypes. The biased scenarios represent the case that the discovery samples are a subset of the typed samples (adapted from Nielsen 2004).

As many population genetic statistics rely on the allele frequency spectrum, they are directly affected by ascertainment bias with different intensities (Clark *et al.* 2005). The most direct impact of ascertainment bias is present for estimators that are directly based on the observed allele frequency spectrum, such as the neutrality test Tajima’s D (Tajima 1989; Ramirez-Soriano and Nielsen 2009) or estimates of heterozygosity (Rogers and Jorde 1996; Clark *et al.* 2005; Albrechtsen *et al.* 2010; Malomane *et al.* 2018; Geibel *et al.* 2021b). For example, Bradbury *et al.* (2011) observed that ascertainment bias decreased expected heterozygosity in Atlantic Cod by up to 30 % the further away the discovery population was.

When considering population differentiation, effects become less predictable, and different ascertainment schemes lead to different results (Dokan *et al.* 2021). The fact that common SNPs across different populations may be rather old variants (Wakeley *et al.* 2001) introduces biases towards lower population subdivision estimates when ascertainment is conducted independently in multiple subpopulations (Nielsen 2004; Dokan *et al.* 2021), or in a third population (Dokan *et al.* 2021). Upward biased population differentiation is also present when ascertainment bias affects the subpopulations differently strong (e.g. ascertainment in only one of the subpopulations; Dokan *et al.* 2021). In contrast, if an ascertainment scheme preferentially selects variants that are common in multiple populations, the subdivision will be overestimated (Dokan *et al.* 2021). Nevertheless, estimators are influenced by ascertainment bias to a different extent. So is F_{ST} less affected as an estimator compared to other distance estimators that are not scaled by overall heterozygosity when the numerator and denominator of F_{ST} are affected in the same direction (Albrechtsen *et al.* 2010; Geibel *et al.* 2021a).

Ascertainment that is performed unbalanced across subpopulations also rotates the principal components of a PCA (McVean 2009; Malomane *et al.* 2018; Dokan *et al.* 2021). The variation within the discovery populations, as well as differentiation between discovery and non-discovery populations, will be overestimated compared to variation within non-discovery populations (Nielsen 2004; Albrechtsen *et al.* 2010; Dokan *et al.* 2021), which has an effect comparable to uneven sampling (McVean 2009).

Common variants are on average older variants that had already time to recombine more often than younger variants (Clark *et al.* 2003; Nielsen and Signorovitch 2003). Ascertainment of medium frequent variants, therefore, results in an SNP panel that is older than an unbiased panel. This, in turn, means an underestimation of LD decay from frequency-independent estimators as $|D'|$ (Nielsen and Signorovitch 2003). Pairwise MAF differences, however, become on average smaller through ascertainment bias. This, in turn, inflates LD estimates by r^2 (Nielsen and Signorovitch 2003; Qanbari 2020), as the upper limit of r^2 is defined by the MAF difference (VanLiere and Rosenberg 2008).

Other than bivariate SNPs, polymorphic markers as microsatellites are less affected by ascertainment bias (Bradbury *et al.* 2011; Lachance and Tishkoff 2013). The same counts for haplotype-based estimators (Lachance and Tishkoff 2013).

To cope with ascertainment bias, it may be advisable to correct the allele frequency spectrum by reverse-engineering the ascertainment process (Nielsen *et al.* 2004; Clark *et al.* 2005; Albrechtsen *et al.* 2010), or account for ascertainment bias in the estimators (Nielsen 2000; Nielsen and Signorovitch 2003; Ramirez-Soriano and Nielsen 2009). Those methods, however, rely on simplified ascertainment schemes and require exact knowledge of the ascertainment process, which commonly conflicts with reality (Albrechtsen *et al.* 2010). More versatile is the attempt to model ascertainment within demographic simulations, as implemented into fastsimcoal2 (Excoffier *et al.* 2013) and used by McTavish and Hillis (2015) to test different combinations of demographic models and ascertainment schemes in cattle for the goodness of fit with observed data. Further, Quinto-Cortés *et al.* (2018) described a comparable method that implemented a Bayesian optimization process to automate the search for the best fitting demographic scenario.

However, with very broad demographic scenarios, these simulations also become too complex. Malomane *et al.* (2018) tested therefore how different filtering strategies affect ascertainment bias.

They identified LD-pruning as a promising approach, as it reduces redundant information of high-MAF SNPs while keeping the information of rare SNPs. Further, **Chapter 3** presents an approach to use imputation based on a sparse WGS reference panel to mitigate the effects of ascertainment bias.

The chicken

Origin and domestication

It is commonly accepted that the wild origin of the domesticated chicken (*Gallus gallus domesticus*) is the red jungle fowl (RJF) whose natural habitats stretch mainly across Southeast Asia. However, the amount of contribution of the five wild *Gallus gallus* subspecies (*G. g. gallus*, *G. g. spadiceus*, *G. g. bankiva*, *G. g. jabouillei*, *G. g. murghi*) is still content of scientific discourse. Some authors argue for multiple independent domestication events, as reviewed by Tixier-Boichard *et al.* (2011). However, the by now largest study on chicken domestication by Wang *et al.* (2020) argues based on 863 sequenced chickens that domestication of chickens was based on *G. g. spadiceus* ~9,500 ± 3,300 years ago with later introgression by the other four subspecies (*G. g. murghi* and *G. g. jabouillei* > *G. g. gallus* > *G. g. bankiva*). Note, however, that wild RJF samples, which are often even sampled in zoological parks, may not be free from hybridization with domestic chickens, as e.g. shown by Mariadassou *et al.* (2021). This could influence the results as well as a sampling bias in the study towards Asian chickens (Lawal and Hanotte 2021). Additionally, later introgression from other *Gallus* species into domesticated chicken populations seems also to have contributed significantly, as e.g. shown for the grey jungle fowl (*Gallus sonneratii*) from India that seems to be the origin of the yellow skin color of domestic chickens (Eriksson *et al.* 2008).

Dispersion across the world strongly followed human migration routes, as recently reviewed by Lawal and Hanotte (2021). The broad diversity of breeds may thereby have been shaped by multiple migration events and a rich crossbreeding history. This is e.g. reported from Europe, where many fancy breeds were developed by crossing imported Asian breeds to local chicken populations in the 19th century (Malomane *et al.* 2019).

Value in farming

The chicken is the agricultural vertebrate species with the most individuals worldwide (FAO 2021b). A strongly increasing trend in the reported number of chickens can be observed especially in Asia since the 1980s (**Figure 1.1**). For 2019, the Food and Agriculture Organization of the United Nations (FAO) reported 25.9 billion chickens worldwide, 15.8 billion (61 %) in Asia, and 5.9 billion (23 %) on the American continent. In contrast, the numbers have been only 2.2 billion on each of the two continents in 1980 (FAO 2021b).

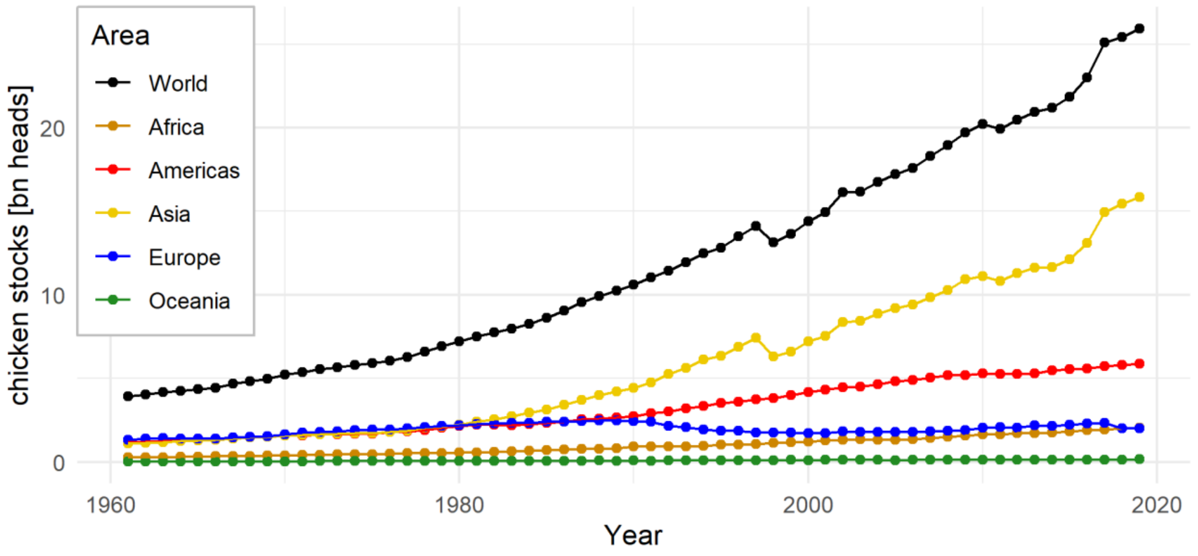

Figure 1.3: Worldwide chicken numbers by continent and year (data source: FAO 2021b).

Commercial chicken breeding is done by nucleus hybrid breeding schemes with strong horizontal and vertical concentrations of the market. Exemplarily for layers, a worldwide egg need of 900 billion eggs/year could be satisfied by a four-line crossing scheme with theoretically only 15,000 purebred grand-grand mothers (Preisinger 2018). This and the high costs of performance testing resulted in currently only four companies sharing the laying hen market (Preisinger 2018). The intensive breeding programs and negative correlations between growth and egg numbers (Willam and Simianer 2011) also led to a strong specialization of commercial lines for egg (white and brown layers) vs. meat production (broilers). Nevertheless, in developing countries backyard chicken farming with native chicken breeds still plays a significant role (e.g. ~50 % in the Philippines in 2005; Chang 2007).

Global chicken diversity

The limitation to few chicken lines in commercial meat and egg production contrasts with a large number of global chicken breeds. The Domestic Animal Diversity Information System (DAD-IS) currently lists 1,823 chicken breeds worldwide with 125 counting as extinct and 524 as at risk in at least one country (FAO 2021a). Breeding goals on a global scale extend the production of animal protein (e.g. game birds or a large diversity of fancy breeds; Crawford 1993). Further, Malomane *et al.* (2019) describe a gradual genetic separation between European and Asian breeds with African and South American breeds clustering in between.

The within-breed diversity of chickens exhibits a decline with genetic distance to the wild populations (Malomane *et al.* 2021) with European populations showing an on average lower diversity than Asian ones (Malomane *et al.* 2019; Malomane *et al.* 2021). The premature assumption that commercial populations generally exhibit very low levels of genetic diversity due to their intensive breeding history can thereby only be confirmed for white layers (Malomane *et al.* 2019). Commercial brown layers show a medium heterozygosity and broilers a rather high heterozygosity (Malomane *et al.* 2019). This is commonly explained by the single-breed origin of white layer lines in contrast to multi-breed origins of brown layers and broilers (Crawford 1993; Malomane *et al.* 2019; Tixier-Boichard 2020).

Genome

The chicken genome consists of 38 autosomes, a Z/W heterosomal sex system, and the mitochondrial genome, in total ~1.2 Gb. The first reference genome was published in 2004 based on a female from a red jungle fowl inbreeding line (International Chicken Genome Sequencing Consortium 2004). The initial build was successively updated through the last years and the current build GRCg6a (Genome Reference Consortium GRCg6a 2018) consists of 32 autosomes, the heterosomes, and the mitochondrial sequence.

A difference of avian genomes to mammalian genomes is the strong decay in chromosome lengths across the genome. Autosomes are therefore often divided into macro- (1-5), intermediate (6-11), and micro-chromosomes (12-38; International Chicken Genome Sequencing Consortium 2004). However, the exact classification varies across publications. The micro-chromosomes show several differences in comparison to macro-chromosomes. This includes elevated recombination rates (International Chicken Genome Sequencing Consortium 2004; Groenen *et al.* 2009; Megens *et al.* 2009), elevated rates of synonymous substitutions, higher GC content and gene density, and lower repeat density (International Chicken Genome Sequencing Consortium 2004). A further feature of the chicken genome is the known bad assembly quality of chromosome 16 due to the major histocompatibility (MHC) complex with a strong repetitive genome content (Solinhac *et al.* 2010; see also **Chapter 5**).

There are currently 23.4 M SNPs and short InDels published on ENSEMBL (**Table 1.1**). However, this seems to be an underestimation of the total number, as we already called >20 M bivariate SNPs just on chr1 – chr28 in our studies (**Chapter 3**). An accurate estimate of the number and length of chicken SVs is not yet available. Although studies identified up to 12,955 SV (Sohrabi *et al.* 2018) after filtering, the studies were all based on arrays or short-read data and limited to a single calling algorithm in a limited set of breeds, most likely lacking from a high number of false positives and low sensitivity at the same time.

For chickens, there exist currently four commercially available SNP arrays. The first array by Groenen *et al.* (2011) contains 60 k SNPs on the Illumina platform that were selected from reduced library sequences of four discovery populations (two broiler lines, a white layer line, and a brown layer line). Kranis *et al.* (2013) created a 580 k Affymetrix array. They used a broader discovery set, including multiple lines of white layers, brown layers, broilers, and inbreeding lines from the Roslin Institute, UK.

The SNPs for the array were further validated in a broad set of fancy breeds. The 55 k Affymetrix array by Liu *et al.* (2019) was developed with the intention to capture the variation of indigenous Chinese chicken breeds while still showing overlap with the previously existing arrays. Further, recently multispecies arrays with the purpose of monitoring small European populations were developed in the scope of the EU project IMAGE (https://www.imageh2020.eu/). The IMAGE001 multispecies array thereby contains ~10 k chicken SNPs (IMAGE 2020).

Aim of the thesis

The previous chapter highlighted the wide usability of genomic markers in livestock sciences. However, the different marker classes and technologies come with their specific properties and problems. Outstanding are especially the ascertainment bias of SNP arrays and the inaccurate SV calling pipelines. Further, the chicken is an excellent model organism in livestock sciences due to its broad diversity of populations. The thesis, therefore, aims in answering the following questions by using chicken data:

Chapter 2 asks which steps in the array design process created the SNP ascertainment bias. The question is answered by remodeling the design process of a commercial SNP array-based on WGS data.

Chapter 3 investigates whether imputation of array data to WGS level allows for *in silico* correction of SNP ascertainment bias.

Chapter 4 then assesses whether a separate SV calling is necessary for genomic studies, or whether potential effects of SV would already be captured by SNPs in strong LD to the SV.

References

Abyzov, A; Urban, AE; Snyder, M; Gerstein, M (2011): CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. In *Genome Res* 21 (6), pp. 974–984. DOI: 10.1101/gr.114876.110.

Akey, JM; Zhang, G; Zhang, K; Jin, L; Shriver, MD (2002): Interrogating a high-density SNP map for signatures of natural selection. In *Genome Res* 12 (12), pp. 1805–1814. DOI: 10.1101/gr.631202.

Albrechtsen, A; Nielsen, FC; Nielsen, R (2010): Ascertainment biases in SNP chips affect measures of population divergence. In *Mol Biol Evol* 27 (11), 2534–2547.

Alexander, DH; Novembre, J; Lange, K (2009): Fast model-based estimation of ancestry in unrelated individuals. In *Genome Res.* 19 (9), pp. 1655–1664. DOI: 10.1101/gr.094052.109.

Andrews, KR; Good, JM; Miller, MR; Luikart, G; Hohenlohe, PA (2016): Harnessing the power of RADseq for ecological and evolutionary genomics. In *Nat Rev Genet* 17 (2), p. 81.

Axelsson, E; Webster, MT; Smith, NGC; Burt, DW; Ellegren, H (2005): Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. In *Genome Res* 15 (1), pp. 120–125. DOI: 10.1101/gr.3021305.

Backhaus, K (2003): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 10., neu bearb. und erw. Aufl. Berlin: Springer (Springer-Lehrbuch).

Belyeu, JR; Chowdhury, M; Brown, J; Pedersen, BS; Cormier, MJ; Quinlan, AR; Layer, RM (2021): Samplot: a platform for structural variant visual validation and automated filtering. In *Genome Biol* 22 (1), p. 161. DOI: 10.1186/s13059-021-02380-5.

Belyeu, JR; Nicholas, TJ; Pedersen, BS; Sasani, TA; Havrilla, JM; Kravitz, SN et al. (2018): SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. In *GigaScience* 7 (7). DOI: 10.1093/gigascience/giy064.

Berry, DP; McClure, MC; Mullen, MP (2014): Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. In *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* 131 (3), pp. 165–172. DOI: 10.1111/jbg.12067.

Bertolotti, AC; Layer, RM; Gundappa, MK; Gallagher, MD; Pehlivanoglu, E; Nome, T et al. (2020): The structural variation landscape in 492 Atlantic salmon genomes. In *Nature Communications* 11 (1), p. 5176. DOI: 10.1038/s41467-020-18972-x.

Boichard, DA; Chung, H; Dassonneville, R; David, X; Eggen, A; Fritz, S et al. (2012): Design of a bovine low-density SNP array optimized for imputation. In *PLoS One* 7 (3), e34130.

Boitard, S; Schlötterer, C; Nolte, V; Pandey, RV; Futschik, A (2012): Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. In *Mol Biol Evol* 29 (9), pp. 2177–2186. DOI: 10.1093/molbev/mss090.

Bortoluzzi, C; Crooijmans, RPMA; Bosse, M; Hiemstra, SJ; Groenen, MAM; Megens, H-J (2018): The effects of recent changes in breeding preferences on maintaining traditional Dutch chicken genomic diversity. In *Heredity* 121 (6), pp. 564–578. DOI: 10.1038/s41437-018-0072-3.

Botstein, D; White, RL; Skolnick, M; Davis, RW (1980): Construction of a genetic linkage map in man using restriction fragment length polymorphisms. In *Am J Hum Genet* 32 (3), pp. 314–331.

Bouwman, AC; Derks, MF; Broekhuijse, ML; Harlizius, B; Veerkamp, RF (2020): Using short read sequencing to characterise balanced reciprocal translocations in pigs. DOI: 10.21203/rs.3.rs-28830/v3.

Bradbury, IR; Hubert, S; Higgins, B; Bowman, S; Paterson, IG; Snelgrove, PVR et al. (2011): Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, Gadus morhua. In *Mol Ecol Res* 11 (s1), pp. 218–225.

Broman, KW; Weber, JL (1999): Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. In *Am J Hum Genet* 65 (6), pp. 1493–1500. DOI: 10.1086/302661.

Brøndum, RF; Guldbrandtsen, B; Sahana, G; Lund, MS; Su, G (2014): Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. In *BMC Genomics* 15 (1), p. 728. DOI: 10.1186/1471-2164-15-728.

Browning, BL; Browning, SR (2009): A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. In *Am J Hum Genet* 84 (2), pp. 210–223. DOI: 10.1016/j.ajhg.2009.01.005.

Browning, BL; Zhou, Y; Browning, SR (2018): A One-Penny Imputed Genome from Next-Generation Reference Panels. In *Am J Hum Genet* 103 (3), pp. 338–348. DOI: 10.1016/j.ajhg.2018.07.015.

Bundesministerium für Ernährung,Landwirtschaft und Verbraucherschutz (Ed.) (2008): Tiergenetische Ressourcen in Deutschland. Nationales Fachprogramm zur Erhaltung und nachhaltigen Nutzung tiergenetischer Ressourcen in Deutschland. Available online at https://www.genres.de/fileadmin/SITE_MASTER/content/Publikationen/TGR__Nat._Fachprogramm.pdf, checked on 9/11/2021.

Carvalho, CMB; Pehlivan, D; Ramocki, MB; Fang, P; Alleva, B; Franco, LM et al. (2013): Replicative mechanisms for CNV formation are error prone. In *Nature Genetics* 45 (11), pp. 1319–1326. DOI: 10.1038/ng.2768.

Chang, H-S (2007): Analysis of the Philippine Chicken Industry: Commercial versus Backyard Sectors. In *Asian Journal of Agriculture and Development* 4 (1362-2016-107848). DOI: 10.22004/ag.econ.165854.

Chen, X; Listman, JB; Slack, FJ; Gelernter, J; Zhao, H (2012): Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples. In *Genet Epidemiol* 36 (6), pp. 549–560. DOI: 10.1002/gepi.21648.

Chowdhury, M; Layer, RM (2020): Learning What a Good Structural Variant Looks Like. In *bioRxiv*. DOI: 10.1101/2020.05.22.111260.

Clark, AG; Hubisz, MJ; Bustamante, CD; Williamson, SH; Nielsen, R (2005): Ascertainment bias in studies of human genome-wide polymorphism. In *Genome Res* 15 (11), pp. 1496–1502.

Clark, AG; Nielsen, R; Signorovitch, J; Matise, TC; Glanowski, S; Heil, J et al. (2003): Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. In *Am J Hum Genet* 73 (2), pp. 285–300.

Conrad, DF; Andrews, TD; Carter, NP; Hurles, ME; Pritchard, JK (2006): A high-resolution survey of deletion polymorphism in the human genome. In *Nature Genetics* 38 (1), pp. 75–81. DOI: 10.1038/ng1697.

Crawford, RD (1993): Poultry genetic resources. evolution, diversity and conservation. In Crawford, RD (Ed.): Poultry breeding and genetics. 2. print. Amsterdam: Elsevier (Developments in animal and veterinary science, 22).

Davey, JW; Hohenlohe, PA; Etter, PD; Boone, JQ; Catchen, JM; Blaxter, ML (2011): Genome-wide genetic marker discovery and genotyping using next-generation sequencing. In *Nat Rev Genet* 12 (7), pp. 499–510. DOI: 10.1038/nrg3012.

Delage, WJ; Thevenon, J; Lemaitre, C (2020): Towards a better understanding of the low recall of insertion variants with short-read based variant callers. In *BMC genomics* 21 (1), p. 762. DOI: 10.1186/s12864-020-07125-5.

Devlin, RH; Nagahama, Y (2002): Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. In *Aquaculture* 208 (3-4), pp. 191–364. DOI: 10.1016/S0044-8486(02)00057-1.

Dokan, K; Kawamura, S; Teshima, KM (2021): Effects of single nucleotide polymorphism ascertainment on population structure inferences. In *G3 (Bethesda, Md.)*. DOI: 10.1093/g3journal/jkab128.

Dröge, P (2015a): DNA-Replikation. Verdopplung der genetischen Information. In Nordheim, A, Knippers, R (Eds.): Molekulare Genetik. With assistance of Dröge, P, Meister, G, Schiebel, E, Vingron, M, Walter, J. 10., vollständig überarbeitete und erweiterte Auflage. Stuttgart, New York: Thieme, pp. 163–197.

Dröge, P (2015b): Mutationen, DNA-Schädigungen und DNA-Reperatur. In Nordheim, A, Knippers, R (Eds.): Molekulare Genetik. With assistance of Dröge, P, Meister, G, Schiebel, E, Vingron, M, Walter, J. 10., vollständig überarbeitete und erweiterte Auflage. Stuttgart, New York: Thieme, pp. 250–284.

Elshire, RJ; Glaubitz, JC; Sun, Q; Poland, JA; Kawamoto, K; Buckler, ES; Mitchell, SE (2011): A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. In *PLoS One* 6 (5), e19379.

Erbe, M; Gredler, B; Seefried, FR; Bapst, B; Simianer, H (2013): A function accounting for training set size and marker density to model the average accuracy of genomic prediction. In *PLoS One* 8 (12), e81046. DOI: 10.1371/journal.pone.0081046.

Erbe, M; Hayes, BJ; Matukumalli, LK; Goswami, S; Bowman, PJ; Reich, CM et al. (2012): Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. In *J. Dairy Sci.* 95 (7), pp. 4114–4129.

Eriksson, J; Larson, G; Gunnarsson, U; Bed'hom, B; Tixier-Boichard, M; Strömstedt, L et al. (2008): Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. In *PLoS Genet* 4 (2), e1000010. DOI: 10.1371/journal.pgen.1000010.

Excoffier, L; Dupanloup, I; Huerta-Sánchez, E; Sousa, VC; Foll, M (2013): Robust demographic inference from genomic and SNP data. In *PLoS Genet* 9 (10), e1003905. DOI: 10.1371/journal.pgen.1003905.

Falconer, DS; Mackay, TFC (1996): Introduction to quantitative genetics. 4th. Noida: Pearson.

Fay, JC; Wu, C-I (2000): Hitchhiking Under Positive Darwinian Selection. In *Genetics* 155 (3), pp. 1405–1413.

Fernández, J; Bennewitz, J (2017): Defining genetic diversity based on genomic tools. In Oldenbroek, K (Ed.): Genomic management of animal genetic diversity. Edited by: Kor Oldenbroek. Wageningen: Wageningen Academic Publishers, pp. 49–79.

Food and Agriculture Orgnization of the United Nations (FAO) (2021a): Domestic Animal Diversity Information System (DAD-IS). Available online at http://www.fao.org/dad-is/en/, checked on 8/18/2021.

Food and Agriculture Orgnization of the United Nations (FAO) (2021b): FAOSTAT. Available online at http://www.fao.org/faostat/en/#home, checked on 8/18/2021.

Fuller, CW; Middendorf, LR; Benner, SA; Church, GM; Harris, T; Huang, X et al. (2009): The challenges of sequencing by synthesis. In *Nat Biotechnol* 27 (11), pp. 1013–1023. DOI: 10.1038/nbt.1585.

Futschik, A; Schlötterer, C (2010): The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. In *Genetics* 186 (1), pp. 207–218. DOI: 10.1534/genetics.110.114397.

Garrison, E; Marth, G (2012): Haplotype-based variant detection from short-read sequencing. Available online at https://arxiv.org/pdf/1207.3907.

Gautier, M; Foucaud, J; Gharbi, K; Cézard, T; Galan, M; Loiseau, A et al. (2013): Estimation of population allele frequencies from next-generation sequencing data. Pool-versus individual-based genotyping. In *Mol Ecol* 22 (14), pp. 3766–3779. DOI: 10.1111/mec.12360.

Geibel, J; Reimer, C; Pook, T; Weigend, S; Weigend, A; Simianer, H (2021a): How imputation can mitigate SNP ascertainment Bias. In *BMC Genomics* 22 (1). DOI: 10.1186/s12864-021-07663-6.

Geibel, J; Reimer, C; Weigend, S; Weigend, A; Pook, T; Simianer, H (2021b): How array design creates SNP ascertainment bias. In *PLoS One* 16 (3), e0245178. DOI: 10.1371/journal.pone.0245178.

Genome Reference Consortium GRCg6a (2018): GRCg6a chicken reference genome. Available online at http://hgdownload.soe.ucsc.edu/goldenPath/galGal6/bigZips/galGal6.fa.gz, checked on 4/9/2019.

Gianola, D; los Campos, G de; Hill, WG; Manfredi, E; Fernando, R (2009): Additive genetic variability and the Bayesian alphabet. In *Genetics* 183 (1), pp. 347–363. DOI: 10.1534/genetics.109.103952.

Graves, JA; Watson, JM (1991): Mammalian sex chromosomes: evolution of organization and function. In *Chromosoma* 101 (2), pp. 63–68. DOI: 10.1007/bf00357055.

Green, RE; Krause, J; Briggs, AW; Maricic, T; Stenzel, U; Kircher, M et al. (2010): A draft sequence of the Neandertal genome. In *Science (New York, N.Y.)* 328 (5979), pp. 710–722. DOI: 10.1126/science.1188021.

Groenen, MAM; Archibald, AL; Uenishi, H; Tuggle, CK; Takeuchi, Y; Rothschild, MF et al. (2012): Analyses of pig genomes provide insight into porcine demography and evolution. In *Nature* 491 (7424), pp. 393–398. DOI: 10.1038/nature11622.

Groenen, MAM; Megens, H-J; Zare, Y; Warren, WC; Hillier, LW; Crooijmans, RPMA et al. (2011): The development and characterization of a 60K SNP chip for chicken. In *BMC Genomics* 12 (1), p. 274.

Groenen, MAM; Wahlberg, P; Foglio, M; Cheng, HH; Megens, H-J; Crooijmans, RPMA et al. (2009): A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. In *Genome Res* 19 (3), pp. 510–519.

Gu, W; Zhang, F; Lupski, JR (2008): Mechanisms for human genomic rearrangements. In *PathoGenetics* 1 (1), p. 4. DOI: 10.1186/1755-8417-1-4.

Gunderson, KL; Steemers, FJ; Lee, G; Mendoza, LG; Chee, MS (2005): A genome-wide scalable SNP genotyping assay using microarray technology. In *Nat Genet* 37 (5), pp. 549–554. DOI: 10.1038/ng1547.

Hansen, S; Faye, O; Sanabani, SS; Faye, M; Böhlken-Fascher, S; Faye, O et al. (2018): Combination random isothermal amplification and nanopore sequencing for rapid identification of the causative agent of an outbreak. In *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* 106, pp. 23–27. DOI: 10.1016/j.jcv.2018.07.001.

Herrero-Medrano, JM; Megens, H-J; Groenen, MAM; Bosse, M; Pérez-Enciso, M; Crooijmans, RPMA (2014): Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. In *BMC Genomics* 15 (1), p. 601. DOI: 10.1186/1471-2164-15-601.

Hickey, JM; Crossa, J; Babu, R; los Campos, G de (2012): Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. In *Crop Science* 52 (2), p. 654. DOI: 10.2135/cropsci2011.07.0358.

Hiendleder, S (2007): Mitochondrial DNA inheritance after SCNT. In *Advances in experimental medicine and biology* 591, pp. 103–116. DOI: 10.1007/978-0-387-37754-4_8.

Hiendleder, S; Lewalski, H; Janke, A (2008): Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. In *Cytogenet Genome Res* 120 (1-2), pp. 150–156. DOI: 10.1159/000118756.

Ho, SS; Urban, AE; Mills, RE (2019): Structural variation in the sequencing era. In *Nature Reviews Genetics*. DOI: 10.1038/s41576-019-0180-9.

Hoff, JL; Decker, JE; Schnabel, RD; Taylor, JF (2017): Candidate lethal haplotypes and causal mutations in Angus cattle. In *BMC Genomics* 18 (1), p. 799. DOI: 10.1186/s12864-017-4196-2.

Howe, KL; Achuthan, P; Allen, J; Allen, J; Alvarez-Jarreta, J; Amode, MR et al. (2021): Ensembl 2021. In *Nucleic Acids Res* 49 (D1), D884-D891. DOI: 10.1093/nar/gkaa942.

Howie, B; Fuchsberger, C; Stephens, M; Marchini, J; Abecasis, GR (2012): Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. In *Nature Genetics* 44 (8), pp. 955–959. DOI: 10.1038/ng.2354.

Howie, B; Marchini, J; Stephens, M (2011): Genotype imputation with thousands of genomes. In *G3 (Bethesda, Md.)* 1 (6), pp. 457–470. DOI: 10.1534/g3.111.001198.

Illumina (2015): BovineHD Genotyping BeadChip. More than 777,000 SNPs that deliver the densest coverage available for the bovine genome. Available online at https://www.illumina.com/documents/products/datasheets/datasheet_bovineHD.pdf, updated on 4/6/2015, checked on 10/30/2020.

Innovative Management of Animal Genetic Resources (IMAGE) (2020): DELIVERABLE D4.5. A standard multi-species chip for genomic assessment of collections. Available online at https://www.imageh2020.eu/deliverable/D4.5_resubmitted_final.pdf, updated on 3/1/2020, checked on 8/17/2021.

International Chicken Genome Sequencing Consortium (2004): Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. In *Nature* 432 (7018), pp. 695–716.

Itsara, A; Wu, H; Smith, JD; Nickerson, DA; Romieu, I; London, SJ; Eichler, EE (2010): De novo rates and selection of large copy number variation. In *Genome Res.* 20 (11), pp. 1469–1481. DOI: 10.1101/gr.107680.110.

Jain, M; Olsen, HE; Paten, B; Akeson, M (2016): The Oxford Nanopore MinION. Delivery of nanopore sequencing to the genomics community. In *Genome Biology* 17 (1), p. 239. DOI: 10.1186/s13059-016-1103-0.

Kimura, M (1968): Evolutionary rate at the molecular level. In *Nature* 217 (5129), pp. 624–626. DOI: 10.1038/217624a0.

Kimura, M (1991): The neutral theory of molecular evolution. A review of recent evidence. In *Idengaku zasshi* 66 (4), pp. 367–386.

Knippers, R (2015): DNA. Träger der genetischen Information. In Nordheim, A, Knippers, R (Eds.): Molekulare Genetik. With assistance of Dröge, P, Meister, G, Schiebel, E, Vingron, M, Walter, J. 10., vollständig überarbeitete und erweiterte Auflage. Stuttgart, New York: Thieme, pp. 29–49.

Koning, D-J de (2016): Meuwissen et al. on Genomic Selection. In *Genetics* 203 (1), pp. 5–7. DOI: 10.1534/genetics.116.189795.

Korkuć, P; Arends, D; Brockmann, GA (2019): Finding the Optimal Imputation Strategy for Small Cattle Populations. In *Frontiers in genetics* 10, p. 52. DOI: 10.3389/fgene.2019.00052.

Kosugi, S; Momozawa, Y; Liu, X; Terao, C; Kubo, M; Kamatani, Y (2019): Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. In *Genome Biology* 20 (1), p. 117. DOI: 10.1186/s13059-019-1720-5.

Kranis, A; Gheyas, AA; Boschiero, C; Turner, F; Le Yu; Smith, S et al. (2013): Development of a high density 600K SNP genotyping array for chicken. In *BMC Genomics* 14 (1), p. 59. DOI: 10.1186/1471-2164-14-59.

Kreiner-Møller, E; Medina-Gomez, C; Uitterlinden, AG; Rivadeneira, F; Estrada, K (2015): Improving accuracy of rare variant imputation with a two-step imputation approach. In *Eur J Hum Genet* 23 (3), pp. 395–400. DOI: 10.1038/ejhg.2014.91.

Kumar, J; Choudhary, AK; Solanki, RK; Pratap, A (2011): Towards marker-assisted selection in pulses: a review. In *Plant Breeding* 130 (3), pp. 297–313. DOI: 10.1111/j.1439-0523.2011.01851.x.

Lachance, J; Tishkoff, SA (2013): SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. In *Bioessays* 35 (9), pp. 780–786.

LaFramboise, T (2009): Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. In *Nucleic acids research* 37 (13), pp. 4181–4193. DOI: 10.1093/nar/gkp552.

Lander, ES; Linton, LM; Birren, B; Nusbaum, C; Zody, MC; Baldwin, J et al. (2001): Initial sequencing and analysis of the human genome. In *Nature* 409 (6822), pp. 860–921. DOI: 10.1038/35057062.

Lawal, RA; Hanotte, O (2021): Domestic chicken diversity: Origin, distribution, and adaptation. In *Anim Genet* 52 (4), pp. 385–394. DOI: 10.1111/age.13091.

Lawson, DJ; van Dorp, L; Falush, D (2018): A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. In *Nat Commun* 9 (1), p. 3258. DOI: 10.1038/s41467-018-05257-7.

Lee, Y-L; Bosse, M; Mullaart, E; Groenen, MAM; Veerkamp, RF; Bouwman, AC (2020): Functional and population genetic features of copy number variations in two dairy cattle populations. In *BMC genomics* 21 (1), p. 89. DOI: 10.1186/s12864-020-6496-1.

Li, H (2013): Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available online at http://arxiv.org/pdf/1303.3997v2, updated on 3/16/2013.

Li, H (2014): Toward better understanding of artifacts in variant calling from high-coverage samples. In *Bioinformatics (Oxford, England)* 30 (20), pp. 2843–2851. DOI: 10.1093/bioinformatics/btu356.

Li, Q; Yu, K (2008): Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. In *Genet Epidemiol* 32 (3), pp. 215–226. DOI: 10.1002/gepi.20296.

Li, Y-C; Korol, AB; Fahima, T; Beiles, A; Nevo, E (2002): Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. In *Mol Ecol* 11 (12), pp. 2453–2465. DOI: 10.1046/j.1365-294x.2002.01643.x.

Lin, P; Hartz, SM; Zhang, Z; Saccone, SF; Wang, J; Tischfield, JA et al. (2010): A new statistic to evaluate imputation reliability. In *PLoS One* 5 (3), e9697. DOI: 10.1371/journal.pone.0009697.

Liu, R; Xing, S; Wang, J; Zheng, M; Cui, H; Crooijmans, RPMA et al. (2019): A new chicken 55K SNP genotyping array. In *BMC genomics* 20 (1), p. 410. DOI: 10.1186/s12864-019-5736-8.

los Campos, G de; Pook, T; Gonzalez-Reymundez, A; Simianer, H; Mias, G; Vazquez, AI (2020): ANOVA-HD: Analysis of variance when both input and output layers are high-dimensional. In *PLoS One* 15 (12), e0243251. DOI: 10.1371/journal.pone.0243251.

Lush, JL (1933): The Bull Index Problem in the Light of Modern Genetics. In *J. Dairy Sci.* 16 (6), pp. 501–522. DOI: 10.3168/jds.S0022-0302(33)93369-X.

Lynch, M (2010): Evolution of the mutation rate. In *Trends in genetics : TIG* 26 (8), pp. 345–352. DOI: 10.1016/j.tig.2010.05.003.

Ma, Y; Ding, X; Qanbari, S; Weigend, S; Zhang, Q; Simianer, H (2015): Properties of different selection signature statistics and a new strategy for combining them. In *Heredity* 115 (5), pp. 426–436. DOI: 10.1038/hdy.2015.42.

MacHugh, DE; Larson, G; Orlando, L (2017): Taming the Past: Ancient DNA and the Study of Animal Domestication. In *Annual Review of Animal Biosciences* 5 (1), pp. 329–351. DOI: 10.1146/annurev-animal-022516-022747.

Mackay, I; Powell, W (2007): Methods for linkage disequilibrium mapping in crops. In *Trends in Plant Science* 12 (2), pp. 57–63. DOI: 10.1016/j.tplants.2006.12.001.

Malinsky, M; Matschiner, M; Svardal, H (2021): Dsuite - Fast D-statistics and related admixture evidence from VCF files. In *Mol Ecol Res* 21 (2), pp. 584–595. DOI: 10.1111/1755-0998.13265.

Malomane, DK; Reimer, C; Weigend, S; Weigend, A; Sharifi, AR; Simianer, H (2018): Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. In *BMC Genomics* 19 (1), p. 22. DOI: 10.1186/s12864-017-4416-9.

Malomane, DK; Simianer, H; Weigend, A; Reimer, C; Schmitt, AO; Weigend, S (2019): The SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high genomic resolution. In *BMC genomics* 20 (1), p. 345. DOI: 10.1186/s12864-019-5727-9.

Malomane, DK; Weigend, S; Schmitt, AO; Weigend, A; Reimer, C; Simianer, H (2021): Genetic diversity in global chicken breeds in relation to their genetic distances to wild populations. In *Genet Sel Evol* 53 (1), p. 36. DOI: 10.1186/s12711-021-00628-z.

Maples, BK; Gravel, S; Kenny, EE; Bustamante, CD (2013): RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. In *American Journal of Human Genetics* 93 (2), pp. 278–288. DOI: 10.1016/j.ajhg.2013.06.020.

Marchini, J; Howie, B; Myers, S; McVean, G; Donnelly, P (2007): A new multipoint method for genome-wide association studies by imputation of genotypes. In *Nat Genet* 39 (7), pp. 906–913.

Mardis, ER (2017): DNA sequencing technologies: 2006-2016. In *Nat Protoc* 12 (2), pp. 213–218. DOI: 10.1038/nprot.2016.182.

Mariadassou, M; Suez, M; Sathyakumar, S; Vignal, A; Arca, M; Nicolas, P et al. (2021): Unraveling the history of the genus Gallus through whole genome sequencing. In *Molecular Phylogenetics and Evolution* 158, p. 107044. DOI: 10.1016/j.ympev.2020.107044.

Matukumalli, LK; Lawley, CT; Schnabel, RD; Taylor, JF; Allan, MF; Heaton, MP et al. (2009): Development and characterization of a high density SNP genotyping assay for cattle. In *PLoS One* 4 (4), e5350.

McCarroll, SA; Hadnott, TN; Perry, GH; Sabeti, PC; Zody, MC; Barrett, JC et al. (2006): Common deletion polymorphisms in the human genome. In *Nature Genetics* 38 (1), pp. 86–92. DOI: 10.1038/ng1696.

McKenna, A; Hanna, M; Banks, E; Sivachenko, A; Cibulskis, K; Kernytsky, A et al. (2010): The Genome Analysis Toolkit. A MapReduce framework for analyzing next-generation DNA sequencing data. In *Genome Res* 20 (9), pp. 1297–1303. DOI: 10.1101/gr.107524.110.

McQuillan, R; Leutenegger, A-L; Abdel-Rahman, R; Franklin, CS; Pericic, M; Barac-Lauc, L et al. (2008): Runs of homozygosity in European populations. In *American Journal of Human Genetics* 83 (3), pp. 359–372. DOI: 10.1016/j.ajhg.2008.08.007.

McTavish, EJ; Hillis, DM (2015): How do SNP ascertainment schemes and population demographics affect inferences about population history? In *BMC Genomics* 16 (1), p. 1.

McVean, G (2009): A genealogical interpretation of principal components analysis. In *PLoS Genet* 5 (10), e1000686.

Megens, H-J; Crooijmans, RPMA; Bastiaansen, JWM; Kerstens, HHD; Coster, A; Jalving, R et al. (2009): Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. In *BMC Genet* 10 (1), p. 86. DOI: 10.1186/1471-2156-10-86.

Meuwissen, T; Hayes, B; Goddard, M (2016): Genomic selection: A paradigm shift in animal breeding. In *Animal Frontiers* 6 (1), pp. 6–14. DOI: 10.2527/af.2016-0002.

Meuwissen, THE (2009): Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. In *Genet Sel Evol* 41 (1), p. 35. DOI: 10.1186/1297-9686-41-35.

Meuwissen, THE; Hayes, BJ; Goddard, ME (2001): Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. In *Genetics* 157 (4), pp. 1819–1829. Available online at https://www.genetics.org/content/157/4/1819.

Meyermans, R; Gorssen, W; Buys, N; Janssens, S (2020): How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. In *BMC Genomics* 21 (1), p. 94. DOI: 10.1186/s12864-020-6463-x.

Miga, KH; Koren, S; Rhie, A; Vollger, MR; Gershman, A; Bzikadze, A et al. (2020): Telomere-to-telomere assembly of a complete human X chromosome. In *Nature* 585 (7823), pp. 79–84. DOI: 10.1038/s41586-020-2547-7.

Money, D; Gardner, K; Migicovsky, Z; Schwaninger, H; Zhong, G-Y; Myles, S (2015): LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. In *G3* 5 (11), p. 2383. DOI: 10.1534/g3.115.021667.

Mullis, K; Faloona, F; Scharf, S; Saiki, R; Horn, G; Erlich, H (1986): Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology* 51 Pt 1, pp. 263–273. DOI: 10.1101/sqb.1986.051.01.032.

Nakamura, K; Oshima, T; Morimoto, T; Ikeda, S; Yoshikawa, H; Shiwa, Y et al. (2011): Sequence-specific error profile of Illumina sequencers. In *Nucleic Acids Res* 39 (13), e90. DOI: 10.1093/nar/gkr344.

National Human Genome Institute (2020): The Cost of Sequencing a Human Genome. Available online at https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, updated on 12/7/2020, checked on 7/19/2021.

Nei, M (1972): Genetic Distance between Populations. In *The American Naturalist* 106 (949), pp. 283–292.

Nielsen, R (2000): Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms. In *Genetics* 154 (2), pp. 931–942. Available online at http://www.genetics.org/content/154/2/931.

Nielsen, R (2004): Population genetic analysis of ascertained SNP data. In *Hum Genomics* 1 (3), p. 1.

Nielsen, R (2005): Molecular signatures of natural selection. In *Annu Rev Genet* 39, pp. 197–218.

Nielsen, R; Hubisz, MJ; Clark, AG (2004): Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. In *Genetics* 168 (4), pp. 2373–2382.

Nielsen, R; Signorovitch, J (2003): Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. In *Theor Popul Biol* 63 (3), pp. 245–255.

Nolte, W; Kalm, E; Krattenmacher, N; Lehner, S; Reents, R; Stock, K et al. (2020): Nutzung der Imputation für den Übergang von der Mikrosatelliten- basierten Abstammungsüberprüfung zur SNP-Genotypisierung. In : 9. Pferdeworkshop. Bad Bevensen, Germany.

Nordheim, A (2015): Struktur eukaryotischer Gene. In Nordheim, A, Knippers, R (Eds.): Molekulare Genetik. With assistance of Dröge, P, Meister, G, Schiebel, E, Vingron, M, Walter, J. 10., vollständig überarbeitete und erweiterte Auflage. Stuttgart, New York: Thieme, pp. 285–303.

Novembre, J (2016): Pritchard, Stephens, and Donnelly on Population Structure. In *Genetics* 204 (2), pp. 391–393. DOI: 10.1534/genetics.116.195164.

Novembre, J; Johnson, T; Bryc, K; Kutalik, Z; Boyko, AR; Auton, A et al. (2008): Genes mirror geography within Europe. In *Nature* 456, 98. DOI: 10.1038/nature07331.

Novembre, J; Ramachandran, S (2011): Perspectives on human population structure at the cusp of the sequencing era. In *Annu Rev Genomics Hum Genet* 12, pp. 245–274.

Ober, U; Ayroles, JF; Stone, EA; Richards, S; Zhu, D; Gibbs, RA et al. (2012): Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. In *PLoS Genet* 8 (5), e1002685. DOI: 10.1371/journal.pgen.1002685.

Orlando, L (2020): Ancient Genomes Reveal Unexpected Horse Domestication and Management Dynamics. In *BioEssays* 42 (1), e1900164. DOI: 10.1002/bies.201900164.

Patterson, N; Moorjani, P; Luo, Y; Mallick, S; Rohland, N; Zhan, Y et al. (2012): Ancient admixture in human history. In *Genetics* 192 (3), pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

Patterson, N; Price, AL; Reich, D (2006): Population structure and eigenanalysis. In *PLoS Genet* 2 (12), e190. DOI: 10.1371/journal.pgen.0020190.

Pausch, H; Aigner, B; Emmerling, R; Edel, C; Götz, K-U; Fries, R (2013): Imputation of high-density genotypes in the Fleckvieh cattle population. In *Genetics, selection, evolution : GSE* 45, p. 3. DOI: 10.1186/1297-9686-45-3.

Perez-Enciso, M; Rincon, JC; Legarra, A (2015): Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. In *Genet Sel Evol* 47, p. 43. DOI: 10.1186/s12711-015-0117-5.

Perini, F (2020): Morphological and genetic characterization of 13 Italian local chicken breeds. In *Acta fytotechn zootechn* 23 (Monothematic Issue), pp. 137–143. DOI: 10.15414/afz.2020.23.mi-fpap.137-143.

Peripolli, E; Reimer, C; Ha, N-T; Geibel, J; Machado, MA; Panetto, João Cláudio do Carmo et al. (2020): Genome-wide detection of signatures of selection in indicine and Brazilian locally adapted taurine cattle breeds using whole-genome re-sequencing data. In *BMC Genomics* 21 (1), p. 624. DOI: 10.1186/s12864-020-07035-6.

Pickrell, JK; Pritchard, JK (2012): Inference of population splits and mixtures from genome-wide allele frequency data. In *PLoS Genet* 8 (11), e1002967. DOI: 10.1371/journal.pgen.1002967.

Pook, T; Mayer, M; Geibel, J; Weigend, S; Cavero, D; Schoen, CC; Simianer, H (2019): Improving Imputation Quality in BEAGLE for Crop and Livestock Data. In *G3*, g3.400798.2019. DOI: 10.1534/g3.119.400798.

Pook, T; Nemri, A; Gonzalez Segovia, EG; Simianer, H; Schoen, C-C (2021): Increasing calling accuracy, coverage, and read depth in sequence data by the use of haplotype blocks. DOI: 10.1101/2021.01.07.425688.

Preisinger, R (2018): Struktur der Legehennenzucht weltweit. In Zentralverband der Deutschen Geflügelwirtschaft e.V.: Geflügeljahrbuch 2019. Schwerpunkt: Biosicherheit und Hygiene. With assistance of Damme, K, Mayer, A. Stuttgart: Ulmer (Geflügeljahrbuch, 2019), pp. 78–85.

Price, AL; Tandon, A; Patterson, N; Barnes, KC; Rafaels, N; Ruczinski, I et al. (2009): Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. In *PLoS Genet* 5 (6), e1000519. DOI: 10.1371/journal.pgen.1000519.

Pritchard, JK; Stephens, M; Donnelly, P (2000): Inference of Population Structure Using Multilocus Genotype Data. In *Genetics* 155 (2), pp. 945–959. DOI: 10.1093/genetics/155.2.945.

Purvis, A; Bromham, L (1997): Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. In *J Mol Evol* 44 (1), pp. 112–119. DOI: 10.1007/pl00006117.

Qanbari, S (2020): On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. In *Frontiers in genetics* 10, p. 1304. DOI: 10.3389/fgene.2019.01304.

Qanbari, S; Rubin, C-J; Maqbool, K; Weigend, S; Weigend, A; Geibel, J et al. (2019): Genetics of adaptation in modern chicken. In *PLoS Genet* 15 (4), e1007989. DOI: 10.1371/journal.pgen.1007989.

Qanbari, S; Simianer, H (2014): Mapping signatures of positive selection in the genome of livestock. In *Livestock Science* 166, pp. 133–143. DOI: 10.1016/j.livsci.2014.05.003.

Quick, J; Loman, NJ; Duraffour, S; Simpson, JT; Severi, E; Cowley, L et al. (2016): Real-time, portable genome sequencing for Ebola surveillance. In *Nature* 530 (7589), pp. 228–232. DOI: 10.1038/nature16996.

Quinto-Cortés, CD; Woerner, AE; Watkins, JC; Hammer, MF (2018): Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. In *Sci Rep* 8 (1), p. 10209. DOI: 10.1038/s41598-018-28539-y.

Ramirez-Soriano, A; Nielsen, R (2009): Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. In *Genetics* 181 (2), pp. 701–710. DOI: 10.1534/genetics.108.094060.

Raudsepp, T; Chowdhary, BP (2015): The Eutherian Pseudoautosomal Region. In *Cytogenet Genome Res* 147 (2-3), pp. 81–94. DOI: 10.1159/000443157.

Reimer, C; Ha, N-T; Sharifi, AR; Geibel, J; Mikkelsen, LF; Schlather, M et al. (2020): Assessing breed integrity of Göttingen Minipigs. In *BMC Genomics* 21 (1), p. 308. DOI: 10.1186/s12864-020-6590-4.

Reynolds, J; Weir, BS; Cockerham, CC (1983): Estimation of the Coancestry Coefficient. Basis for a Short-Term Genetic Distance. In *Genetics* 105 (3), pp. 767–779. Available online at http://www.genetics.org/content/genetics/105/3/767.full.pdf.

Rhoads, A; Au, KF (2015): PacBio Sequencing and Its Applications. In *Genomics, Proteomics & Bioinformatics* 13 (5), pp. 278–289. DOI: 10.1016/j.gpb.2015.08.002.

Rogers, AR; Jorde, LB (1996): Ascertainment bias in estimates of average heterozygosity. In *Am J Hum Genet* 58 (5), pp. 1033–1041.

Roshyara, NR; Scholz, M (2015): Impact of genetic similarity on imputation accuracy. In *BMC Genet* 16 (1), p. 90. DOI: 10.1186/s12863-015-0248-2.

Ross, MG; Russ, C; Costello, M; Hollinger, A; Lennon, NJ; Hegarty, R et al. (2013): Characterizing and measuring bias in sequence data. In *Genome Biol* 14 (5), R51. DOI: 10.1186/gb-2013-14-5-r51.

Rowan, TN; Hoff, JL; Crum, TE; Taylor, JF; Schnabel, RD; Decker, JE (2019): A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. In *Genetics Selection Evolution* 51 (1), p. 77. DOI: 10.1186/s12711-019-0519-x.

Sabeti, PC; Reich, DE; Higgins, JM; Levine, HZP; Richter, DJ; Schaffner, SF et al. (2002): Detecting recent positive selection in the human genome from haplotype structure. In *Nature* 419, 832 EP -. DOI: 10.1038/nature01140.

Saitou, N; Nei, M (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. In *mbe* 4 (4), pp. 406–425. DOI: 10.1093/oxfordjournals.molbev.a040454.

Sanger, F; Nicklen, S; Coulson, AR (1977): DNA sequencing with chain-terminating inhibitors. In *Proc Natl Acad Sci U S A* 74 (12), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

Sargolzaei, M; Chesnais, JP; Schenkel, FS (2014): A new approach for efficient genotype imputation using information from relatives. In *BMC Genomics* 15 (1), p. 478. DOI: 10.1186/1471-2164-15-478.

Schaeffer, LR (1991): C. R. Henderson: Contributions to Predicting Genetic Merit. In *J. Dairy Sci.* 74 (11), pp. 4052–4066. DOI: 10.3168/jds.S0022-0302(91)78601-3.

Schaeffer, LR (2006): Strategy for applying genome-wide selection in dairy cattle. In *Journal of animal breeding and genetics* 123 (4), pp. 218–223. DOI: 10.1111/j.1439-0388.2006.00595.x.

Schäler, J; Krüger, B; Thaller, G; Hinrichs, D (2020): Comparison of ancestral, partial, and genomic inbreeding in a local pig breed to achieve genetic diversity. In *Conservation Genet Resour* 12 (1), pp. 77–86. DOI: 10.1007/s12686-018-1057-5.

Schlötterer, C; Tobler, R; Kofler, R; Nolte, V (2014): Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. In *Nat Rev Genet* 15 (11), pp. 749–763.

Smeds, L; Kawakami, T; Burri, R; Bolivar, P; Husby, A; Qvarnström, A et al. (2014): Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. In *Nat Commun* 5, 5448 EP -. DOI: 10.1038/ncomms6448.

Smith, JM; Haigh, J (1974): The hitch-hiking effect of a favourable gene. In *Genet Res* 23 (1), pp. 23–35. DOI: 10.1017/s0016672300014634.

Sohrabi, SS; Mohammadabadi, M; Wu, D-D; Esmailizadeh, A (2018): Detection of breed-specific copy number variations in domestic chicken genome. In *Genome* 61 (1), pp. 7–14. DOI: 10.1139/gen-2017-0016.

Solinhac, R; Leroux, S; Galkina, S; Chazara, O; Feve, K; Vignoles, F et al. (2010): Integrative mapping analysis of chicken microchromosome 16 organization. In *BMC Genomics* 11 (1), p. 616. DOI: 10.1186/1471-2164-11-616.

Steemers, FJ; Chang, W; Lee, G; Barker, DL; Shen, R; Gunderson, KL (2006): Whole-genome genotyping with the single-base extension assay. In *Nat Methods* 3 (1), pp. 31–33. DOI: 10.1038/nmeth842.

Stevens, L (1997): Sex chromosomes and sex determining mechanisms in birds. In *Science progress* 80 (Pt 3), pp. 197–216.

Sved, JA (1971): Linkage disequilibrium and homozygosity of chromosome segments in finite populations. In *Theor Popul Biol* 2 (2), pp. 125–141. DOI: 10.1016/0040-5809(71)90011-6.

Tajima, F (1989): Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. In *Genetics* 123 (3), pp. 585–595.

Tixier-Boichard, M (2020): From the jungle fowl to highly performing chickens: are we reaching limits? In *World Poultry Sci J* 76 (1), pp. 2–17. DOI: 10.1080/00439339.2020.1729676.

Tixier-Boichard, M; Bed'hom, B; Rognon, X (2011): Chicken domestication. From archeology to genomics. In *Comptes rendus biologies* 334 (3), pp. 197–204.

Unterseer, S; Bauer, E; Haberer, G; Seidel, M; Knaak, C; Ouzunova, M et al. (2014): A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. In *BMC Genomics* 15 (1), p. 823.

Upadhyay, M; Bortoluzzi, C; Barbato, M; Ajmone-Marsan, P; Colli, L; Ginja, C et al. (2019): Deciphering the patterns of genetic admixture and diversity in southern European cattle using genome-wide SNPs. In *Evolutionary applications* 12 (5), pp. 951–963. DOI: 10.1111/eva.12770.

van Binsbergen, R; Bink, MC; Calus, MP; van Eeuwijk, FA; Hayes, BJ; Hulsegge, I; Veerkamp, RF (2014): Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. In *Genet Sel Evol* 46 (1), p. 41. DOI: 10.1186/1297-9686-46-41.

van Binsbergen, R; Calus, MPL; Bink, MCAM; van Eeuwijk, FA; Schrooten, C; Veerkamp, RF (2015): Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. In *Genet Sel Evol* 47 (1), p. 71. DOI: 10.1186/s12711-015-0149-x.

van der Auwera, GA; Carneiro, MO; Hartl, C; Poplin, R; Del Angel, G; Levy-Moonshine, A et al. (2013): From FastQ data to high confidence variant calls. The Genome Analysis Toolkit best practices pipeline. In *Current protocols in bioinformatics* 43 (1), 11.10.1-11.10.33. DOI: 10.1002/0471250953.bi1110s43.

VanLiere, JM; Rosenberg, NA (2008): Mathematical properties of the r2 measure of linkage disequilibrium. In *Theoretical population biology* 74 (1), pp. 130–137.

VanRaden, PM (2007): Genomic measures of relationship and inbreeding. In *INTERBULL bulletin* (37), p. 33. Available online at https://journal.interbull.org/index.php/ib/article/view/981/972.

VanRaden, PM (2008): Efficient Methods to Compute Genomic Predictions. In *J. Dairy Sci.* 91 (11), pp. 4414–4423. DOI: 10.3168/jds.2007-0980.

VanRaden, PM; Null, DJ; Sargolzaei, M; Wiggans, GR; Tooker, ME; Cole, JB et al. (2013): Genomic imputation and evaluation using high-density Holstein genotypes. In *J. Dairy Sci.* 96 (1), pp. 668–678. DOI: 10.3168/jds.2012-5702.

VanRaden, PM; Olson, KM; Null, DJ; Hutchison, JL (2011): Harmful recessive effects on fertility detected by absence of homozygous haplotypes. In *J. Dairy Sci.* 94 (12), pp. 6153–6161. DOI: 10.3168/jds.2011-4624.

Visscher, PM; Brown, MA; McCarthy, MI; Yang, J (2012): Five years of GWAS discovery. In *Am J Hum Genet* 90 (1), pp. 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

Vitti, JJ; Grossman, SR; Sabeti, PC (2013): Detecting natural selection in genomic data. In *Annu Rev Genet* 47, pp. 97–120. DOI: 10.1146/annurev-genet-111212-133526.

Wakchaure, R; Ganguly, S (2015): Marker Assisted Selection (MAS) in Animal Breeding: A Review. In *J Drug Metab Toxicol* 06 (05). DOI: 10.4172/2157-7609.1000e127.

Wakeley, J; Nielsen, R; Liu-Cordero, SN; Ardlie, K (2001): The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. In *Am J Hum Genet* 69 (6), pp. 1332–1347. DOI: 10.1086/324521.

Wan, L; Sun, K; Ding, Q; Cui, Y; Li, M; Wen, Y et al. (2009): Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. In *Nucleic Acids Res* 37 (17), e117. DOI: 10.1093/nar/gkp559.

Wang, K; Li, M; Hadley, D; Liu, R; Glessner, J; Grant, SFA et al. (2007): PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. In *Genome Res* 17 (11), pp. 1665–1674. DOI: 10.1101/gr.6861907.

Wang, M-S; Thakur, M; Peng, M-S; Jiang, Y; Frantz, LAF; Li, M et al. (2020): 863 genomes reveal the origin and domestication of chicken. In *Cell Res* 30 (8), pp. 693–701. DOI: 10.1038/s41422-020-0349-y.

Watson, JD; Crick, FH (1953): Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. In *Nature* 171 (4356), pp. 737–738. DOI: 10.1038/171737a0.

Weir, BS (1996): Genetic data analysis. Methods for discrete population genetic data. 2. ed., [rev. and expanded]. Sunderland, Mass.: Sinauer.

Weir, BS; Cockerham, CC (1984): Estimating F-statistics for the analysis of population structure. In *Evolution*, pp. 1358–1370.

Wenger, AM; Peluso, P; Rowell, WJ; Chang, P-C; Hall, RJ; Concepcion, GT et al. (2019): Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. In *Nat Biotechnol* 37 (10), pp. 1155–1162. DOI: 10.1038/s41587-019-0217-9.

Wick, RR; Judd, LM; Holt, KE (2019): Performance of neural network basecalling tools for Oxford Nanopore sequencing. In *Genome Biol* 20 (1), p. 129. DOI: 10.1186/s13059-019-1727-y.

Willam, A; Simianer, H (2011): Tierzucht. Grundwissen Bachelor ; 54 Tabellen. Stuttgart: Ulmer (UTB Argrarwissenschaften, 3526).

Wobbe, M; Reinhardt, F; Stock, KF; Reents, R (2019): Wege zum angemessenen Umgang mit WFFS und anderen genetischen Eigenschaften. Available online at https://www.vit.de/fileadmin/DE/Zuchtwertschaetzung/InformationsartikelWFFS.vit20190412.pdf, updated on 4/12/2019, checked on 8/16/2021.

Wright, S (1949): The genetical structure of populations. In *Ann Eugen* 15 (1), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.

Ye, S; Yuan, X; Huang, S; Zhang, H; Chen, Z; Li, J et al. (2019): Comparison of genotype imputation strategies using a combined reference panel for chicken population. In *Animal : an international journal of animal bioscience* 13 (6), pp. 1119–1126. DOI: 10.1017/S1751731118002860.

Yoder, AD; Tiley, GP (2021): The challenge and promise of estimating the de novo mutation rate from whole-genome comparisons among closely related individuals. In *Mol Ecol*. DOI: 10.1111/mec.16007.

Zarate, S; Carroll, A; Mahmoud, M; Krasheninina, O; Jun, G; Salerno, WJ et al. (2020): Parliament2: Accurate structural variant calling at scale. In *GigaScience* 9 (12). DOI: 10.1093/gigascience/giaa145.

Zhang, Z; Xiao, X; Zhou, W; Zhu, D; Amos, CI (2021): False positive findings during genome-wide association studies with imputation: Influence of allele frequency and imputation accuracy. In *Hum Mol Genet*. DOI: 10.1093/hmg/ddab203.

Chapter 2

How Array Design creates SNP Ascertainment Bias

Johannes Geibel[12], Christian Reimer[12], Steffen Weigend[23], Annett Weigend[3], Torsten Pook[12], Henner Simianer[12]

[1] University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[2] University of Goettingen, Center for Integrated Breeding Research, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[3] Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, Höltystrasse 10, 31535 Neustadt-Mariensee, Germany

Published in *PLoS One* 16(3): e0245178
https://doi.org/10.1371/journal.pone.0245178

Abstract

Single nucleotide polymorphisms (SNPs), genotyped with arrays, have become a widely used marker type in population genetic analyses over the last 10 years. However, compared to whole genome re-sequencing data, arrays are known to lack a substantial proportion of globally rare variants and tend to be biased towards variants present in populations involved in the development process of the respective array. This affects population genetic estimators and is known as SNP ascertainment bias. We investigated factors contributing to ascertainment bias in array development by redesigning the Axiom™ Genome-Wide Chicken Array *in silico* and evaluating changes in allele frequency spectra and heterozygosity estimates in a stepwise manner. A sequential reduction of rare alleles during the development process was shown. This was mainly caused by the identification of SNPs in a limited set of populations and a within-population selection of common SNPs when aiming for equidistant spacing. These effects were shown to be less severe with a larger discovery panel. Additionally, a generally massive overestimation of expected heterozygosity for the ascertained SNP sets was shown. This overestimation was 24 % higher for populations involved in the discovery process than not involved populations in case of the original array. The same was observed after the SNP discovery step in the redesign. However, an unequal contribution of populations during the SNP selection can mask this effect but also adds uncertainty. Finally, we make suggestions for the design of specialized arrays for large scale projects where whole genome re-sequencing techniques are still too expensive.

Introduction

Starting in the first decade of this century, the possibility of cost-efficiently genotyping high numbers of Single Nucleotide Polymorphisms (SNP) for many individuals in parallel via SNP arrays led to an increase in their usage for population genetic analyses in humans (Novembre *et al.* 2008; Patterson *et al.* 2012), model species (Laurie *et al.* 2007; Platt *et al.* 2010), plants (Travis *et al.* 2015; Mayer *et al.* 2017) and livestock (Muir *et al.* 2008; Gibbs *et al.* 2009; Kijas *et al.* 2009; Gautier *et al.* 2010; Qanbari *et al.* 2010; McTavish *et al.* 2013; Malomane *et al.* 2019).

Various SNP arrays exist for humans (Perkel 2008), plants (Unterseer *et al.* 2014; Singh *et al.* 2015) and all major livestock species (Matukumalli *et al.* 2009; Ramos *et al.* 2009; Groenen *et al.* 2011; Boichard *et al.* 2012; Kranis *et al.* 2013; Tosser-Klopp *et al.* 2014; Sandenbergh *et al.* 2016). SNP numbers within these arrays range from 10 k SNPs (Boichard *et al.* 2012) over approximately 50 k (Matukumalli *et al.* 2009; Groenen *et al.* 2011; Tosser-Klopp *et al.* 2014; Singh *et al.* 2015) up to 600 k (Kranis *et al.* 2013; Unterseer *et al.* 2014). The design process of every array has an initial step of SNP discovery in common, where SNPs are identified from existing databases and/or from a small set of sequenced individuals. SNPs are then selected based on different quality criteria like minor allele frequency (MAF) thresholds and platform specific design scores (Fan *et al.* 2010). Additional criteria like equidistant spacing over the genome (Kranis *et al.* 2013), overrepresentation of some areas like chromosomal ends to increase imputation accuracy (Boichard *et al.* 2012) or genic regions (Kranis *et al.* 2013), or increased overrepresentation of high MAF SNPs (Matukumalli *et al.* 2009) are applied dependent on the design intentions. In the end, draft arrays are validated either on the set of populations used for the SNP discovery itself (Ramos *et al.* 2009) and/or on a broad set of individuals from different populations (Fan *et al.* 2010; Kranis *et al.* 2013).

In contrast to whole genome re-sequencing (WGS) data, SNP arrays often show a clear underrepresentation of SNPs with extreme allele frequencies (Nielsen 2004). As population genetic statistics are mostly based on estimates of allele frequencies, this context leads to biased population genetic estimators (Nielsen 2004; Clark *et al.* 2005) and is known as SNP ascertainment bias.

The absence of rare alleles is mainly driven by two factors in the array design process where SNPs are selected (ascertained) based on different requirements and decisions (Eller 2001). The first factor is a relatively small panel of individuals being used for discovery of SNPs, leading to a large proportion of globally rare variants not being selected, since they appear monomorphic in the discovery panel (Nielsen and Signorovitch 2003; Clark *et al.* 2005). The second factor is the across population use of arrays. Arrays are developed based on the variation within the discovery panel, thus missing variation present in distantly related individuals or populations (Eller 2001; Nielsen 2004). This second source of bias was shown to be of relatively high importance for livestock studies, where arrays are usually developed for large commercial breeds and later used to genotype diverse sets of local breeds all over the world (McTavish and Hillis 2015; Malomane *et al.* 2018).

Besides different strategies to minimize the impact of ascertainment bias (Lachance and Tishkoff 2013; Malomane *et al.* 2018), there are some attempts to correct the allele frequency spectrum via Bayesian methods (Nielsen and Signorovitch 2003; Nielsen 2004; Nielsen *et al.* 2004). However, those corrections highly rely on detailed statistical assumptions of the ascertainment process (Guillot and Foll 2009; Albrechtsen *et al.* 2010) or take a variety of ascertainment processes and demographic patterns into account to model evolutionary scenarios which are then compared to real world data (McTavish and Hillis 2015; Quinto-Cortés *et al.* 2018). However, those methods are currently only tested for corrections of the first source of ascertainment bias, the small discovery panel (Nielsen and Signorovitch 2003; Nielsen 2004; Nielsen *et al.* 2004). Additionally, detailed information on the design process is limited in practice (Albrechtsen *et al.* 2010) and the complexity of the processes makes statistical models for the corrections inaccurate.

Agricultural species such as chickens often show a complex domestication history, and therefore allow for few prior assumptions on ascertainment bias. Domestic chickens are assumed to originate from red jungle fowl (*Gallus gallus*) ancestors in Southeast Asia (West and Zhou 1988; Lawal *et al.* 2020), represented by the five subspecies *G. g. gallus*, *G. g. spadiceus*, *G. g. murghi*, *G. g. bankiva* and *G. g. jabouillei* (Tixier-Boichard *et al.* 2011). Additionally, some hybridization events with other *Gallus* species (e.g. grey jungle fowl; *Gallus sonneratii) have been suggested (Eriksson et al. 2008; Lawal et al. 2020).* The diversity of today's local breeds of chickens in Europe originates from chickens that reached the continent about 3000 years ago via a northern and a southern route, followed by selection and crossing with Asian chicken breeds introduced in the 19th century (Tixier-Boichard *et al.* 2011). While commercial white layers were derived solely by intensive directional selection of a single breed, the White Leghorn, commercial brown layers are derived from a broader genetic basis (e.g. Rhode Island Red, New Hampshire, Barred Plymouth Rock). Commercial broilers are derived by cross-breeding of paternal lines (e.g. White Cornish) with maternal lines which descend from a comparable basis as brown layers (e.g. White Plymouth Rock) (Crawford 1993). For more detailed information on chicken ancestry we refer to Lawal et al. (2020) and for a comprehensive overview on diversity and population structure of domesticated chickens to Malomane et al. (2019).

Given the complexity of modern array design processes and the chicken population structure, this study aims at highlighting the mechanisms which promote the bias by illustrating the effects of the different steps of the array design process on the allele frequency spectrum, using real data in a typical setting from livestock sciences. For this purpose, the design process of the Axiom™ Genome-Wide Chicken Array (Kranis *et al.* 2013) was simulated in a set of diverse chicken WGS data. Allele frequency spectra as well as expected heterozygosity (H_{exp}) were compared to the WGS data and the SNPs of the Axiom™ Genome-Wide Chicken Array. Finally, some recommendations are made to design an array for monitoring genetic diversity.

Material and methods

Ethics approval and consent to participate

DNA samples were taken from a data base established during the project AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; https://aviandiv.fli.de/) and later extended by samples of the project SYNBREED (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de). Blood sampling was done in strict accordance to the German animal welfare regulations, with written consent of the animal owners and was approved by the at the according times ethics responsible persons of the Friedrich-Loeffler-Institut. According to German animal welfare regulations, notice was given to the responsible governmental institution, the Lower Saxony State Office for Consumer Protection and Food Safety (33.9-42502-05-10A064).

Populations and sequencing

The analysis is based on WGS data of a diverse set of 46 commercial, non-commercial and wild chicken populations, sampled within the framework of the projects AVIANDIV (www.aviandiv.fli.de) and SYNBREED (www.synbreed.tum.de). Commercial brown (BL) and white layer (WL) populations consist of 25 individually re-sequenced animals each, while the two commercial broiler lines (BR1 and BR2) include 20 individually sequenced animals each. For 41 populations, pooled DNA from 9 - 11 animals per population was sequenced, while *Gallus varius* (green jungle fowl; GV) samples of only two animals were sequenced as a pool. More detailed information about the samples can be found in **S1 File** and two previously published papers, from Malomane *et al.* (2018) and Qanbari *et al.* (2019). Coverage was between 7X and 10X for the individual sequences, while DNA pools were sequenced with 15X to 70X coverage. Sequencing was conducted on Illumina HiSeq machines at the Helmholtz Zentrum, German Research Center for Environmental Health in Munich, Germany.

Raw data preparation and SNP calling

Sequences were aligned to the reference genome Gallus_gallus-5.0 (UCSC 2016; Warren *et al.* 2017) and the SNP calling was conducted according to GATK Best Practices guidelines (DePristo *et al.* 2011; van der Auwera *et al.* 2013). BWA-MEM 0.7.12 (Li 2013) was used for the alignment step, duplicates were marked using Picard Tools 2.0.1 (Broad Institute 2015) MarkDuplicatesWithMateCigar and base qualities were recalibrated with GATK 3.7 (McKenna *et al.* 2010) BaseQualityRecalibrator. The set of known SNPs, necessary for base quality score recalibration, was downloaded from ENSEMBL release 87 (ENSEMBL 2016). SNPs were called for all samples separately using the GATK 3.7 HaplotypeCaller and later on simultaneously genotyped across samples with GATK 3.7 GenotypeGVCFs. Due to computational limitations, the ploidy parameter of HaplotypeCaller was set to two instead of the higher true ploidy of the pooled sequences. By this, slightly less rare alleles were called. However, effects of this limitation are negligible (**S2 File; S1 Fig**). Note that allele frequencies were estimated from the ratio of allelic depth by total depth.

SNP filtering was conducted using GATK 3.7 VariantRecalibrator, which filtered the called SNPs by a machine learning approach (use of a Gaussian mixture model), which uses both a set of previous known (low confidence needed) and a set of highly reliable (assumed to be true) variants as training sources (Broad Institute 2018). The source for known SNPs (prior 2) provided to VariantRecalibrator was again ENSEMBL (release 87) and the SNPs of the Axiom™ Genome-Wide Chicken Array were defined as true training set (prior 15). The algorithm was trained on the quality parameters DP, QD, FS, SOR, MQ and MQRankSum. Filters were set to recover 99 % of the training SNPs in the filtered set, which resulted in a Transition/Transversion ratio of 2.52 for known SNPs, and a Transition/Transversion ratio of 2.26 for novel SNPs. Only biallelic autosomal SNPs were used in all further analyses.

Identification of the ancestral allele

Ancestral alleles were defined using allele frequency information from the three wild populations *Gallus gallus gallus* (GG), *Gallus gallus spadiceus* (GS) and *Gallus varius* (GV) by an approach comparable to Rocha *et al.* (2014). It was assumed that the *Gallus gallus* and *Gallus varius* species emerged from a common ancestor and *Gallus gallus* later split into *Gallus gallus gallus* and *Gallus gallus spadiceus* subspecies. Additionally, assuming neutral molecular evolution (Kimura 1991), the ancestral allele was most likely the major allele within those three populations, when weighting the allele frequency of *Gallus varius* twice. This procedure assigned the ancestral status to the reference allele for 86 % of the SNPs and to the alternative allele for 14 % of the SNPs. The change in the allele frequency spectrum was only relevant for the interval from 0.95 – 1.00, which was reduced by 111,851 SNPs (0.39 % of all SNPs) when switching from alternative to derived allele frequency (**S2 File; S2 Fig**).

Reference Sets

Three different reference sets were defined as follows: the **unfiltered WGS** SNPs (28.5 M SNPs), SNPs filtered using GATK 3.7 (McKenna *et al.* 2010) VariantRecalibrator (20.9 M SNPs; **filtered WGS**) and **array SNPs** (540 k SNPs), which are the intersection of the unfiltered SNPs and the SNPs of the Axiom™ Genome-Wide Chicken Array. The separate use of unfiltered and filtered WGS SNPs was done to assess the effect of filtering (especially the use of an ascertained SNP set as the true set) on ascertainment bias.

Redesigning the SNP Array

The process of redesigning the array *in silico* is briefly shown in **Figure 2.1** and explained in more detail in the following. For the design process, the populations were divided into four groups:

1) Discovery populations (8)
2) Validation populations (19)
3) Application populations (18)
4) Outgroup (1)

For SNP discovery, firstly the four commercial lines (commercial white layers, WL; commercial brown layers, BL and the two commercial broiler lines, BR1 and BR2) were used. The set was then extended by additionally selecting those populations that were closest related to each of the commercial populations based on pairwise Nei's standard genetic distance (Nei 1972). As the two broiler populations were closest related (**S3 Fig**), the next two closest populations were chosen. This resulted in the inclusion of White Leghorn (LE), Rhode Island Red (RI), Marans (MR) and Rumpless Araucana (AR). Note that the commercial populations are closely related to the populations used as discovery populations for the development of the Axiom™ Genome-Wide Chicken Array (2013) with exception of some inbred lines from the Roslin Institute in Edinburgh of which we do not know the genetic origin. The discovery set used for the original array (2013) additionally consisted of more animals from multiple layer and broiler lines than ours. Further, the discovery set had to be split into broilers (BR1, BR2, MR, AR) and layers (WL, LE, BL, RI) for the equal spacing step. From the remaining populations, 19 were randomly chosen for validation of previously discovered SNPs (validation populations), 18 populations (which were not included in the array development) were used as a case study for an application of the array (application populations), and *Gallus varius* as a different species was defined as outgroup. The interested reader can find all underlying pairwise Nei's standard genetic distances (Nei 1972) in **S3 File** and additionally pairwise F_{ST} values (Wright 1949) in **S4 File**.

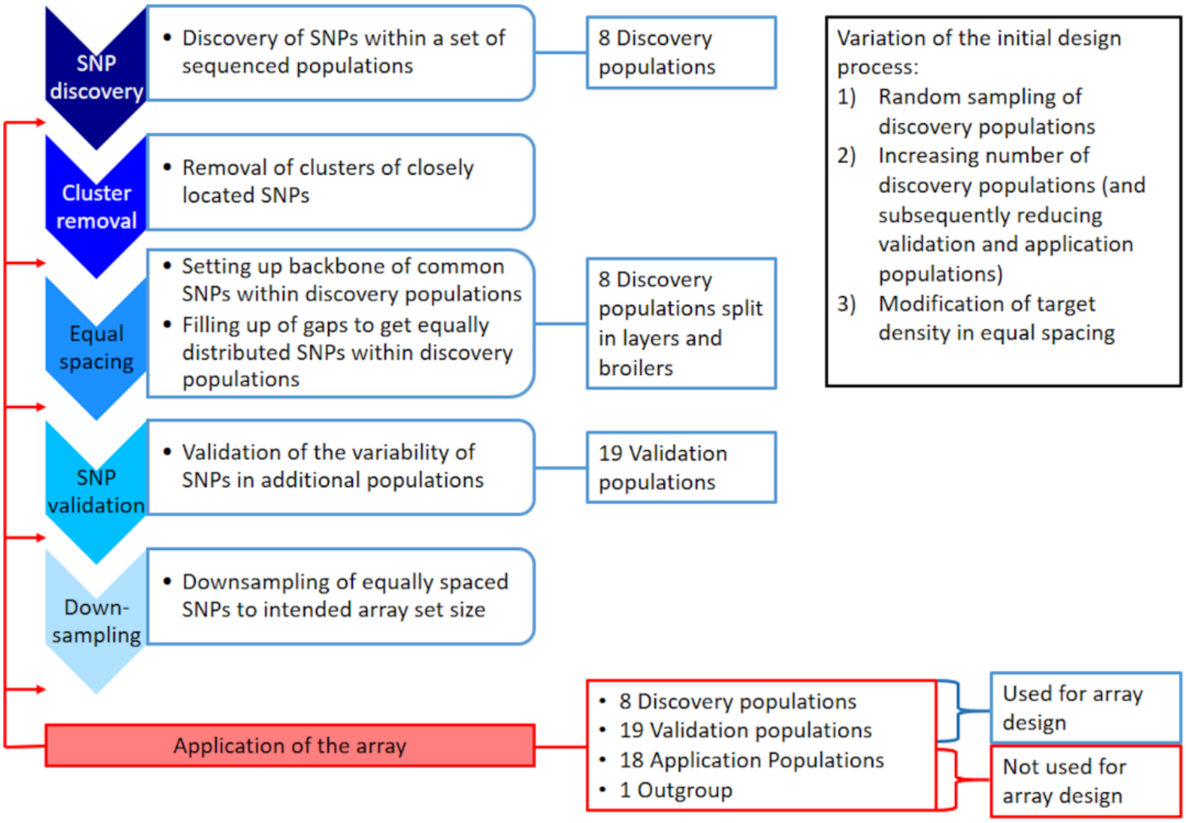
SNP discovery
• Discovery of SNPs within a set of sequenced populations
8 Discovery populations
Variation of the initial design process:
1) Random sampling of discovery populations
2) Increasing number of discovery populations (and subsequently reducing validation and application populations)
3) Modification of target density in equal spacing
Cluster removal
• Removal of clusters of closely located SNPs
Equal spacing
• Setting up backbone of common SNPs within discovery populations
• Filling up of gaps to get equally distributed SNPs within discovery populations
8 Discovery populations split in layers and broilers
SNP validation
• Validation of the variability of SNPs in additional populations
19 Validation populations
Down-sampling
• Downsampling of equally spaced SNPs to intended array set size
Application of the array
• 8 Discovery populations
• 19 Validation populations
• 18 Application Populations
• 1 Outgroup
Used for array design
Not used for array design

Figure 2.1: Flow chart of the array redesign process. The steps of redesigning the array (blue) are described in more detail in the text. Application of the array (red) was done after each subsequent step to assess the effects of the according step on the frequency spectrum.

Based on the unfiltered SNP set, the sampling of the SNPs for an approximately 600 k sized array was remodeled *in silico* in five consecutive steps according to the design process of the original array which was described by Kranis *et al.* (2013), starting from the unfiltered SNP set:

1) **SNP discovery → 10.9 M SNPs**

   Discovery of SNPs fulfilling basic criteria (quality ≥ 60; MAF ≥ 0.05; coverage ≤ mean + three standard deviations) within the discovery populations.

2) **Cluster removal → 8.8 M SNPs**

   SNP clusters were defined as SNPs with less than 4 bp invariant sites at one side of a SNP and less than 10 bp invariant sites at the other side of the SNP within the discovery populations. Those SNPs were removed, which is justified rather technically to enable probe binding, but could also lead to an overrepresentation of conserved regions compared to highly variable regions of the genome.

3) **Equal spacing → 2.1 M SNPs**

   Reduction of SNPs to achieve approximately equidistant spacing between variable SNPs within discovery populations based on genetic distances. This algorithm was modeled according to Kranis et al. (2013) and followed a two-step procedure. The first step was setting up an initial backbone of common SNPs (three sub-steps). It started with selecting SNPs which segregated in all discovery populations (MAF within each population > 0) while requiring a minimal distance of 2 kb, resulting in about 8 k SNPs. This was complemented by a backbone of SNPs which segregated in all layer populations and another one of SNPs which segregated in all broiler populations. Note that Kranis et al. (2013) additionally constructed a backbone from a group of inbred lines for which no comparable samples were available for this study. In the second step, the algorithm iterated over all single populations and filled in potential gaps between backbone SNPs which are variable within the according population. This was done by choosing the SNPs closest to equidistant positions within the gap while aiming for a predefined local target density of 667 segregating SNPs/cM (linkage map taken from Groenen *et al.* 2009). See **S4 Fig** for the detailed contribution of additional SNPs from each sub-step of the algorithm.

4) **SNP validation → 1.7 M SNPs**

   Removing SNPs (~ 20 %) which were not variable in at least 8 of the 19 validation populations. This step would in reality be done by genotyping with preliminary test arrays and therefore allows the use of a broader set of populations than the discovery step.

5) **Downsampling → 580 k SNPs**

   Downsampling of SNPs comparable to step 3, but without adding the broiler/ layer specific backbones and instead keeping all exonic SNPs (annotation using Ensembl VEP 89.7; McLaren *et al.* 2016). Additionally, the target density in broiler lines was set as three times the target density of the layer lines. The increased target density in broilers is intended to account for lower levels of linkage disequilibrium in these lines.

Variation of the design process

The whole design process was repeated 50 times with populations being randomly assigned to be discovery, validation or application populations, while the *Gallus varius* population was always kept as the outgroup. In this process, the number of populations per group was the same as in the previous scenario.

To assess the impact of the number of discovery populations on the design process, the number of discovery populations was varied in additional runs from 4 to 40 randomly chosen populations (while assigning the remaining populations, except *Gallus varius,* to validation and application groups of equal size) with 20 random replicates for each number of discovery populations. In a last scenario, equal spacing was varied with respect to the target density (33 – 3333 SNPs/cM) with 20 independent population groupings for each target density, with or without the initial backbone. As the number of SNPs from the backbone was constant, the increase of the target density led to a higher number of SNPs chosen by the algorithm due to the equal spacing itself and hence the relative influence of the fixed number of common backbone SNPs decreased.

Analyses of the results

Per-locus-allele frequencies for individually sequenced populations were estimated from genotypes, whereas the estimation for the sequenced DNA-pools was based on the allelic depth. Influences on the allele frequency spectra were examined by comparing density estimates of derived allele frequency spectra (unfolded frequency spectrum). Further H$_{\text{exp}}$, the expected heterozygosity assuming Hardy Weinberg frequencies of the genotypes, for the different populations were used as summary statistics of the within population allele frequency spectra and calculated as in equation (2.1), where $p_{ref;l}$ denotes the frequency of the reference allele at locus l and L the total number of loci.

$$H_{exp} = \frac{\sum_{l} 2p_{ref;l}(1-p_{ref;l})}{L} \tag{2.1}$$

Deviations in the estimation of H$_{\text{exp}}$ from the various SNP sets were quantified as differences between the H$_{\text{exp}}$ calculated from the respective SNP set and the H$_{\text{exp}}$ calculated from the filtered WGS SNPs relative to the H$_{\text{exp}}$ from the filtered WGS SNPs, further called overestimation of H$_{\text{exp}}$ (OHE; equation (2.2)), which was calculated per population.

$$OHE = \frac{H_{exp;\,SNP\ set} - H_{exp;\ filtered\ WGS\ SNPs}}{H_{exp;\ filtered\ WGS\ SNPs}} \tag{2.2}$$

An OHE of zero means that the estimates are equal, while an OHE of one describes doubling of the unbiased estimate.

The effects of the population group assignments on the OHE of the random population assignments were evaluated by pairwise comparisons of least square means (LSMEANS; calculated with the R package emmeans (R Core Team 2017; Lenth 2019) by using Tukey correction for multiple pairwise contrasts) of the population groups. An underlying mixed linear model for the estimation of LSMEANS was fitted using the R package lme4 (Bates *et al.* 2015) as shown in equation (2.3), where the OHE depended on an overall mean μ, the fixed effect of the population group $popG_i$ (i can be discovery-, validation-, application- or outgroup), a random effect for the j^th^ repetition of random population grouping ($rep_j \sim N(0, I\sigma_{rep}^2)$) and a random error $e_{ijk} \sim N(0, I\sigma_e^2)$. The procedure is comparable to simple pairwise comparisons of group means, the correction by the repetition only reduces the error variance and thus decreases the confidence intervals.

$$OHE_{ijk} = \mu + popG_i + rep_j + e_{ijk} \tag{2.3}$$

Results

Numbers of SNPs

The SNP calling identified 28.5 M biallelic autosomal SNPs from which 20.9 M SNPs passed GATK's filtering procedure. 540 k SNPs from the unfiltered WGS SNP set are also mapped on the original Axiom™ Genome-Wide 580 k Chicken Array. The remodeling of the array according to the design process of the original array returned 10.9 M SNPs from the discovery step, which were reduced to approximately 580 k in steps as described. Numbers of identified SNPs for the additional runs differed depending on the populations and settings used and are listed in **S1 Table**. It has to be noted that the different sub-steps of the equal spacing algorithm contributed with different amounts of SNPs (**S4 Fig**). Especially the much higher contribution of SNPs which were segregating in all broiler populations compared to SNPs segregating in all layer populations in the remodeling with populations chosen comparable to the original array was remarkable. This is due to closer relationships between the broiler populations and their generally higher heterozygosity. Additional information about the identified number of SNPs depending on the number of discovery populations and target density as well as information about the share of SNPs of different random runs can be found in **S5 – S7 Figs**.

Underrepresentation of rare SNPs

A clear underrepresentation of rare SNPs in all ascertained SNP sets compared to WGS is evident from the allele frequency spectra (**Figure 2.2**). Major changes in the allele frequency spectra during the array development process were observed after the SNP discovery step and the equal spacing step. The SNP discovery led to an underrepresentation of rare SNPs compared to sequence data, which was intensified by the equal spacing step (**Figure 2.2**). The process finally resulted in a spectrum which was comparable to the spectrum of the original array, albeit slightly more right skewed. Randomly choosing populations as discovery populations confirmed the shape of the first remodeling, where the population groups were chosen according to the original array (Kranis *et al.* 2013). As major changes in the spectra mainly occurred after the SNP discovery and equal spacing, further results will concentrate on those steps.

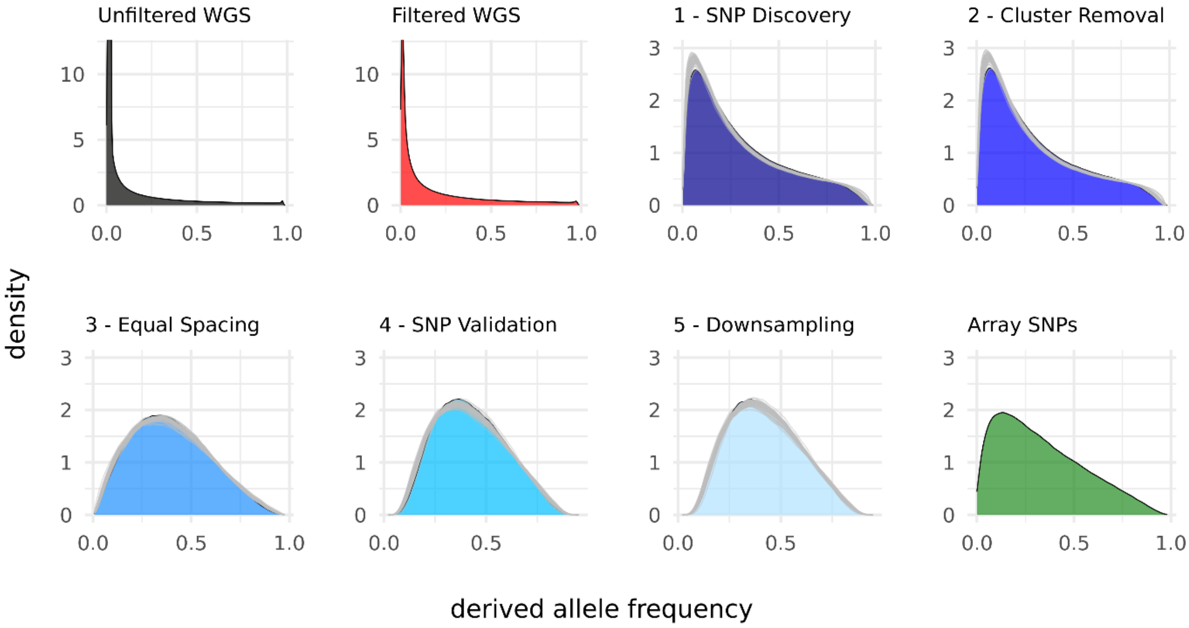

Figure 2.2: Derived allele frequency spectra for the different SNP sets. For the remodeled sets, areas show the modelling according to the original array (2013) while grey lines represent the 50 random population groupings.

The allele frequency spectra (**Figure 2.3**) within discovery populations, compared to the spectra over all populations, clearly showed the cutoff from the MAF 0.05 filter. Furthermore, the allele frequency spectra of the discovery populations revealed a higher share of common SNPs than the overall spectra after equal spacing. In contrast, the spectra within validation- and application populations showed less pronounced peaks after the discovery step and the outgroup (*Gallus varius*) revealed fixation of most SNPs variable in the discovery populations.

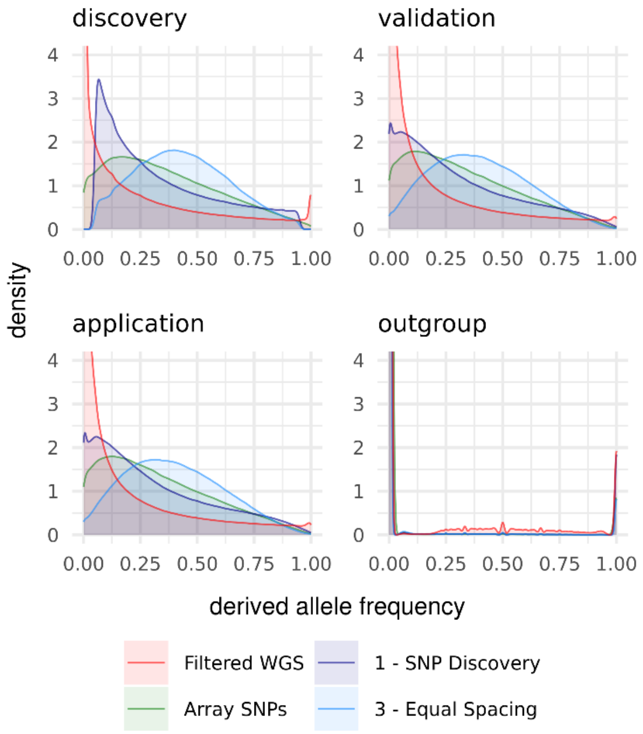

Figure 2.3: Derived allele frequency spectra within the population groups.

Influence of number of discovery populations and target density on allele frequency spectra

Not surprisingly, an increased number of discovery populations resulted in a higher number of rare alleles after the discovery step, and thus an allele frequency spectrum with a more pronounced peak of rare alleles (**Figure 2.4 A**). Apparently, the shift of the allele frequency spectrum after the equal spacing step was dependent on the number of discovery populations, as an increase in the number of discovery populations shifted the allele frequency spectra towards a higher proportion of alleles with a low derived allele frequency. With an increasing number of discovery populations, the shape of the allele frequency spectra got closer to the spectrum of the original array.

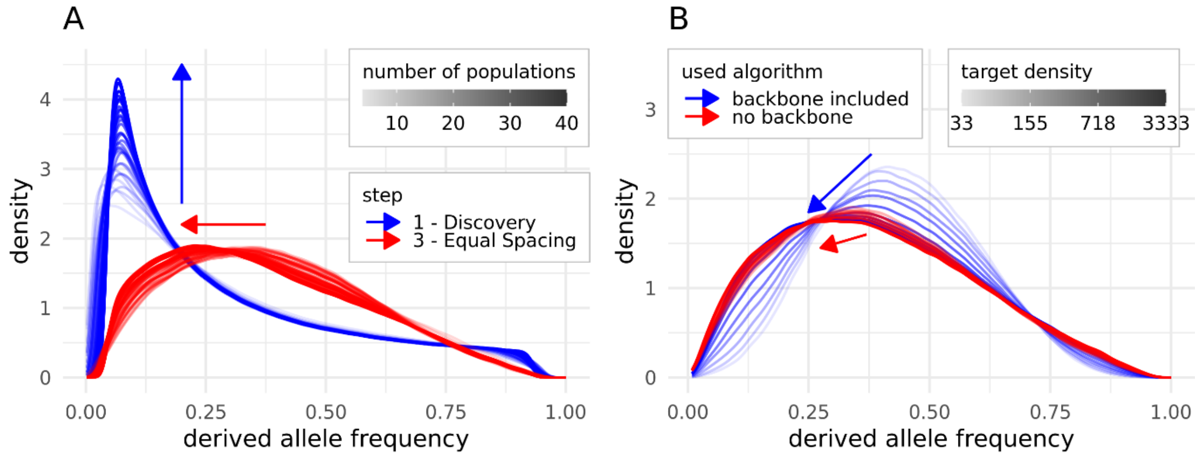

Figure 2.4: Impact of a varying number of discovery populations (A) or target density (B) on the derived allele frequency spectrum. For A, blue indicates the spectra after the discovery step and red after the equal spacing step. For B, only the equal spacing step is shown and blue indicates that the algorithm including the initial backbone, while red shows the results without the backbone included in the algorithm. Different numbers of populations in the discovery set (4 to 40) or the increase in the target density are indicated by an intensifying color gradient and only one representative and randomly picked run per population number/ target density is shown. As the differences in the color gradients are hard to distinguish, arrows in the respective color are indicating the shift of the spectra with increasing numbers of discovery populations.

A very low target density, indicating that SNPs were mostly called due to being common backbone SNPs, resulted in an allele frequency spectrum with the majority of alleles having a MAF of around 0.5 (**Figure 2.4 B**). Increasing the target density for the equal spacing and thus reducing the influence of the initial backbone of common SNPs shifted the peak of the allele frequency spectrum left towards a higher proportion of alleles with small derived allele frequencies. Using only the backbone SNPs common over all discovery populations and thus calling SNPs mostly by the equal spacing procedure resulted, independently from the target density, in a spectrum similar to the one obtained with a high target density with backbone (**Figure 2.4 B**).

Overestimation of H_{exp}

Figure 2.5 shows the H_{exp} of different SNP sets by population. The H_{exp} obtained from the filtered WGS SNPs were slightly higher than from the unfiltered WGS SNPs. H_{exp} obtained from the ascertained SNP sets showed an even more pronounced overestimation together with an increase during the design steps. In general, the correlations between the H_{exp} obtained in the different SNP sets were relatively high (≥ 0.95; **S2 Table**). Especially the H_{exp} of the two WGS SNP sets showed a nearly perfect correlation of > 0.99, which led to an almost constant OHE of -0.23 (**Table 2.1**) for the unfiltered WGS SNPs. As already recognizable from the H_{exp} themselves, the OHE was positive for all ascertained SNP sets (0.66 – 1.29), which at the same time showed a slightly reduced correlation to the filtered WGS SNP set (0.95 – 0.97). Comparable to the allele frequency spectra, the most pronounced increase of the OHE was caused by the SNP discovery and followed by the equal spacing step (OHE increased by 0.66), while the OHE from the original array SNPs (1.41; **Figure 2.5**; **Table 2.1**) laid in the range covered by the remodeling steps.

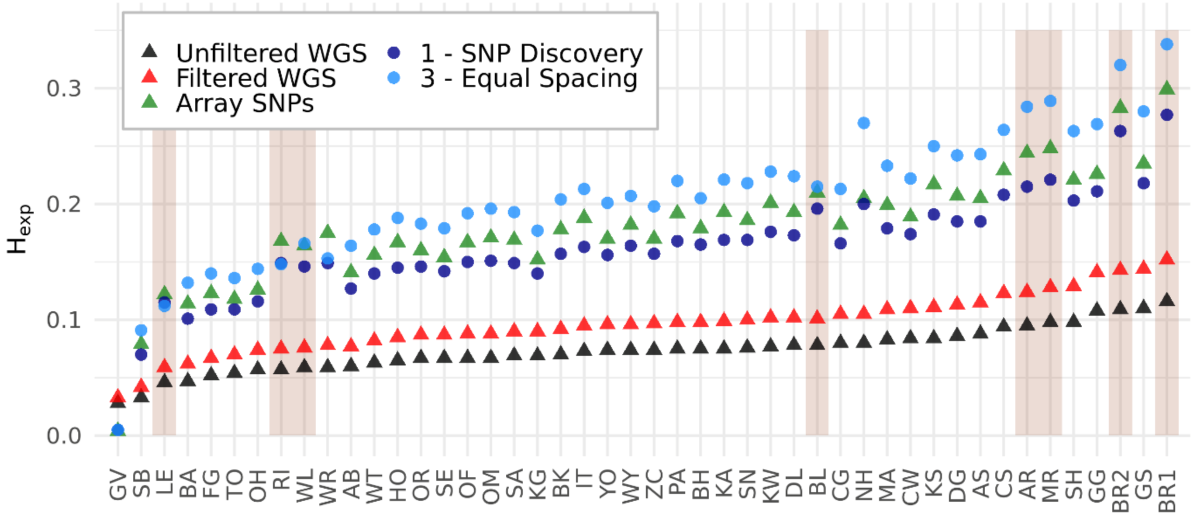

Figure 2.5: Expected Heterozygosity (Hexp) by population and SNP set. Populations are ordered by the H_{exp} of the unfiltered WGS SNP set. Only the reference sets and relevant steps of the array design are shown. Discovery populations are shaded with a darker background.

Table 2.1: OHE of the SNP sets from the first run

Populations	Unfiltered WGS	Array SNPs	1 – SNP discovery	2 - cluster removal	3 – equal spacing	4 – validation	5 – downsampling
All	-0.23 ± 0.01	0.84 ± 0.30	0.66 ± 0.26	0.66 ± 0.26	1.09 ± 0.32	1.29 ± 0.35	1.27 ± 0.34
Discovery	-0.23 ± 0.00	1.05 ± 0.10	0.86 ± 0.10	0.86 ± 0.10	1.15 ± 0.15	1.28 ± 0.17	1.32 ± 0.13
Validation	-0.23 ± 0.00	0.87 ± 0.13	0.68 ± 0.10	0.67 ± 0.10	1.15 ± 0.13	1.36 ± 0.14	1.33 ± 0.12
Application	-0.23 ± 0.00	0.83 ± 0.12	0.64 ± 0.07	0.63 ± 0.07	1.10 ± 0.11	1.33 ± 0.14	1.30 ± 0.15
Outgroup	-0.17	-0.88	-0.85	-0.86	-0.85	-0.85	-0.84

Mean OHE ± standard deviation.
An OHE of zero means no bias and an OHE of 1 means doubling the H_{exp}.

Averaging the OHE within the population groups revealed a 30 % higher OHE of the discovery populations compared to validation and application populations after the discovery step. The equal spacing step reduced this difference to an only 1 % larger OHE for discovery populations, while it came with a substantial increase of the variance of OHE, which was larger for the discovery populations than validation and application populations. The validation step then increased the OHE of the validation populations more than the OHE of discovery and application populations. This stronger OHE of discovery populations was also apparent within the array SNPs (24 % higher). In contrast to the other populations, the outgroup showed an underestimation of the H_{exp}, resulting in an OHE of < -0.84 for all ascertained SNP sets (**Figure 2.5**; **Table 2.1**).

A closer look on the contribution of the sub-steps during the equal spacing step revealed that 62 % of the SNPs which were preserved during equal spacing were variable in all of the four closely related broiler populations (BR1, BR2, MR, AR; maximum pairwise Nei's distance of 0.06 and FST of 0.17 in the filtered SNP set), while only 3 % of the SNPs were retained due to being variable in all of the four less closely related layer populations (WL, LE, BL, RI; maximum pairwise Nei's distance of 0.15 and FST of 0.48 in the filtered SNP set). The first population used to fill in the gaps in the backbone (WL) contributed 17 % of the SNPs, while the other populations contributed < 8 %.

Table 2.2: OHE of the SNP sets out of the 50 random population groupings

Populations	1 – SNP discovery	2 – cluster removal	3 – equal spacing	4 – validation	5 – down-sampling
Discovery	$0.76_{\pm 0.004}^{a}$	$0.75_{\pm 0.004}^{a}$	$1.13_{\pm 0.006}^{a}$	$1.28_{\pm 0.006}^{b}$	$1.33_{\pm 0.007}^{a}$
Validation	$0.61_{\pm 0.003}^{b}$	$0.60_{\pm 0.003}^{b}$	$1.11_{\pm 0.004}^{b}$	$1.29_{\pm 0.004}^{b}$	$1.29_{\pm 0.005}^{b}$
Application	$0.61_{\pm 0.003}^{b}$	$0.61_{\pm 0.003}^{b}$	$1.12_{\pm 0.004}^{ab}$	$1.35_{\pm 0.004}^{a}$	$1.34_{\pm 0.004}^{a}$
Outgroup	$-0.85_{\pm 0.008}^{c}$	$-0.86_{\pm 0.008}^{c}$	$-0.85_{\pm 0.015}^{c}$	$-0.84_{\pm 0.017}^{c}$	$-0.84_{\pm 0.019}^{c}$

LSMEANS for OHE ± standard error.
An OHE of zero means no bias and an OHE of 1 means doubling the H_{exp}.
Different lowercase letters within columns indicate significant differences to the 5 % level.

These findings were supported by the 50 random groupings (**S8 Fig**). The LSMEANS (**Table 2.2**) of the population groups revealed 24 % larger OHE for discovery populations than for validation and application populations after discovery and cluster removal step, which was decreased to a numerically insignificant difference after the equal spacing step. Interestingly, and in contrast to the findings from the first remodeling, SNP validation led to a significantly higher OHE (5 % larger) for application populations than discovery and validation populations.

Influence of number of discovery populations and target density on H_{exp}

Figure 2.6 A shows that increasing the number of discovery populations reduces the median OHE of discovery populations after SNP discovery while not affecting the OHE of validation and application populations. Equal spacing (**Figure 2.6 B**) removed the average difference of OHE between the

different population groups. Due to the limited number of populations in the complete set, the number of validation populations had to be reduced with more populations in the discovery set. This led to an increasing impact of individual validation populations on the ascertainment. The OHE of validation populations therefore increased with a high number of discovery populations (**S9 D Fig**), comparable to the higher OHE of discovery populations for a small number of discovery populations. In our case, the biased array for validation populations was therefore obtained with a combination of 30 populations in the discovery set and 7 populations in the validation set. However, the least biased array for discovery and application populations was the array with the maximum number of discovery populations (40).

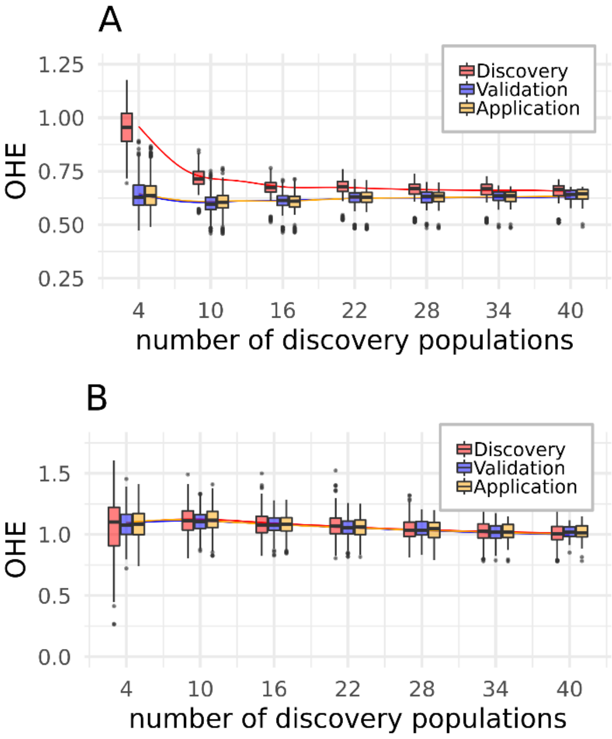

Figure 2.6: Relation of the OHE as a function of the number of discovery populations. A - discovery, B - equal spacing. While the number of discovery populations was varied from 4 to 40 by increments of one, the Boxplots are only shown for a subset of the number of discovery populations to avoid a crowded figure. The smoothing lines, which show the trend, are calculated from all observations. Plots for all five steps can be found in **S9 Fig.**

In the equal spacing step, using only backbone SNPs resulted in a higher OHE for discovery than for non- discovery populations. Increasing the target density and thus increasing the proportion of SNPs due to the equal spacing part of the algorithm reduced the difference in OHE between the population groups (**Figure 2.7 A**). If the SNPs from the initial backbone were not used, no difference of OHE between discovery and non- discovery populations was present, regardless of the target density (**Figure 2.7 B**).

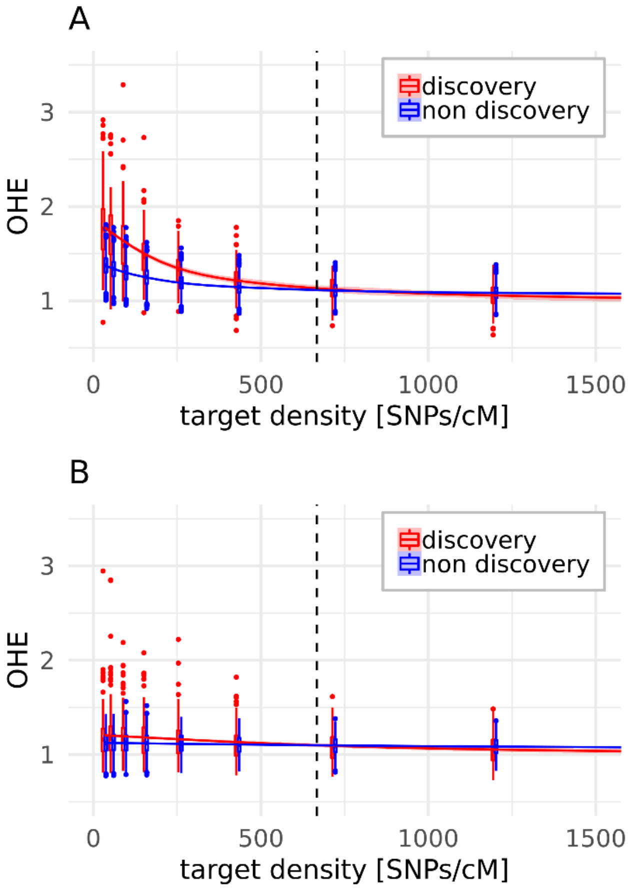

Figure 2.7: OHE after equal spacing (step 3) by target density in SNPs/cM and population group. The smoothing lines show the trend and the dashed lines the target density of 667 SNPs/cM, used for the remodeling according to the original array (Kranis *et al.* 2013). The algorithm was run including the initial backbone SNPs (A) or not including them (B). Gallus varius is not included, as it is constantly underestimated.

Discussion

In this study we used a uniquely diverse collection of sequenced wild, commercial and non-commercial chicken populations, mainly based on samples of the Synbreed Chicken Diversity Panel (Malomane *et al.* 2019). Parts of our set were also involved in the development process of the Axiom™ Genome-Wide 580 k Chicken Array (Kranis *et al.* 2013). This offered an excellent possibility for assessing the impact of ascertainment bias on real data in a complex scenario. In general, results derived from this study should therefore be transferable to other species. However, domestic chickens show a rich history of hybridization and crossbreeding events (Malomane *et al.* 2019). The effects of using a discovery set closely related to the commercial populations and distributing the discovery set randomly across the spectrum of populations were therefore comparably small in this study. Special patterns of population structure e.g. the stronger differentiation in cattle due to the two subspecies *Bos taurus* and *Bos indicus* (Hiendleder *et al.* 2008) accompanied by limiting the discovery set to one of the two clades, should increase the impact of population structure dependent ascertainment bias.

Potential impacts of the SNP calling pipeline

As the state of the art pipeline of GATK relies on a supervised machine learning approach for filtering the SNP calls, which needs a highly reliable set of known SNPs, we started with examining potential impacts of the filtering procedure on ascertainment bias. The number of rare variants was slightly reduced by the filtering procedure and thus increased estimates of H_{exp} were obtained in the filtered WGS set. As rare variants have a higher risk to be discarded as sequencing errors (Heslot *et al.* 2013), this reduction is expected when applying quality filters. However, a clear assessment of correctly and falsely filtered variants is not possible here and one has to balance this tradeoff based on the study purpose.

Another source of ascertainment bias could be the use of array SNPs as training set for GATK RecalibrateVariants, which potentially leads to discarding rare variants more likely if they are not present in the discovery populations of the used array. As the correlation between the H_{exp} of the unfiltered and filtered WGS SNPs was nearly one, this source seems to be negligible and the use of array variants as a highly reliable training set seems to be unproblematic.

Due to computational limitations, we had to assume a ploidy of two for pools during the SNP calling process, which resulted in a minimal reduction of rare alleles. However, this effect was shown to have a very minor impact on the findings of this study (**S2 File**). Nevertheless, pooled sequencing itself can slightly bias allele frequency estimates compared to individual sequencing (Futschik and Schlötterer 2010; Chen *et al.* 2012; Schlötterer *et al.* 2014; Wang *et al.* 2016). As all frequency estimates for single SNPs were taken from the same data source throughout the study, this does not affect our results. However, estimates for the magnitude of the ascertainment bias for single populations have to be understood rather relative to our gold standard than as absolute values.

General impact over all groups

The general reduction of rare alleles in array data compared to WGS data and the resulting overestimation of H_{exp} supports findings of previous studies (Nielsen 2004; Clark *et al.* 2005; Albrechtsen *et al.* 2010; Malomane *et al.* 2018). This reduction of rare alleles was mainly seen at steps where selection was explicitly biased towards high MAF alleles (MAF filter for quality control in discovery step and use of common alleles for the backbone in the equal spacing step) and/ or was applied to a small number of populations (small discovery set vs. small validation set). Thereby, the strongest shifts of the allele frequency spectra and increases of H_{exp} are observed after SNP discovery and equal spacing. Both, cluster removal and second downsampling had almost no effect on the allele frequency spectra and H_{exp}, while the validation step slightly decreased the share of rare SNPs.

The discovery step had the strongest impact on discovery populations, when a small set of discovery populations was used (**Figure 2.6 A**). Similarly, the influence of the validation step on validation populations was strongest in case of a small number of validation populations (**S9 D Fig**). A balancing of these two groups of samples is therefore necessary, if the number of available DNA samples for array development is limited. Instead of using separate populations for discovery and validation, we rather suggest to space the discovery set across all available populations and validate test arrays on additional samples of the same populations.

If the equal spacing step contains a preselection of SNPs based on their variability within population groups, the bias is stronger towards high MAF SNPs and thus yields a higher OHE. This effect was reduced by increasing the target density and thus selecting relatively more SNPs due to the equal spacing instead of common occurrence.

Differences between groups

If allele frequency spectra are changed in the same way for all populations and are therefore biasing heterozygosity estimates to the same extent, findings for between population comparisons will be little affected. Ascertainment bias then is only of importance if one compares populations based on different arrays, and corrections of the allele frequency spectrum as reviewed by Nielsen (2004) should be possible. As correlations between H_{exp} of ascertained SNP sets and unfiltered/ filtered WGS SNP sets were consistently high (> 0.94), arrays designed in the way as performed in this paper should mostly be suitable for robust and cost efficient analyses. Biasedness of estimates could be reduced even more by considering filter strategies according to Malomane *et al.* (2018).

However, we could show that the bias acts with different extent on different population groups (population structure dependent bias) and therefore changes ranking of populations and can affect conclusions. This population structure dependent bias was already shown to have severe impact on findings from SNP arrays. For example, Bradbury *et al.* (2011) found a demographic decline up to an approximately 30 % lower H_{exp} for Atlantic cod based on the distance to the sampling location of the discovery panel and McTravish and Hillis (2015) showed strong deviations between simulated and observed polymorphisms for different combinations of migration and ascertainment scenarios on simulated cattle populations. In concordance with this, populations which are closely related to the discovery populations of the original array in our study on average showed a 24 % higher OHE than validation and application populations for the original array.

This population structure dependent bias was mainly introduced by the initial discovery step. It was also observed in the random population groupings, but to a slightly different extent. The difference in overestimation decreased with an increase in the number of discovery populations (**Figure 2.6**) and was smallest if the discovery populations showed minimum distance to the application and validation populations (results not shown). Comparable observations were already made by Frascaroli *et al.* (2013) which found very small ascertainment bias for European elite maize lines when using a SNP panel discovered in a combination of a maize diversity set and inbred lines, but strong ascertainment bias when using SNPs which were discovered in American elite lines. Therefore, we suggest to ideally choose an array where the discovery panel does span the scope of populations it will be applied to, and by this covers the existing variation in a most representative way, or to design such an array for oneself if it does not exist.

The equal spacing step lowered the difference in mean OHE between population groups in most of our remodeling scenarios, but obviously not in case of the original array. In the remodeling, we saw this difference only with a low target density and thus calling SNPs in the equal spacing step mainly due to being common over many populations (**Figure 2.7 A**). However, the equal spacing step also increased the variance of OHE in the discovery panel, meaning that the OHE was increased more for some of the discovery populations than for others, thus causing more uncertainty for resulting effects. This effect is driven by the unequal contribution of variable SNPs to the chosen SNP set by the different populations during the equal spacing step (**S3 Fig**). The equal spacing step increases the OHE for some of the discovery populations, while it decreases it for others, and hence it does not remove the population structure dependent bias. This means that the knowledge of which discovery populations were used is not sufficient to draw conclusions regarding a possible ascertainment bias, since their relative contribution varies through the described pipeline.

Outgroup

Gallus varius as an outgroup showed a different behavior than all other populations. It already exhibited the lowest H_{exp} in the unfiltered WGS SNP set, which was most likely driven by the small number of only two samples in the pool, and showed less upward bias of H_{exp} in the filtered WGS SNP set than all other populations. The *Gallus varius* sequence reads on average showed weak Phred-scaled mapping quality scores of 19 (1.3 % probability of misalignment), while the mean quality scores of the other populations ranged from 25 (0.3 %) to 28 (0.1 %). Variation, only present in *Gallus varius*, will therefore be more likely missed due to misplacement of the reads or discarded as possible sequencing errors. Additionally, every ascertained SNP set showed an OHE for *Gallus varius* of < -0.84, as variation being present only in *Gallus varius* was not found in *Gallus gallus* discovery panels and, vice versa, variants from *Gallus gallus* were not variable in *Gallus varius* (**Figure 2.3**). This demonstrates that arrays should not be used if different species (even closely related ones) are included in the research project. Even sequence based estimates can be slightly biased, if the reference genome does not fit properly.

Potential impact on other breeding applications

In general, we cannot infer the impact on breeding applications which require phenotypic data (e.g. genomic selection (Meuwissen *et al.* 2001) or genome wide association studies (Goddard and Hayes 2009)) and/or individually sequenced or genotyped individuals (e.g. linkage disequilibrium decay (Qanbari *et al.* 2014) or runs of homozygosity analyses (Peripolli *et al.* 2017)) from this study. However, literature highlights the increased power of high MAF SNPs to capture/ detect effects which are caused by common variants due to stronger linkage disequilibrium and higher levels of variance explained. Therefore, increasing MAF in a first instance increases prediction accuracy when the number of SNPs is limited (Perez-Enciso *et al.* 2015) and therefore some SNP ascertainment schemes intentionally bias the used SNPs towards high MAF within the desired populations (Matukumalli *et al.* 2009). The switch to WGS data, and therefore the additional inclusion of rare alleles, is then expected to increase the possibility of capturing the effects of rare alleles (Druet *et al.* 2014; Perez-Enciso *et al.* 2015; Wainschtein *et al.* 2019). However, the increase in efficiency by higher numbers of SNPs levels off when going towards WGS data (Ober *et al.* 2012). Nevertheless, we would expect negative impacts of ascertainment bias due to the across population use of the arrays. When biasing the genotyped variation towards the discovery population, the variability in populations, which are less related to the discovery populations, is less increased or even reduced, and arrays therefore become less valuable in non-target populations. Slight effects of this were demonstrated by simulation (Perez-Enciso *et al.* 2015) and we can clearly support these findings by the levels of differences in the genotyped heterozygosity which we observed in this study. For the effect of ascertainment bias on a broader set of applications, we further refer the interested reader to studies which specifically address those issues (e.g. Nielsen 2004; Lachance and Tishkoff 2013; Qanbari *et al.* 2014; Malomane *et al.* 2018; Quinto-Cortés *et al.* 2018).

Further recommendations for future studies

We showed that existing arrays come with a large potential for ascertainment bias which is barely predictable due to a diverse set of promoting factors. Strongly decreasing costs for WGS and increasing availability of powerful computing resources therefore promote an intensified use of WGS for population genetic analyses, especially when diverse populations are included in the studies. However, costs and computational effort will still be substantial for large scale projects. Possible cost effective alternatives could be reduced library sequencing approaches like Genotyping-by-Sequencing (Elshire *et al.* 2011; Heslot *et al.* 2013), even though such methods introduce other problems related to the use of restriction enzymes which are reviewed by Andrews *et al.* (2016).

For the purpose of monitoring genetic diversity in a large set of small non-commercial populations, the development of a specialized new array for cost effective high throughput genotyping could be still a good option. For the design of such an array, unbiasedness would thereby be represented by a random draw of the total variation within the target populations. As this is only a theoretical possibility, the practical solution closest to unbiasedness one can achieve would be a random draw form the SNPs present in the discovery set. It is thereby crucial to extend the discovery set in a way which represents the total variability over all populations as balanced as possible. The use of publicly available sequences can be helpful to reach this goal. The ascertainment of the SNPs should then be done preferably over a large set of highly diverse populations covering a wide spectrum of the diversity within a species available populations instead of biasing the process towards common alleles by performing within population ascertainment.

Acknowledgments

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

Author's contributions

JG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing

CR: Conceptualization, Investigation, Writing – review & editing

SW: Funding acquisition, Supervision, Writing – review & editing

AW: Data curation

TP: Methodology, Software, Writing – review & editing

HS: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing

Supplementary

The supplementary material can be accessed via the original publication (https://doi.org/10.1371/journal.pone.0245178).

S1 File: Abbreviations for breeds and accession numbers for the sequencing data.

S2 File: Supplementary methods.

S3 File: Pairwise Nei's standard genetic distances.

S4 File: Pairwise F_{ST} values.

S1 Fig: Pooled vs. ploidy two calling.

S2 Fig: Alternative allele frequency spectrum for SNPs where reference allele is not ancestral allele (A) and alternative (B) vs. derived (C) allele frequency spectrum of all SNPs.

S3 Fig: UPGMA tree based on pairwise Nei's standard genetic distances.

S4 Fig: Cumulative number of SNPs in million retained during the equal spacing step.

S5 Fig: SNPs retained by number of populations in the discovery set.

S6 Fig: Number of SNPs retained after the equal spacing step by target density.

S7 Fig: Relative amount of SNPs shared by a specific number of runs from 50 random population assignments.

S8 Fig: OHE for discovery validation and application populations after the five steps of array design.

S9 Fig: Relation of the OHE as a function of the number of discovery populations.

S1 Table: Number of SNPs from the remodeling processes.

S2 Table: Pearson correlations between the H_{exp} of the different SNP sets.

References

Albrechtsen, A; Nielsen, FC; Nielsen, R (2010): Ascertainment biases in SNP chips affect measures of population divergence. In *Mol Biol Evol* 27 (11), 2534–2547.

Andrews, KR; Good, JM; Miller, MR; Luikart, G; Hohenlohe, PA (2016): Harnessing the power of RADseq for ecological and evolutionary genomics. In *Nat Rev Genet* 17 (2), p. 81.

Bates, D; Mächler, M; Bolker, B; Walker, S (2015): Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software* 67 (1), pp. 1–48. DOI: 10.18637/jss.v067.i01.

Boichard, DA; Chung, H; Dassonneville, R; David, X; Eggen, A; Fritz, S et al. (2012): Design of a bovine low-density SNP array optimized for imputation. In *PLoS One* 7 (3), e34130.

Bradbury, IR; Hubert, S; Higgins, B; Bowman, S; Paterson, IG; Snelgrove, PVR et al. (2011): Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, Gadus morhua. In *Mol Ecol Res* 11 (s1), pp. 218–225.

Broad Institute (2015): Picard Tools 2.0.1. Available online at https://broadinstitute.github.io/picard/.

Broad Institute (2018): GATK User Guide. Available online at https://software.broadinstitute.org/gatk/documentation/, checked on 3/20/2018.

Chen, X; Listman, JB; Slack, FJ; Gelernter, J; Zhao, H (2012): Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples. In *Genet Epidemiol* 36 (6), pp. 549–560. DOI: 10.1002/gepi.21648.

Clark, AG; Hubisz, MJ; Bustamante, CD; Williamson, SH; Nielsen, R (2005): Ascertainment bias in studies of human genome-wide polymorphism. In *Genome Res* 15 (11), pp. 1496–1502.

Crawford, RD (1993): Poultry genetic resources. evolution, diversity and conservation. In Crawford, RD (Ed.): Poultry breeding and genetics. 2. print. Amsterdam: Elsevier (Developments in animal and veterinary science, 22).

DePristo, MA; Banks, E; Poplin, R; Garimella, KV; Maguire, JR; Hartl, C et al. (2011): A framework for variation discovery and genotyping using next-generation DNA sequencing data. In *Nat Genet* 43, 491. DOI: 10.1038/ng.806.

Druet, T; Macleod, IM; Hayes, BJ (2014): Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. In *Heredity* 112 (1), pp. 39–47. DOI: 10.1038/hdy.2013.13.

Eller, E (2001): Effects of Ascertainment Bias on Recovering Human Demographic History. In *Human Biology* 73 (3), pp. 411–427. DOI: 10.1353/hub.2001.0034.

Elshire, RJ; Glaubitz, JC; Sun, Q; Poland, JA; Kawamoto, K; Buckler, ES; Mitchell, SE (2011): A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. In *PLoS One* 6 (5), e19379.

ENSEMBL (2016): Chicken Germline SNP and INDELS. Available online at http://e87.ensembl.org/Gallus_gallus, checked on 1/6/2017.

Eriksson, J; Larson, G; Gunnarsson, U; Bed'hom, B; Tixier-Boichard, M; Strömstedt, L et al. (2008): Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. In *PLoS Genet* 4 (2), e1000010. DOI: 10.1371/journal.pgen.1000010.

Fan, B; Du, Z-Q; Gorbach, DM; Rothschild, MF (2010): Development and application of high-density SNP arrays in genomic studies of domestic animals. In *Asian-Australas J Anim Sci* 23 (7), pp. 833–847.

Frascaroli, E; Schrag, TA; Melchinger, AE (2013): Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. In *Theoretical and Applied Genetics* 126 (1), pp. 133–141. DOI: 10.1007/s00122-012-1968-6.

Futschik, A; Schlötterer, C (2010): The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. In *Genetics* 186 (1), pp. 207–218. DOI: 10.1534/genetics.110.114397.

Gautier, M; Laloe, D; Moazami-Goudarzi, K (2010): Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. In *PLoS One* 5 (9). DOI: 10.1371/journal.pone.0013038.

Gibbs, RA; Taylor, JF; van Tassell, CP; Barendse, W; Eversole, KA; Gill, CA et al. (2009): Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. In *Science* 324 (5926), pp. 528–532. DOI: 10.1126/science.1167936.

Goddard, ME; Hayes, BJ (2009): Mapping genes for complex traits in domestic animals and their use in breeding programmes. In *Nat Rev Genet* 10 (6), pp. 381–391. DOI: 10.1038/nrg2575.

Groenen, MAM; Megens, H-J; Zare, Y; Warren, WC; Hillier, LW; Crooijmans, RPMA et al. (2011): The development and characterization of a 60K SNP chip for chicken. In *BMC Genomics* 12 (1), p. 274.

Groenen, MAM; Wahlberg, P; Foglio, M; Cheng, HH; Megens, H-J; Crooijmans, RPMA et al. (2009): A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. In *Genome Res* 19 (3), pp. 510–519.

Guillot, G; Foll, M (2009): Correcting for ascertainment bias in the inference of population structure. In *Bioinformatics* 25 (4), pp. 552–554. DOI: 10.1093/bioinformatics/btn665.

Heslot, N; Rutkoski, J; Poland, JA; Jannink, J-L; Sorrells, ME (2013): Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. In *PLoS One* 8 (9), e74612.

Hiendleder, S; Lewalski, H; Janke, A (2008): Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. In *Cytogenet Genome Res* 120 (1-2), pp. 150–156. DOI: 10.1159/000118756.

Kijas, JW; Townley, D; Dalrymple, BP; Heaton, MP; Maddox, JF; McGrath, A et al. (2009): A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. In *PLoS One* 4 (3), e4668.

Kimura, M (1991): The neutral theory of molecular evolution. A review of recent evidence. In *Idengaku zasshi* 66 (4), pp. 367–386.

Kranis, A; Gheyas, AA; Boschiero, C; Turner, F; Le Yu; Smith, S et al. (2013): Development of a high density 600K SNP genotyping array for chicken. In *BMC Genomics* 14 (1), p. 59. DOI: 10.1186/1471-2164-14-59.

Lachance, J; Tishkoff, SA (2013): SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. In *Bioessays* 35 (9), pp. 780–786.

Laurie, CC; Nickerson, DA; Anderson, AD; Weir, BS; Livingston, RJ; Dean, MD et al. (2007): Linkage Disequilibrium in Wild Mice. In *PLoS Genet* 3 (8), e144. DOI: 10.1371/journal.pgen.0030144.

Lawal, RA; Martin, SH; Vanmechelen, K; Vereijken, A; Silva, P; Al-Atiyat, RM et al. (2020): The wild species genome ancestry of domestic chickens. In *BMC Biology* 18 (1), p. 13. DOI: 10.1186/s12915-020-0738-1.

Lenth, R (2019): emmeans: Estimated Marginal Means, aka Least-Squares Means. Available online at https://CRAN.R-project.org/package=emmeans.

Li, H (2013): Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available online at http://arxiv.org/pdf/1303.3997v2, updated on 3/16/2013.

Malomane, DK; Reimer, C; Weigend, S; Weigend, A; Sharifi, AR; Simianer, H (2018): Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. In *BMC Genomics* 19 (1), p. 22. DOI: 10.1186/s12864-017-4416-9.

Malomane, DK; Simianer, H; Weigend, A; Reimer, C; Schmitt, AO; Weigend, S (2019): The SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high genomic resolution. In *BMC genomics* 20 (1), p. 345. DOI: 10.1186/s12864-019-5727-9.

Matukumalli, LK; Lawley, CT; Schnabel, RD; Taylor, JF; Allan, MF; Heaton, MP et al. (2009): Development and characterization of a high density SNP genotyping assay for cattle. In *PLoS One* 4 (4), e5350.

Mayer, M; Unterseer, S; Bauer, E; Leon, N de; Ordas, B; Schön, C-C (2017): Is there an optimum level of diversity in utilization of genetic resources? In *Theor Appl Genet* 130 (11), pp. 2283–2295. DOI: 10.1007/s00122-017-2959-4.

McKenna, A; Hanna, M; Banks, E; Sivachenko, A; Cibulskis, K; Kernytsky, A et al. (2010): The Genome Analysis Toolkit. A MapReduce framework for analyzing next-generation DNA sequencing data. In *Genome Res* 20 (9), pp. 1297–1303. DOI: 10.1101/gr.107524.110.

McLaren, W; Gil, L; Hunt, SE; Riat, HS; Ritchie, GRS; Thormann, A et al. (2016): The Ensembl Variant Effect Predictor. In *Genome Biol* 17 (1), p. 122. DOI: 10.1186/s13059-016-0974-4.

McTavish, EJ; Decker, JE; Schnabel, RD; Taylor, JF; Hillis, DM (2013): New World cattle show ancestry from multiple independent domestication events. In *Proc Natl Acad Sci* 110 (15), E1398-E1406.

McTavish, EJ; Hillis, DM (2015): How do SNP ascertainment schemes and population demographics affect inferences about population history? In *BMC Genomics* 16 (1), p. 1.

Meuwissen, THE; Hayes, BJ; Goddard, ME (2001): Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. In *Genetics* 157 (4), pp. 1819–1829. Available online at https://www.genetics.org/content/157/4/1819.

Muir, WM; Wong, GK-S; Zhang, Y; Wang, J; Groenen, MAM; Crooijmans, RPMA et al. (2008): Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. In *Proc Natl Acad Sci* 105 (45), pp. 17312–17317. DOI: 10.1073/pnas.0806569105.

Nei, M (1972): Genetic Distance between Populations. In *The American Naturalist* 106 (949), pp. 283–292.

Nielsen, R (2004): Population genetic analysis of ascertained SNP data. In *Hum Genomics* 1 (3), p. 1.

Nielsen, R; Hubisz, MJ; Clark, AG (2004): Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. In *Genetics* 168 (4), pp. 2373–2382.

Nielsen, R; Signorovitch, J (2003): Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. In *Theor Popul Biol* 63 (3), pp. 245–255.

Novembre, J; Johnson, T; Bryc, K; Kutalik, Z; Boyko, AR; Auton, A et al. (2008): Genes mirror geography within Europe. In *Nature* 456, 98. DOI: 10.1038/nature07331.

Ober, U; Ayroles, JF; Stone, EA; Richards, S; Zhu, D; Gibbs, RA et al. (2012): Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. In *PLoS Genet* 8 (5), e1002685. DOI: 10.1371/journal.pgen.1002685.

Patterson, N; Moorjani, P; Luo, Y; Mallick, S; Rohland, N; Zhan, Y et al. (2012): Ancient admixture in human history. In *Genetics* 192 (3), pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

Perez-Enciso, M; Rincon, JC; Legarra, A (2015): Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. In *Genet Sel Evol* 47, p. 43. DOI: 10.1186/s12711-015-0117-5.

Peripolli, E; Munari, DP; Silva, M. V. G. B.; Lima, ALF; Irgang, R; Baldi, F (2017): Runs of homozygosity: current knowledge and applications in livestock. In *Anim Genet* 48 (3), pp. 255–271. DOI: 10.1111/age.12526.

Perkel, J (2008): SNP genotyping. Six technologies that keyed a revolution. In *Nature Methods* 5, 447. DOI: 10.1038/nmeth0508-447.

Platt, A; Horton, M; Huang, YS; Li, Y; Anastasio, AE; Mulyati, NW et al. (2010): The Scale of Population Structure in Arabidopsis thaliana. In *PLoS Genet* 6 (2), e1000843. DOI: 10.1371/journal.pgen.1000843.

Qanbari, S; Pausch, H; Jansen, S; Somel, M; Strom, T-M; Fries, R et al. (2014): Classic selective sweeps revealed by massive sequencing in cattle. In *PLoS Genet* 10 (2), e1004148. DOI: 10.1371/journal.pgen.1004148.

Qanbari, S; Pimentel, EC; Tetens, J; Thaller, G; Lichtner, P; Sharifi, AR; Simianer, H (2010): A genome‐wide scan for signatures of recent selection in Holstein cattle. In *Animal genetics* 41 (4), pp. 377–389.

Qanbari, S; Rubin, C-J; Maqbool, K; Weigend, S; Weigend, A; Geibel, J et al. (2019): Genetics of adaptation in modern chicken. In *PLoS Genet* 15 (4), e1007989. DOI: 10.1371/journal.pgen.1007989.

Quinto-Cortés, CD; Woerner, AE; Watkins, JC; Hammer, MF (2018): Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. In *Sci Rep* 8 (1), p. 10209. DOI: 10.1038/s41598-018-28539-y.

R Core Team (2017): R. A Language and Environment for Statistical Computing. Version 3.4.1. Vienna, Austria. Available online at https://www.R-project.org/.

Ramos, AM; Crooijmans, RPMA; Affara, NA; Amaral, AJ; Archibald, AL; Beever, JE et al. (2009): Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. In *PLoS One* 4 (8), e6524.

Rocha, D; Billerey, C; Samson, F; Boichard, D; Boussaha, M (2014): Identification of the putative ancestral allele of bovine single-nucleotide polymorphisms. In *J Anim Breed Genet* 131 (6), pp. 483–486. DOI: 10.1111/jbg.12095.

Sandenbergh, L; Cloete, SW; Roodt-Wilding, R; Snyman, MA; Bester-van der Merwe, AE (2016): Evaluation of the OvineSNP50 chip for use in four South African sheep breeds. In *S Afr J Anim Sci* 46 (1), pp. 89–93.

Schlötterer, C; Tobler, R; Kofler, R; Nolte, V (2014): Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. In *Nat Rev Genet* 15 (11), pp. 749–763.

Singh, N; Jayaswal, PK; Panda, K; Mandal, P; Kumar, V; Singh, B et al. (2015): Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. In *Sci Rep* 5, p. 11600.

Tixier-Boichard, M; Bed'hom, B; Rognon, X (2011): Chicken domestication. From archeology to genomics. In *Comptes rendus biologies* 334 (3), pp. 197–204.

Tosser-Klopp, G; Bardou, P; Bouchez, O; Cabau, C; Crooijmans, RPMA; Dong, Y et al. (2014): Design and characterization of a 52K SNP chip for goats. In *PLoS One* 9 (1), e86227.

Travis, AJ; Norton, GJ; Datta, S; Sarma, R; Dasgupta, T; Savio, FL et al. (2015): Assessing the genetic diversity of rice originating from Bangladesh, Assam and West Bengal. In *Rice* 8 (1), p. 35. DOI: 10.1186/s12284-015-0068-z.

UCSC (2016): Reference Genome Gallus gallus 5.0. Available online at http://hgdownload.soe.ucsc.edu/goldenPath/galGal5/bigZips/galGal5.fa.gz, checked on 10/25/2016.

Unterseer, S; Bauer, E; Haberer, G; Seidel, M; Knaak, C; Ouzunova, M et al. (2014): A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. In *BMC Genomics* 15 (1), p. 823.

van der Auwera, GA; Carneiro, MO; Hartl, C; Poplin, R; Del Angel, G; Levy-Moonshine, A et al. (2013): From FastQ data to high confidence variant calls. The Genome Analysis Toolkit best practices pipeline. In *Current protocols in bioinformatics* 43 (1), 11.10.1-11.10.33. DOI: 10.1002/0471250953.bi1110s43.

Wainschtein, P; Jain, DP; Yengo, L; Zheng, Z; Cupples, LA; Shadyab, AH et al. (2019): Recovery of trait heritability from whole genome sequence data. In *bioRxiv*. DOI: 10.1101/588020.

Wang, J; Skoog, T; Einarsdottir, E; Kaartokallio, T; Laivuori, H; Grauers, A et al. (2016): Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples. In *Sci Rep* 6 (1), p. 33256. DOI: 10.1038/srep33256.

Warren, WC; Hillier, LW; Tomlinson, C; Minx, P; Kremitzki, M; Graves, TA et al. (2017): A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. In *G3* 7 (1), pp. 109–117. DOI: 10.1534/g3.116.035923.

West, B; Zhou, B-X (1988): Did chickens go north? New evidence for domestication. In *Journal of archaeological science* 15 (5), pp. 515–533.

Wright, S (1949): The genetical structure of populations. In *Ann Eugen* 15 (1), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.

Chapter 3

How Imputation Can Mitigate SNP Ascertainment Bias

Johannes Geibel[12], Christian Reimer[12], Torsten Pook[12], Steffen Weigend[23], Annett Weigend[3], Henner Simianer[12]

[1] University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[2] University of Goettingen, Center for Integrated Breeding Research, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[3] Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, Höltystrasse 10, 31535 Neustadt-Mariensee, Germany

Published in *BMC Genomics* 22, 340
https://doi.org/10.1186/s12864-021-07663-6

Abstract

Background

Population genetic studies based on genotyped single nucleotide polymorphisms (SNPs) are influenced by a non-random selection of the SNPs included in the used genotyping arrays. The resulting bias in the estimation of allele frequency spectra and population genetics parameters like heterozygosity and genetic distances relative to whole genome sequencing (WGS) data is known as SNP ascertainment bias. Full correction for this bias requires detailed knowledge of the array design process, which is often not available in practice. This study suggests an alternative approach to mitigate ascertainment bias of a large set of genotyped individuals by using information of a small set of sequenced individuals via imputation without the need for prior knowledge on the array design.

Results

The strategy was first tested by simulating additional ascertainment bias with a set of 1,566 chickens from 74 populations that were genotyped for the positions of the Affymetrix Axiom™ 580k Genome-Wide Chicken Array. Imputation accuracy was shown to be consistently higher for populations used for SNP discovery during the simulated array design process. Reference sets of at least one individual per population in the study set led to a strong correction of ascertainment bias for estimates of expected and observed heterozygosity, Wright's Fixation Index and Nei's Standard Genetic Distance. In contrast, unbalanced reference sets (overrepresentation of populations compared to the study set) introduced a new bias towards the reference populations. Finally, the array genotypes were imputed to WGS by utilization of reference sets of 74 individuals (one per population) to 98 individuals (additional commercial chickens) and compared with a mixture of individually and pooled sequenced populations. The imputation reduced the slope between heterozygosity estimates of array data and WGS data from 1.94 to 1.26 when using the smaller balanced reference panel and to 1.44 when using the larger but unbalanced reference panel. This generally supported the results from simulation but was less favorable, advocating for a larger reference panel when imputing to WGS.

Conclusions

The results highlight the potential of using imputation for mitigation of SNP ascertainment bias but also underline the need for unbiased reference sets.

Keywords

SNP ascertainment bias; imputation; chickens; population genetics

Background

To perform cost- and computationally efficient, many of the population genetic studies of the last 10 years for humans (Novembre *et al.* 2008; Patterson *et al.* 2012), as well as for model- (Laurie *et al.* 2007; Platt *et al.* 2010) and agricultural species (Muir *et al.* 2008; Gibbs *et al.* 2009; Travis *et al.* 2015; Mayer *et al.* 2017) were based on single nucleotide polymorphisms (SNP), which were genotyped by commercially available SNP arrays. Those arrays are based on a non-random selection (ascertainment) of SNPs, and come with a bias relative to whole genome re-sequencing (WGS) data, widely known as SNP Ascertainment Bias (Clark *et al.* 2005; Albrechtsen *et al.* 2010; Lachance and Tishkoff 2013).

To design an array, SNPs initially need to be discovered in a finite set of sequenced individuals, the discovery panel. The chance to discover globally common SNPs is higher in this finite set of individuals than the chance to discover globally rare SNPs. This results in allele frequency spectra of arrays showing a shift towards common SNPs as compared to allele frequency spectra of WGS, which typically contain a high share of rare SNPs (Nielsen 2004). Additionally, the discovery panel is typically not a random sample from the global population of a species, but over-represents individuals from more intensively researched populations, e.g. humans of Yoruban, Japanese, Chinese and European descent (The International HapMap Project 2003), commercially bred taurine cattle breeds (Matukumalli *et al.* 2009) or commercial layer and broiler chicken lines (Kranis *et al.* 2013). SNPs that are common in those discovery populations are not necessarily globally common. As a consequence, allele frequency spectra of discovery populations are systematically skewed towards higher minor allele frequencies (MAF) than those of non-discovery populations (Nielsen 2004; Geibel *et al.* 2021). In extreme cases, e.g. when used for samples of other species, this can result in a lack of variable and thus informative SNPs on the array and therefore a shift of the frequency spectrum towards rare variants (Geibel *et al.* 2021).

The shift in the allele frequency spectra has an effect on population genetic estimators that depend on the allele frequency estimates. Exemplarily, the shift in allele frequencies towards common variants leads to an systematic overestimation of the heterozygosity of populations (Malomane *et al.* 2018; Geibel *et al.* 2021). The relative effect is stronger for populations that were part of the discovery set compared to populations that were not part of the discovery set (Geibel *et al.* 2021). Since commercially used breeds tend to be overrepresented in discovery sets (Matukumalli *et al.* 2009; Kranis *et al.* 2013), their diversity thus tends to be overestimated compared to non-commercial breeds not included in the discovery set (Geibel *et al.* 2021). Systematic differences in allele frequency spectra further increase estimates of genetic distances between populations which were part of the discovery set and those which were not (Albrechtsen *et al.* 2010).

The complex interaction between the size of the discovery panel and its restriction to a subset of populations makes it difficult to predict or outright correct for the effect of SNP ascertainment bias.

Further, attempts to implement bias-reduced estimators require strong assumptions on the design process of the used SNP array (Nielsen 2004), which is often not public knowledge or too complicated to be remodeled (Nielsen *et al.* 2004; Quinto-Cortés *et al.* 2018). Malomane *et al.* (2018) therefore screened different raw data filtering strategies on mitigation of ascertainment bias in SNP data and identified linkage pruning to result in slightly decreasing ascertainment bias. Due to strongly decreasing sequencing costs and the complexity of the ascertainment bias correction strategies, more and more studies started using WGS data for population genetic analysis during the last years (Qanbari *et al.* 2014; Qanbari *et al.* 2015; Lawal *et al.* 2018; Qanbari *et al.* 2019; Peripolli *et al.* 2020). However, costs for broad WGS based studies are still rather high, resulting in large-scale collaborations such as the 1000 Genomes Project (Auton *et al.* 2015), the 1000 Bull Genomes Project (Hayes and Daetwyler 2019), or the 1001 Arabidopsis Genomes Project (Alonso-Blanco *et al.* 2016).

A commonly used method to *in silico* increase the resolution of SNP data sets is imputation (Marchini and Howie 2010). Over the years a variety of imputation approaches (Li and Stephens 2003; Marchini *et al.* 2007; Howie *et al.* 2009; Delaneau *et al.* 2012; Sargolzaei *et al.* 2014; Money *et al.* 2015; Browning *et al.* 2018) have been proposed that utilize linkage, pedigree, and haplotype information. To increase the marker density, an additional reference panel of individuals that were genotyped/sequenced by the intended resolution is required to additionally infer information from SNPs missing on the respective lower density study set.

Imputation-based studies mostly either used a reference panel of the same population as the study set itself (Pausch *et al.* 2013; Heidaritabar *et al.* 2016; van den Berg *et al.* 2019) or utilized large global reference panels such as the 1000 Genomes (Huang *et al.* 2012; Artigas *et al.* 2015; Auton *et al.* 2015) or 1000 Bull genomes (Raymond *et al.* 2018; Hayes and Daetwyler 2019) projects. Especially for admixed or small endangered populations, the use of additional distantly related populations in the reference panel was investigated. On one hand, Brøndum *et al.* (2014), Ye *et al.* (2019) and Rowan *et al.* (2019) identified multi-breed reference panels to increase imputation accuracy especially in admixed breeds and for low frequent alleles when imputing from high-density genotypes to sequence data. On the other hand, Berry *et al.* (2014) observed that smaller within breed reference panels (140 - 688 reference cattle individuals per breed) performed always superior compared to the combined across breed reference panel when imputing from low density to high-density array genotypes. Korkuć *et al.* (2019) showed that adding 100 to 500 Holstein cattle sequences to a reference panel of 30 German Black Pied cattle significantly decreased the imputation accuracy in comparison to the pure panel when imputing from array to sequence data. Adding the same numbers of a multi-breed reference panel only outperformed the pure panel when at least 300 reference animals were added. Pook *et al.* (Pook *et al.* 2019) investigated the inclusion of chicken populations to the reference set which were differently distantly related to the study set. While error rates generally decreased for rare alleles, the inclusion of distantly related populations slightly increased error rates for previously good imputed SNPs. Overall, the ideal setup of a reference panel seems to be highly dependent on the application with positive effects for some, but also potential harm in other cases.

In this context, the current study aims at assessing the influence of a study design on SNP ascertainment bias, which uses a small number of sequenced chickens (the reference set) to *in silico* correct SNP ascertainment bias in a broad multi-population set of genotyped chickens (the study set) by imputation to sequence level. The general idea behind this design is to allow for a large sample size, which reduces sampling bias while keeping sequencing costs affordable as most individuals will only be genotyped. We, therefore, assessed the potential effects of this design by imputing *in silico* created low-density array data to high-density array data, and by imputing real high-density data to WGS data.

Material and Methods

Data

Three different sets of genomic data were used for this study:

Set 1: Individual sequence data of 68 chickens from 68 different populations, sequenced within the scope of the EU project Innovative Management of Animal Genetic Resources (IMAGE; www.imageh2020.eu) (Bortoluzzi *et al.* 2020). They were complemented by 25 sequences (17 + 8) from two commercial white layer lines, 25 sequences (19 + 6) from two commercial brown layer lines, and 40 sequences (20 each) from two commercial broiler lines (Qanbari *et al.* 2019). In total 158 sequences from 74 populations.

Set 2: Pooled sequence data from 37 populations (9-11 chickens per population) (Malomane *et al.* 2018). All except 4 chickens from two populations were part of set 3.

Set 3: Genotypes of 1,566 chickens from 74 populations, either genotyped (sub-set of the Synbreed Chicken Diversity Panel; SCDP) (Malomane *et al.* 2019) with the Affymetrix Axiom™ 580k Genome-Wide Chicken Array (Kranis *et al.* 2013), or complemented from set 1.

The intersection of the used data sets is shown in **Figure 3.1** and accession information of the raw data per sample can be found in Supplementary File 1. All three data sets came with their own characteristics. While individual sequences are considered to be the gold standard throughout this study, genotypes of the Affymetrix Axiom™ 580k Genome-Wide Chicken Array (Kranis *et al.* 2013) are biased towards variation which is common in the commercial chicken lines (Geibel *et al.* 2021) and pooled sequences only allow for an estimate of population allele frequencies and show a slight bias due to sample size and coverage (Supplementary File 2) (Futschik and Schlötterer 2010; Schlötterer *et al.* 2014).

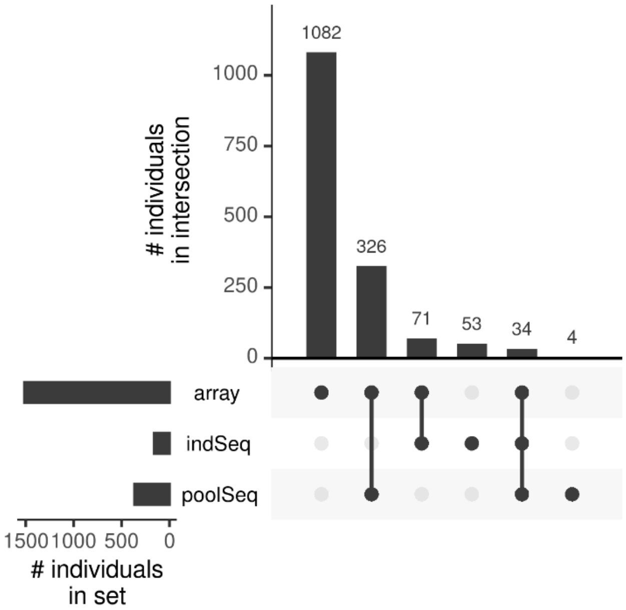

Figure 3.1: UpSet plot showing the distinct intersections of chickens between the used sequencing/ genotyping technologies. The left bar plot contains the total number of individuals that were genotyped (array), individually sequenced (indSeq), or pooled sequenced (poolSeq). The upper bar plot contains the number of individuals within each distinct intersection, indicated by the connected points below.

Calling of WGS SNPs and generation of genotype set

Alignment of the raw sequencing reads against the latest chicken reference genome GRCg6a (Genome Reference Consortium GRCg6a 2018) and SNP calling was conducted for individual and pooled sequenced data following GATK best practices (DePristo *et al.* 2011; van der Auwera *et al.* 2013). As the Affymetrix Axiom™ 580k Genome-Wide Chicken Array (Kranis *et al.* 2013) does not contain enough SNPs on chromosomes 30 – 33 for imputation (and chromosome 29 is not annotated in the reference genome), only up to chromosome 28 was used. This resulted in 20,829,081 biallelic SNPs on chromosomes 1 - 28 which were used in further analyses. Additionally, all individual sequences were genotyped for the positions of the Affymetrix Axiom™ 580k Genome-Wide Chicken Array (Kranis *et al.* 2013).

To ensure compatibility between Array- and WGS data, the genotypes of the Synbreed Chicken Diversity panel were lifted over from galGal5 to galGal6 and corrected for switches of reference and alternate alleles. Only SNPs with known autosomal position, call rates > 0.95 and genotype recall rates > 0.95 were further considered. MAF filters were later used when subsampling the different sets and thus not considered in this step. Further, missing genotypes were imputed using Beagle 5.0 (Browning *et al.* 2018) with ne=1000 (Pook *et al.* 2019) and the genetic map taken from Groenen *et al.* (2009). This resulted in a final set of 1,566 animals from 74 populations (18 - 37 animals per population) and 462,549 autosomal SNPs, further referred to as the **genotype set**.

As Malomane *et al.* (2018) described LD-based pruning as an effective filtering strategy to minimize the impact of ascertainment bias in SNP array data, the genotype set was additionally LD pruned using plink 1.9 (Chang *et al.* 2015) with --indep 50 5 2 flag. This reduced the genotype set to 136,755 SNPs (30 %) and will be referred to as **pruned genotype set**.

The description of the detailed pipeline can be found in Supplementary File 2.

Analyses based on simulation of ascertainment bias within the genotype set

A first comparison was based solely on the 15,868 SNPs of chromosome 10 of the genotype set which allowed for a high number of repetitions while still being based on a sufficiently sized chromosome. To simulate an ascertainment bias of known strength, an even more strongly biased array was designed *in silico* from the genotype set for each of the 74 populations (further called discovery populations) by using only SNPs with MAF > 0.05 within the according discovery population. This simulates the limitation to common variants in the discovery samples, which is the main reason for the ascertainment bias. Then, reference samples for imputation were chosen in five different ways with ten different numbers of reference samples and three repetitions per sampling:

1) allPop_74_740: Equally distributed across all populations by sampling one to ten chickens per population (74 - 740 reference samples).
2) randSamp_5_50: 5, 10, …, 50 randomly sampled chickens (5-50 reference samples).
3) randPop_5_50: Five chickens from each of one to ten randomly sampled populations (5 - 50 reference samples).
4) minPop_5_50: Five chickens from each of one to ten populations which were closest related to the discovery population, based on Nei’s Distance (Nei 1972; 5 - 50 reference samples).
5) maxPop_5_50: Five chickens from each of one to ten populations which were most distantly related to the discovery population, based on Nei’s Distance (Nei 1972; 5 - 50 reference samples).

This resulted in 2,200 repetitions of *in silico* array development and re-imputation per sampling strategy. The reference set was formed by sub-setting the total genotype matrix to SNPs with MAF > 0.01 within the reference samples and the reference samples chosen via the above-mentioned strategies. Imputation of the *in silico* arrays to the reference set was performed by running Beagle 5.0

(Browning *et al.* 2018) with ne=1000 (Pook *et al.* 2019), the genetic distances taken from Groenen *et al.* (2009) and the according reference set. The schematic workflow can be found in **Figure 3.2**.

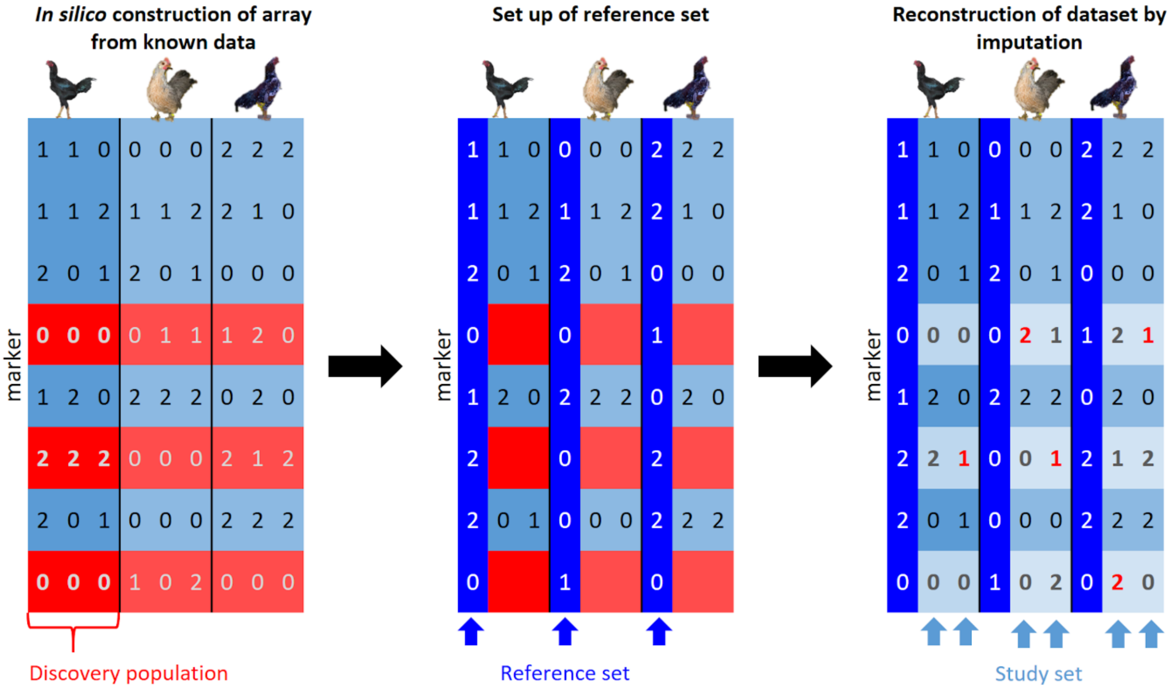

Figure 3.2: Schematic representation of the workflow of creating and re-imputing the *in silico* arrays. The starting point was a 0/1/2 coded marker matrix with SNPs in rows and individuals in columns (different populations separated by vertical lines). In a first step, an array (light blue rows) was constructed *in silico* from known data by setting all SNPs to missing which were invariable (MAF < 0.05, red rows) in the discovery population (first three columns). In a second step, a reference set (dark blue columns) was set up from animals for which complete knowledge of all SNPs was assumed. This Reference set was then used in a third step to impute the missing SNPs in the study set using Beagle 5.0 and resulting in a certain amount of imputation errors (red numbers).

Analyses were then based on comparisons between the *in silico* ascertained and later imputed sets and the genotype set, which was considered as the 'true' set for those comparisons.

Imputation of genotype set to sequence level

After the initial tests of the imputation strategies by the *in silico* designed arrays, we imputed the complete genotype set to sequence level, using the available individual sequences as the reference panel. In the first run, one reference sample per sequenced population was chosen (74 reference samples; 74_1perLine) which is equivalent to the first scenario allPop_74 of the *in silico* array imputation. As we had more than one sequenced individual for the commercial lines, the number of reference samples for the commercial lines was subsequently increased to five reference samples per line (up to 98 reference samples; 98_5perLine). Finally, we used all available individually sequenced animals as reference samples (158 reference samples; 158_all), which resulted in a strong imbalance towards the two broiler lines (20 reference samples per broiler line).

Parameter settings in Beagle were further tweaked by increasing the window parameter to 200 cM to ensure enough overlap between reference and study SNPs. This was needed as we observed low assembly quality and insufficient coverage of the array on the small chromosomes. Analyses were then based on comparisons between the genotype set, the pruned set or the imputed sets and the gold standard, the WGS data.

Comparison of population genetic estimators

Ascertainment bias shows its primary effect on the allele frequency spectrum. As populations are affected differently, we first concentrated on two heterozygosity estimates: expected (H_E) and observed (H_O) heterozygosity, which summarize per-population allele frequency spectra. We additionally included two allele frequency dependent distance measurements: Wright's fixation index (F_{ST}) (Wright 1949) and Nei's distance (D) (Nei 1972).

H_O, as the proportion of heterozygous genotypes in a population, could only be calculated when the genotypic status of a population was known (individual sequences or genotypes). In contrast, H_E could also be calculated from pooled sequences which allow the estimation of allele frequencies (p). Thereby, H_O and H_E (equation (3.1)) are calculated as average over all loci (l = 1, …, L).

$$H_E = \frac{\sum_l 2p_l(1-p_l)}{L} \tag{3.1}$$

As pooled sequence data comes with a slight but systematic underestimation of H_E (Futschik and Schlötterer 2010; Supplementary File 2), H_E for pooled sequences was multiplied with the correction factor $\frac{n}{n-1}$, introduced by Futschik and Schlötterer (2010), where n is the number of haplotypes in the pool. This partially corrected the H_E estimates for the bias introduced by pooled sequencing (Supplementary File 2).

D was calculated as given by equation (3.2), where D_{xy} accounts for the genetic distance between populations X and Y, while x_{il} and y_{il} represent the frequency of the ith allele at the lth locus in population X and Y, respectively.

$$D_{xy} = -\ln\left(\frac{\sum_l \sum_i x_{il} y_{il}}{\sqrt{\sum_l \sum_i x_{il}^2 \sum_l \sum_i y_{il}^2}}\right) \tag{3.2}$$

Pairwise F_{ST} values between populations X and Y were estimated using equation (3.3), where HT_l accounts for the H_E within the total population at locus l and $\overline{HS_l}$ for the mean H_E within the two subpopulations at locus l (Wright 1949).

$$F_{ST} = \frac{\sum_l (HT_l - \overline{HS_l})}{\sum_l HT_l} \tag{3.3}$$

D and F_{ST} both show a downward bias that is comparable to HE when estimated from pooled data (Supplementary File 2). The effect of ascertainment bias is much larger than the effect of pooling for D. In contrast, F_{ST} is generally robust against the effects of ascertainment bias when a sufficiently large discovery panel was used for array development (Albrechtsen *et al.* 2010). Therefore, it shows underestimation when calculated from pooled sequence data, which is larger than the effect of ascertainment bias (Supplementary File 2). We therefore could not dissect the effects of the two biases in the comparisons on sequence level and did not include F_{ST} there.

Having no ascertainment bias would mean that estimates of a respective set would lie on the line of identity (diagonal) when regressing the set against the true values. The magnitude of the bias can therefore be defined as the distance of the estimates to that line. We therefore regressed the estimates from biased data (y_{ij}) on the unbiased ones (x_{ij}) while fitting group specific intercepts ($group_i$) as well as group-specific slopes ($group_i \times \beta_i$) and a random error ($\epsilon_{ij}, \epsilon \sim N(0, \mathbf{I}\sigma_e^2)$) as in equation (3.4).

$$y_{ij} = group_i + group_i \times \beta_i x_{ij} + \epsilon_{ij} \tag{3.4}$$

The definition of a group describes for within-population estimators (e.g. H_E) whether a population was used for SNP discovery (discovery population), samples from that population were used as reference set (reference population) or none of both (application population). Note that in scenarios where reference individuals were present for every population, we only divided them into discovery and application populations. For between population estimators (F_{ST}, D), a group describes the according combination of the two involved population groups. Differences of the estimated slopes from one and the correlation between heterozygosity and distance estimates from biased and true set within groups were used as indicators for the magnitude of bias and random estimation error.

To get a measure for a fixed estimation error, we also calculated the mean overestimation across populations (j = 1 ... J) as in equation (3.5).

$$\textit{mean}\ \text{overestimation} = \frac{\sum_{j} \frac{\textit{biased estimate}_j - \textit{true estimate}_j}{\textit{true estimate}_j}}{J} \tag{3.5}$$

Note, that we had more than one (pooled) sequenced chicken for only 45 populations. Comparisons of population estimates on sequence level are therefore limited to 45 populations out of the 74 populations which were used as study and reference set for the imputation process.

Assessment of imputation accuracy

Assessment of imputation accuracy was done by using Pearson correlation (r) between true and imputed genotypes (Hickey *et al.* 2012; Berry *et al.* 2014) for the *in silico* designed arrays. Pearson correlation puts a higher relative weight on imputation errors in rare alleles than plain comparison of allele- or genotype concordance rates (Hickey *et al.* 2012). In case of the imputation to sequence level, we used leave-one-out validation to assess per-animal imputation accuracy. However, the leave-one-out validation in our case shows a slightly downward biased accuracy estimate for the non-commercial samples (**Figure S 11**, Supplementary File 2). For validation, the only sequenced sample of those populations was the test sample, which had to be removed from the reference set. Therefore, no closely related sample to the test sample remained in the reference set and the accuracy was subsequently underestimated. We additionally used the internal Beagle quality measure, the dosage r-squared (DR2) (Browning and Browning 2009) to evaluate per-SNP imputation accuracy. This, however, only shows the theoretical imputation accuracy and cannot capture biases due to biased reference sets.

Results

In silico array to genotype

As expected, the *in silico* ascertained sets showed a strong overestimation of the H_E for nearly all populations in all cases. The overestimation was much stronger for populations used for SNP discovery (**Figure 3.3 A**). Imputation using an equal number of reference samples per population (scenario allPop_74_740) massively decreased this bias (**Figure 3.3 B**). The correction became stronger with an increasing number of reference populations.

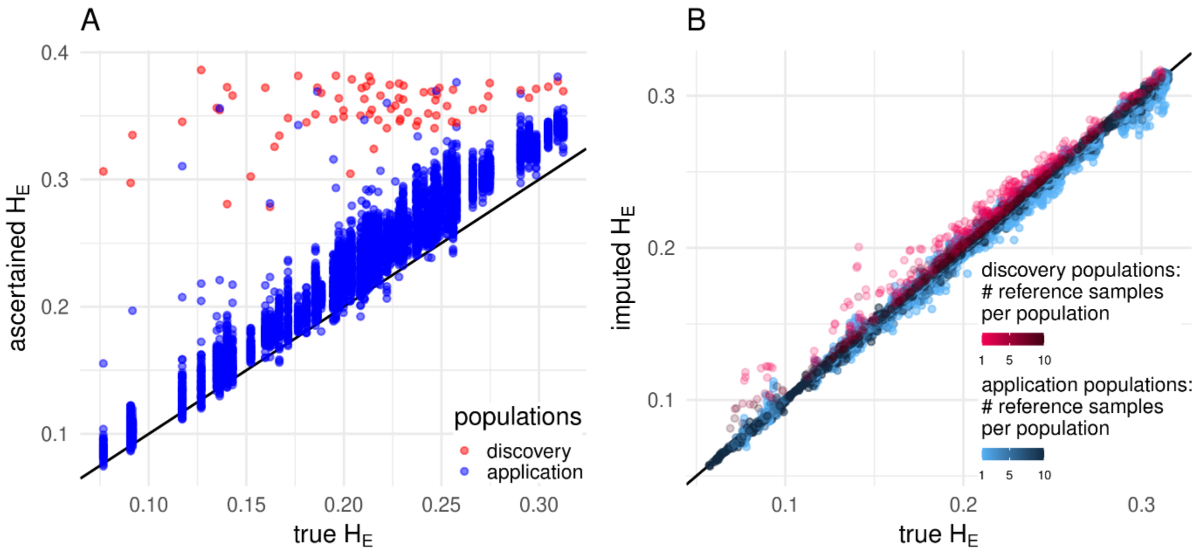

Figure 3.3: True H_E vs. ascertained H_E (A) and imputed H_E (B) by population group. For the imputed case, the strategy of using the same number of reference samples per population (allPop_74_740) is shown, an increase in the number of reference samples per population (1-10) is marked by an increasing color gradient and the line of identity is marked by a solid black line.

To get an impression on the strength of the correction and the needed size of the reference panel, **Figure 3.4** compares the correlation by population group, the slope for the within-group regression of the true H_E and H_O vs. the ascertained/ imputed cases and mean overestimation for strategy allPop_74_740. It shows that the effects of ascertainment bias were stronger for H_E than for H_O. Imputation when using the reference set with just one individual per population corrects the initially much lower correlation within population group to > 0.99. While slope and mean overestimation are also pushed promptly towards the intended values of one and zero respectively for the non-discovery populations, there remains a small bias for the discovery populations, which decreases with an increasing number of reference samples.

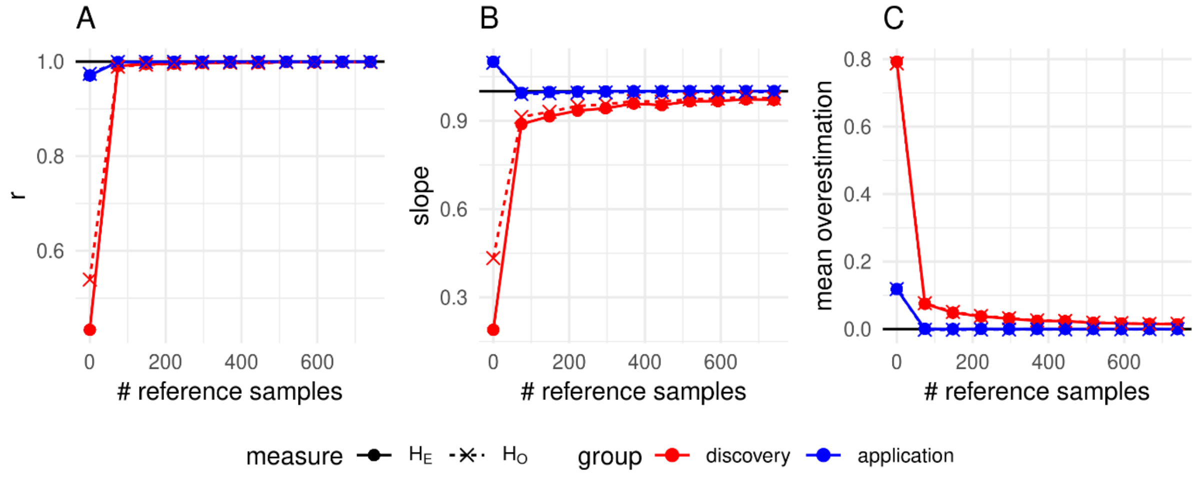

Figure 3.4: Development of correlation within population group (A), slope (B) and mean overestimation (C) of the regression lines for the two heterozygosity estimates when distributing the reference samples equally across all populations (allPop_74_740). The intended value for unbiasedness and minimum variance is marked as dense black horizontal line. Note that the case without imputation is consistent with zero reference samples.

The effects were observed in a comparable manner for the other imputation strategies (**Figure S 3**). Due to smaller reference panels, the correction effect of the imputation was generally worse than for strategy allPop_74_740. Interestingly, when limiting the reference samples to a small number of populations (strategies randPop_5_50, minPop_5_50, maxPop_5_50), we observed a newly introduced bias towards the reference populations (**Figure S 3**). This effect was strongest for strategy maxPop_5_50, where we chose the reference populations with a maximum distance from the discovery population. However, increasing the number of reference samples minimized the bias of reference and discovery populations with all strategies.

The effects of ascertainment bias were less pronounced in the distance measurements (D and F_{ST}; **Figure S 4**) than in the heterozygosity estimates. The bias was thereby only of numerical relevance, when estimating the distances between populations which belong to differently strongly biased population groups and was partly increased for some population groups by imputation with unbalanced reference samples (**Figure S 5**). Note that F_{ST} was, all in all, less affected than D.

The reduction of ascertainment bias was accompanied by high per-animal imputation accuracies (r). Strategy allPop_74 (one reference individual per population) resulted in a median imputation accuracy of 0.94. Increasing the number of reference individuals subsequently increased the accuracy up to 0.99 for 10 reference individuals per population (allPop_740). The accuracy was consistently higher for individuals which were part of the discovery population (**Figure 3.5**). Accuracies were lower for the other strategies, mainly due to a maximum number of 50 reference individuals, which are fewer than the 74 of allPop_74. Detailed results for imputation accuracy can be found in **Figure S 2** and Supplementary File 2.

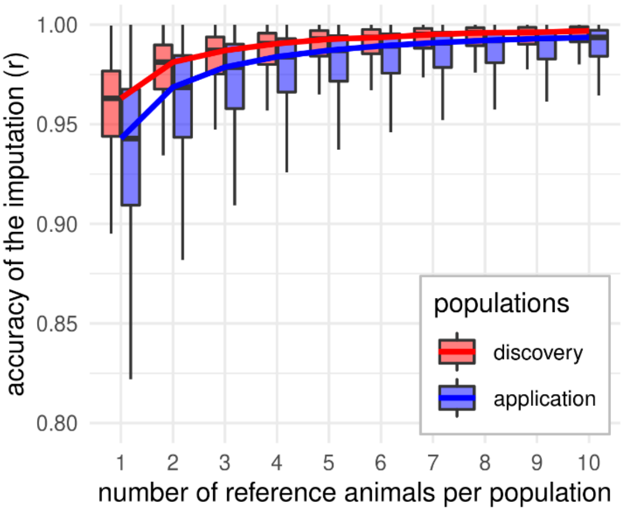

Figure 3.5: Development of the per-animal imputation accuracy for the *in silico* array to genotype set imputation with an increasing number of reference animals per population. Individuals are grouped on whether they belong to the population used for SNP discovery or not and reference individuals were chosen as in scenario allPop_74_740. The lines show the trend of the median and outliers are not shown in the plot as they do not add valuable information due to the high number of repetitions.

Genotype to sequence

The effect of imputation to WGS on ascertainment bias of H_E is shown in **Figure 3.6**. Given the situation that we cannot completely exclude pooling bias for the pooled sequenced samples (Supplementary File 2), only the effect on the individually sequenced samples can be discussed with adequate reliability. While the regression of array-based H_E estimates on sequence-based H_E estimates showed a slope of 1.94 for the individually sequenced populations, the linkage pruning slightly reduced this slope to 1.71. The clearly best result was achieved with imputation to WGS (slope = 1.26; 74_1perLine; **Figure 3.6 A**). This effect was also observed when considering all samples. However, note that there is also a slight effect of the remaining pooling bias, which cannot be separated from ascertainment bias for the pooled sequenced populations. Slightly increasing the reference panel (**Figure 3.6 B**) up to five samples per commercial line (98_5perLine) does not show any effect, while using all commercial samples in the reference panel (158_all) and thereby clearly biasing the reference panel towards the broiler samples increases HE again for all samples (slope = 1.44).

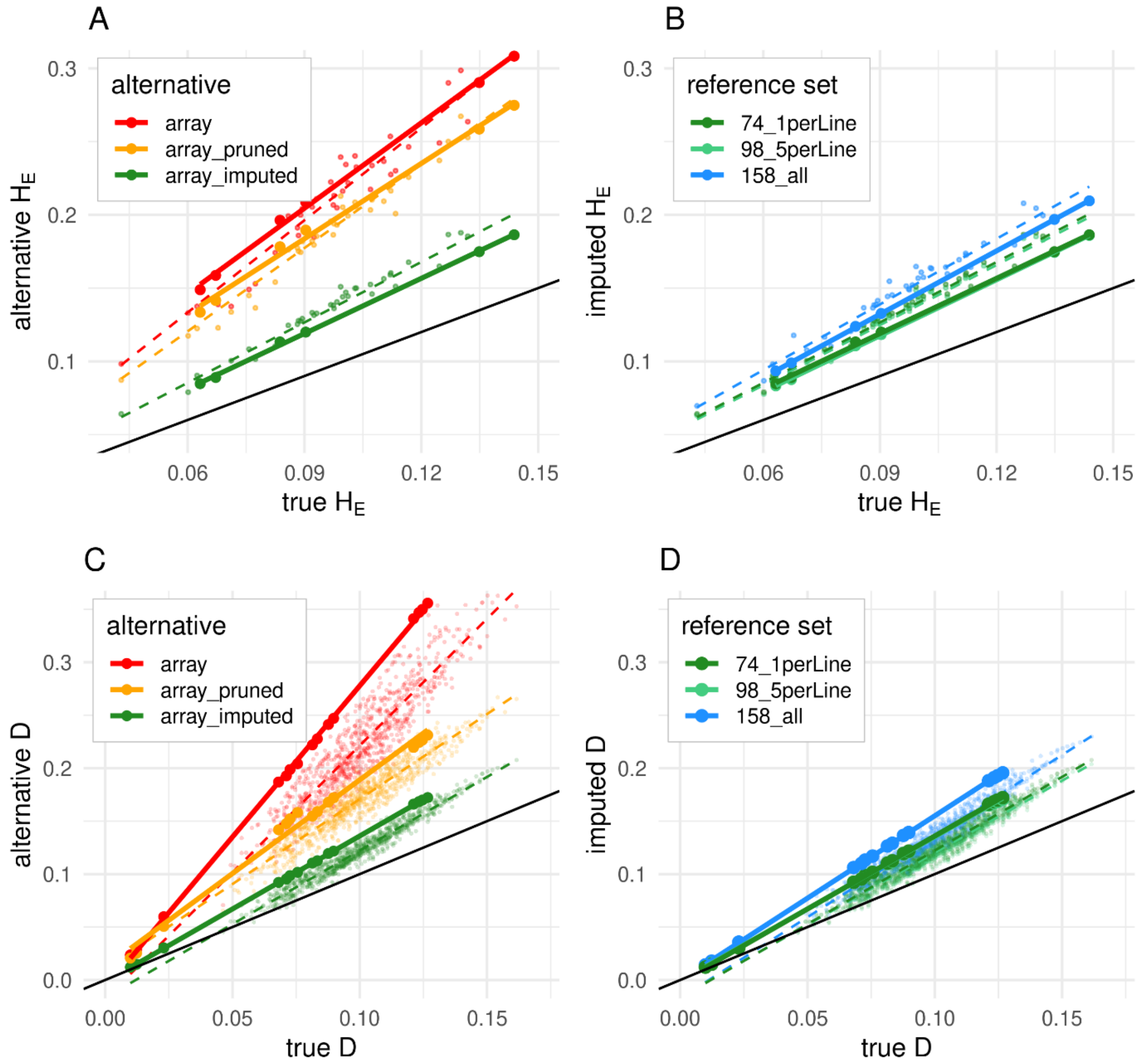

Figure 3.6: Effect of different correction strategies on ascertainment bias for expected heterozygosity (H_E; A + B) and for Nei's standard genetic distance (D; C + D). A + C – uncorrected array, linkage pruned array and imputed array (reference set 74_1perLine) based vs. sequence-based H_E/ D. B + D – array imputed with different reference sets vs. sequence-based H_E/ D. The solid black line represents the line of identity, the solid colored lines are regression lines within the individually sequenced populations (larger points) and the dashed lines regression lines within all populations which include individually and pooled (small points) sequenced populations. Note that there is also an effect of pooled sequencing which affects the 'true' values of the pooled sequenced populations.

The results for Nei's standard genetic distance (D;**Figure 3.6**) showed the same pattern as the results for HE. The slope for distances between individually sequenced populations decreased from 2.86 (array) and 1.77 (array_pruned) to 1.38 (imputed, 74_1perLine). The unbalanced reference panel 158_all then again increased the slope to 1.56. The correlation for all distances, besides being also influenced by pooling bias and therefore being a rough estimate, was increased from 0.93 (array) respectively 0.95 (array_pruned) to 0.98 (all reference sets).

The overall imputation accuracy was lower than the one obtained for *in silico* array to array imputation. Increasing the number of commercial reference samples only resulted in increased imputation accuracies for the commercial samples. See **Supplementary File 2**, **Table S 1**, **Figure S 6**, **Figure S 7** and **Figure S 11** for details.

Discussion

Overall performance of the correction method

Imputation of SNP data sets from lower to higher density is a commonly used technique to either increase the resolution of data sets (Pausch *et al.* 2013; Heidaritabar *et al.* 2016; Raymond *et al.* 2018) or make them comparable across different platforms (Al-Tassan *et al.* 2015; Bouwman *et al.* 2018). The according studies mostly use a relatively homogeneous study set and a closely related and large reference set (Pausch *et al.* 2013; Heidaritabar *et al.* 2016). However, studies exist which investigate the effect of increasing the reference set to a multi-population reference set to use an increased number of reference haplotypes (Berry *et al.* 2014; Brøndum *et al.* 2014; Korkuć *et al.* 2019; Rowan *et al.* 2019; Ye *et al.* 2019). To our knowledge, we here present the first study that investigates the use of a relatively small and diverse reference set on a large and diverse study set to correct for a genotyping platform-specific bias, the SNP ascertainment bias.

This approach intends that single imputation errors do not harm, if the mean across the genome, presented by different population genetic estimators, shows unbiased results with minimum variance. Therefore, imputation to WGS level using a comparably small reference panel can be used to correct for the ascertainment bias of commercial arrays.

Especially the *in silico* ascertained SNP arrays showed that even a very small reference panel consisting of one individual of each population showed very good results for all investigated estimators (e.g. correlation between biased H_E and true H_E of initially < 0.5 for the discovery populations increased to > 0.99; **Figure 3.4**; **Figure S 4**) and became better with an increasing number of reference populations. The results were less beneficial for the real WGS data, but also showed a strong decrease of the slope towards one. From the imputed *in silico* arrays, we could additionally realize a fast closing of the gap of the stronger overestimation of heterozygosity within discovery populations and the less severe overestimation in non-discovery populations. This also seemed to be the case when imputing to WGS level where we observed that the slope within the commercial populations (closely related to discovery populations of the real array) decreased more than the slope within all populations due to imputation. However, this observation in the WGS data has to be regarded with caution, as we additionally identified a non-negligible bias due to pooled sequencing which interfered with the assessment of ascertainment bias and which was, in our study, confounded with the difference between commercial populations (sequenced individually) and non-commercial populations (sequenced as pools).

The use of WGS information via imputation also consistently showed better results in regard of reduction of ascertainment bias than using linkage pruned array SNPs which was reported to be an effective filtering strategy for ascertainment bias mitigation by Malomane *et al.* (2018).

Generally, the effect of imputation on the investigated estimators was shown to be comparable across estimators, regardless of their initial reaction to ascertainment bias. An interesting side observation was that F_{ST} did not show any ascertainment bias on the real array data (**Figure S 10)** when calculated in the form of summing the numerator across SNPs and dividing by the sum of the denominator as calculated in this study. F_{ST} was only affected when used to estimate differentiation between the discovery- and non-discovery populations in the simulated array data, whose heterozygosity estimates were affected by ascertainment bias to a different degree. This strongly supports the findings of Albrechtsen *et al.* (2010), who showed F_{ST} to be relatively robust against the effects of ascertainment bias.

We also investigated the effect of differently sized and constructed reference sets for imputation. Generally, larger reference sets increased the accuracy of imputation and thus decreased the ascertainment bias more than smaller reference sets. The best results were achieved when the reference set was as evenly distributed across the study set as possible. When reference populations were closely related to the discovery population, reduction in imputation quality and increase in ascertainment bias were less severe in case of unbalanced reference sets than if distantly related reference populations were used. This suggests that variation within study- and reference set needs to show enough overlap to achieve sufficient imputation accuracy and therefore reduction of ascertainment bias.

Results from literature suggest that multi-breed reference panels generally increase imputation accuracy especially for rare variants and within admixed populations (Brøndum *et al.* 2014; Rowan *et al.* 2019; Ye *et al.* 2019). Additionally, Rowan *et al.* (2019) argue that they do not seem to introduce variation at a relevant scale for markers for which the breeds are actually fixed. However, some studies also showed that strongly unbalanced reference sets can reduce imputation accuracy (Berry *et al.* 2014; Korkuć *et al.* 2019). In this study, including additional reference samples in a biased way when going from reference set 74_1perLine to 158_all increased the effects of ascertainment bias on H_E and D. Additionally, only the commercial populations, for which we increased the number of reference samples, showed a gain in per-animal imputation accuracy (**Figure S 11).** However, theoretical imputation accuracies rather increased than decreased (**Figure S 6**; **Table S 1**) for previously poorly imputed SNPs. The increase in accuracies for poorly imputed SNPs supports the findings of Brøndum

(2014), Rowan *et al.* (2019) and Ye *et al.* (2019) that multi-breed reference panels rather help in getting better imputation results. However, the missing gain in per-animal accuracy for non-commercial populations together with the observed bias in the leave-one-out validation for our sparse reference set highlights the still existing need for closely related individuals as shown by Berry *et al.* (Berry *et al.* 2014), Korkuć *et al.* (Korkuć *et al.* 2019) and Pook *et al.* (Pook *et al.* 2019). The worsening effect on bias correction, however, highlights the main reason for ascertainment bias. One can only identify variation which is present in the investigated samples. When developing an array, one observes the variation in the discovery set, while in our case we observed variation in the reference set used for imputation. An overrepresentation of certain populations in the reference set biases estimators towards variation present in those populations. Besides the aforementioned effects in the imputations to WGS, we also observed this by an increasing bias for the unbalanced reference sets in the *in silico* array imputations (**Figure S 3**, **Figure S 5**). Therefore, it is crucial to use a reference set for imputation which covers the intended range of variation.

Besides the previously described effects of imputation on ascertainment bias, we also identified an effect of array design on imputation accuracy. Discovery populations show higher imputation accuracies than non-discovery populations (**Figure 3.5**). As markers on arrays are more representative for discovery populations than non-discovery populations, relatively more of the genetic variability in discovery populations is explained by the array and imputation is more accurate on average.

Conclusion

The problem to which we provide at least a partial solution is that relevant population genetic parameters are systematically biased through the design process of SNP arrays. Imputation was able to mitigate this SNP ascertainment bias in our samples for all studied estimators (H_E, H_O, F_{ST}, D), measured as correlation, average relative difference and slope of the regression line when comparing the biased estimators to the according gold standard. The effect was already present when using a very small reference set of only one sequenced individual per population. Imputation also performed better than simple filtering strategies based on the array data alone. However, when using imputation for ascertainment bias reduction care has to be taken in designing an evenly spaced reference panel to not introduce a new bias towards variation present in the reference panel while missing variants of other populations. We also suggest using a larger reference panel than the one which was available for this study to achieve better results. Additionally, we observed an effect of array design on imputation accuracy as discovery populations showed a higher imputation accuracy than non-discovery populations. This should be taken into account when designing studies based on imputed SNPs by choosing an appropriate genotyping array for the intended study populations.

Abbreviations

D: Nei's Standard Genetic Distance; F_{ST}: Wright's Fixation Index; H_E: expected heterozygosity; H_O: observed heterozygosity; MAF: minor allele frequency; r: Pearson Correlation; SNP: single nucleotide polymorphism; WGS: whole genome sequencing

Declarations

Ethics approval and consent to participate

The study did not involve new treatment of animals as only published data was used. DNA samples for all already published raw data were taken from a data base established during the project AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; www.aviandiv.fli.de) and later extended by samples of the project SYNBREED (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de). Blood sampling was done in strict accordance to the German animal welfare regulations, with written consent of the animal owners and was approved by the at the according times ethics responsible persons of the Friedrich-Loeffler-Institut. According to German animal welfare regulations, notice was given to the responsible institution, the Lower Saxony State Office for Consumer Protection and Food Safety (33.9-42502-05-10A064).

Consent for publication

Not applicable

Availability of data and materials

Raw sequencing and genotyping data were previously published by different studies. The repository information for each sample can be found in Supplementary File 1. All datasets generated by analyses during this study from the raw data are additionally available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

The positions of JG and TP were partly funded by the project IMAGE - Innovative Management of Animal Genetic Resources (www.imageh2020.eu, funded by the EU Horizon 2020 research and innovation program No 677353). The sampling, genotyping and sequencing of the chicken populations involved funding by the EC project AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; www.aviandiv.fli.de), the German Federal Ministry of Education and Research (BMBF) via the

SYNBREED project (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de) and the EU project IMAGE - Innovative Management of Animal Genetic Resources (www.imageh2020.eu, funded by the EU Horizon 2020 research and innovation program No 677353). Further, the publication fees were covered by the Open Access Publication Funds of the University of Goettingen. Open Access funding enabled and organized by Projekt DEAL.

Authors' contributions

JG conceptualized and designed the study, analyzed and interpreted the data, wrote the initial draft and revised the manuscript. CR and TP substantially contributed to design of the study, interpretation of the data and revision of the manuscript. SW substantially contributed to acquisition and interpretation of the data and revised the manuscript. AW substantially contributed to acquisition and curation of the data. HS substantially contributed to conception and design of the study, interpretation of the data and revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgment

Not applicable

Supplementary

The supplementary files can be accessed via the original publication (https://doi.org/10.1186/s12864-021-07663-6).

Supplementary File 1: Accession Information of raw data per sample

Supplementary File 2: Supplementary Methods

Figure S 1: Recall rates for samples which were genotyped as well as sequenced per SNP (A; B) and per animal (C; D); before (A; C; red) and after (B; D; blue) correction of potential reference allele switches in the genotype data.

Figure S 2: Development of the per-animal imputation accuracy with an increasing number of reference animals per population.

Figure S 3: Development of correlations within population group (r), slope and mean overestimation of the regression lines for H_E and H_O estimates and different reference panel strategies

Figure S 4: Development of correlation within population group (A), slope (B) and intercept (C) of the regression lines for D and F_{ST} when distributing the reference samples equally over all populations (allPop_74_740).

Figure S 5: Development of correlation within population group (r), slope and mean overestimation of the regression lines for Nei's Distance (D) and F_{ST} estimates and different reference panel strategies.

Figure S 6: Distribution of DR2 values by chromosome and reference set.

Figure S 7: Two-dimensional distributions of DR2 values vs. MAF by chromosome when imputed with the reference set 74_1perLine.

Figure S 8: Effect of pooled sequencing and the correction factor of Futschik and Schlötterer (2010) on expected heterozygosity (HE) and ascertainment bias.

Figure S 9: Effect of pooled sequencing on the expression of the ascertainment bias in Nei's standard genetic distance (D).

Figure S 10: Effect of pooled sequencing on the expression of the ascertainment bias in Wright's fixation index (F_{ST}).

Figure S 11: Per animal imputation accuracies (r) for the array to sequence imputation from leave-one-out validation.

Table S 1: Quantiles of theoretical imputation accuracies (DR2) by reference set

References

Albrechtsen, A; Nielsen, FC; Nielsen, R (2010): Ascertainment biases in SNP chips affect measures of population divergence. In *Mol Biol Evol* 27 (11), 2534–2547.

Alonso-Blanco, C; Andrade, J; Becker, C; Bemm, F; Bergelson, J; Borgwardt, KM et al. (2016): 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. In *Cell* 166 (2), pp. 481–491. DOI: 10.1016/j.cell.2016.05.063.

Al-Tassan, NA; Whiffin, N; Hosking, FJ; Palles, C; Farrington, SM; Dobbins, SE et al. (2015): A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. In *Sci Rep* 5 (1), p. 10442. DOI: 10.1038/srep10442.

Artigas, MS; Wain, LV; Miller, S; Kheirallah, AK; Huffman, JE; Ntalla, I et al. (2015): Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. In *Nat Commun* 6 (1), p. 8658. DOI: 10.1038/ncomms9658.

Auton, A; Abecasis, GR; Altshuler, DM; Durbin, RM; Bentley, DR; Chakravarti, A et al. (2015): A global reference for human genetic variation. In *Nature* 526 (7571), pp. 68–74. DOI: 10.1038/nature15393.

Berry, DP; McClure, MC; Mullen, MP (2014): Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. In *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* 131 (3), pp. 165–172. DOI: 10.1111/jbg.12067.

Bortoluzzi, C; Megens, H-J; Bosse, M; Derks, MFL; Dibbits, B; Laport, K et al. (2020): Parallel Genetic Origin of Foot Feathering in Birds. In *Mol Biol Evol* 37 (9), pp. 2465–2476. DOI: 10.1093/molbev/msaa092.

Bouwman, AC; Daetwyler, HD; Chamberlain, AJ; Ponce, CH; Sargolzaei, M; Schenkel, FS et al. (2018): Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. In *Nat Genet* 50 (3), pp. 362–367. DOI: 10.1038/s41588-018-0056-5.

Brøndum, RF; Guldbrandtsen, B; Sahana, G; Lund, MS; Su, G (2014): Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. In *BMC Genomics* 15 (1), p. 728. DOI: 10.1186/1471-2164-15-728.

Browning, BL; Browning, SR (2009): A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. In *Am J Hum Genet* 84 (2), pp. 210–223. DOI: 10.1016/j.ajhg.2009.01.005.

Browning, BL; Zhou, Y; Browning, SR (2018): A One-Penny Imputed Genome from Next-Generation Reference Panels. In *Am J Hum Genet* 103 (3), pp. 338–348. DOI: 10.1016/j.ajhg.2018.07.015.

Chang, CC; Chow, CC; Tellier, LC; Vattikuti, S; Purcell, SM; Lee, JJ (2015): Second-generation PLINK. Rising to the challenge of larger and richer datasets. In *GigaScience* 4, p. 7. DOI: 10.1186/s13742-015-0047-8.

Clark, AG; Hubisz, MJ; Bustamante, CD; Williamson, SH; Nielsen, R (2005): Ascertainment bias in studies of human genome-wide polymorphism. In *Genome Res* 15 (11), pp. 1496–1502.

Delaneau, O; Marchini, J; Zagury, J-F (2012): A linear complexity phasing method for thousands of genomes. In *Nature Methods* 9 (2), pp. 179–181.

DePristo, MA; Banks, E; Poplin, R; Garimella, KV; Maguire, JR; Hartl, C et al. (2011): A framework for variation discovery and genotyping using next-generation DNA sequencing data. In *Nat Genet* 43, 491. DOI: 10.1038/ng.806.

Futschik, A; Schlötterer, C (2010): The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. In *Genetics* 186 (1), pp. 207–218. DOI: 10.1534/genetics.110.114397.

Geibel, J; Reimer, C; Weigend, S; Weigend, A; Pook, T; Simianer, H (2021): How array design creates SNP ascertainment bias. In *PLoS One* 16 (3), e0245178. DOI: 10.1371/journal.pone.0245178.

Genome Reference Consortium GRCg6a (2018): GRCg6a chicken reference genome. Available online at http://hgdownload.soe.ucsc.edu/goldenPath/galGal6/bigZips/galGal6.fa.gz, checked on 4/9/2019.

Gibbs, RA; Taylor, JF; van Tassell, CP; Barendse, W; Eversole, KA; Gill, CA et al. (2009): Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. In *Science* 324 (5926), pp. 528–532. DOI: 10.1126/science.1167936.

Groenen, MAM; Wahlberg, P; Foglio, M; Cheng, HH; Megens, H-J; Crooijmans, RPMA et al. (2009): A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. In *Genome Res* 19 (3), pp. 510–519.

Hayes, BJ; Daetwyler, HD (2019): 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. In *Annual Review of Animal Biosciences* 7 (1), pp. 89–102. DOI: 10.1146/annurev-animal-020518-115024.

Heidaritabar, M; Calus, MPL; Megens, H-J; Vereijken, A; Groenen, MAM; Bastiaansen, JWM (2016): Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. In *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* 133 (3), pp. 167–179. DOI: 10.1111/jbg.12199.

Hickey, JM; Crossa, J; Babu, R; los Campos, G de (2012): Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. In *Crop Science* 52 (2), p. 654. DOI: 10.2135/cropsci2011.07.0358.

Howie, BN; Donnelly, P; Marchini, J (2009): A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. In *PLoS Genet* 5 (6), pp. 1–15. DOI: 10.1371/journal.pgen.1000529.

Huang, J; Ellinghaus, D; Franke, A; Howie, B; Li, Y (2012): 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. In *European Journal Of Human Genetics* 20 (7), pp. 801–805. DOI: 10.1038/ejhg.2012.3.

Korkuć, P; Arends, D; Brockmann, GA (2019): Finding the Optimal Imputation Strategy for Small Cattle Populations. In *Frontiers in genetics* 10, p. 52. DOI: 10.3389/fgene.2019.00052.

Kranis, A; Gheyas, AA; Boschiero, C; Turner, F; Le Yu; Smith, S et al. (2013): Development of a high density 600K SNP genotyping array for chicken. In *BMC Genomics* 14 (1), p. 59. DOI: 10.1186/1471-2164-14-59.

Lachance, J; Tishkoff, SA (2013): SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. In *Bioessays* 35 (9), pp. 780–786.

Laurie, CC; Nickerson, DA; Anderson, AD; Weir, BS; Livingston, RJ; Dean, MD et al. (2007): Linkage Disequilibrium in Wild Mice. In *PLoS Genet* 3 (8), e144. DOI: 10.1371/journal.pgen.0030144.

Lawal, RA; Al-Atiyat, RM; Aljumaah, RS; Silva, P; Mwacharo, JM; Hanotte, O (2018): Whole-Genome Resequencing of Red Junglefowl and Indigenous Village Chicken Reveal New Insights on the Genome Dynamics of the Species. In *Frontiers in genetics* 9, p. 264. DOI: 10.3389/fgene.2018.00264.

Li, N; Stephens, M (2003): Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. In *Genetics* 165 (4), pp. 2213–2233.

Malomane, DK; Reimer, C; Weigend, S; Weigend, A; Sharifi, AR; Simianer, H (2018): Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. In *BMC Genomics* 19 (1), p. 22. DOI: 10.1186/s12864-017-4416-9.

Malomane, DK; Simianer, H; Weigend, A; Reimer, C; Schmitt, AO; Weigend, S (2019): The SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high genomic resolution. In *BMC genomics* 20 (1), p. 345. DOI: 10.1186/s12864-019-5727-9.

Marchini, J; Howie, B (2010): Genotype imputation for genome-wide association studies. In *Nat Rev Genet* 11, 499 EP -. DOI: 10.1038/nrg2796.

Marchini, J; Howie, B; Myers, S; McVean, G; Donnelly, P (2007): A new multipoint method for genome-wide association studies by imputation of genotypes. In *Nat Genet* 39 (7), pp. 906–913.

Matukumalli, LK; Lawley, CT; Schnabel, RD; Taylor, JF; Allan, MF; Heaton, MP et al. (2009): Development and characterization of a high density SNP genotyping assay for cattle. In *PLoS One* 4 (4), e5350.

Mayer, M; Unterseer, S; Bauer, E; Leon, N de; Ordas, B; Schön, C-C (2017): Is there an optimum level of diversity in utilization of genetic resources? In *Theor Appl Genet* 130 (11), pp. 2283–2295. DOI: 10.1007/s00122-017-2959-4.

Money, D; Gardner, K; Migicovsky, Z; Schwaninger, H; Zhong, G-Y; Myles, S (2015): LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. In *G3* 5 (11), p. 2383. DOI: 10.1534/g3.115.021667.

Muir, WM; Wong, GK-S; Zhang, Y; Wang, J; Groenen, MAM; Crooijmans, RPMA et al. (2008): Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. In *Proc Natl Acad Sci* 105 (45), pp. 17312–17317. DOI: 10.1073/pnas.0806569105.

Nei, M (1972): Genetic Distance between Populations. In *The American Naturalist* 106 (949), pp. 283–292.

Nielsen, R (2004): Population genetic analysis of ascertained SNP data. In *Hum Genomics* 1 (3), p. 1.

Nielsen, R; Hubisz, MJ; Clark, AG (2004): Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. In *Genetics* 168 (4), pp. 2373–2382.

Novembre, J; Johnson, T; Bryc, K; Kutalik, Z; Boyko, AR; Auton, A et al. (2008): Genes mirror geography within Europe. In *Nature* 456, 98. DOI: 10.1038/nature07331.

Patterson, N; Moorjani, P; Luo, Y; Mallick, S; Rohland, N; Zhan, Y et al. (2012): Ancient admixture in human history. In *Genetics* 192 (3), pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

Pausch, H; Aigner, B; Emmerling, R; Edel, C; Götz, K-U; Fries, R (2013): Imputation of high-density genotypes in the Fleckvieh cattle population. In *Genetics, selection, evolution : GSE* 45, p. 3. DOI: 10.1186/1297-9686-45-3.

Peripolli, E; Reimer, C; Ha, N-T; Geibel, J; Machado, MA; Panetto, João Cláudio do Carmo et al. (2020): Genome-wide detection of signatures of selection in indicine and Brazilian locally adapted taurine cattle breeds using whole-genome re-sequencing data. In *BMC Genomics* 21 (1), p. 624. DOI: 10.1186/s12864-020-07035-6.

Platt, A; Horton, M; Huang, YS; Li, Y; Anastasio, AE; Mulyati, NW et al. (2010): The Scale of Population Structure in Arabidopsis thaliana. In *PLoS Genet* 6 (2), e1000843. DOI: 10.1371/journal.pgen.1000843.

Pook, T; Mayer, M; Geibel, J; Weigend, S; Cavero, D; Schoen, CC; Simianer, H (2019): Improving Imputation Quality in BEAGLE for Crop and Livestock Data. In *G3*, g3.400798.2019. DOI: 10.1534/g3.119.400798.

Qanbari, S; Pausch, H; Jansen, S; Somel, M; Strom, T-M; Fries, R et al. (2014): Classic selective sweeps revealed by massive sequencing in cattle. In *PLoS Genet* 10 (2), e1004148. DOI: 10.1371/journal.pgen.1004148.

Qanbari, S; Rubin, C-J; Maqbool, K; Weigend, S; Weigend, A; Geibel, J et al. (2019): Genetics of adaptation in modern chicken. In *PLoS Genet* 15 (4), e1007989. DOI: 10.1371/journal.pgen.1007989.

Qanbari, S; Seidel, M; Strom, T-M; Mayer, KFX; Preisinger, R; Simianer, H (2015): Parallel Selection Revealed by Population Sequencing in Chicken. In *Genome Biol Evol* 7 (12), pp. 3299–3306. DOI: 10.1093/gbe/evv222.

Quinto-Cortés, CD; Woerner, AE; Watkins, JC; Hammer, MF (2018): Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. In *Sci Rep* 8 (1), p. 10209. DOI: 10.1038/s41598-018-28539-y.

Raymond, B; Bouwman, AC; Schrooten, C; Houwing-Duistermaat, J; Veerkamp, RF (2018): Utility of whole-genome sequence data for across-breed genomic prediction. In *Genetics Selection Evolution* 50 (1), p. 27. DOI: 10.1186/s12711-018-0396-8.

Rowan, TN; Hoff, JL; Crum, TE; Taylor, JF; Schnabel, RD; Decker, JE (2019): A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. In *Genetics Selection Evolution* 51 (1), p. 77. DOI: 10.1186/s12711-019-0519-x.

Sargolzaei, M; Chesnais, JP; Schenkel, FS (2014): A new approach for efficient genotype imputation using information from relatives. In *BMC Genomics* 15 (1), p. 478. DOI: 10.1186/1471-2164-15-478.

Schlötterer, C; Tobler, R; Kofler, R; Nolte, V (2014): Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. In *Nat Rev Genet* 15 (11), pp. 749–763.

The International HapMap Project (2003). In *Nature* 426 (6968), pp. 789–796.

Travis, AJ; Norton, GJ; Datta, S; Sarma, R; Dasgupta, T; Savio, FL et al. (2015): Assessing the genetic diversity of rice originating from Bangladesh, Assam and West Bengal. In *Rice* 8 (1), p. 35. DOI: 10.1186/s12284-015-0068-z.

van den Berg, S; Vandenplas, J; van Eeuwijk, FA; Bouwman, AC; Lopes, MS; Veerkamp, RF (2019): Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. In *Genetics Selection Evolution* 51 (1), p. 2. DOI: 10.1186/s12711-019-0445-y.

van der Auwera, GA; Carneiro, MO; Hartl, C; Poplin, R; Del Angel, G; Levy-Moonshine, A et al. (2013): From FastQ data to high confidence variant calls. The Genome Analysis Toolkit best practices pipeline. In *Current protocols in bioinformatics* 43 (1), 11.10.1-11.10.33. DOI: 10.1002/0471250953.bi1110s43.

Wright, S (1949): The genetical structure of populations. In *Ann Eugen* 15 (1), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.

Ye, S; Yuan, X; Huang, S; Zhang, H; Chen, Z; Li, J et al. (2019): Comparison of genotype imputation strategies using a combined reference panel for chicken population. In *Animal : an international journal of animal bioscience* 13 (6), pp. 1119–1126. DOI: 10.1017/S1751731118002860.

Chapter 4

Assessment of linkage disequilibrium patterns between structural variants and single nucleotide polymorphisms in three commercial chicken populations

Johannes Geibel[12], Nora Paulina Praefke[12], Steffen Weigend[23], Henner Simianer[12], Christian Reimer[12]

[1] University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[2] University of Goettingen, Center for Integrated Breeding Research, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[3] Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, Höltystrasse 10, 31535 Neustadt-Mariensee, Germany

Submitted to BMC Genomics and available as preprint on Research Square
https://doi.org/10.21203/rs.3.rs-861830/v1

Abstract

Background

Structural variants (SV) are causative for some prominent phenotypic traits of livestock as different comb types in chickens or color patterns in pigs. Their effects on production traits are also increasingly studied. Nevertheless, accurately calling SV remains challenging. It is therefore of interest, whether close-by single nucleotide polymorphisms (SNPs) are in strong linkage disequilibrium (LD) with SVs and can serve as markers. Literature comes to different conclusions on whether SVs are in LD to SNPs on the same level as SNPs to other SNPs. The present study aimed to generate a precise SV callset from whole-genome short-read sequencing (WGS) data for three commercial chicken populations and to evaluate LD patterns between the called SVs and surrounding SNPs. It is thereby the first study that assessed LD between SVs and SNPs in chickens.

Results

The final callset consisted of 12,294,329 bivariate SNPs, 4,301 deletions (DEL), 224 duplications (DUP), 218 inversions (INV) and 117 translocation breakpoints (BND). While average LD between DELs and SNPs was at the same level as between SNPs and SNPs, LD between other SVs and SNPs was strongly reduced (DUP: 40 %, INV: 27 %, BND: 19 % of between-SNP LD). A main factor for the reduced LD was the presence of local minor allele frequency differences, which accounted for 50 % of the difference between SNP – SNP and DUP – SNP LD. This was potentially accompanied by lower genotyping accuracies for DUP, INV and BND compared with SNPs and DELs. An evaluation of the presence of tag SNPs (SNP in highest LD to the variant of interest) further revealed DELs to be slightly less tagged by WGS SNPs than WGS SNPs by other SNPs. This difference, however, was no longer present when reducing the pool of potential tag SNPs to SNPs located on four different chicken genotyping arrays.

Conclusions

The results implied that genomic variance due to DELs in the chicken populations studied can be captured by different SNP marker sets as good as variance from WGS SNPs, whereas separate SV calling might be advisable for DUP, INV, and BND effects.

Keywords

Chickens; single nucleotide polymorphisms; structural variants; linkage disequilibrium

Background

A type of genomic variation that affects large regions of the genome is caused by structural variants (SV). SVs can alter the total genome size by deleting (deletions, DEL), duplicating (duplications, DUP) or inserting (insertions, INS) longer stretches of DNA (unbalanced SV). Those SVs are often referred to as copy number variations (CNV). In contrast, inversions (INV) and translocations (TRA) do not affect the length of the genome (balanced SV) (Ho *et al.* 2019). Especially unbalanced SVs are assumed to come with a strong functional impact on the phenotype, e.g. by strong deleterious effects of DELs which can remove complete genes (Feuk *et al.* 2006) or by DUPs that increase numbers of cis-regulatory elements (Feuk *et al.* 2006; Lee *et al.* 2021). SVs and complex combinations of multiple SVs are also known to be causative for some of the most prominent phenotypic breed characteristics of livestock breeds as walnut- and rose comb in chickens (Imsland *et al.* 2012) or belted color patterns and dominant-white color in pigs (Rubin *et al.* 2012).

The power for detection of SVs of certain types and sizes, however, is highly technology-dependent in various aspects (Ho *et al.* 2019). During the last two decades, technologies evolved that increased the resolution and accuracy of SV detection at the submicroscopic level. Array-based comparative genomic hybridization (aCGH) allowed the detection of long CNVs >35 kb (Feuk *et al.* 2006). The development and increased use of microarrays led to technologies that either detect DELs from characteristics of population-level single nucleotide polymorphism (SNP) genotypes (Conrad *et al.* 2006; McCarroll *et al.* 2006) or utilized signal intensity information (Wang *et al.* 2007). The increasing availability of short-read sequences during the last decade led to the development of multiple SV detection algorithms which use read depth distributions (Abyzov *et al.* 2011; Ho *et al.* 2019) and/ or information from split reads and insert size distributions of paired-end reads, potentially combined with local assembly procedures (Rausch *et al.* 2012; Layer *et al.* 2014; Chen *et al.* 2016; Ho *et al.* 2019). However, short-read-based methods still come with a variety of limitations due to the short read sizes which highly vary between the algorithms (Escaramís *et al.* 2015; Ho *et al.* 2019) and especially a general deficit in calling INS (Delage *et al.* 2020). Therefore, current state-of-the-art methods nowadays utilize the information of PacBio or Nanopore long-read sequencing or linked-read technologies as HI-C (Sedlazeck *et al.* 2018), but the availability of these types of sequencing data is still very limited for the majority of intensively researched livestock species.

Other than for SVs, the use of SNPs has become routine over the last two decades. Therefore, large whole-genome-sequencing (WGS) reference panels (Auton *et al.* 2015; Hayes and Daetwyler 2019) and collections of individuals, which were genotyped by microarrays and phenotyped in routine breeding programs or during large-scale research projects (Malomane *et al.* 2019), exist. Given the complexity of SV detection, it is of interest to know which part of the effects of SVs on the phenotype is already captured by potential linkage disequilibrium (LD) between the SV of interest and nearby SNPs. Strong LD would allow for the inclusion of those effects in e.g. genomic prediction without the need for a separate SV analysis.

LD between two variants can be measured using a variety of estimators (reviewed e.g. by Qanbari 2020), of which the squared correlation of haplotypes (r^2) is probably the most prominent one. It can be interpreted as the amount of information of a variant that is captured by another one. However, its upper limit is defined by the difference in minor allele frequency (ΔMAF) between the two variants (VanLiere and Rosenberg 2008). The overall strength of LD is highly population depended and closely linked to the effective population size (Qanbari 2020). LD thereby shows a characteristic decay pattern of mean LD by distance. However, for many applications as genome-wide association studies (GWAS), the interest is more in the maximum observed LD of a causal variant to a close-by so-called tag SNP, which can capture the effect as a marker genotype.

By now, a bunch of studies has addressed the question of LD between SVs and surrounding SNPs in humans with contrasting results. Generally, common DELs were shown to be in good LD to SNPs by most of the studies (Hinds *et al.* 2006; McCarroll *et al.* 2006; Cooper *et al.* 2008; Conrad *et al.* 2010; Mills *et al.* 2011), but some found this LD to be weaker than SNP – SNP LD (Redon *et al.* 2006; Kato *et al.* 2009). Literature additionally suggests, that rare DELs are weaker tagged (tag SNP is SNP with highest LD to the variant within a defined distance) than common DELs (McCarroll *et al.* 2008; Conrad *et al.* 2010) and DUP were in weaker LD to SNPs than DELs (Kato *et al.* 2009; Conrad *et al.* 2010; Sudmant *et al.* 2015). It was additionally shown that the availability of tag SNPs for SVs depends on the SNP panel used (WGS vs. different arrays) (Cooper *et al.* 2008; Conrad *et al.* 2010; Mills *et al.* 2011). A further effect that was found is the location of the SV on the genome. Regions of segmental duplications are known to trigger recurrent SV formation by non-allelic homologous recombination and therefore lead to SV hotspots (Gu *et al.* 2008; Ho *et al.* 2019). A closer look at those regions by Locke *et al.* (2006) found very few of those CNV to be tagged by surrounding SNPs.

Reduced LD between SNPs and SVs can have diverse reasons. A main factor is the increased possibility of the occurrence of recurrent mutations in regions of low sequence complexity by non-allelic homologous recombination (NAHR; Gu *et al.* 2008). SVs from recurrent mutational events then show reduced LD to variants from a unique mutational event (Locke *et al.* 2006; McCarroll *et al.* 2006). LD between SNPs and SVs may further be decreased by different selectional properties of SNPs and SV (Berger *et al.* 2015), MAF differences between SVs and SNPs (VanLiere and Rosenberg 2008), or ascertainment of SNPs for arrays that excludes regions of high structural complexity due to technical reasons (Lee *et al.* 2020). Additionally, known problems with SV calling accuracy (Ho *et al.* 2019) may lead to a high share of false-positive SV calls and therefore on average low LD to more accurately called SNPs.

For livestock, results on SV – SNP LD are very rare, even though a high number of publications targeted SV. Based on a GWAS on 26,362 Holstein dairy cattle 50 k genotypes, Xu *et al.* (2014) found a quarter of CNVs that were significantly associated with milk traits not being tagged by adjacent SNPs. The same was observed by Lee *et al.* (2020) who investigated functional and population genetic features of CNV regions in two dairy cattle breeds, also called from a 50 k SNP array. They identified a weak linkage between CNV regions and SNPs, which was slightly stronger between DELs and SNPs than between DUPs and SNPs. Wang *et al.* (2015) included a local LD analysis around CNVs (called from SNP arrays) that were significantly associated with production traits in pigs. Four out of eight significantly associated CNVs overlapped haploblocks of non-significant SNPs, but only one CNV was found 300 kb downstream of significantly associated SNPs. Note that this, however, may also have been an artifact of a much stronger correction for multiple testing in SNPs than in CNVs.

In chickens, a variety of studies investigated CNVs on a quantitative basis. The studies either used aCGH (Wang *et al.* 2010; Wang *et al.* 2012; Crooijmans *et al.* 2013; Tian *et al.* 2013; Han *et al.* 2014), utilized signal information of SNP arrays (Jia *et al.* 2013; Zhang *et al.* 2014; Rao *et al.* 2016; Gorla *et al.* 2017; Strillacci *et al.* 2017; Lin *et al.* 2018) via PennCNV (Wang *et al.* 2007), or read depth information of short-read sequences (Fan *et al.* 2013; Yan *et al.* 2015; Sohrabi *et al.* 2018; Seol *et al.* 2019; Weng *et al.* 2020). There were only three studies that also included non-CNV SVs (Kerstens *et al.* 2011; Fan *et al.* 2013; Weng *et al.* 2020). None of the studies analyzed the LD patterns of the variants.

Aim of the study

This is the first study that assessed SV – SNP LD in chickens to investigate the usefulness of SNP markers in capturing SV-based genomic variance. We, therefore, identified SVs from paired-end short-read sequences in three commercial chicken populations (white layers, brown layers, broilers), thoroughly described the SV callset, and assessed the strength of LD between those SVs and SNPs. We also identified major reasons for some existing differences to SNP – SNP LD and evaluated the performance of four available SNP arrays to tag SVs.

Results

Calling results and description of variants

For the study, paired-end short-read sequences of 90 chickens from three populations (25 commercial white layers, WL; 25 commercial brown layers, BL; 40 commercial broiler chickens, BR) were used. The raw data was first published by Qanbari *et al.* (2019) who described the studied populations in more detail. SNP genotypes were retrieved from a previous study (Geibel *et al.* 2021a). SVs were called by a consensus calling approach, which used three paired-end and split-read-based tools, followed by a strict filtering procedure that further utilized read-depth and SNP information. Finally, the remaining SV calls were visually checked by evaluating samplots (Belyeu *et al.* 2021) for each variant, the merged SNP and SV set was phased, and missing genotypes were imputed. The filtering procedure retained 12,294,329 bivariate SNPs, 4,301 DELs, 224 DUPs, 218 INVs, and 117 translocation breakpoints (break ends; BND) on chromosomes 1 - 33. Note that all INS were filtered out due to missing support by at least two variant callers.

Figure 4.1 A shows the length distribution of the called SVs. DELs were on average shortest with a median of 443 bp and a maximum of 67,037 bp. DUPs (median = 12,285 bp; maximum = 778,041 bp) were larger than DELs and INVs were largest (median = 25,643 bp; maximum = 5,795,187 bp). BNDs only indicate translocation breakpoints and, therefore, do not come with length information. The called SNPs in total accounted for 1.28 % of the autosomal reference genome length, while DELs covered 0.35 %, DUPs 0.39 %, and INVs 2.80 % of the chicken genome. The distributions by individuals can be found in **Figure 4.1 B**. We additionally checked how much of the autosomal reference genome is homozygously deleted in the chickens. This number varied from 0.045 % (135 kb) to 0.076 % (727 kb) with BL showing a larger size of homozygously deleted reference genome than WL and BR (**Figure 4.1 C**).

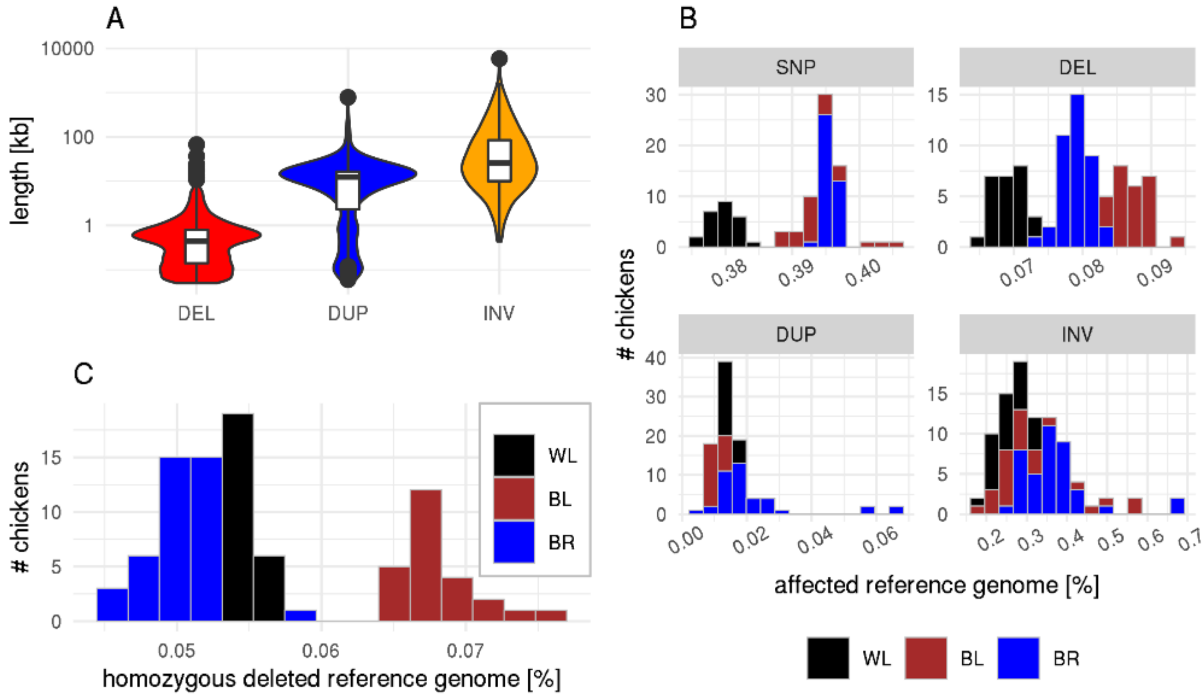

Figure 4.1: Length distribution of SVs (A), percent of affected autosomal reference genome by individual and variation type (B), and percent of homozygously deleted reference genome by individuals (C). The size in B is calculated as the average between the haplotypes of an individual affected by the non-reference allele. Note the log-scaled y-axis in A. Per-breed bars in the histograms are stacked on each other.

We further checked for chromosome-wise differences in the number of called variants by regressing the relative number of called variants per chromosome on the relative chromosome length (**Figure S 3**). SNPs did not show any difference to the line of identity (slope = 1.00, p = 1.00), while DELs (slope = 1.28, p = 1.4e-4) and INVs (slope = 1.39, p = 6.1e-9) showed a significant bias towards larger chromosomes. DUPs (slope = 1.13, p = 0.34) and BNDs (slope = 1.14, p = 0.17) also showed a numerical bias towards larger chromosomes, which, however, was not significant. Note that the R^2 value of the model was comparably small with 0.39.

Distributions of minor allele frequencies (MAF; **Figure 4.2**) revealed a slight (DEL) to strong (DUP) shift towards rare variants compared with SNPs for DELs and DUPs, while INVs and BNDs showed a slight shift towards more common variants.

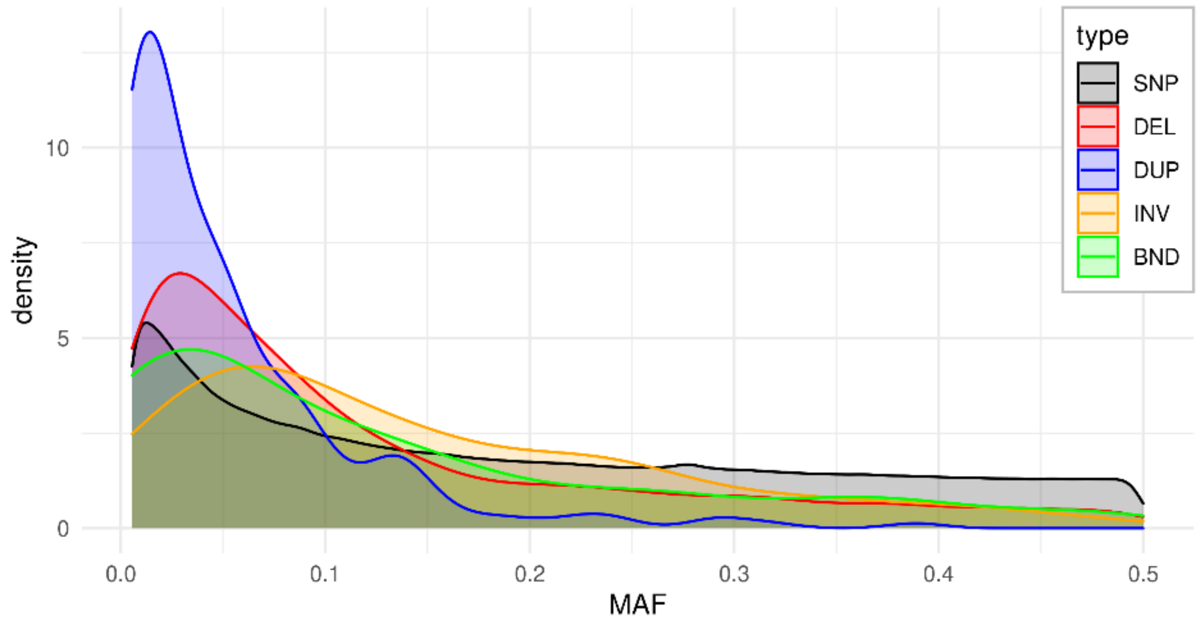

Figure 4.2: Distribution of minor allele frequency (MAF) across all samples by variant type.

Variant effect predictions of Ensembl-vep (McLaren *et al.* 2016) classified 98.48 % of the impacts of SNPs on genes as MODIFIER, 1.14 % as LOW, 0.37 as MODERATE and only 0.01 % as HIGH. DEL impacts were classified only in 0.41 % of the cases other than MODIFIER (MODERATE = 0.01 %; HIGH = 0.40 %), while DUP impacts were classified as HIGH in 9.95 % of the cases (MODIFIER = 90.05 %). In contrast, INV and BND impacts were completely classified as MODIFIER. Further results of VEP are summarized in **Figure S 4**.

LD decay

To assess the information content of SNPs on SVs, we calculated the LD between SVs and all bivariate SNPs up to 100 kb apart from the breakpoints as squared haplotype correlation (r^2). Note, that SNPs that were located on SVs were excluded from the analysis, as their calls may be directly influenced by the SV. To get a baseline for comparisons, we also calculated the SNP – SNP LD within this distance.

Mean SNP – SNP r^2 was highest in WL (0.51 within 500 bp), followed by BL (0.41) and BR (0.26). The DEL – SNP LD decay curve follows closely the pattern of the SNP – SNP LD decay (**Figure 4.3**). Even though the level of LD was strongly reduced for the other variant types, a slight decay curve with increasing distance was still noticeable. Due to the small number of called DUPs in WL, the decay curve strongly fluctuated in this population. However, BR and BL gave some evidence that the DUP – SNP and INV – SNP decay curves were comparable, while BND – SNP decay came with a slightly lower level of LD.

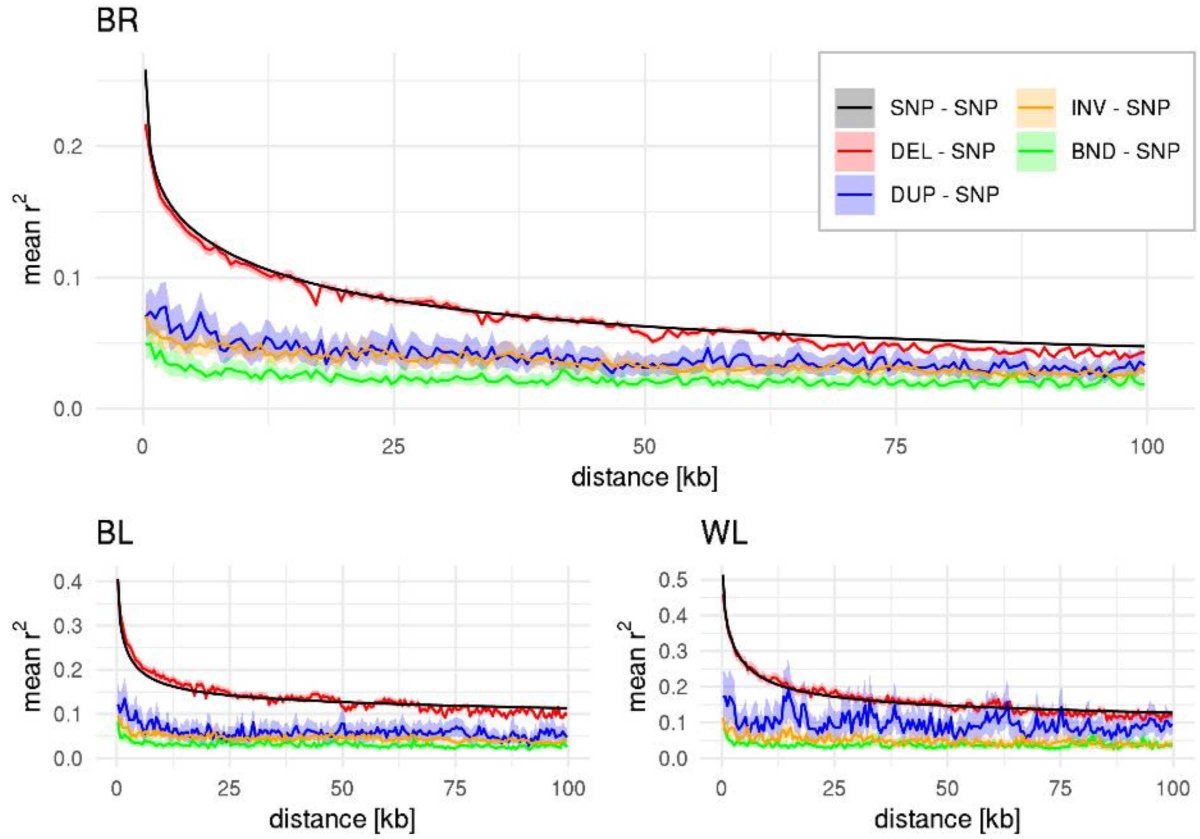

Figure 4.3: LD decay in the broiler (BR), brown layer (BL) and white layer (WL) chickens. The LD is presented as mean $\boldsymbol{r^2}$ in 500 bp distance bins and the shaded areas represent Bonferroni-corrected 95 % bootstrap confidence intervals. For SNP – SNP distance bins with > 1M $\boldsymbol{r^2}$ values, no confidence intervals were estimated.

To quantify the difference in LD between variants and populations and account for the population-specific level of LD, we expressed the mean LD in the 500 bp bins relative to the SNP – SNP LD and further averaged those values for the first 10 bins (**Table 4.1**). This revealed comparable values within variants and across populations of less than 12 % difference. Across all populations, DEL – SNP LD was on the same level as SNP – SNP LD, while DUP – SNP LD was ~40 %, INV – SNP ~27 % and BND – SNP ~19 % of SNP – SNP LD within 5 kb distance. Note that the relative r^2 was not necessarily constant across the complete range of 100 kb (**Figure S 5**).

Table 4.1: SV – SNP r^2/r_S^2 relative to the SNP – SNP r^2/r_S^2

Type	All			BR			BL			WL		
	r^{2*}	r_S^{2**}	Δ^{***}	r^{2*}	r_S^{2**}	Δ^{***}	r^{2*}	r_S^{2**}	Δ^{***}	r^{2*}	r_S^{2*}	Δ^{***}
DEL – SNP	100.1 ± 6.1	98.8 ± 4.3	-1.3	95.4 ± 4.1	94.2 ± 2.5	-1.2	107.0 ± 3.2	103.2 ± 1.2	-3.8	98.1 ± 3.4	98.8 ± 2.5	0.7
DUP – SNP	39.9 ± 6.8	68.2 ± 8.9	28.3	39.5 ± 5.8	66.7 ± 4.3	27.2	41.1 ± 7.0	65.6 ± 9.9	24.5	39.1 ± 8.1	72.3 ± 10.5	33.2
INV – SNP	26.8 ± 5.2	46.0 ± 4.3	19.2	32.6 ± 2.4	46.8 ± 1.7	14.2	26.0 ± 2.4	50.1 ± 3.4	24.1	21.6 ± 3.1	50.1 ± 6.2	28.5
BND – SNP	18.5 ± 3.6	46.9 ± 5.4	28.4	22.4 ± 2.3	50.4 ± 3.7	28.0	18.0 ± 1.9	44.6 ± 5.1	26.6	15.3 ± 2.0	45.5 ± 5.7	30.3

* Means of the first ten 500 bp bins relative to the SNP – SNP r^2 [%] ± standard deviations [%]

** Means of the first ten 500 bp bins relative to the SNP – SNP r_S^2 [%] ± standard deviations [%]

*** Difference between relative r^2 and relative r_S^2

Effect of allele frequency

Figure 4.2 revealed differences in the MAF spectra of the variant types. We therefore further evaluated local MAF differences (ΔMAF) within-population by comparing ΔMAF for the SNP – SNP and SV – SNP pairs within 5 kb distance. This revealed elevated ΔMAF for DUP – SNP, INV – SNP, and BND – SNP pairs compared to SNP – SNP and DEL – SNP pairs in BL and WL (**Figure S 7**, **Figure S 8**), but not in BR (**Figure S 6**). As the upper bound of r^2 directly depends on ΔMAF (VanLiere and Rosenberg 2008), we investigated which part of the observed differences in the LD decay curves is due to the observed allele frequency differences. For this, we used the standardized squared correlation coefficient (r_S^2), which expresses r^2 as the proportion of the maximum possible r^2 given ΔMAF of the two variants (VanLiere and Rosenberg 2008) and thereby excludes effects of different allele frequencies on r^2. Mean r_S^2 values (**Figure S 1**) were generally higher than mean r^2 values (**Figure 4.3**) due to the removal of the allele-frequency-dependent component. While the r_S^2 values of DEL – SNP relative to the SNP – SNP values (**Table 4.1**) were on a comparable level of > 94 % as the relative r^2 values (-3.8 % to +0.7 %), the relative r_S^2 values of DUPs, INVs and BNDs were between 14 % and 33 % higher than the according relative r^2 values. The relative r_S^2 values for the complete range of 100 kb are shown in **Figure S 9**.

Absence of homozygous SV genotypes

During the investigation of the reasons for the lower level of LD between non-DEL SVs and SNPs, we realized a strong absence of homozygous calls for DUPs, INVs, and BNDs, but not for DEL (exemplarily demonstrated for BR in **Figure 4.4 A**). To check whether this deviation is due to small variant allele frequencies, we calculated the deviation to Hardy-Weinberg-Equilibrium (HWE) and tested those for significance, using a Haldane Exact test under usage of the R package HardyWeinberg 1.7.2

(Graffelman 2015) (exemplarily shown for BR in **Figure 4.4 B**). Homozygous DEL calls deviated into positive as well as into negative direction from the HWE. Homozygous calls for the other SV classes instead nearly exclusively deviated into a negative direction for all populations and only negative deviations were significant.

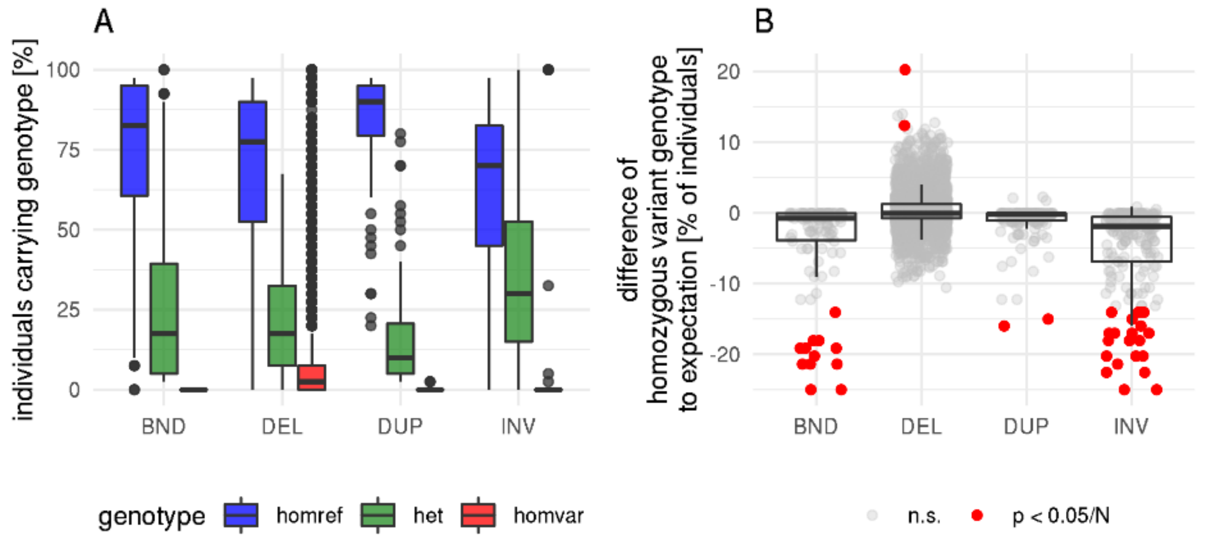

Figure 4.4: Percentage of individuals carrying SV genotype (A) and deviations of homozygous variant genotypes from the Hardy-Weinberg-Expectation (B) in the broiler population for each called SV. Deviations from HWE were tested by a Haldane Exact test under usage of the R package HardyWeinberg 1.7.2 (Graffelman 2015). Bonferroni correction of the p values was applied within SV class. Homref – homozygous for the reference allele; het – heterozygous; homvar – homozygous for the variant allele; n.s. – not significant. Comparable figures for WL and BL can be found in **Figure S 11** and **Figure S 12**.

We tried to tackle the effect of this problem by correlating the 0/1/2 coded SNP genotypes with a coverage-dependent measure of copy number for DELs and DUPs, the Duphold Flanking Fold Change (DHFFC; Pedersen and Quinlan 2019). However, as the DHFFC was also used for filtering, the results of this are potentially confounded and are only part of the supplementary material (Supplementary File 1).

Taggability

Theoretically, one SNP in strong LD to the variant of interest would be enough to serve as a marker that (partly) captures the effect of the variant for, e.g., GWAS or genomic selection as tag SNP. We, therefore, investigated the presence of potential tag SNPs close to the variants of interest. The used measure was the maximum observed r^2 between a variant of interest and a pool of potential tag SNPs within a certain distance (r_{tag}^2). Nearly all variants in all variant classes came with at least one variable SNP within proximity of 10 kb (**Figure S 14**). Mean r_{tag}^2 for all variants and populations showed an asymptotic trend with identifying the best tag SNP within 10 kb for most of the variants in all three populations (**Figure 4.5**). Only mean r_{tag}^2 of DUPs in BR was continuously growing until 100 kb distance (**Figure 4.5**). Mean r_{tag}^2 for SNPs only reached ~0.9 within 100 kb in all three populations, meaning that some SNPs were not in full phase to any other SNP. Mean r_{tag}^2 was slightly reduced for DELs and strongly for DUPs, INVs and BNDs compared to SNPs (**Figure 4.5**).

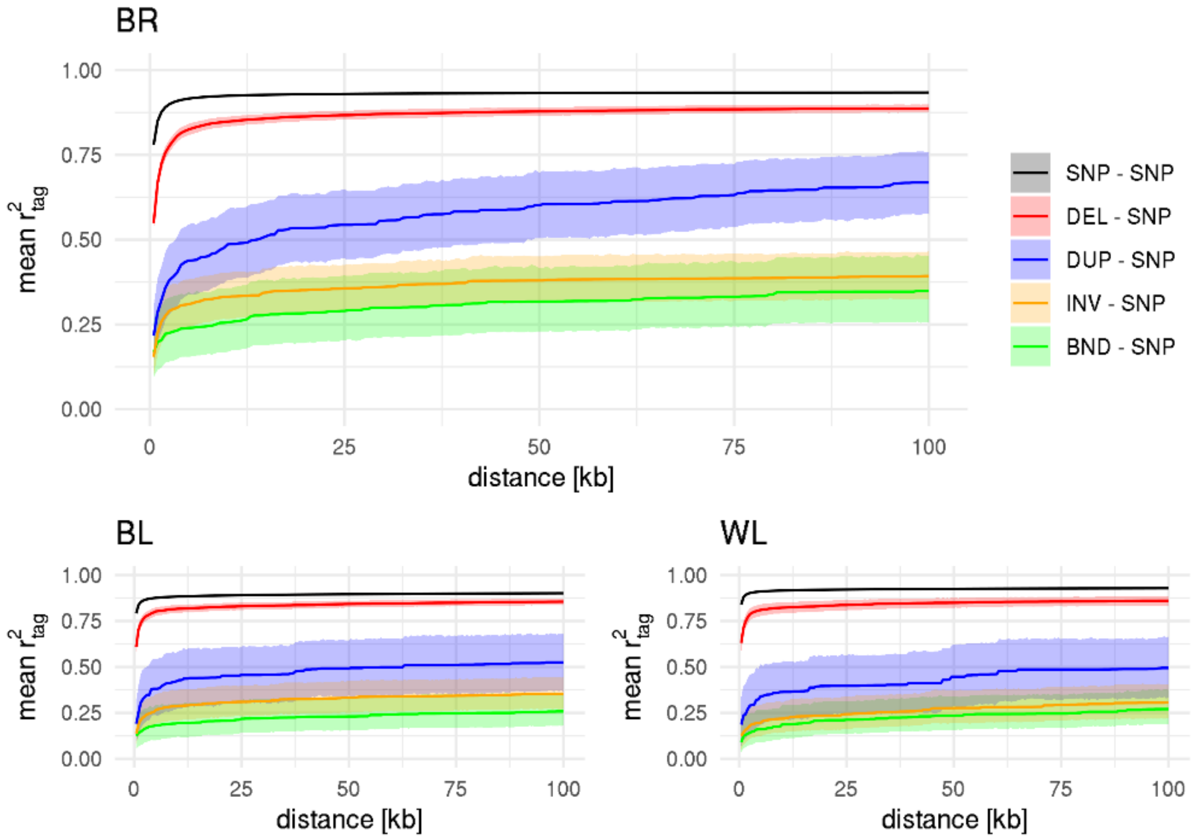

Figure 4.5: Mean taggability for broiler (BR), brown layer (BL), and white layer (WL) chickens. Taggability ($\boldsymbol{r_{tag}^2}$) was calculated as the maximum $\boldsymbol{r^2}$ value up to a certain distance from the variant of interest. Means across variants are presented as lines while the shaded area represents the Bonferroni-corrected 95 % bootstrap confidence intervals.

We additionally defined a variant as tagged if $r_{tag}^2 > 0.75$ and evaluated shares of accordingly tagged variants. While more than 85 % of the SNPs were tagged in BR within 10 kb, this number was slightly smaller for DELs (>75 %). More than 25 % of the DUPs were tagged within 10 kb distance and 50 % within 100 kb, while less than 15 % of INVs and BNDs were tagged. The tendency is the same in the two layer populations, but the absolute numbers slightly deviate. As a maximum value of a sample is not independent of the number of sampled values, we also checked the number of present potential tag SNPs within 5 kb distance to the variant of interest. Interestingly, SNPs were surrounded by significantly more close variable SNPs on average than SVs in all three populations (**Table 4.2**). This difference was still present when regarding only tag SNPs ($r^2 > 0.75$).

Table 4.2: Median number of variable SNPs within 5 kb distance to variants of interest

Variant	BR		BL		WL	
	All	$r^2 \geq 0.75$	all	$r^2 \geq 0.75$	all	$r^2 \geq 0.75$
SNP	140 ^{a}	7 ^{a}	85 ^{a}	5 ^{a}	73 ^{a}	9 ^{a}
DEL	70 ^{d}	5 ^{b}	41 ^{c}	4 ^{b}	38 ^{c}	6 ^{b}
DUP	78 ^{cd}	4 ^{b}	48 ^{bc}	3 ^{ab}	31 ^{c}	4 ^{ab}
INV	90 ^{c}	3 ^{b}	49 ^{bc}	5 ^{ab}	42 ^{bc}	11 ^{ab}
BND	119 ^{b}	6 ^{ab}	59 ^{b}	1 ^{ab}	61 ^{b}	1 ^{ab}

Different lowercase letters within columns account for significantly different medians at the significance level of 0.05 (Bonferroni-corrected pairwise Wilcoxon rank-sum test)

In practice, the interest of researchers and breeding companies may not be the taggability of SVs by WGS SNPs, but by array SNPs. Those come with a different allele frequency spectrum and lower resolution than WGS SNPs, which influences the LD patterns (Qanbari 2020). However, they are often available for a huge number of phenotyped individuals due to their use in routine breeding programs. We, therefore, evaluated the potential performance of four publically available chicken genotyping arrays with resolutions of 600 k (Kranis *et al.* 2013), 60 k (Groenen *et al.* 2011), 55 k (Liu *et al.* 2019), and 10 k (IMAGE 2020).

The availability of variable SNPs close to the variants of interest was strongly dependent on the resolution of the arrays. While the 600 k array had a variable array SNP within 15 kb for more than 90 % of the variants in all three populations, the 60 k and the 55 k array came with a slight shift of this dependency of having a variable array SNP for >80 % of the variants at 50 kb and >90 % at 100 kb (**Figure S 16**). The 10 k array, however, contained no variable array SNP for 50 % of the variants within 100 kb. A non-random difference in SNP density by variant type is not present for any array. The reduced density compared to WGS also reduced the taggability. Mean r_{tag}^2 values for SNPs and DELs reached between 0.06 for BR and the 10 k array and 0.65 for WL and the 600 k array within 100 kb distance (**Figure S 15**). Interestingly, DELs seem to be slightly stronger tagged than SNPs in BL and WL (**Figure S 15**), while the other variant types were tagged by maximally 50 % of the level which was reached in SNPs and DELs. The results are comparable when checking the proportion of variants with r_{tag}^2 > 0.75 (**Figure S 17**). 40 % of the WGS SNPs and even 45 % of DELs were tagged with more than r_{tag}^2 > 0.75 by a SNP of the 600 k array in WL. In contrast, less than 1 % of SNPs and DELs were tagged by a SNP of the 10k array in BR.

Discussion

Strong LD between genomic markers and causal genomic variants is the fundamental requirement of methods like genomic prediction (Meuwissen *et al.* 2001) and GWAS (Visscher *et al.* 2012). A stringent evaluation of LD between SNP marker panels and potentially causal SVs of different classes is therefore of strong interest for researchers and practical breeders, especially as the strength of this LD is discussed differently in literature (e.g. Hinds *et al.* 2006; McCarroll *et al.* 2006; Redon *et al.* 2006; Cooper *et al.* 2008; McCarroll *et al.* 2008; Kato *et al.* 2009; Conrad *et al.* 2010; Mills *et al.* 2011; Sudmant *et al.* 2015; Lee *et al.* 2020). We here present the first study that performed this evaluation in chickens.

Implications from the SV calling pipeline

The median sequencing coverage of the samples (5 – 17 X) was comparably low for SV discovery. Despite the fact that the sequencing depth differed between layers and broilers, results were similar for all three populations. An effect of the sequencing depth on the results is therefore unlikely, as the results could be repeated across sequencing depths.

The SV calling approach was intended to return highly accurate variant calls, therefore prioritizing precision over sensitivity. This especially required the exclusion of regions with unusually high coverage, as they may be artefacts of inaccurate read mapping in regions of low sequence complexity (Li 2014). As those regions are known to be hot spots for SV formation by non-allelic homologous recombination (NAHR) (Locke *et al.* 2006; Gu *et al.* 2008; Bickhart and Liu 2014; Sudmant *et al.* 2015), we expect to have missed a significant proportion of SVs, especially multi-copy DUP. Further, there was a missing overlap between DELLY and MANTA at INS calling, resulting in no INS calls. A generally weak power in INS calling from short reads is expected, though (Delage *et al.* 2020). Those two problems highlight the need for long-read sequencing data for future studies, which should allow for improved resolution of complex regions and comes with improved abilities for INS calling (Sedlazeck *et al.* 2018; Ho *et al.* 2019). The limitations of the calling approach and the resulting characteristics of the callset need to be considered when comparing our results to SV callsets that were derived by different approaches and therefore probably capturing SVs with different properties.

We further identified a lack of homozygous calls of DUPs, INVs, and BNDs with regard to HWE (**Figure 4.4**, **Figure S 11**, **Figure S 12**). One possible reason may be a deleterious load and therefore purifying selection on those variants. While literature highlights the deleterious potential of DELs, INVs, and BNDs (Feuk *et al.* 2006; Bouwman *et al.* 2020), DUPs are rather considered positive by increasing gene expression (Feuk *et al.* 2006; Lee *et al.* 2021). In our case, DELs rather show a slight excess of homozygotes than an expected lack under purifying selection (**Figure 4.4**, **Figure S 11**, **Figure S 12**). The lack of homozygous calls was instead present for DUPs, INVs, and BNDs. Additionally, VEP impact predictions classified 99.6 % of the DEL impacts as MODIFIER and only 0.4 % as HIGH, while DUP impacts were classified as HIGH in 10 % of the cases. The discrepancy with literature for DELs may partly be due to past inbreeding in the populations (Qanbari *et al.* 2019; Talebi *et al.* 2020), which resulted in small effective population sizes (Qanbari *et al.* 2010) and therefore may have purged strongly deleterious DELs (Bortoluzzi *et al.* 2020; Kyriazis *et al.* 2020). Purging of deleterious DELs may, together with limitations of the used SV callers, also be a reason for the relatively short sizes of the called DELs. Nevertheless, as none of the INVs and BNDs had predicted impacts besides MODIFIER, a second reason seems to be more likely: There may be deficits of the genotypers in accurately distinguishing between heterozygous and homozygous calls of DUPs, INVs, and BNDs.

LD decay results

The overall levels of SNP – SNP LD within the populations reflect the knowledge from the literature (Qanbari *et al.* 2010; Qanbari 2020) and the different levels of variability (BR > BL > WL) (Qanbari *et al.* 2019; Geibel *et al.* 2021b). This resulted in WL having the strongest overall level of LD and BR the weakest. Besides that and if not especially indicated differently, results were the same for all three populations throughout the following sections.

The DEL – SNP LD, all in all, was on the same level as SNP – SNP LD. This implies good predictability of DEL effects by SNP call sets and is in accordance with the majority of the existing studies (Hinds *et al.* 2006; McCarroll *et al.* 2006; Cooper *et al.* 2008; Conrad *et al.* 2010; Mills *et al.* 2011). Studies that found DEL – SNP LD to be on a reduced level compared to SNP – SNP LD mostly performed the DEL calling from SNP arrays, which implies low breakpoint resolution (Lee *et al.* 2020). It is also common to merge CNV to copy number variable regions (CNVR) in SNP array or read-depth-based studies (Lee *et al.* 2020). Therefore, a CNVR can reflect multiple mutation events and not only a single variant, resulting in reduced LD to bivariate SNPs, an effect we do not expect to be present in our data due to the more precise variant definition.

The level of DUP – SNP LD was strongly reduced compared to SNP – SNP LD and DEL – SNP LD, which is in accordance with the existing studies (Kato *et al.* 2009; Conrad *et al.* 2010; Sudmant *et al.* 2015; Lee *et al.* 2020). However, levels of ~40 % of the SNP – SNP LD (**Table 4.1**) were higher than what was found e.g. by Lee *et al.* (2020), who found DUP – SNP LD to be ~20 % of SNP – SNP LD in two dairy cattle populations. A main factor of DUP – SNP LD being reduced compared to SNP – SNP LD may be due to the lower allele frequencies of DUP in our callset (**Figure 4.2**) and therefore increased local ΔMAF (**Figure S 7**, **Figure S 8**) in BL and WL. Removing the ΔMAF dependent part of LD by expressing LD as r_S^2 increased the relative r^2 of 30 % to a relative r_S^2 of 68 % of the SNP – SNP r_S^2 (+28 %, **Table 4.1**). This means that local differences in the allele frequency spectra between SNPs and DUP account for ~50 % of the difference between SNP – SNP LD and DUP – SNP LD

A second cause for reduced DUP – SNP LD could be a higher rate of genotyping errors in DUP. In fact, we identified a significant reduction of homozygous DUP calls compared to HWE (**Figure 4.4**, **Figure S 11**, **Figure S 12**) as already discussed above. The potential genotyping inaccuracy may additionally be supported by the, admittedly subjective, observation of the two assessors during the visual filtering step that DUP came with less clear support than DEL. This, however, resulted only in a moderately reduced inter-observer reliability of 94 % in DUP compared to 97 % in DEL (**Supplementary file 3**).

A further possibility of reduced DUP – SNP LD may be the occurrence of multi-copy CNVs (mCNVs) (Locke *et al.* 2006; Sudmant *et al.* 2015) in our callset. DUP in the callset may partly represent CNVs that occur with different copy numbers and are therefore multi- instead of bivariate variants. This reduces the linkage to bivariate SNPs. We saw slight support for the occurrence of some mCNV in the callset e.g. by some high DHFFC values. However, mCNVs are known to cluster in special regions of the genome (Sudmant *et al.* 2015) due to non-allelic homologous recombination (NAHR) as a formation mechanism (Gu *et al.* 2008; Hastings *et al.* 2009). Note that NAHR can also occur recurrently (Gu *et al.* 2008), resulting in variants that are called bivariate but stem from multiple mutation events. As those clusters should result in high-coverage regions, which we removed in the filtering step, we do not expect a higher number of mCNV and recurrent mutations in our callset.

We also evaluated the linkage between SNPs and INV/ BND and found low levels of LD (26.8 % and 18.5 % of SNP – SNP LD). The reduced LD in our study is again partly due to local allele frequency differences (**Figure S 6** - **Figure S 8**) as for DUP. Relative r_s^2 values were therefore 14 % to 30 % higher than relative r^2 values (**Table 4.1**). However, r_s^2 values for INV – SNP and BND – SNP were still only ~50 % of SNP – SNP r_s^2. The remaining gap may partly be due to genotyping problems. We identified the lack of homozygous calls for INVs and BNDs (**Figure 4.4**, **Figure S 11**, **Figure S 12**) as for DUPs. In combination with the missing ability to use coverage information for filtering, we would trust the INV and BND genotypes least in our callset. In contrast to our results, Sudmant *et al.* (2015) found INV to be in good LD to SNPs in a very accurate callset from 2,504 human genomes, which further supports that the accuracy of INV calls was low in our study.

Taggability

The analysis of taggability revealed comparable patterns as the LD decay. A high fraction of SNPs and DEL was tagged by close-by WGS SNPs in all three populations (**Figure 4.5**; **Figure S 14**), while only a small fraction of DUPs, INVs, and BNDs was tagged. However, in contrast to the decay patterns, SNPs on average were tagged slightly stronger than DEL, and between 5 % and 10 % more SNPs were tagged with r_{tag}^2 > 0.75 than DEL. A reason for the higher taggability of SNPs compared to DEL, while the LD decay does not differ, may be the reduced SNP density around DELs (**Table 4.2**), as the chance for higher maximum values increases with the number of SNPs in the region of interest. In contrast, DELs were tagged slightly better by array SNPs than WGS SNPs by array SNPs. In the case of array SNPs, no locally increased density was present, as array design aims at an equidistant spacing of markers across the genome (Kranis *et al.* 2013). This resulted in no difference between the taggability of SNPs and DELs by array SNPs. Potential issues of excluding SNPs in complex regions during array design as suggested by Lee *et al.* (2020) as a reason for reduced CNV – SNP LD, were not observed in this study, as we excluded SVs in those regions due to a minor calling accuracy. Using array SNPs to tag the WGS variants further revealed a strong need for dense marker maps to provide good tag SNPs, as only the 600 k array could provide tag SNPs with r_{tag}^2 > 0.75 for more than 25 % of SNPs and DEL. This may largely explain why e.g. Xu *et al.* (2014) found a quarter of CNVs that were significantly associated with milk traits in Holstein cattle to be not tagged by SNPs of a 50 k array. It suggests that this is not solely due to the nature of CNV but that they also missed a comparable fraction of effects, which are caused by SNPs.

The concept of taggability is especially relevant for GWAS, where phenotype-marker associations are tested for each marker separately. The strength of the LD between marker and causal variant then directly influences the power of the GWAS. However, the absence of single tag SNPs does not imply that the effect of an SV cannot be captured by a longer haplotype. Methods that utilize effects of multiple SNP at once (e.g. ridge regression best linear unbiased prediction (Meuwissen *et al.* 2001)), of which each can explain a slightly different fraction of the variance of the causal variant, may be more robust in this sense. Additionally, imputation of known SVs would probably be a way to overcome the issue of low taggability and needs further investigation.

Conclusions

We evaluated LD patterns between a comprehensive SV callset and surrounding SNPs in three commercial chicken populations. We found DEL – SNP LD to be on the same level as SNP – SNP LD, while DUP – SNP, INV – SNP, and BND – SNP LD were strongly reduced. This was in accordance with the availability of tag SNPs for a high share of SNPs and DELs, while tag SNPs for DUPs were rare and mostly missing for INVs and BNDs. Different arrays came with a density-dependent ability to tag WGS SNPs and SVs but did not show strong systematic differences compared with taggability by WGS SNPs. The main reason for existing differences in SNP – SNP and DUP/INV/BND – SNP LD in our study was due to local MAF differences. Those accounted for ~50 % of this difference in the strength of LD. This implies that genomic variance due to DELs in the chicken populations studied can be captured by different SNP marker sets as good as variance from WGS SNPs, whereas separate SV calling might be advisable for DUP, INV, and BND effects.

Material and Methods

Data

The study used WGS data of 25 white layers, 25 brown layers, and 40 broiler chickens. The raw data was first published by Qanbari *et al.* (2019), which contains more information about the samples. Chickens were paired-end sequenced with a median coverage between 5 X and 17 X, read length of 100 bp (WL + BL) or 126 bp (BR), and insert sizes of ~400 bp. Basic quality statistics can be found in **Supplementary file 2** as MultiQC report (Ewels *et al.* 2016).

Population integrity was controlled using principal component analysis in plink 1.9 (Purcell *et al.* 2007). The SNPs were first LD pruned by setting the --indep-pairwise flag to sliding windows of 50 kb, a stepsize of five SNPs and an r^2 of 0.5. Based on the pruned SNPs, plink extracted then 90 prime components. Results for the first four prime components and the variance explained can be found in **Figure S 18**. The first two prime components, which in total accounted for 33.2 % of the total variance, clearly separated broilers, white- and brown layers. The two broiler subpopulations were only slightly separated by the second prime component and clearly by the third, which accounted for 4.5 % of the total variance. The fourth component started splitting one of the broiler populations. We assumed this to be sufficiently closely related to consider the two broiler subpopulations as a combined population for further analyses.

Variant Calling Pipeline

Alignment on the reference genome galGal6/ GRGC6a and SNP calling were conducted in a previous study (Geibel *et al.* 2021a) following GATK best practices pipeline (McKenna *et al.* 2010). The SNPs needed for this study were then extracted from the old callset using bcftools (Li 2011) and the duplicate-marked and base quality score recalibrated BAM files were used as starting point for the SV calling process.

SV calling was conducted following a consensus calling approach. SVs were first separately called per individual and then genotyped on population-level by running Delly 0.8.5 (Rausch *et al.* 2012), Manta 1.6.0 (Chen *et al.* 2016), and a combination of Lumpy 0.2.13 (Layer *et al.* 2014) and Svtyper 0.7.0 (Chiang *et al.* 2015) in parallel on the complete set. The genotyping results of the three calling pipelines were then merged using SURVIVOR 1.0.7 (Jeffares *et al.* 2017) and allowing for breakpoint differences of 1000 bp. This resulted in 95,478 raw SV calls.

Additionally, read depth profiles for all samples in 100 bp windows were generated using Mosdepth 0.2.9 (Pedersen and Quinlan 2018) and SVs were annotated with Dupholds (version 0.2.1) (Pedersen and Quinlan 2019) flanking fold change (DHFFC) and the SNP genotype calls located on the SV.

The merged callset was then filtered based on the following parameters:

1) Caller overlap: At least two of the three callers needed to support the variant.
2) Genotype concordance: The genotype that was supported by two out of the three callers was considered as the consensus genotype. Genotypes without the necessary support were set to missing for later re-imputation. If more than two samples did not have the necessary genotype concordance support for an SV, the complete SV was removed from the data set.
3) Removal of high coverage regions: Local coverage was extracted by Mosdepth 0.2.9 (Pedersen and Quinlan 2018) in 100 bp windows. If windows exceeded a threshold of twice the average coverage across all samples (expected value for a fixed DUP) plus two standard deviations, they were classified as unusually highly covered. Unusually highly covered regions were further merged if they were less than 1000 bp apart from each other. SVs with breakpoint confidence intervals falling in such a region were removed from the data set.
4) Difference to flanking coverage: DELs and DUPs calls were checked for non-consistent coverage changes relative to the flanking coverage by evaluating the Duphold Flanking Fold Change (DHFFC) (Pedersen and Quinlan 2019). DELs were considered as wrong genotypes when heterozygotes were not between 0.1 and 0.9 and homozygous DEL genotypes not smaller than 0.25. Heterozygous DUPs had to be >1.1 and homozygous DUPs >1.5. DELs/DUPs with more than one error or more than 10 % wrong genotypes were filtered. Otherwise, the putatively wrong DEL/DUP genotypes were set to missing for later re-imputation.
5) Support by SNP calls on DELs: SNP calls need to be homozygous on heterozygous DELs and missing on homozygous DELs. We, therefore, calculated for each DEL genotype the relative number of wrong SNP genotypes (e.g. one error by five total SNPs on the DEL = 0.1). If the sum of those error rates across samples exceeded two or 50 % of the number of samples that were at least heterozygous for the DEL, the DEL was filtered. Otherwise, the putatively wrong DEL genotypes were set to missing for later re-imputation.

This resulted in 5,600 SVs (4,831 DELs; 253 DUPs; 346 INVs; 170 BNDs; 94.1 % filtered). No INS remained, as Lumpy does not call INS and there was no overlap between Delly and Manta. Samplot 1.0.19 (Belyeu *et al.* 2021) was then used to generate quality control plots for each SV that passed the previous filtering step. The quality plots were visually screened by two separate observers comparable to the workflow implemented in SV-plaudit (Belyeu *et al.* 2018), but implemented locally by using image-sorter2 (https://github.com/Nestak2/image-sorter2). The SVs needed to be scored as ‘pass’ by each of the two observers to be further used (**Supplementary file 3**). By this, a further 6.9 % of the SVs (3.5 % of DEL, 11.1 % of DUP, 36.0 % of INV, and 30.8 % of BND) were removed. The removed SVs were mainly in regions with complex mapping patterns.

The final SV callset (4,301 DEL, 224 DUP, 218 INV, 117 BND) was then merged with the SNP callset (12,294,329 bivariate autosomal SNPs). The samples were phased and missing genotypes were imputed by beagle 5.0 (Browning *et al.* 2018) with default settings besides reducing 'ne' to 10,000 (Pook *et al.* 2019). Functional consequences were annotated by ensembl-vep (McLaren *et al.* 2016) using the release 100 GRGC6a annotation files.

Estimation of LD

LD between two loci with a maximum distance of 100 kb was initially estimated from phased haplotypes as follows:

$$r_{AB}^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_B (1 - p_A)(1 - p_B)} \tag{4.1}$$

Where p_A and p_B account for the alternative allele frequencies at the two loci and p_{AB} for the according haplotype frequency. To control for allele frequency deviations that influence the maximum possible r^2, we further scaled r^2 by the maximum possible r^2 given ΔMAF ($r_S^2 = r^2 / r_{\max|\Delta MAF}^2$) where $r_{\max|\Delta MAF}^2$ was derived as described by VanLiere and Rosenberg (2008). As we realized a problem with calling of homozygous DUP, we additionally estimated LD as squared Pearson Correlation between 0/1/2 coded SNP genotypes and the Duphold Flanking Fold Change (DHFFC) (Pedersen and Quinlan 2019) as a measure for the relative reference genome coverage at DEL and DUP (due to possible confounding only part of **Supplementary file 1**). LD decay was then summarized in means of 500 bp bins between the variants.

Bonferroni corrected bootstrap confidence intervals for the LD decay were estimated by resampling the r^2 values within each bin 100,000 times with replacement. As tests showed confidence intervals for SNP - SNP LD being < 0.001 due to the huge number of underlying values, we decided to skip estimation of confidence intervals for bins with > 1M r^2 values.

A tag SNP was defined as the SNP with the highest r^2 to the variant of interest within a certain distance (r_{tag}^2). The taggability of variant classes was then investigated by comparing means of r_{tag}^2 and shares of variants with $r_{tag}^2 > 0.75$. Additionally to the taggability by WGS SNPs, we compared the taggability by SNPs of four commercially available SNP arrays. The 600 k Affymetrix Axiom chicken genotyping array (Kranis *et al.* 2013), a 60 k Illumina Bead Chip (Groenen *et al.* 2011), a 55 k Affymetrix genotyping array (Liu *et al.* 2019), and the IMAGE_001 multispecies array, which contains 10 k chicken-specific SNPs on an Affymetrix genotyping array (IMAGE 2020). The annotation files were lifted over to the reference genome galGal6/GRGC6a by the UCSC (Kent *et al.* 2002) liftOver tool under usage of the according chain files and the overlaps with the variable WGS SNPs were defined as pools of potential Array tag SNPs.

Workflow

The complete pipeline was set up in snakemake 5.3.0 (Köster and Rahmann 2012) and the according scripts including the snakefile with all used parameters as well as the dependency analytics graph (DAG) and the rulegraph of the pipeline can be found on Zenodo (https://doi.org/10.5281/zenodo.5770348).

Declarations

Ethics approval and consent to participate

The study did not involve new treatment of animals as only published data was used. DNA samples for all already published raw data were taken from a database established during the project AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; https://aviandiv.fli.de/) and later extended by samples of the project SYNBREED (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de). Blood sampling was done in strict accordance to the German animal welfare regulations, with written consent of the animal owners and was approved by the at the according times ethics responsible persons of the Friedrich-Loeffler-Institut. According to German animal welfare regulations, notice was given to the responsible institution, the Lower Saxony State Office for Consumer Protection and Food Safety (33.9-42502-05-10A064).

Consent for publication

Not applicable

Availability of data and materials

The raw fastq files were already published by Qanbari et al. (Qanbari *et al.* 2019) and can be accessed via the ENA project PRJEB30270 (https://www.ebi.ac.uk/ena/browser/view/PRJEB30270). The snakemake workflow together with all scripts and supplementary files can be found on Zenodo (https://doi.org/10.5281/zenodo.5770348). Intermediate results are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge support by the Open Access Publication Funds of the University of Goettingen.

Authors' contributions

JG designed the study, analyzed the data, interpreted the results, and wrote the initial manuscript. NPP substantially contributed to the data analysis. SW contributed to the acquisition of the data and revision of the manuscript. HS and CR contributed to design of the study, interpretation of the results, and substantially revised the manuscript. All authors have read and approved the submitted manuscript.

Acknowledgments

Not applicable

Supplementary

The supplementary material can be accessed via the preprint (https://doi.org/10.21203/rs.3.rs-861830/v1).

Supplementary file 1: Supplementary results, tables and figures.

Supplementary file 2: MultiQC report.

Supplementary file 3: Observer concordance of the visual filtering step.

References

Abyzov, A; Urban, AE; Snyder, M; Gerstein, M (2011): CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. In *Genome Res* 21 (6), pp. 974–984. DOI: 10.1101/gr.114876.110.

Auton, A; Abecasis, GR; Altshuler, DM; Durbin, RM; Bentley, DR; Chakravarti, A et al. (2015): A global reference for human genetic variation. In *Nature* 526 (7571), pp. 68–74. DOI: 10.1038/nature15393.

Belyeu, JR; Chowdhury, M; Brown, J; Pedersen, BS; Cormier, MJ; Quinlan, AR; Layer, RM (2021): Samplot: a platform for structural variant visual validation and automated filtering. In *Genome Biol* 22 (1), p. 161. DOI: 10.1186/s13059-021-02380-5.

Belyeu, JR; Nicholas, TJ; Pedersen, BS; Sasani, TA; Havrilla, JM; Kravitz, SN et al. (2018): SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. In *GigaScience* 7 (7). DOI: 10.1093/gigascience/giy064.

Berger, S; Schlather, M; los Campos, G de; Weigend, S; Preisinger, R; Erbe, M; Simianer, H (2015): A Scale-Corrected Comparison of Linkage Disequilibrium Levels between Genic and Non-Genic Regions. In *PLOS ONE* 10 (10), e0141216. DOI: 10.1371/journal.pone.0141216.

Bickhart, DM; Liu, GE (2014): The challenges and importance of structural variation detection in livestock. In *Frontiers in genetics* 5, p. 37. DOI: 10.3389/fgene.2014.00037.

Bortoluzzi, C; Bosse, M; Derks, MFL; Crooijmans, RPMA; Groenen, MAM; Megens, H-J (2020): The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. In *Evolutionary applications* 13 (2), pp. 330–341. DOI: 10.1111/eva.12872.

Bouwman, AC; Derks, MF; Broekhuijse, ML; Harlizius, B; Veerkamp, RF (2020): Using short read sequencing to characterise balanced reciprocal translocations in pigs. DOI: 10.21203/rs.3.rs-28830/v3.

Browning, BL; Zhou, Y; Browning, SR (2018): A One-Penny Imputed Genome from Next-Generation Reference Panels. In *Am J Hum Genet* 103 (3), pp. 338–348. DOI: 10.1016/j.ajhg.2018.07.015.

Chen, X; Schulz-Trieglaff, O; Shaw, R; Barnes, B; Schlesinger, F; Källberg, M et al. (2016): Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. In *Bioinformatics (Oxford, England)* 32 (8), pp. 1220–1222. DOI: 10.1093/bioinformatics/btv710.

Chiang, C; Layer, RM; Faust, GG; Lindberg, MR; Rose, DB; Garrison, EP et al. (2015): SpeedSeq: ultra-fast personal genome analysis and interpretation. In *Nature Methods* 12 (10), pp. 966–968. DOI: 10.1038/nmeth.3505.

Conrad, DF; Andrews, TD; Carter, NP; Hurles, ME; Pritchard, JK (2006): A high-resolution survey of deletion polymorphism in the human genome. In *Nature Genetics* 38 (1), pp. 75–81. DOI: 10.1038/ng1697.

Conrad, DF; Pinto, D; Redon, R; Feuk, L; Gokcumen, O; Zhang, Y et al. (2010): Origins and functional impact of copy number variation in the human genome. In *Nature* 464 (7289), pp. 704–712. DOI: 10.1038/nature08516.

Cooper, GM; Zerr, T; Kidd, JM; Eichler, EE; Nickerson, DA (2008): Systematic assessment of copy number variant detection via genome-wide SNP genotyping. In *Nature Genetics* 40 (10), pp. 1199–1203. DOI: 10.1038/ng.236.

Crooijmans, RP; Fife, MS; Fitzgerald, TW; Strickland, S; Cheng, HH; Kaiser, P et al. (2013): Large scale variation in DNA copy number in chicken breeds. In *BMC genomics* 14 (1), p. 398. DOI: 10.1186/1471-2164-14-398.

Delage, WJ; Thevenon, J; Lemaitre, C (2020): Towards a better understanding of the low recall of insertion variants with short-read based variant callers. In *BMC genomics* 21 (1), p. 762. DOI: 10.1186/s12864-020-07125-5.

Escaramís, G; Docampo, E; Rabionet, R (2015): A decade of structural variants: description, history and methods to detect structural variation. In *Briefings in functional genomics* 14 (5), pp. 305–314. DOI: 10.1093/bfgp/elv014.

Ewels, P; Magnusson, M; Lundin, S; Käller, M (2016): MultiQC: summarize analysis results for multiple tools and samples in a single report. In *Bioinformatics (Oxford, England)* 32 (19), pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354.

Fan, W-L; Ng, CS; Chen, C-F; Lu, M-YJ; Chen, Y-H; Liu, C-J et al. (2013): Genome-wide patterns of genetic variation in two domestic chickens. In *Genome biology and evolution* 5 (7), pp. 1376–1392. DOI: 10.1093/gbe/evt097.

Feuk, L; Carson, AR; Scherer, SW (2006): Structural variation in the human genome. In *Nature reviews. Genetics* 7 (2), pp. 85–97. DOI: 10.1038/nrg1767.

Geibel, J; Reimer, C; Pook, T; Weigend, S; Weigend, A; Simianer, H (2021a): How imputation can mitigate SNP ascertainment Bias. In *BMC Genomics* 22 (1). DOI: 10.1186/s12864-021-07663-6.

Geibel, J; Reimer, C; Weigend, S; Weigend, A; Pook, T; Simianer, H (2021b): How array design creates SNP ascertainment bias. In *PLoS One* 16 (3), e0245178. DOI: 10.1371/journal.pone.0245178.

Gorla, E; Cozzi, MC; Román-Ponce, SI; Ruiz López, FJ; Vega-Murillo, VE; Cerolini, S et al. (2017): Genomic variability in Mexican chicken population using copy number variants. In *BMC Genet* 18 (1), p. 61. DOI: 10.1186/s12863-017-0524-4.

Graffelman, J (2015): Exploring Diallelic Genetic Markers: The HardyWeinberg Package. In *Journal of Statistical Software* 64 (3), pp. 1–23. Available online at https://www.jstatsoft.org/v64/i03/.

Groenen, MAM; Megens, H-J; Zare, Y; Warren, WC; Hillier, LW; Crooijmans, RPMA et al. (2011): The development and characterization of a 60K SNP chip for chicken. In *BMC genomics* 12 (1), p. 274. DOI: 10.1186/1471-2164-12-274.

Gu, W; Zhang, F; Lupski, JR (2008): Mechanisms for human genomic rearrangements. In *PathoGenetics* 1 (1), p. 4. DOI: 10.1186/1755-8417-1-4.

Han, R; Yang, P; Tian, Y; Wang, D; Zhang, Z; Wang, L et al. (2014): Identification and functional characterization of copy number variations in diverse chicken breeds. In *BMC genomics* 15, p. 934. DOI: 10.1186/1471-2164-15-934.

Hastings, PJ; Lupski, JR; Rosenberg, SM; Ira, G (2009): Mechanisms of change in gene copy number. In *Nature Reviews Genetics* 10 (8), pp. 551–564. DOI: 10.1038/nrg2593.

Hayes, BJ; Daetwyler, HD (2019): 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. In *Annual Review of Animal Biosciences* 7 (1), pp. 89–102. DOI: 10.1146/annurev-animal-020518-115024.

Hinds, DA; Kloek, AP; Jen, M; Chen, X; Frazer, KA (2006): Common deletions and SNPs are in linkage disequilibrium in the human genome. In *Nature Genetics* 38 (1), pp. 82–85. DOI: 10.1038/ng1695.

Ho, SS; Urban, AE; Mills, RE (2019): Structural variation in the sequencing era. In *Nature Reviews Genetics*. DOI: 10.1038/s41576-019-0180-9.

Imsland, F; Feng, C; Boije, H; Bed'hom, B; Fillon, V; Dorshorst, B et al. (2012): The Rose-comb mutation in chickens constitutes a structural rearrangement causing both altered comb morphology and defective sperm motility. In *PLoS genetics* 8 (6), e1002775. DOI: 10.1371/journal.pgen.1002775.

Innovative Management of Animal Genetic Resources (IMAGE) (2020): DELIVERABLE D4.5. A standard multi-species chip for genomic assessment of collections. Available online at https://www.imageh2020.eu/deliverable/D4.5_resubmitted_final.pdf, updated on 3/1/2020, checked on 8/17/2021.

Jeffares, DC; Jolly, C; Hoti, M; Speed, D; Shaw, L; Rallis, C et al. (2017): Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. In *Nature Communications* 8, p. 14061. DOI: 10.1038/ncomms14061.

Jia, X; Chen, S; Zhou, H; Li, D; Liu, W; Yang, N (2013): Copy number variations identified in the chicken using a 60K SNP BeadChip. In *Animal genetics* 44 (3), pp. 276–284. DOI: 10.1111/age.12009.

Kato, M; Kawaguchi, T; Ishikawa, S; Umeda, T; Nakamichi, R; Shapero, MH et al. (2009): Population-genetic nature of copy number variations in the human genome. In *Hum Mol Genet* 19 (5), pp. 761–773. DOI: 10.1093/hmg/ddp541.

Kent, WJ; Sugnet, CW; Furey, TS; Roskin, KM; Pringle, TH; Zahler, AM; Haussler, D (2002): The human genome browser at UCSC. In *Genome Res* 12 (6), pp. 996–1006. DOI: 10.1101/gr.229102.

Kerstens, HHD; Crooijmans, RP; Dibbits, BW; Vereijken, A; Okimoto, R; Groenen, M am (2011): Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries. In *BMC genomics* 12 (1), p. 94. DOI: 10.1186/1471-2164-12-94.

Köster, J; Rahmann, S (2012): Snakemake--a scalable bioinformatics workflow engine. In *Bioinformatics (Oxford, England)* 28 (19), pp. 2520–2522. DOI: 10.1093/bioinformatics/bts480.

Kranis, A; Gheyas, AA; Boschiero, C; Turner, F; Le Yu; Smith, S et al. (2013): Development of a high density 600K SNP genotyping array for chicken. In *BMC Genomics* 14 (1), p. 59. DOI: 10.1186/1471-2164-14-59.

Kyriazis, CC; Wayne, RK; Lohmueller, KE (2020): Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. In *Evolution Letters* n/a (n/a). DOI: 10.1002/evl3.209.

Layer, RM; Chiang, C; Quinlan, AR; Hall, IM (2014): LUMPY: a probabilistic framework for structural variant discovery. In *Genome Biology* 15 (6), R84. DOI: 10.1186/gb-2014-15-6-r84.

Lee, Y-L; Bosse, M; Mullaart, E; Groenen, MAM; Veerkamp, RF; Bouwman, AC (2020): Functional and population genetic features of copy number variations in two dairy cattle populations. In *BMC genomics* 21 (1), p. 89. DOI: 10.1186/s12864-020-6496-1.

Lee, Y-L; Takeda, H; Costa Monteiro Moreira, G; Karim, L; Mullaart, E; Coppieters, W et al. (2021): A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. In *PLoS Genet* 17 (7), e1009331. DOI: 10.1371/journal.pgen.1009331.

Li, H (2011): A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. In *Bioinformatics (Oxford, England)* 27 (21), pp. 2987–2993. DOI: 10.1093/bioinformatics/btr509.

Li, H (2014): Toward better understanding of artifacts in variant calling from high-coverage samples. In *Bioinformatics (Oxford, England)* 30 (20), pp. 2843–2851. DOI: 10.1093/bioinformatics/btu356.

Lin, S; Lin, X; Zhang, Z; Jiang, M; Rao, Y; Nie, Q; Zhang, X (2018): Copy Number Variation in SOX6 Contributes to Chicken Muscle Development. In *Genes* 9 (1). DOI: 10.3390/genes9010042.

Liu, R; Xing, S; Wang, J; Zheng, M; Cui, H; Crooijmans, RPMA et al. (2019): A new chicken 55K SNP genotyping array. In *BMC genomics* 20 (1), p. 410. DOI: 10.1186/s12864-019-5736-8.

Locke, DP; Sharp, AJ; McCarroll, SA; McGrath, SD; Newman, TL; Cheng, Z et al. (2006): Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome. In *The American Journal of Human Genetics* 79 (2), pp. 275–290. DOI: 10.1086/505653.

Malomane, DK; Simianer, H; Weigend, A; Reimer, C; Schmitt, AO; Weigend, S (2019): The SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high genomic resolution. In *BMC genomics* 20 (1), p. 345. DOI: 10.1186/s12864-019-5727-9.

McCarroll, SA; Hadnott, TN; Perry, GH; Sabeti, PC; Zody, MC; Barrett, JC et al. (2006): Common deletion polymorphisms in the human genome. In *Nature Genetics* 38 (1), pp. 86–92. DOI: 10.1038/ng1696.

McCarroll, SA; Kuruvilla, FG; Korn, JM; Cawley, S; Nemesh, J; Wysoker, A et al. (2008): Integrated detection and population-genetic analysis of SNPs and copy number variation. In *Nature Genetics* 40 (10), pp. 1166–1174. DOI: 10.1038/ng.238.

McKenna, A; Hanna, M; Banks, E; Sivachenko, A; Cibulskis, K; Kernytsky, A et al. (2010): The Genome Analysis Toolkit. A MapReduce framework for analyzing next-generation DNA sequencing data. In *Genome Res* 20 (9), pp. 1297–1303. DOI: 10.1101/gr.107524.110.

McLaren, W; Gil, L; Hunt, SE; Riat, HS; Ritchie, GRS; Thormann, A et al. (2016): The Ensembl Variant Effect Predictor. In *Genome Biol* 17 (1), p. 122. DOI: 10.1186/s13059-016-0974-4.

Meuwissen, TH; Hayes, BJ; Goddard, ME (2001): Prediction of total genetic value using genome-wide dense marker maps. In *Genetics* 157 (4), pp. 1819–1829.

Mills, RE; Walter, K; Stewart, C; Handsaker, RE; Chen, K; Alkan, C et al. (2011): Mapping copy number variation by population-scale genome sequencing. In *Nature* 470 (7332), pp. 59–65. DOI: 10.1038/nature09708.

Pedersen, BS; Quinlan, AR (2018): Mosdepth: quick coverage calculation for genomes and exomes. In *Bioinformatics (Oxford, England)* 34 (5), pp. 867–868. DOI: 10.1093/bioinformatics/btx699.

Pedersen, BS; Quinlan, AR (2019): Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. In *GigaScience* 8 (4), giz040. DOI: 10.1093/gigascience/giz040.

Pook, T; Mayer, M; Geibel, J; Weigend, S; Cavero, D; Schoen, CC; Simianer, H (2019): Improving Imputation Quality in BEAGLE for Crop and Livestock Data. In *G3*, g3.400798.2019. DOI: 10.1534/g3.119.400798.

Purcell, S; Neale, B; Todd-Brown, K; Thomas, L; Ferreira, MAR; Bender, D et al. (2007): PLINK: a tool set for whole-genome association and population-based linkage analyses. In *Am J Hum Genet* 81 (3), pp. 559–575. DOI: 10.1086/519795.

Qanbari, S (2020): On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. In *Frontiers in genetics* 10, p. 1304. DOI: 10.3389/fgene.2019.01304.

Qanbari, S; Hansen, M; Weigend, S; Preisinger, R; Simianer, H (2010): Linkage disequilibrium reveals different demographic history in egg laying chickens. In *BMC Genet* 11, p. 103. DOI: 10.1186/1471-2156-11-103.

Qanbari, S; Rubin, C-J; Maqbool, K; Weigend, S; Weigend, A; Geibel, J et al. (2019): Genetics of adaptation in modern chicken. In *PLoS Genet* 15 (4), e1007989. DOI: 10.1371/journal.pgen.1007989.

Rao, YS; Li, J; Zhang, R; Lin, XR; Xu, JG; Xie, L et al. (2016): Copy number variation identification and analysis of the chicken genome using a 60K SNP BeadChip. In *ps* 95 (8), pp. 1750–1756. DOI: 10.3382/ps/pew136.

Rausch, T; Zichner, T; Schlattl, A; Stütz, AM; Benes, V; Korbel, JO (2012): DELLY: structural variant discovery by integrated paired-end and split-read analysis. In *Bioinformatics (Oxford, England)* 28 (18), i333-i339. DOI: 10.1093/bioinformatics/bts378.

Redon, R; Ishikawa, S; Fitch, KR; Feuk, L; Perry, GH; Andrews, TD et al. (2006): Global variation in copy number in the human genome. In *Nature* 444 (7118), pp. 444–454. DOI: 10.1038/nature05329.

Rubin, C-J; Megens, H-J; Martinez Barrio, A; Maqbool, K; Sayyab, S; Schwochow, D et al. (2012): Strong signatures of selection in the domestic pig genome. In *Proceedings of the National Academy of Sciences of the United States of America* 109 (48), pp. 19529–19536. DOI: 10.1073/pnas.1217149109.

Sedlazeck, FJ; Lee, H; Darby, CA; Schatz, MC (2018): Piercing the dark matter. Bioinformatics of long-range sequencing and mapping. In *Nature Reviews Genetics* 19 (6), pp. 329–346. DOI: 10.1038/s41576-018-0003-4.

Seol, D; Ko, BJ; Kim, B; Chai, H-H; Lim, D; Kim, H (2019): Identification of Copy Number Variation in Domestic Chicken Using Whole-Genome Sequencing Reveals Evidence of Selection in the Genome. In *Animals : an open access journal from MDPI* 9 (10). DOI: 10.3390/ani9100809.

Sohrabi, SS; Mohammadabadi, M; Wu, D-D; Esmailizadeh, A (2018): Detection of breed-specific copy number variations in domestic chicken genome. In *Genome* 61 (1), pp. 7–14. DOI: 10.1139/gen-2017-0016.

Strillacci, MG; Cozzi, MC; Gorla, E; Mosca, F; Schiavini, F; Román-Ponce, SI et al. (2017): Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. In *Animal : an international journal of animal bioscience* 11 (5), pp. 737–745. DOI: 10.1017/S1751731116002135.

Sudmant, PH; Rausch, T; Gardner, EJ; Handsaker, RE; Abyzov, A; Huddleston, J et al. (2015): An integrated map of structural variation in 2,504 human genomes. In *Nature* 526 (7571), pp. 75–81. DOI: 10.1038/nature15394.

Talebi, R; Szmatoła, T; Mészáros, G; Qanbari, S (2020): Runs of Homozygosity in Modern Chicken Revealed by Sequence Data. In *G3 (Bethesda, Md.)* 10 (12), pp. 4615–4623. DOI: 10.1534/g3.120.401860.

Tian, M; Wang, Y; Gu, X; Feng, C; Fang, S; Hu, X; Li, N (2013): Copy number variants in locally raised Chinese chicken genomes determined using array comparative genomic hybridization. In *BMC genomics* 14, p. 262. DOI: 10.1186/1471-2164-14-262.

VanLiere, JM; Rosenberg, NA (2008): Mathematical properties of the r2 measure of linkage disequilibrium. In *Theoretical population biology* 74 (1), pp. 130–137.

Visscher, PM; Brown, MA; McCarthy, MI; Yang, J (2012): Five years of GWAS discovery. In *Am J Hum Genet* 90 (1), pp. 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

Wang, K; Li, M; Hadley, D; Liu, R; Glessner, J; Grant, SFA et al. (2007): PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. In *Genome Res* 17 (11), pp. 1665–1674. DOI: 10.1101/gr.6861907.

Wang, L; Xu, L; Liu, X; Zhang, T; Li, N; Hay, EH et al. (2015): Copy number variation-based genome wide association study reveals additional variants contributing to meat quality in Swine. In *Scientific Reports* 5 (1), p. 12535. DOI: 10.1038/srep12535.

Wang, X; Nahashon, S; Feaster, TK; Bohannon-Stewart, A; Adefope, N (2010): An initial map of chromosomal segmental copy number variations in the chicken. In *BMC genomics* 11, p. 351. DOI: 10.1186/1471-2164-11-351.

Wang, Y; Gu, X; Feng, C; Song, C; Hu, X; Li, N (2012): A genome-wide survey of copy number variation regions in various chicken breeds by array comparative genomic hybridization method. In *Animal genetics* 43 (3), pp. 282–289. DOI: 10.1111/j.1365-2052.2011.02308.x.

Weng, Z; Xu, Y; Li, W; Chen, J; Zhong, M; Zhong, F et al. (2020): Genomic variations and signatures of selection in Wuhua yellow chicken. In *PLoS One* 15 (10), e0241137. DOI: 10.1371/journal.pone.0241137.

Xu, L; Cole, JB; Bickhart, DM; Hou, Y; Song, J; VanRaden, PM et al. (2014): Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. In *BMC genomics* 15 (1), p. 683. DOI: 10.1186/1471-2164-15-683.

Yan, Y; Yang, N; Cheng, HH; Song, J; Qu, L (2015): Genome-wide identification of copy number variations between two chicken lines that differ in genetic resistance to Marek's disease. In *BMC genomics* 16, p. 843. DOI: 10.1186/s12864-015-2080-5.

Zhang, H; Du, Z-Q; Dong, J-Q; Wang, H-X; Shi, H-Y; Wang, N et al. (2014): Detection of genome-wide copy number variations in two chicken lines divergently selected for abdominal fat content. In *BMC genomics* 15, p. 517. DOI: 10.1186/1471-2164-15-517.

Chapter 5

General Discussion

Prevalence of genomic data in livestock sciences

Due to the fast implementation of genomic breeding programs in livestock breeding, especially in dairy cattle breeding, cheap commercial SNP arrays have been on the market relatively early for the major livestock species (e.g. Matukumalli *et al.* 2009; Ramos *et al.* 2009; Groenen *et al.* 2011; McCue *et al.* 2012; Tosser-Klopp *et al.* 2014). This and the fund of large datasets of genotypes and phenotypes from routine breeding programs led to a huge interest from researchers in livestock sciences in utilizing them, which led to by now 1,409 publications listed in Web of Science^{TM}, with the main interest in cattle (**Figure 5.1**). The strongly decreasing sequencing costs (NHGRI 2020) and development of standard pipelines for variant calling (van der Auwera *et al.* 2013) led to an even stronger increasing trend for whole-genome-sequencing (WGS) based studies since 2013, with 2020 being the first year with more WGS- than array-based studies (**Figure 5.1**) in livestock sciences. The high number of publications implies that SNP-arrays and WGS are strongly used in livestock research. The number may even be underestimating the true number, as search terms probably did not capture all writing options or miss publications that are directly denoted to a breed name without using the species name.

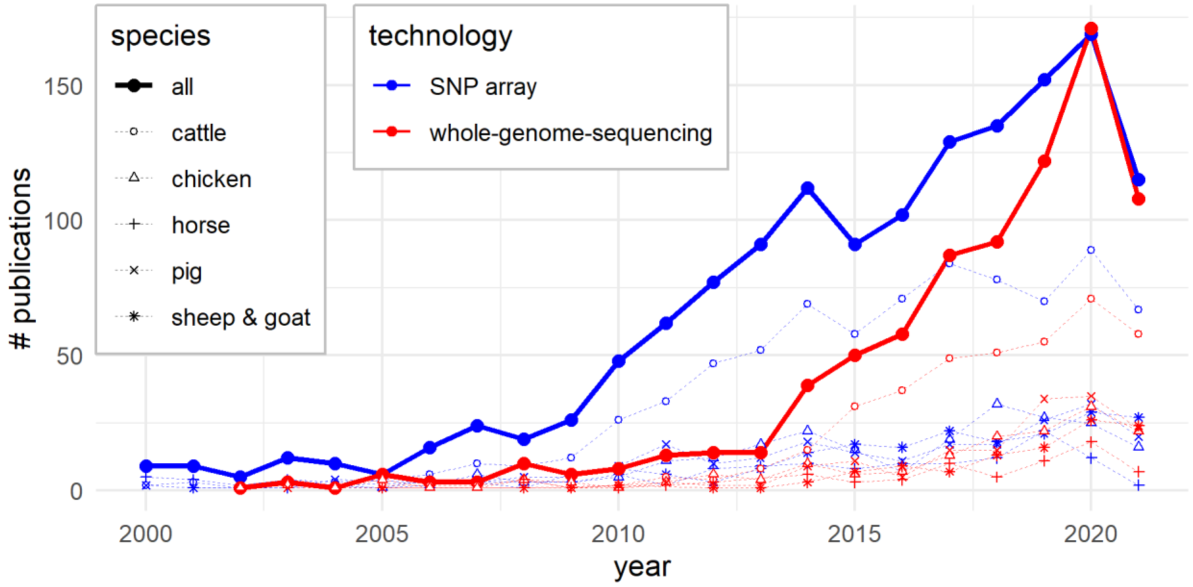

Figure 5.1: Numbers of publications from livestock sciences that utilized genomic technologies by year. The numbers were derived by a Web of Science^{TM} search (https://www.webofscience.com; 03.09.2021) for the search terms "SNP array"/ "SNP chip" (1,469) and "whole-genome-sequencing" (809), respectively. Results were restricted to the major livestock species and the categories "Genetics Heredity", "Veterinary Sciences", and "Agriculture Dairy Animal Science" to exclude publications targeting microbiology and comparable topics. Results before the year 2000 were excluded from the graph.

The probably largest gains in knowledge about human genomics of the recent years were derived through the thousand genomes project (Huang *et al.* 2012). The largest data sets in livestock sciences are proprietary, as owned by breeding companies. Nevertheless, a comparable attempt to generate a huge collection of sequenced animals in livestock sciences is the 1,000 Bull Genomes project (Hayes and Daetwyler 2019). This already enabled some huge meta-GWAS studies (e.g. Bouwman *et al.* 2018), that allow combining results from proprietary phenotypes and imputed genotypes while restricting the access of phenotypic information to confidants of the companies. Projects at that scale are currently missing for other livestock species. However, in chickens, the amount of publically available genomic data is steadily increasing. Malomane *et al.* (2019), e.g., published a dataset of 3,235 genotyped chickens from 162 populations, and Wang *et al.* (2021) utilized 868 chicken sequences that are available via different projects. This scattered availability of partly public chicken data should be a welcoming situation for a consortium as in humans or cattle to enable according research in chickens.

Strength and impact of ascertainment bias in livestock genomics

As shown in the previous subchapter, a large number of genomic studies in livestock are based on SNP arrays. Assessment on how this may have impacted findings is therefore important.

The focus of ascertainment schemes on commercially important populations in livestock sciences (Matukumalli *et al.* 2009; Kranis *et al.* 2013) should have led to a high prevalence of biased results. Quantitative assessment of ascertainment bias, however, was rarely done by now. This may be due to the need for broad sets of sequenced samples for a direct comparison between array and WGS data. By now, only the overestimation of array-based LD in cattle (Qanbari *et al.* 2014) and chicken (Qanbari 2020) was shown, and Malomane *et al.* (2018) explored the effect on expected heterozygosity (H_{exp}), F_{ST} and prime component analysis (PCA) in chickens. Further, some conclusions can be drawn from **Chapter 2** and **Chapter 3**, which should widely overlap with the results of Malomane *et al.* (2018), as the used samples show a strong overlap.

In **Chapter 2**, H_{exp} calculated from array data was on average overestimated by 84 % compared to H_{exp} from WGS data in chickens. H_{exp} thereby always depends on the total number of variants. This means, that more and diverse populations in the callset lead to an increased number of invariable SNPs within populations and a reduced H_{exp}. H_{exp} of a population should therefore rather be regarded relative to other populations than absolute. However, the overestimation was ~20 % stronger for populations that are genetically comparable to the discovery populations of the original array than for other populations, highlighting the uneven ascertainment scheme. Comparable was observed by Herrero-Medrano *et al.* (2014), who compared heterozygosity estimates from array data and WGS in mainly European and some Asian pigs. They found a strong correlation between array- and WGS-heterozygosity for most European breeds, but not for Asian breeds and breeds with likely Asian introgression. This suggests that the Porcine SNP60 BeadChip (Ramos *et al.* 2009) misses Asian SNP variation.

The intensity of ascertainment bias should be different across livestock species. It should generally be lower in chickens than in cattle. Commercial chicken breeds are scattered across the global diversity spectrum (Malomane *et al.* 2019), and ascertainment was done in broilers, brown layers, and white layers together with SNP validation in a diversity set (Kranis *et al.* 2013). However, note that the validation step did not show a relevant impact on the ascertainment bias in **Chapter 2**. Cattle, in contrast, suffers from the strong division into the two clades of taurine and indicine cattle lineages, accompanied by initial ascertainment only in taurine cattle (Matukumalli *et al.* 2009), which was only later extended by indicine SNPs (Utsunomiya *et al.* 2019). As no studies directly assess the prevalence of ascertainment bias in cattle, this needs to be done by evaluating auxiliary information. Imputation accuracy strongly depends on the availability of variable markers in the study set. Due to ascertainment bias, this is commonly higher for populations closely related to the discovery populations. Thus, in **Chapter 3**, we found that imputation accuracy is higher for discovery populations than for non-discovery populations in our chicken data set. By implication, this can be used to roughly compare marker panels of comparable density for their bias. As earlier cattle SNP arrays only used taurine discovery sets (Matukumalli *et al.* 2009), while later arrays also included indicine samples (Utsunomiya *et al.* 2019), those second-generation arrays commonly performed better for imputation of indicine or crossbred samples (e.g. Boison *et al.* 2015; Toro Ospina *et al.* 2021). This suggests a strong prevalence of ascertainment bias for indicine cattle at least in the arrays without indicine-specific SNPs. Further, McTavish and Hillis (2015) used a comparison between simulations of different cattle demographic scenarios and ascertainment schemes with empirical BovineSNP50 bead chip (Matukumalli *et al.* 2009) data. They found a scenario, which strongly upward biases heterozygosity in taurine cattle while underestimating heterozygosity in indicine cattle, to most likely reflect the empirical situation.

As shown by Dokan *et al.* (2021), the effect of ascertainment bias on F_{ST} in direction and intensity highly depends on underlying demographic scenarios and sampling schemes. In the simulations by McTavish and Hillis (2015), scenarios with ascertainment in taurine cattle reduced F_{ST} between indicine and taurine cattle up to -30 % depending on the intensity of isolation between taurine and indicine cattle. In contrast, Malomane *et al.* (2018) found overwhelmingly upward biased F_{ST} values in chickens. Note, however, that the empirical chicken data in **Chapter 4** did not show an ascertainment bias for F_{ST} in chickens larger than the pooling bias. The discrepancy to Malomane *et al.* (2018) may be due to different ways of averaging F_{ST} across loci. While Malomane *et al.* (2018) calculated F_{ST} as mean across loci, we divided the sum of the numerator by the sum of the denominator as initially suggested by Wright (1949) and later Weir and Cockerham (1984). This seems to be more robust against ascertainment bias (unpublished observations).

Possibilities to deal with ascertainment bias

Given the prevalence of ascertainment bias, the question arises of how to deal with it. The first intention would probably be to switch to WGS, as e.g. proposed by Qanbari and Simianer (2014), and which is already done for an increasing number of projects (**Figure 5.1**). This, however, may only be feasible for larger groups or consortia, as costs, as well as computational needs for sequencing, are still magnitudes higher than for genotyping. Attempts to reduce costs for sequencing have been e.g. pooled sequencing (Futschik and Schlötterer 2010) or reduced representation library methods (Davey *et al.* 2011). Pooled sequencing was used to utilize a complete flowcell for sequencing multiple samples from a population and by this decrease the sequencing costs. Pooled sequencing, however, comes with a series of biases (supplementary material of **Chapter 3**; Futschik and Schlötterer 2010; Boitard *et al.* 2012; Chen *et al.* 2012) and problems due to a dramatic increase in the computational need for accurate variant calling (supplementary methods of **Chapter 2**). Further, pooled sequencing only rudimentary allows analyses that go beyond the estimation of allele frequencies (e.g. short-distance LD analyses by physical linkage; Feder *et al.* 2012). Nowadays, the use of pooled sequencing should be pointless in most cases, as barcoding techniques allow the use of flow cells for multiple samples in parallel. This allows methods that sequence a share of samples with very low depth and then impute variants by utilizing populations-wide haplotype information (Pook *et al.* 2021). This, however, is still limited to homozygous populations that do not exist in livestock (Pook *et al.* 2021). Another way of reducing the sequencing need is to only sequence a subset of the genome, e.g. by sheering the genome through restriction enzymes and then sequencing only the beginning and end of the sheered fragments (genotyping by sequencing, GBS; Elshire *et al.* 2011). SNP discovery is then influenced by the prevalence of restriction sites and the choice of the restriction enzyme (Davey *et al.* 2011). It further comes with highly skewed genome coverage, causing further problems in SNP calling (Beissinger *et al.* 2013).

Despite the opportunities of low-coverage sequencing, the current development in routine breeding programs, which require phenotyping and genotyping on a large scale, is to rather extend the sample size while decreasing marker density by the use of low-density SNP arrays (e.g. Rensing *et al.* 2017). For the within-breed genomic selection, a certain ascertainment bias is actually intended. The main purpose here is to find a balance between genotyping costs, defined by the number of needed markers on an array and the number of genotyped and phenotyped individuals, and the prediction accuracy to maximize genetic gain per time and costs. With a limited SNP panel, high minor allele frequency (MAF) SNPs, potentially even biased due to their LD to QTLs, generally have a higher effect on prediction accuracy than low-MAF SNPs in the first instance (Perez-Enciso *et al.* 2015). Further, variability of SNPs is in this situation only relevant for the breed of interest. The use of specialized arrays will in those situations probably stay the best option for larger breeding programs, as long as low-coverage sequencing methods do not compete in terms of accuracy. However, note that MAF and LD of the SNPs on those arrays may change over time and reevaluation of the arrays may become necessary. Further, those specialized arrays cannot be used in other breeds and the breed of interest needs to be large enough that the design of a specialized array is economically efficient.

The situation is different in livestock population genomics. Here, the accurate and unbiased representation of all populations on arrays is of major interest. In a long term, the broad use of WGS data will probably become the standard for those cases, especially as the design process of an array will not be able to consider all future use cases of the array. Nevertheless, by now this is often not feasible on a logistic and financial basis. As we showed in **Chapter 2** that a broad and large discovery panel is the key factor in limiting ascertainment bias, the design or selection of an appropriate array needs to take this primarily into account. With this in mind, it is critical that detailed information about the discovery panel is publicly available for each array.

Further, as arrays will never be free from some amount of ascertainment bias, the robustness of methods needs to be evaluated alongside the effects they show. **Chapter 3** e.g. revealed that FST was more robust against ascertainment bias in our setting than Nei's distance, and would therefore probably be the better choice to express population differentiation in this setting. This information is, however, not available for many methods and may also differ between different scenarios. As WGS data for a direct evaluation of estimator robustness may not be available oftentimes, evaluations, as proposed in Chapter 3 that add additional known bias onto array data, may therefore give first impressions on the behavior of the estimator under ascertainment bias. Further, modern simulation software such as MoBPS (Pook *et al.* 2020) allows the evaluation based on different simulated scenarios (e.g. done for F_{ST} by Dokan *et al.* 2021).

It may be further advisable to correct the SNP data for ascertainment bias. Methods as proposed by Nielsen *et al.* (2004), however, require detailed knowledge on the ascertainment process and should therefore be unfeasible in most cases. Malomane *et al.* (2018) investigated the effect of filtering strategies and identified e.g. LD pruning to reduce ascertainment bias, while strict MAF filters increased it. However, the filtering of SNP data should always be critically questioned. The effects may be different in certain situations. So might LD pruning, which certainly has a positive effect e.g. on F_{ST}, hinder the accurate identification of rare haplotypes for haplotype-based methods (comparable to what was shown for identification of runs of homozygosity by Meyermans *et al.* 2020).

In **Chapter 3**, we further proposed to use imputation to WGS as *in silico* correction of SNP data. Our results were promising and certainly suggest this for future use. However, access to a large reference panel, which is evenly spaced across the intended range of populations, is thereby crucial. An unbalanced reference panel can have the same effect as an unbalanced discovery panel and introduces its own bias. Further, we performed imputation in this case from 600 k to WGS. Imputation from arrays of lower density may result in reduced accuracy and worse results.

Inclusion of structural variants into genomic studies

As already indicated in **Chapter 1** and **Chapter 4**, SVs are causal for some prominent qualitative and quantitative traits of livestock. By now, detailed knowledge on the amount of genomic variance explained by SVs in livestock is missing. This has its major reason in low precision and recall rates of array- and short-read-based SV calling algorithms. The problem can be highlighted by the results of **Chapter 4**, where we called 95,478 raw SVs that were reduced to 4,860 (~5 %) by the, admittedly strict, filtering procedure. This is common across studies and to a high share influenced by mapping problems in repetitive regions. So attributed Bertolotti *et al.* (2020) an overall false discovery rate (FDR) of 91 % in SV calling for Atlantic salmon to an FDR of 99.2 % in regions of complex mapping patterns and 85 % in the rest of the genome. Algorithms are thereby better suited to identify DELs than other variants, explaining why 88 % of our called variants were DEL. When evaluating the share of chicken genomes affected by the called variants (**Figure 4.1 B**), the included SNP still affected most of the genome (~0.37 % – 0.41 %). INV showed a large range between 0.1 % and 0.7 %, most likely affected by single long INVs, as the overall number was low with a skewed length distribution (**Figure 4.1 A**) and small MAFs (**Figure 4.2**). DELs (~ 0.08 %) and DUPs (~0.01 %) covered less of the genome. This is certainly an underestimation of the total length covered, as lots of SV are expected to be associated with regions of low sequence complexity, which we had to exclude due to short-read mapping problems. Those results argue for both, the high prevalence and thereby possibly associated effects of SV on phenotypes and the need to use long-read sequencing data to get comprehensive insights into the SV landscape of chickens.

In **Chapter 4**, we also evaluated whether SNPs could serve as markers for SV effects. As SNP-DEL LD was as strong as SNP-SNP LD in nearly all regarded senses, at least DEL effects should already be captured in SNP-based studies. The overall LD, however, depended on the density of the SNP panel. Low-density arrays therefore of course capture less of the total DEL variance than WGS sets. LD between SNPs and the other SV was strongly reduced, advocating for separate SV calling if those effects are also of interest. Interestingly, the main reason for this reduction was due to larger MAF differences. They may indicate different mutation or selection patterns and require further research.

A practical consideration for the inclusion of SVs in studies is that SV classification is harder than SNP classification, as they commonly span an interval. This may cause problems for the use of SV in tools that rely on exact marker positions, but do not consider endpoints of the SV. The calling tools additionally often show a low breakpoint accuracy. Merging of SV calls for later genotyping across samples within caller as well as between callers, as needed for consensus calling approaches, therefore involves a relaxed definition of the breakpoint confidence interval (usually ~1000 bp used). This potentially leads to merging of different (overlapping) variants. The problem is especially prevalent in read-depth-based studies, which commonly define CNV regions rather than exact CNV (e.g. Lee *et al.* 2020). The prevalence of this merging of different SVs is likely one explanation of reduced LD between SNPs and SV, as already discussed in **Chapter 4**, but may also mask SV-phenotype associations in GWAS which needs to be considered when including SVs into studies.

Combinability of genomic data across studies

A general problem in genomic studies that stood out during this project is the sometimes rare combinability of existing datasets. Some research questions require large sample sizes and costs for the generation of such large genomic data sets commonly go beyond the budget of a project. This can either be solved by forming large consortia as for the 1000 genomes project (Auton *et al.* 2015) or the 1000 bull genomes project (Hayes and Daetwyler 2019). Another possibility is to combine publically available data sets, whose availability is steadily increasing through the open data politics of the larger scientific journals (e.g. **Chapter 3**; Wang *et al.* 2021). In both cases, but also within projects, data sets are oftentimes based on different technologies and therefore come with different marker maps. This requires imputation to bring them onto the same scale for analysis. Accurate imputation results, however, require a good overlap of the marker maps.

Considering the requirements of imputation during design processes of arrays as by Boichard *et al.* (2012), however, is not necessarily part of the array design process. When comparing the four available chicken arrays (**Figure 5.2**), the very small overlap between the two first genotyping platforms by Groenen *et al.* (2011) and Kranis *et al.* (2013) of only 2,180 SNPs (3.9 % of the 60 k SNPs) stands out. As Liu *et al.* (2019) explicitly included SNPs of the previously existing arrays to ensure overlap, it overlaps with the 600 k array by 24,164 SNPs (46.5 %) and with the 60 k array by 6,310 SNPs (12.1 %). The IMAGE multispecies array (IMAGE 2020) was designed to allow for cheap genomic characterization of European gene bank samples for later use in research projects. It overlaps with the (no longer commercially available) 50 k array by 6,751 SNPs (73.2 %) and with the 600 k array by only 1,043 SNPs (11.3 %). While the inclusion of SNPs of both previously existing arrays by Liu *et al.* (2019) allows a combination with other data sets by imputation, the minimal overlap between the 10 k / 60 k array and the 600 k array will most likely require the additional use of a sequence-based reference panel to impute them onto the same scale.

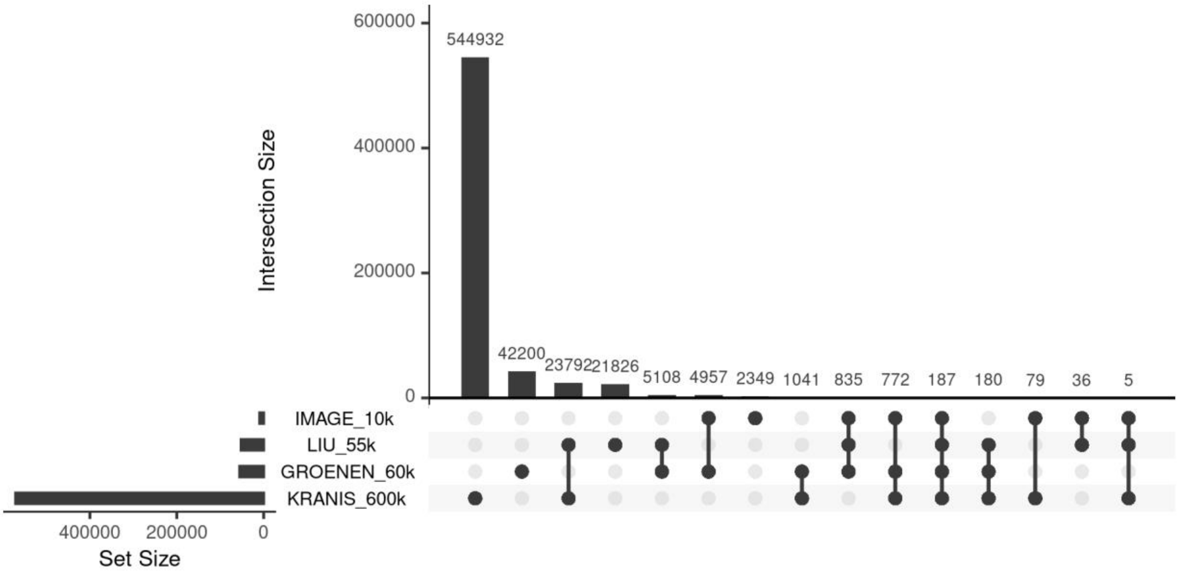

Figure 5.2: Overlap between the four chicken SNP arrays. KRANIS_600k = 600 k Affymetrix array (Kranis *et al.* 2013); GROENEN_60k = 60 k Illumina Bead Chip (Groenen *et al.* 2011); LIU_55k = 55 k Affymetrix genotyping array (Liu *et al.* 2019); IMAGE_10k = 10 k Affymetrix genotyping array (IMAGE_001 multispecies array; IMAGE 2020).

Besides the necessary overlap between the marker maps, the reference panel plays an important role in imputation. As shown in **Chapter 3** and already noted before, representative reference sets are crucial for the results of imputation. Currently, no global reference panel as the 1000 genomes project (Auton *et al.* 2015) or 1000 bull genomes project (Hayes and Daetwyler 2019) is available for chickens. A joint effort in combining the currently strongly scattered and only partly public datasets to such a panel is, therefore, necessary to further merge array-based study results on the scale.

A, rather practical, problem when combining study results is the switch between versions of reference genomes and marker coding. The first one can be done easily by liftOver tools (e.g. from UCSC; Kent *et al.* 2002), if the according chain files exist. The switch between A/B coding of the arrays and reference-based coding, as needed to combine datasets from different platforms, requires the according annotation files of the providers. If they are available, translation is also relatively easy. However, care should be taken if plink (Purcell *et al.* 2007) up to version 1.9 was used for data curation. Plink by default silently introduces a major allele coding and deletes all information of the original A/B coding (Chang *et al.* 2015), making the translation to reference-based coding nearly impossible. As many researchers are not aware of this problem, a certain share of public genotyping data should be affected by this problem. We experienced this as well in **Chapter 3**. A back-translation was possible in this case, as a certain share of chickens in the analysis was sequenced as well as genotyped, which allowed identifying the wrongly coded SNPs by low recall rates of homozygous SNPs (supplementary methods of **Chapter 3**). However, this is unlikely to be the case for the overwhelming majority of studies and underscores the need to share true raw data rather than curated datasets to avoid the accumulation of such unidentifiable error sources.

The genomic technology of the future

Current developments show two contrasting future trends. The first one is related to large intensively bred populations. In those populations, the trend is to increase the number of genotyped individuals and by this the number of phenotypes in the training population, which is needed for accurate breeding value estimation. This is in cattle e.g. done by the use of (low-density) arrays in combination with genotyping or even sequencing key ancestors on a higher density to enable imputation (Rensing *et al.* 2017). Another way is to utilize low coverage sequencing to identify haplotype information for imputation to achieve the same goal (Pook *et al.* 2021).

The other trend, mainly focused by research, is to increase knowledge about, by now hardly accessible, genomic regions through the utilization of long-read sequencing technologies as nanopore and PacBio sequencing. This is necessary, as short reads cannot resolve repetitive and complex genomic regions. This leads to strongly increased coverage of repetitive regions with low mapping quality (Li 2014), as prevalent in **Chapter 4**. Further, assemblies are commonly bad in those regions. A prominent example is the highly variable chr16 of the chicken genome that contains the major histocompatibility complex (MHC). It is currently the worst assembled chicken autosome with a 149 kb unplaced scaffold and a 502 kb gap at the beginning of the chromosome. The bad assembly quality and a high amount of repetitive regions lead to unreliable mapping results for short reads (**Figure 5.3**). This is similar for chr25, which is known for a high amount of repetitive elements (Masabanda *et al.* 2004). Long read sequencing technologies are hereby assumed to better resolve those regions and thereby allow the calling of SNPs as well as SVs in those regions (Sedlazeck *et al.* 2018). However, as the comparison between Illumina paired-end and PacBio HiFi reads in **Figure 5.3** shows, mapping is only partially improved in those really complex regions.

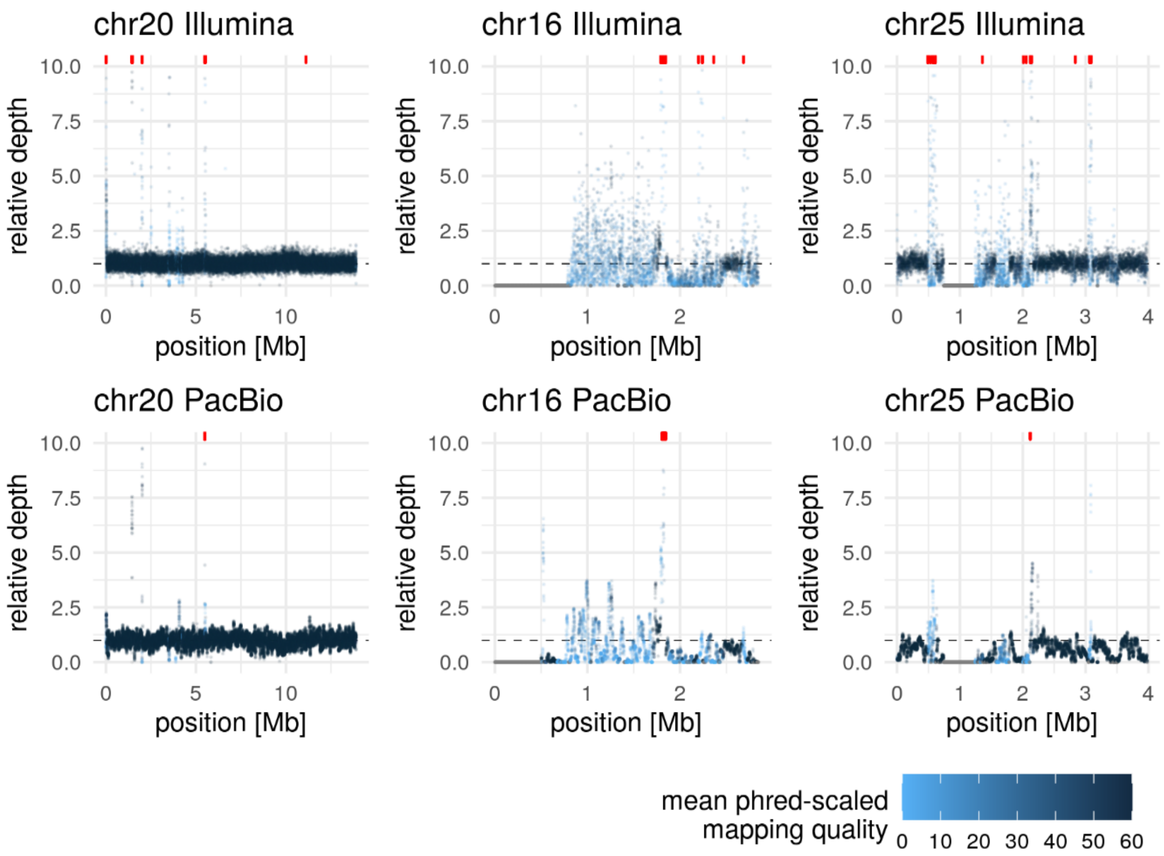

Figure 5.3: Comparison of performance of Illumina and PacBio sequencing reads in resolving complex genomic regions. The graph exemplarily compares the well-assembled chr20 to two chromosomes that are known for complex rearrangements (chr16 and chr25) for one white layer chicken (unpublished data). 151 bp paired-end Illumina reads (NovaSeq 600, PCR-free library) were mapped to the reference GRCG6a using bwa-mem (Li 2013) while >10 kb PacBio HiFi reads (Sequel II) were mapped using minimap2 (Li 2018). Depth of coverage was calculated in 500 bp bins by mosdepth (Pedersen and Quinlan 2018) and set in relation to the median depth on chr20. The mean phred-scaled mapping quality for the bins was further calculated by bedtools (Quinlan and Hall 2010). Bins that exceed the y-axis-cutoff of 10 are indicated by red ticks and the median of chr20 by a dashed horizontal line.

One reason for the only partly improved mappability may be the still incomplete and erroneous reference genome GRGC6a. Recently, two additional reference genomes were published by the Vertebrate Genome Project (Rhie *et al.* 2021). The reference bGalGal1.mat.broiler.GRCg7b is based on a female chicken from a maternal commercial broiler line (NCBI 2021a) and bGalGal1.pat.whiteleghornlayer.GRCg7w on a female chicken from a paternal white leghorn layer line (NCBI 2021b). **Figure 5.4** compares the chromosome lengths of the three reference genomes. While some chromosomes are slightly shorter on the new assemblies than for GRCg6a, the sex chromosomes were enlarged for GRCg7b, while they are not available yet for GRCg7w. The new references further provide assemblies for all in GRCg6a missing autosomes. Interestingly, they also report a 39th autosome pair, while publications by now only reported 38 pairs (e.g. International Chicken Genome Sequencing

Consortium 2004; Schmid *et al.* 2015). As, by now, no official description of the new reference genomes besides the database entries exists and the assemblies are in the status of "high-quality draft assembly" (Vertebrate Genomes Project 2021), this development needs further observation.

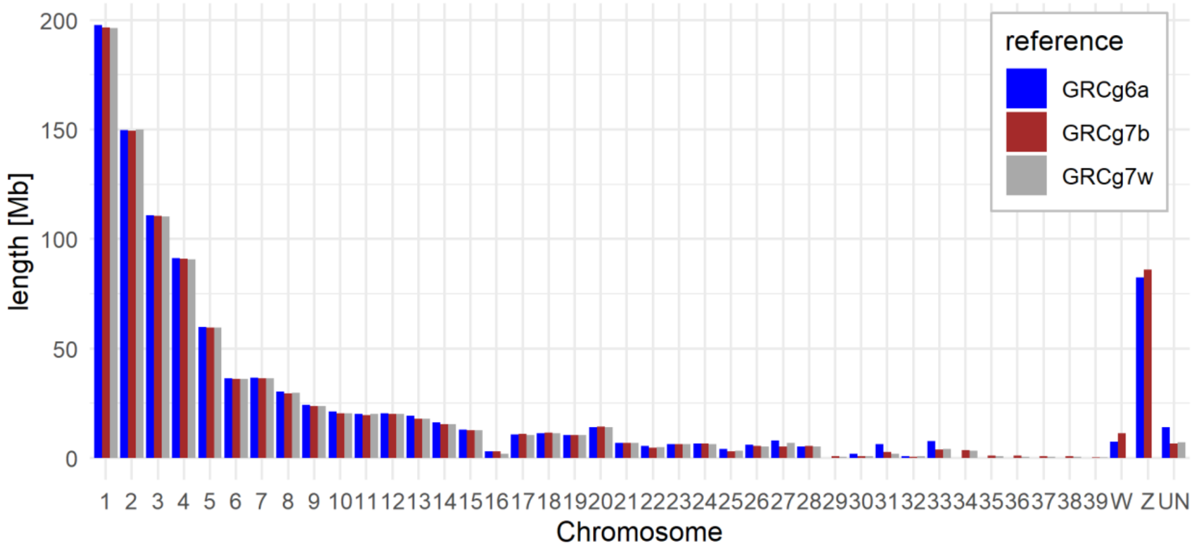

Figure 5.4: The assembly length of the two latest chicken reference genome builds. GRCg6a is based on a Red Jungle Fowl inbred chicken, GRCg7b on a chicken from a commercial maternal broiler line, and GRC7w on a chicken from a paternal white leghorn layer line. Chromosome UN: unplaced scaffolds (data sources: NCBI 2018, NCBI 2021a, and NCBI 2021b).

A further reason for bad mapping may be the presence of SV. Recent projects, therefore, aim in allowing variability in the reference genome through a graph-based representation, e.g. in cattle by Crysnanto and Pausch (2020). Those graph genomes seem to improve mapping quality (Crysnanto and Pausch 2020) and additionally show impressive results for SV discovery (Crysnanto *et al.* 2021). Future studies on the usage of graph-genomes in chickens are therefore indispensable. A first step in this direction was taken by a very recent study by Wang *et al.* (2021). They reported the first chicken pan-genome based on GRCg6a and iterative local reassembly of 664 chicken short-read sequences, which has the same intention as a graph genome but slightly differs in the methodology. They announced the discovery of ~66.5 Mb of additional sequences and 4,063 new genes. However, as newly predicted non-reference genes showed a strongly reduced transcript abundance in comparison with the reference genes (19.4 % vs. 90.6 %) for various transcriptomic datasets, the study seems to lack from a bad short-read assembly quality. The usage of long-read sequencing data seems therefore to be advised for the creation of a high-quality graph genome in chickens.

General conclusion

This study investigated the properties of different marker technologies, namely SNP arrays and WGS in chicken genomics. The main part thereby handled SNP ascertainment bias. We could confirm that SNP ascertainment bias is present in chicken array data. Further, the selection of the discovery panel and possible intentional overrepresentation of SNPs with higher MAF had the main effect on the creation of ascertainment bias, even during a complex array design process. We further showed that imputation to WGS may be a possibility to *in silico* mitigate SNP ascertainment bias. For this, an evenly distributed reference panel is crucial.

To test whether SNPs are also able to capture the effects of SVs, we investigated LD patterns between different SNP marker panels and SVs. This showed that DEL-effects can be captured as well as other SNP-effects by different SNP marker panels, while for non-DEL SV a separate SV calling is necessary. Some difficulties in the SV calling process further pointed out some weaknesses of short-read based SNP calling, which requests long-read sequencing for more accurate results in the future.

The work indicated the broad usability of SNP markers from SNP arrays and WGS in chicken genomics but also highlights the need to carefully consider the shortcomings of the underlying technologies in the design and discussion of studies.

Acknowledgments, funding, and ethics statement

Blood sampling of the white layer chicken was done at the Friedrich-Loeffler-Institut, Mariensee, in strict accordance to the German animal welfare regulations, with written consent of the animal owners and was approved by the at the according times ethics responsible persons of the Friedrich-Loeffler-Institut. According to German animal welfare regulations, notice was given to the responsible institution, the Lower Saxony State Office for Consumer Protection and Food Safety (33.9-42502-05-10A064).

The sequencing of the chicken was funded within the scope of the project IMAGE (www.imageh2020.eu; European Union's Horizon 2020 research and innovation programme under grant agreement No 677353). The Illumina sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, Sweden. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Further, the PacBio sequencing was performed at NGI/ Uppsala Genome Center that has been funded by RFI/ VR and Science for Life Laboratory, Sweden. It was further supported by the National Genomics Infrastructure(NGI)/ Uppsala Genome

Center and UPPMAX by providing assistance in massive parallel sequencing and computational infrastructure.

References

Auton, A; Abecasis, GR; Altshuler, DM; Durbin, RM; Bentley, DR; Chakravarti, A et al. (2015): A global reference for human genetic variation. In *Nature* 526 (7571), pp. 68–74. DOI: 10.1038/nature15393.

Beissinger, TM; Hirsch, CN; Sekhon, RS; Foerster, JM; Johnson, JM; Muttoni, G et al. (2013): Marker density and read depth for genotyping populations using genotyping-by-sequencing. In *Genetics* 193 (4), pp. 1073–1081. DOI: 10.1534/genetics.112.147710.

Bertolotti, AC; Layer, RM; Gundappa, MK; Gallagher, MD; Pehlivanoglu, E; Nome, T et al. (2020): The structural variation landscape in 492 Atlantic salmon genomes. In *Nature Communications* 11 (1), p. 5176. DOI: 10.1038/s41467-020-18972-x.

Boichard, DA; Chung, H; Dassonneville, R; David, X; Eggen, A; Fritz, S et al. (2012): Design of a bovine low-density SNP array optimized for imputation. In *PLoS One* 7 (3), e34130.

Boison, SA; Santos, DJA; Utsunomiya, AHT; Carvalheiro, R; Neves, HHR; O'Brien, AMP et al. (2015): Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (Bos indicus) dairy cattle: Comparison of commercially available SNP chips. In *J. Dairy Sci.* 98 (7), pp. 4969–4989. DOI: 10.3168/jds.2014-9213.

Boitard, S; Schlötterer, C; Nolte, V; Pandey, RV; Futschik, A (2012): Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. In *Mol Biol Evol* 29 (9), pp. 2177–2186. DOI: 10.1093/molbev/mss090.

Bouwman, AC; Daetwyler, HD; Chamberlain, AJ; Ponce, CH; Sargolzaei, M; Schenkel, FS et al. (2018): Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. In *Nat Genet* 50 (3), pp. 362–367. DOI: 10.1038/s41588-018-0056-5.

Chang, CC; Chow, CC; Tellier, LC; Vattikuti, S; Purcell, SM; Lee, JJ (2015): Second-generation PLINK. Rising to the challenge of larger and richer datasets. In *GigaScience* 4, p. 7. DOI: 10.1186/s13742-015-0047-8.

Chen, X; Listman, JB; Slack, FJ; Gelernter, J; Zhao, H (2012): Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples. In *Genet Epidemiol* 36 (6), pp. 549–560. DOI: 10.1002/gepi.21648.

Crysnanto, D; Leonard, AS; Fang, Z-H; Pausch, H (2021): Novel functional sequences uncovered through a bovine multi-assembly graph. In *bioRxiv*, 2021.01.08.425845. DOI: 10.1101/2021.01.08.425845.

Crysnanto, D; Pausch, H (2020): Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. In *Genome Biol* 21 (1), p. 184. DOI: 10.1186/s13059-020-02105-0.

Davey, JW; Hohenlohe, PA; Etter, PD; Boone, JQ; Catchen, JM; Blaxter, ML (2011): Genome-wide genetic marker discovery and genotyping using next-generation sequencing. In *Nat Rev Genet* 12 (7), pp. 499–510. DOI: 10.1038/nrg3012.

Dokan, K; Kawamura, S; Teshima, KM (2021): Effects of single nucleotide polymorphism ascertainment on population structure inferences. In *G3 (Bethesda, Md.)*. DOI: 10.1093/g3journal/jkab128.

Elshire, RJ; Glaubitz, JC; Sun, Q; Poland, JA; Kawamoto, K; Buckler, ES; Mitchell, SE (2011): A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. In *PLoS One* 6 (5), e19379.

Feder, AF; Petrov, DA; Bergland, AO (2012): LDx: Estimation of Linkage Disequilibrium from High-Throughput Pooled Resequencing Data. In *PLoS One* 7 (11), pp. 1–7. DOI: 10.1371/journal.pone.0048588.

Futschik, A; Schlötterer, C (2010): The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. In *Genetics* 186 (1), pp. 207–218. DOI: 10.1534/genetics.110.114397.

Groenen, MAM; Megens, H-J; Zare, Y; Warren, WC; Hillier, LW; Crooijmans, RPMA et al. (2011): The development and characterization of a 60K SNP chip for chicken. In *BMC genomics* 12 (1), p. 274. DOI: 10.1186/1471-2164-12-274.

Hayes, BJ; Daetwyler, HD (2019): 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. In *Annual Review of Animal Biosciences* 7 (1), pp. 89–102. DOI: 10.1146/annurev-animal-020518-115024.

Herrero-Medrano, JM; Megens, H-J; Groenen, MAM; Bosse, M; Pérez-Enciso, M; Crooijmans, RPMA (2014): Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. In *BMC Genomics* 15 (1), p. 601. DOI: 10.1186/1471-2164-15-601.

Huang, J; Ellinghaus, D; Franke, A; Howie, B; Li, Y (2012): 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. In *European Journal Of Human Genetics* 20 (7), pp. 801–805. DOI: 10.1038/ejhg.2012.3.

Innovative Management of Animal Genetic Resources (IMAGE) (2020): DELIVERABLE D4.5. A standard multi-species chip for genomic assessment of collections. Available online at https://www.imageh2020.eu/deliverable/D4.5_resubmitted_final.pdf, updated on 3/1/2020, checked on 8/17/2021.

International Chicken Genome Sequencing Consortium (2004): Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. In *Nature* 432 (7018), pp. 695–716.

Kent, WJ; Sugnet, CW; Furey, TS; Roskin, KM; Pringle, TH; Zahler, AM; Haussler, D (2002): The human genome browser at UCSC. In *Genome Res* 12 (6), pp. 996–1006. DOI: 10.1101/gr.229102.

Kranis, A; Gheyas, AA; Boschiero, C; Turner, F; Le Yu; Smith, S et al. (2013): Development of a high density 600K SNP genotyping array for chicken. In *BMC Genomics* 14 (1), p. 59. DOI: 10.1186/1471-2164-14-59.

Lee, Y-L; Bosse, M; Mullaart, E; Groenen, MAM; Veerkamp, RF; Bouwman, AC (2020): Functional and population genetic features of copy number variations in two dairy cattle populations. In *BMC genomics* 21 (1), p. 89. DOI: 10.1186/s12864-020-6496-1.

Li, H (2013): Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available online at http://arxiv.org/pdf/1303.3997v2, updated on 3/16/2013.

Li, H (2014): Toward better understanding of artifacts in variant calling from high-coverage samples. In *Bioinformatics (Oxford, England)* 30 (20), pp. 2843–2851. DOI: 10.1093/bioinformatics/btu356.

Li, H (2018): Minimap2. Pairwise alignment for nucleotide sequences. In *Bioinformatics (Oxford, England)* 34 (18), pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

Liu, R; Xing, S; Wang, J; Zheng, M; Cui, H; Crooijmans, RPMA et al. (2019): A new chicken 55K SNP genotyping array. In *BMC genomics* 20 (1), p. 410. DOI: 10.1186/s12864-019-5736-8.

Malomane, DK; Reimer, C; Weigend, S; Weigend, A; Sharifi, AR; Simianer, H (2018): Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. In *BMC Genomics* 19 (1), p. 22. DOI: 10.1186/s12864-017-4416-9.

Malomane, DK; Simianer, H; Weigend, A; Reimer, C; Schmitt, AO; Weigend, S (2019): The SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high genomic resolution. In *BMC genomics* 20 (1), p. 345. DOI: 10.1186/s12864-019-5727-9.

Masabanda, JS; Burt, DW; O'Brien, PCM; Vignal, A; Fillon, V; Walsh, PS et al. (2004): Molecular cytogenetic definition of the chicken genome: the first complete avian karyotype. In *Genetics* 166 (3), pp. 1367–1373. DOI: 10.1534/genetics.166.3.1367.

Matukumalli, LK; Lawley, CT; Schnabel, RD; Taylor, JF; Allan, MF; Heaton, MP et al. (2009): Development and characterization of a high density SNP genotyping assay for cattle. In *PLoS One* 4 (4), e5350.

McCue, ME; Bannasch, DL; Petersen, JL; Gurr, J; Bailey, E; Binns, MM et al. (2012): A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. In *PLoS Genet* 8 (1), e1002451. DOI: 10.1371/journal.pgen.1002451.

McTavish, EJ; Hillis, DM (2015): How do SNP ascertainment schemes and population demographics affect inferences about population history? In *BMC Genomics* 16 (1), p. 1.

Meyermans, R; Gorssen, W; Buys, N; Janssens, S (2020): How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. In *BMC Genomics* 21 (1), p. 94. DOI: 10.1186/s12864-020-6463-x.

National Center for Biotechnology Information (NCBI) (2018): GRCg6a. Available online at https://www.ncbi.nlm.nih.gov/assembly/GCF_000002315.6#/st, updated on 3/27/2018, checked on 8/17/2021.

National Center for Biotechnology Information (NCBI) (2021a): bGalGal1.mat.broiler.GRCg7b. bGalGal1 maternal broiler GRCg7b. Available online at https://www.ncbi.nlm.nih.gov/assembly/GCF_016699485.2#/st, updated on 1/19/2021, checked on 8/17/2021.

National Center for Biotechnology Information (NCBI) (2021b): bGalGal1.pat.whiteleghornlayer.GRCg7w. bGalGal1 paternal white leghorn layer GRCg7w. Available online at https://www.ncbi.nlm.nih.gov/assembly/GCF_016700215.1/#/st, updated on 1/19/2021, checked on 8/17/2021.

National Human Genome Institute (2020): The Cost of Sequencing a Human Genome. Available online at https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, updated on 12/7/2020, checked on 7/19/2021.

Nielsen, R; Hubisz, MJ; Clark, AG (2004): Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. In *Genetics* 168 (4), pp. 2373–2382.

Pedersen, BS; Quinlan, AR (2018): Mosdepth: quick coverage calculation for genomes and exomes. In *Bioinformatics (Oxford, England)* 34 (5), pp. 867–868. DOI: 10.1093/bioinformatics/btx699.

Perez-Enciso, M; Rincon, JC; Legarra, A (2015): Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. In *Genet Sel Evol* 47, p. 43. DOI: 10.1186/s12711-015-0117-5.

Pook, T; Nemri, A; Gonzalez Segovia, EG; Simianer, H; Schoen, C-C (2021): Increasing calling accuracy, coverage, and read depth in sequence data by the use of haplotype blocks. DOI: 10.1101/2021.01.07.425688.

Pook, T; Schlather, M; Simianer, H (2020): MoBPS - Modular Breeding Program Simulator. In *G3 (Bethesda, Md.)* 10 (6), pp. 1915–1918. DOI: 10.1534/g3.120.401193.

Purcell, S; Neale, B; Todd-Brown, K; Thomas, L; Ferreira, MAR; Bender, D et al. (2007): PLINK: a tool set for whole-genome association and population-based linkage analyses. In *Am J Hum Genet* 81 (3), pp. 559–575. DOI: 10.1086/519795.

Qanbari, S (2020): On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. In *Frontiers in genetics* 10, p. 1304. DOI: 10.3389/fgene.2019.01304.

Qanbari, S; Pausch, H; Jansen, S; Somel, M; Strom, T-M; Fries, R et al. (2014): Classic selective sweeps revealed by massive sequencing in cattle. In *PLoS Genet* 10 (2), e1004148. DOI: 10.1371/journal.pgen.1004148.

Qanbari, S; Simianer, H (2014): Mapping signatures of positive selection in the genome of livestock. In *Livestock Science* 166, pp. 133–143. DOI: 10.1016/j.livsci.2014.05.003.

Quinlan, AR; Hall, IM (2010): BEDTools: a flexible suite of utilities for comparing genomic features. In *Bioinformatics (Oxford, England)* 26 (6), pp. 841–842. DOI: 10.1093/bioinformatics/btq033.

Ramos, AM; Crooijmans, RPMA; Affara, NA; Amaral, AJ; Archibald, AL; Beever, JE et al. (2009): Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. In *PLoS One* 4 (8), e6524.

Rensing, S; Alkhoder, H; Kubitz, C; Schierenbeck, S; Segelke, D (2017): Cow Reference Population. Benefit for Genomic Evaluation System and Farmers. In *INTERBULL bulletin* (51), pp. 63–66.

Rhie, A; McCarthy, SA; Fedrigo, O; Damas, J; Formenti, G; Koren, S et al. (2021): Towards complete and error-free genome assemblies of all vertebrate species. In *Nature* 592 (7856), pp. 737–746. DOI: 10.1038/s41586-021-03451-0.

Schmid, M; Smith, J; Burt, DW; Aken, BL; Antin, PB; Archibald, AL et al. (2015): Third Report on Chicken Genes and Chromosomes 2015. In *Cytogenet Genome Res* 145 (2), pp. 78–179. DOI: 10.1159/000430927.

Sedlazeck, FJ; Lee, H; Darby, CA; Schatz, MC (2018): Piercing the dark matter. Bioinformatics of long-range sequencing and mapping. In *Nature Reviews Genetics* 19 (6), pp. 329–346. DOI: 10.1038/s41576-018-0003-4.

Toro Ospina, AM; Aguilar, I; Vargas de Oliveira, MH; Cruz dos Santos Correia, LE; Vercesi Filho, AE; Albuquerque, LG; Vasconcelos Silva, JA de, II (2021): Assessing the accuracy of imputation in the Gyr breed using different SNP panels. In *Genome*, pp. 1–7. DOI: 10.1139/gen-2020-0081.

Tosser-Klopp, G; Bardou, P; Bouchez, O; Cabau, C; Crooijmans, RPMA; Dong, Y et al. (2014): Design and characterization of a 52K SNP chip for goats. In *PLoS One* 9 (1), e86227.

Utsunomiya, YT; Milanesi, M; Fortes, MRS; Porto-Neto, LR; Utsunomiya, ATH; Silva, M. V. G. B. et al. (2019): Genomic clues of the evolutionary history of Bos indicus cattle. In *Anim Genet* 50 (6), pp. 557–568. DOI: 10.1111/age.12836.

van der Auwera, GA; Carneiro, MO; Hartl, C; Poplin, R; Del Angel, G; Levy-Moonshine, A et al. (2013): From FastQ data to high confidence variant calls. The Genome Analysis Toolkit best practices pipeline. In *Current protocols in bioinformatics* 43 (1), 11.10.1-11.10.33. DOI: 10.1002/0471250953.bi1110s43.

Vertebrate Genomes Project (2021): GenomeArk. Available online at https://vgp.github.io/genomeark/, checked on 8/24/2021.

Wang, K; Hu, H; Tian, Y; Li, J; Scheben, A; Zhang, C et al. (2021): The chicken pan-genome reveals gene content variation and a promoter region deletion in IGF2BP1 affecting body size. In *Mol Biol Evol*. DOI: 10.1093/molbev/msab231.

Weir, BS; Cockerham, CC (1984): Estimating F-statistics for the analysis of population structure. In *Evolution*, pp. 1358–1370.

Wright, S (1949): The genetical structure of populations. In *Ann Eugen* 15 (1), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.

Appendix

Acknowledgements

I would like to thank

Prof. Dr. Henner Simianer for being my first supervisor and shaping my understanding of breeding and genetics to a large extent, starting from my first 'Tierzucht' lecture up to this thesis. Further, I am very grateful that he always allowed me to follow my interests through that time and to develop and implement my own ideas.

Apl. Prof. Dr. Steffen Weigend for being my second supervisor and all the information about chicken diversity he shared with me.

Prof. Dr. Timothy M Beissinger for agreeing to be my third supervisor even before he arrived to Göttingen in person.

Mrs. Döring for being the best and most helpful secretary ever, even in situations she strictly believes that she will go into jail for.

All the present and former members of the **Animal Breeding and Genetics Group**. Especially **Christian** and **Torsten** for always listening to my (scientific) problems, even when I just needed to think loudly. Also especially **Reza** for all his Iranian stories and for using his special influence to change the weather in London.

Annett Weigend for patiently sending me one sample list after the other, while I tried to merge the different chicken sets.

My Family for supporting me through the time and always understanding when I was poorly reachable.

Curriculum Vitae

Personal Information

Name	Geibel
Surname	Johannes
Date of birth	02.10.1989
Place of birth	Freising, D
E-mail	johannes.geibel@uni-goettingen.de

Education

Since 01.08.2016	Doctoral program agriculture (PAG) Georg-August-Universität Göttingen Topic: “SNP Ascertainment Bias” Supervisors: Henner Simianer, Steffen Weigend, Timothy Beissinger
01.10.2014 - 07.07.2016	M.Sc. Agrar – Animal Sciences Georg-August-Universität Göttingen Thesis: “Influence of Subspecies-Specific Genotypes on Growth- and Carcase-Traits in an Australian *Bos taurus - Bos indicus* Composite Breed” Supervisors: Henner Simianer, Stefan Hiendleder
01.10.2011 - 10.10.2014	B.Sc. Agrar – Animal Sciences Georg-August-Universität Göttingen Thesis: “Ausprägung des Geschlechtsdimorphismus bei verschiedenen Hühnerrassen” Supervisors: Henner Simianer, Ahmad Reza Sharifi
01.08.2008 - 31.07.2011	Vocational training as “Pferdewirt – Zucht und Haltung” Sächsische Gestütsverwaltung - Landgestüt Moritzburg Awarded with “Graf-von-Lehndorff-Plakette in Bronze”
2008 - 2011	Vocational school: Berufliches Schulzentrum für Agrarwirtschaft und Ernährung Dresden
2000 - 2008	Gymnasium: Martin-Andersen-Nexö-Gymnasium Dresden, Abitur
1996 - 2000	Primary School

Work Experience

Since 01.08.2016	Georg-August-Universität Göttingen Department of Animal Sciences Center for Integrated Breeding Research (CiBreed) Animal Breeding and Genetics Group PhD Student
15.06.2016 - 31.07.2016	Georg-August-Universität Göttingen Department of Animal Sciences Animal Breeding and Genetics Group Student Assistant
01.08.2011 - 30.09.2011	Sächsische Gestütsverwaltung - Landgestüt Moritzburg Gestütswärter

Stay Abroad

15.03.2016 - 24.04.2016	Research stay during the Master Thesis in the group of Prof. Dr. Stefan Hiendleder, University of Adelaide, School of Animal and Veterinary Sciences, JS Davies Epigenetics and Genetics Group, Australia

Publications

Geibel J, Praefke NP, Weigend S, Simianer H, Reimer C **(2021)** Linkage patterns between structural variants and SNP in three chicken breeds. **Under Review at BMC Genomics, preprint available on Research Square. DOI: 10.21203/rs.3.rs-861830/v1**

Pook T, Reimer C, Freudenberg A, Büttgen L, **Geibel J**, Ganesan A, Ha NT, Schlather M, Mikkelsen LF, Simianer H **(2021)** The Modular Breeding Program Simulator (MoBPS) allows efficient simulation of complex breeding programs. **Accepted at Animal Production Science**

Geibel J, Reimer C, Pook T, Weigend S, Weigend A, Simianer H **(2021)** How Imputation can Mitigate SNP Ascertainment Bias. BMC Genomics 22 (1): 340. DOI: 10.1186/s12864-021-07663-6

Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H **(2021)** How Array Design creates SNP Ascertainment Bias. PLoS ONE 16(3):e0245178. DOI: 10.1371/journal.pone.0245178

Peripolli E, Reimer C, Ha N-T, **Geibel J,** Machado MA, do Carmo Panetto JC, do Egito AA, Baldi F, Simianer H, da Silva MVG **(2020)** Genome-wide detection of signatures of selection in indicine and Brazilian locally adapted taurine cattle breeds using whole-genome re-sequencing data. BMC Genomics 21 (1). DOI: 10.1186/s12864-020-07035-6

Büttgen L, **Geibel J,** Simianer H, Pook T **(2020)** Simulation Study on the Integration of Health Traits in Horse Breeding Programs. Animals 10 (7). DOI: 10.3390/ani10071153

Pook T, Mayer M, **Geibel J,** Weigend S, Cavero D, Schoen CC, Simianer H. **(2020)** Improving imputation quality in BEAGLE for crop and livestock data. G3: GenesGenomes Genetics 10 (1). DOI: 10.1534/g3.119.400798

Reimer C, Ha N-T, Sharifi A. R, **Geibel J,** Mikkelsen LF, Schlather M, Weigend S, Simianer H **(2020)** Assessing breed integrity of Göttingen Minipigs. BMC Genomics 21 (1). DOI: 10.1186/s12864-020-6590-4

Qanbari S, Rubin C-J, Maqbool K, Weigend S, Weigend A, **Geibel J,** Kerje S, Wurmser C, Townsend Peterson A, Lehr Brisbin Jr. I, Preisinger R, Fries R, Simianer H, Andersson L **(2019)** Genetics of adaptation in modern chicken. PLOS Genetics, DOI: 10.1371/journal.pgen.1007989

Scientific Talks

Reimer C, **Geibel J,** Sharifi AR, Simianer H **(2020)** Using Pool-Sequencing to Elucidate the Genetic Background of Seizures in the Göttingen Minipig. Abstract book. Page 224. International Congress on Quantitative Genetics (ICQG6), Virtual Conference.

Reimer C, **Geibel J,** Pook T, Weigend S, Simianer H **(2020)** Employing trio information to assess CNV detection performance in array data of Göttingen Minipigs. Book of abstracts No. 26. Session 34. Page 384. Annual Meeting of EAAP

Büttgen L, **Geibel J,** Simianer H, Pook T **(2020)** Simulation study for the integration of health traits in horse breeding programs. Book of abstracts No. 26. Session 30. Page 339. Annual Meeting of EAAP

Büttgen L, **Geibel J,** Simianer H, Pook T **(2020)** Simulationsstudie zur Integration von Gesundheitsmerkmalen in Pferdezuchtprogramme. 9. Pferde-Workshop der DGfZ. Bad Bevensen Germany

Geibel J, Pook T, Weigend S, Weigend A, Simianer H **(2019)** Wie Imputing SNP Ascertainment Bias reduzieren kann und wann es ihn verschärft. Vortragstagung der DGfZ und GfT. Gießen Germany 12.09.2019

Geibel J, Weigend S, Weigend A, Reimer C, Pook T, Simianer H **(2018)** Auswirkungen des Array Design Prozesses auf die Überschätzung der Heterozygotie. Vortragstagung der DGfZ und GfT. Bonn Germany 13.09.2018

Reimer C, Ha N-T, Sharifi AR, Weigend S, **Geibel J,** Simianer H **(2018)** Analyses of the breed integrity of the Goettingen Minipig using pool-sequencing. Proceedings 11th World Congress of Genetics Applied to Livestock Production Auckland New Zealand.

Reimer C, **Geibel J,** Weigend S, Simianer H **(2017).** Analyse der Rassenintegrität des Göttinger Miniaturschweins anhand gepoolter Sequenzdaten. Tagungsband der Vortragstagung der Deutschen Gesellschaft für Züchtungskunde und der Gesellschaft für Tierzuchtwissenschaften. Deutschland Hohenheim.

Geibel J, Weigend S, Weigend A, Reimer C, Pook T, Simianer H **(2017).** Array Design und SNP Ascertainment Bias. Tagungsband der Vortragstagung der Deutschen Gesellschaft für Züchtungskunde und der Gesellschaft für Tierzuchtwissenschaften. Deutschland Hohenheim.

Geibel J, Weigend S, Weigend A, Sharifi AR, Simianer H **(2016).** Patterns of sexual size dimorphism in various chicken breeds. Book of Abstracts No. 22. 67th Annual Meeting of the European Association for Animal Production. Belfast UK. Wageningen Academic Publishers P. 124

Geibel J, Simianer H, Weigend S, Weigend A, Sharifi AR **(2015)**. Ausprägung des Geschlechtsdimorphismus bei verschiedenen Hühnerrassen. Tagungsband der Vortragstagung der Deutschen Gesellschaft für Züchtungskunde und der Gesellschaft für Tierzuchtwissenschaften. Deutschland Berlin.

Scientific Posters

Geibel J, Hansen S, Abdelwahed A, Böhlken-Fascher S, Sharifi AR, Czerny C-P, Simianer H **(2020)** Assessing Nanopore Sequencing for Detection of Large Structural Variation in the Goettingen Minipig. IMAGE Project Meeting Madrid

Geibel J, Hansen S, Abdelwahed A, Böhlken-Fascher S, Sharifi AR, Czerny C-P, Simianer H **(2019)** Assessing Nanopore Sequencing for Detection of Large Structural Variation in the Goettingen Minipig. CiBreed Workshop Göttingen

Geibel J, Hansen S, Abdelwahed A, Böhlken-Fascher S, Sharifi AR, Czerny C-P, Simianer H **(2019)** Assessing Nanopore Sequencing for Detection of Large Structural Variation in the Goettingen Minipig. Book of Abstracts EAAP 2019 Gent

Pook T, Mayer M, **Geibel J,** Schoen CC, Simianer H **(2019)** Improving imputation quality in BEAGLE for crop data. PLANT2030 Statusseminar. Potsdam Germany 13.-15.03.2019

Geibel J, Weigend S, Weigend A, Reimer C, Pook T, Simianer H **(2018)** Array design and SNP ascertainment bias. Population Evolutionary and Quantitative Genetics Conference. Madison USA 13.-16.05.2018

Geibel J, Weigend S, Weigend A, Reimer C, Pook T, Simianer H **(2018)** Array design and SNP ascertainment bias. World Congress on Genetics Applied to Livestock Production. Auckland New Zealand 13.02.2018

Supervised Bachelor Theses

Guerndt E **(2021)** Aufbau einer Genreserve des Göttinger Miniaturschweinebestandes in Relliehausen. Supervisors: Reimer C, **Geibel J**

Beerepoot MD **(2021)** Vergleich direkter genomischer Zuchtwerte für Milchleistungs- und Exterieurmerkmale mit deren phänotypischen Ausprägungen. Supervisors: Reimer C, **Geibel J**

Schulze Vels M **(2017)** Analyse des Vorkommens verschiedener Hengstlinien zur Körung beim Deutschen Reitpferd seit 1990. Supervisors: Simianer H, **Geibel J**

Horst N **(2017)** Auswertung der Leistungsprüfungen für Ponys, Kleinpferde und sonstige Rassen der Jahre 2014 – 2016. Supervisors: Simianer H, **Geibel J**

Wollrath A **(2016)** Schätzung genetischer Komponenten aus den Ergebnissen des Louisdor Preises. Supervisors: Simianer H, **Geibel J**

Supervised Master Theses

Matteikat W **(2019)** Zusammenhang zwischen Turniersporterfolgen und dem Ergebnis der Zuchtstutenprüfung sowie der Zuchtbucheintragung beim Holsteiner Pferd. Supervisors: Tetens J, **Geibel J**

Gickel JM **(2018)** Kartierung von Geschlechtsdimorphismen im Synbreed Chicken Diversity Panel. Supervisors: Simianer H, **Geibel J**

Teaching Experiences

B.Agr.0009 Grundlagen der Nutztierwissenschaften II

B.Agr.0325 Nutztierzüchtung

M.Agr.0068 Quantitativ- genetische Methoden der Tierzucht

M.Agr.0076 Statistische Nutztiergenetik

M.Agr.0126/ M.iPAB.0001 Quantitative Genetics and population genetics

M.Agr.0138/ M.iPAB.0013 Selection theory, design and optimization of breeding programs

M.iPAB.0016 Applied effective R programming in animal breeding and genetics

Additional Courses

10.02.2020 – 14.02.2020	Genome Assembly using Oxford Nanopore Sequencing, Physalia Courses Berlin, D
09.12.2019 – 13.12.2019	Course on characterization, management and exploitation of genomic diversity in animals, Wageningen University & Research, NL
08.06.2018	Competence in Research Integrity, University of Goettingen, D
05.03.2018 – 09.03.2018	GFT Statistikkurs „Nutzung der Statistiksoftware R in der Tierzucht und Tierhaltung“, University of Hohenheim, D
11.09.2017 – 15.09.2017	Genomic Data Visualization and Interpretation, Physalia Courses Berlin, D

Declaration of academic integrity

I confirm that I have composed the present scientific thesis independently using no other sources and resources than those stated. I have accepted the assistance of third parties only in a scope that is scientifically justifiable and compliant with the legal statutes of the examinations. In particular, I have completed all parts of the dissertation myself. I have neither, nor will I, accept unauthorized outside assistance either free of charge or subject to a fee. Furthermore, I have not applied for an equivalent doctoral examination elsewhere and submitted the present thesis as a whole or in parts at another university. I am aware of the fact that untruthfulness with respect to the above declaration repeals the admission to complete the doctoral studies and/or subsequently entitles termination of the doctoral process or withdrawal of the title attained.

Candidate's signature: ______________________________

(Johannes Geibel)