RESEARCH ARTICLE

# How array design creates SNP ascertainment bias

**Johannes Geibel**[1,2]*, **Christian Reimer**[1,2], **Steffen Weigend**[2,3], **Annett Weigend**[3], **Torsten Pook**[1,2], **Henner Simianer**[1,2]

**1** Department of Animal Sciences, Animal Breeding and Genetics Group, University of Goettingen, Göttingen, Germany, **2** Center for Integrated Breeding Research, University of Goettingen, Göttingen, Germany, **3** Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, Neustadt-Mariensee, Germany

* johannes.geibel@uni-goettingen.de

**Citation:** 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. https://doi.org/10.1371/journal.pone.0245178

**Editor:** Francesc Calafell, Universitat Pompeu Fabra, SPAIN

**Received:** September 9, 2020

**Accepted:** December 22, 2020

**Published:** March 30, 2021

**Data Availability Statement:** Additional to the sequences already published by Malomane et al. [29] or Qanbari et al. [41], sequencing reads of Gallus varius (GV), Rhode Island Red (RI) and White Rock (WR) pools were uploaded to the European Nucleotide Sequence Database (ENA; https://www.ebi.ac.uk/ena/data/view/PRJEB31780). The study accession numbers of former and newly submitted reads can be found in S1 File. Note that Malomane et al. [29] already published a subset of the individually sequenced white layers, published in total by Qanbari et al.

## 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 [1,2], model species [3,4], plants [5,6] and livestock [7–13].

Various SNP arrays exist for humans [14], plants [15,16] and all major livestock species [17–23]. SNP numbers within these arrays range from 10 k SNPs [20] over approximately 50 k

[41]. The raw SNP callset was additionally deposited on Zenodo (10.5281/zenodo.4320456).

**Funding:** The sampling and sequencing of the chicken populations was funded by the EC project AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; https://aviandiv.fli.de/) and the German Federal Ministry of Education and Research (BMBF) via the SYNBREED project (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de). The authors further acknowledge support by the Open Access Publication Funds of the University of Goettingen.

**Competing interests:** The authors have declared that no competing interests exist.

[16,17,19,22] up to 600 k [15,21]. 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 [24]. Additional criteria like equidistant spacing over the genome [21], overrepresentation of some areas like chromosomal ends to increase imputation accuracy [20] or genic regions [21], or increased overrepresentation of high MAF SNPs [17] 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 [18] and/or on a broad set of individuals from different populations [21,24].

In contrast to whole genome re-sequencing (WGS) data, SNP arrays often show a clear underrepresentation of SNPs with extreme allele frequencies [25]. As population genetic statistics are mostly based on estimates of allele frequencies, this context leads to biased population genetic estimators [25,26] 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 [27]. 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 [26,28]. 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 [25,27]. 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 [29,30].

Besides different strategies to minimize the impact of ascertainment bias [30,31], there are some attempts to correct the allele frequency spectrum via Bayesian methods [25,28,32]. However, those corrections highly rely on detailed statistical assumptions of the ascertainment process [33,34] or take a variety of ascertainment processes and demographic patterns into account to model evolutionary scenarios which are then compared to real world data [29,35]. However, those methods are currently only tested for corrections of the first source of ascertainment bias, the small discovery panel [25,28,32]. Additionally, detailed information on the design process is limited in practice [34] 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 [36,37], represented by the five subspecies *G. g. gallus*, *G. g. spadiceus*, *G. g. murghi*, *G. g. bankiva* and *G. g. jabouillei* [38]. Additionally, some hybridization events with other *Gallus* species (e.g. grey jungle fowl; *Gallus sonneratii*) have been suggested [37,39]. 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 [38]. 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) [40]. For more detailed information on chicken ancestry we refer to Lawal *et al.* [37] and for a comprehensive overview on diversity and population structure of domesticated chickens to Malomane *et al.* [13].

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 [21] 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 (https://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.* [30] and Qanbari *et al.* [41]. 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 [42,43] and the SNP calling was conducted according to GATK Best Practices guidelines [44,45]. BWA-MEM 0.7.12 [46] was used for the alignment step, duplicates were marked using Picard Tools 2.0.1 [47] MarkDuplicatesWithMateCigar and base qualities were recalibrated with GATK 3.7 [48] BaseQualityRecalibrator. The set of known SNPs, necessary for base quality score recalibration, was downloaded from ENSEMBL release 87 [49]. 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 [50]. 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.* [51]. 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 [52], 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 [48] 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 Fig 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 [53]. As the two broiler populations were closest related (S3 Fig), the next two closest populations were

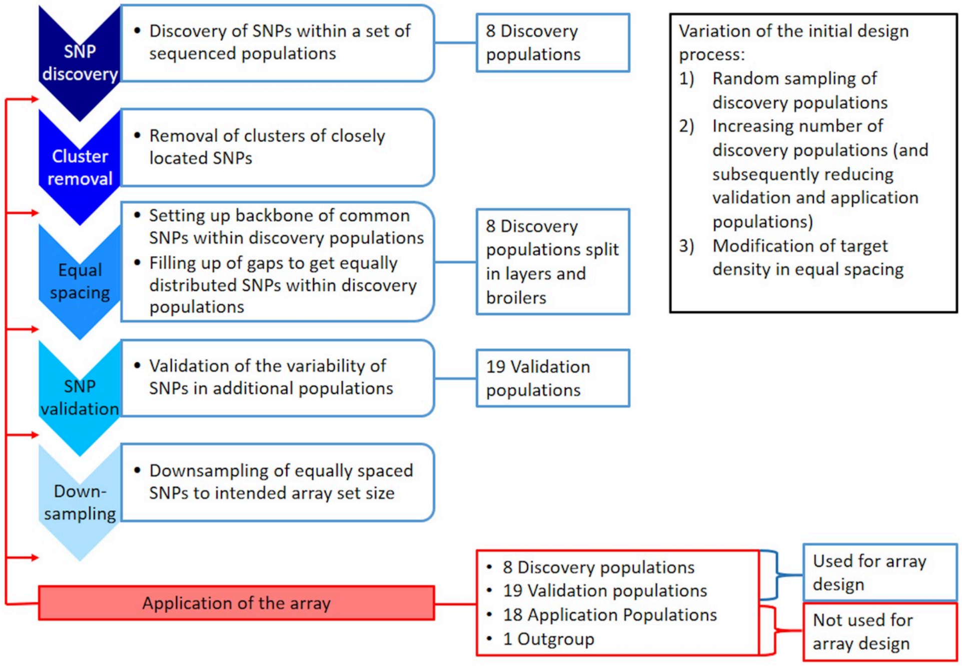
SNP discovery
• Discovery of SNPs within a set of sequenced populations
8 Discovery populations
Cluster removal
• Removal of clusters of closely located SNPs
Equal spacing
• Setting up backbone of common SNPs within discovery populations
• Filling up of gaps to get equally distributed SNPs within discovery 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
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

**Fig 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.

https://doi.org/10.1371/journal.pone.0245178.g001

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 [21] 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 [21] 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 [53] in S3 File and additionally pairwise $F_{ST}$ values [54] in S4 File.

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.* [21], starting from the unfiltered SNP set:

1. **SNP discovery → 10.9 M SNPs**
   Discovery of SNPs fulfilling basic criteria (quality $\geq 60$; MAF $\geq 0.05$; coverage $\leq$ 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*. [21] 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*. [21] 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 [56]). 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; [57]). 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 $H_{exp}$, 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 Eq (1), where $p_{ref;l}$ denotes the frequency of the reference allele at locus $l$ and $L$ the total number of loci.

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

Deviations in the estimation of $H_{exp}$ from the various SNP sets were quantified as differences between the $H_{exp}$ calculated from the respective SNP set and the $H_{exp}$ calculated from the filtered WGS SNPs relative to the $H_{exp}$ from the filtered WGS SNPs, further called overestimation of $H_{exp}$ (OHE; Eq (2)), which was calculated per population.

$$OHE = \frac{H_{\exp;\ \mathrm{SNP\ set}} - H_{\exp;\ \mathrm{filtered\ WGS\ SNPs}}}{H_{\exp;\ \mathrm{filtered\ WGS\ SNPs}}} \tag{2}$$

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 [58,59] 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 [60] as shown in Eq (3), where the OHE depended on an overall mean $\mu$, 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 ($rep_j \tilde{} N(0, I\sigma^2_{rep})$) and a random error $e_{ijk} \tilde{} N(0, I\sigma^2_e)$. 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.

$$OHE_{ijk} = \mu + popG_i + rep_j + e_{ijk} \tag{3}$$

## 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

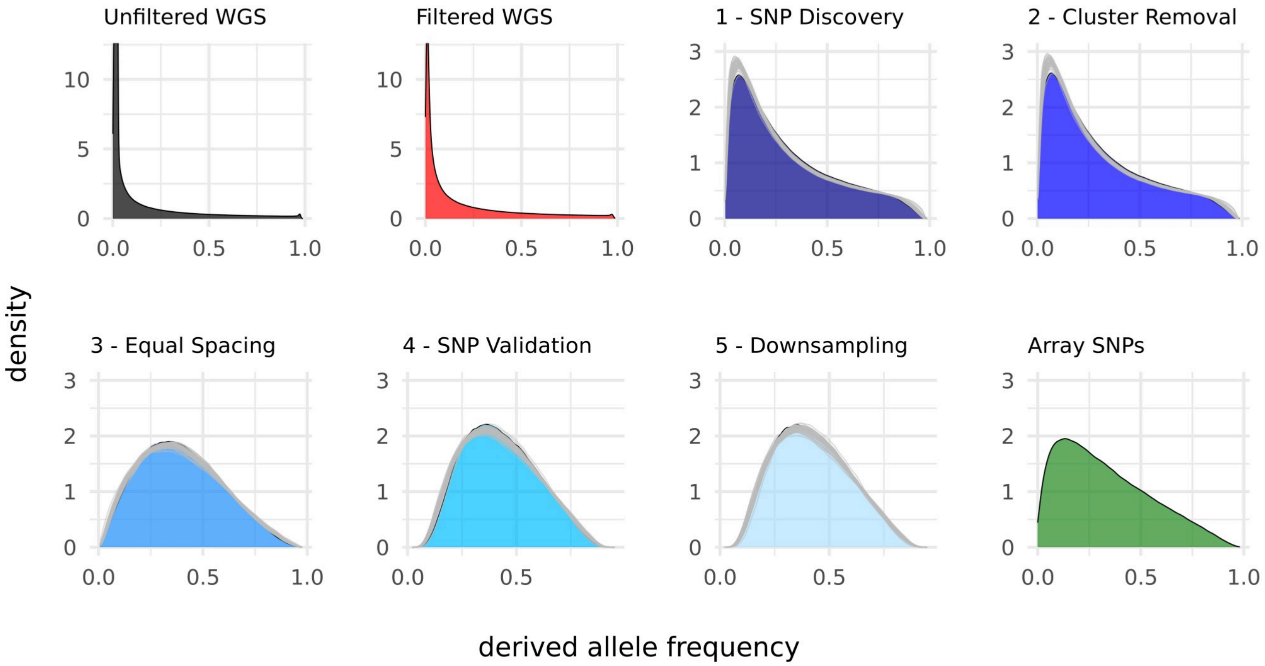

**Fig 2. Derived allele frequency spectra for the different SNP sets.** For the remodeled sets, areas show the modelling according to the original array [21] while grey lines represent the 50 random population groupings.

https://doi.org/10.1371/journal.pone.0245178.g002

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 (Fig 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 (Fig 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 [21]. As major changes in the spectra mainly occurred after the SNP discovery and equal spacing, further results will concentrate on those steps.

The allele frequency spectra (Fig 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.

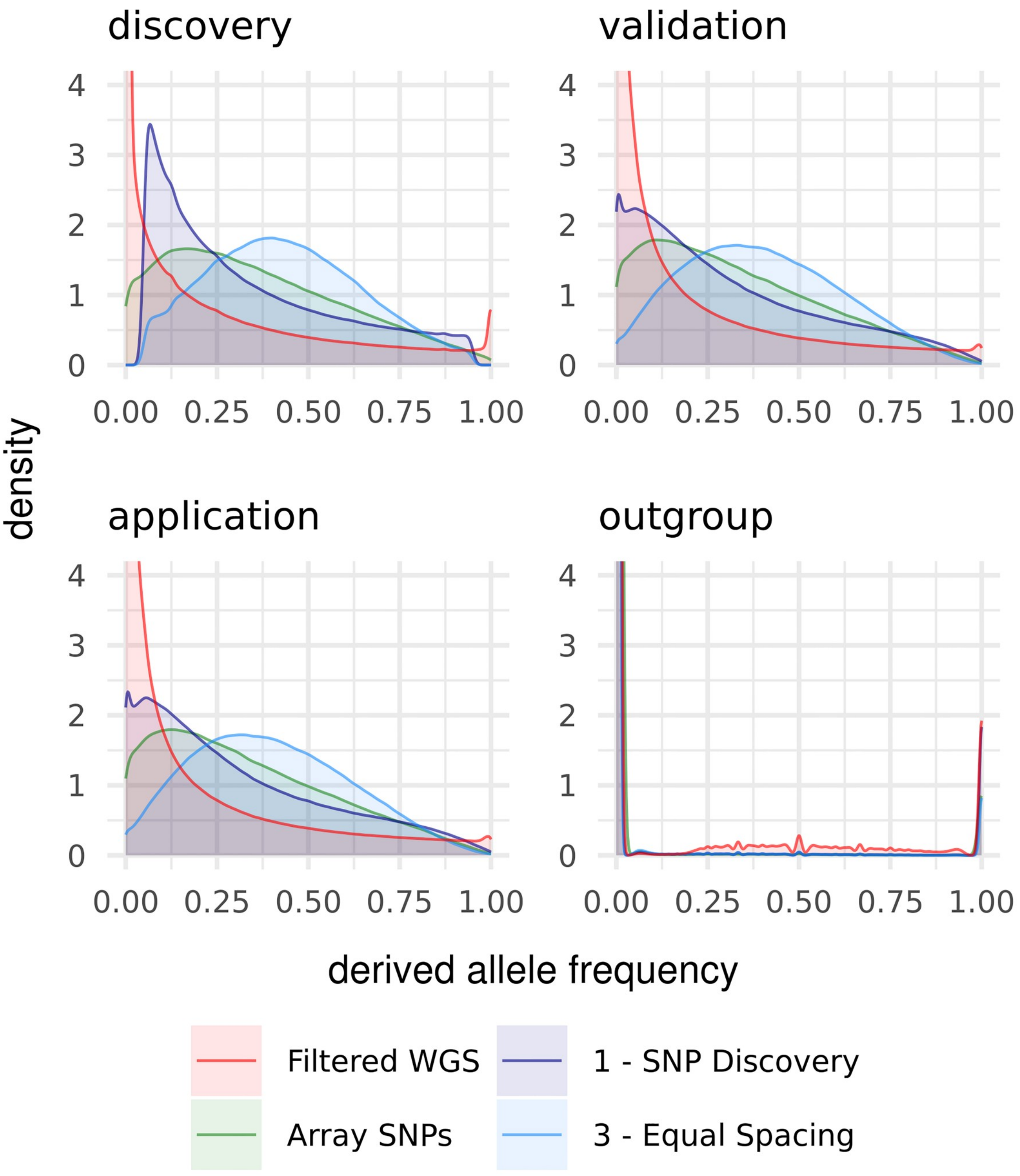

**Fig 3. Derived allele frequency spectra within the population groups.**

https://doi.org/10.1371/journal.pone.0245178.g003

### 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 (Fig 4A). 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

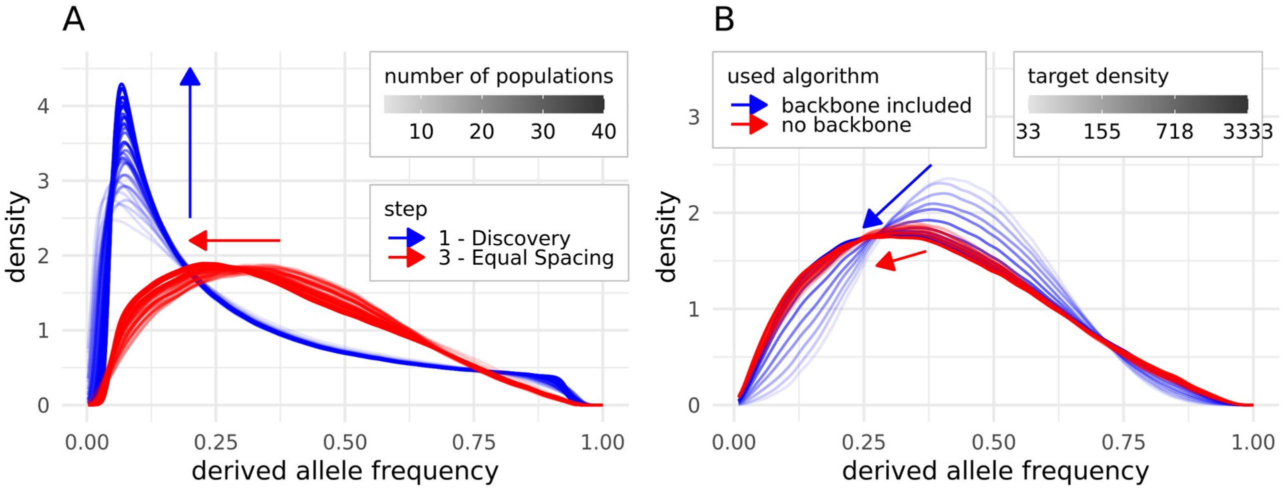

**Fig 4. Impact of a varying number of discovery populations (A) or target density (B) on the derived allele frequency spectrum.** For A, blue indicates the spectra after the discovery step and red after the equal spacing step. For B, only the equal spacing step is shown and blue indicates that the algorithm 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.

https://doi.org/10.1371/journal.pone.0245178.g004

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.

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 (Fig 4B). 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 (Fig 4B).

## Overestimation of $H_{exp}$

Fig 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 ($\geq 0.95$; 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 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

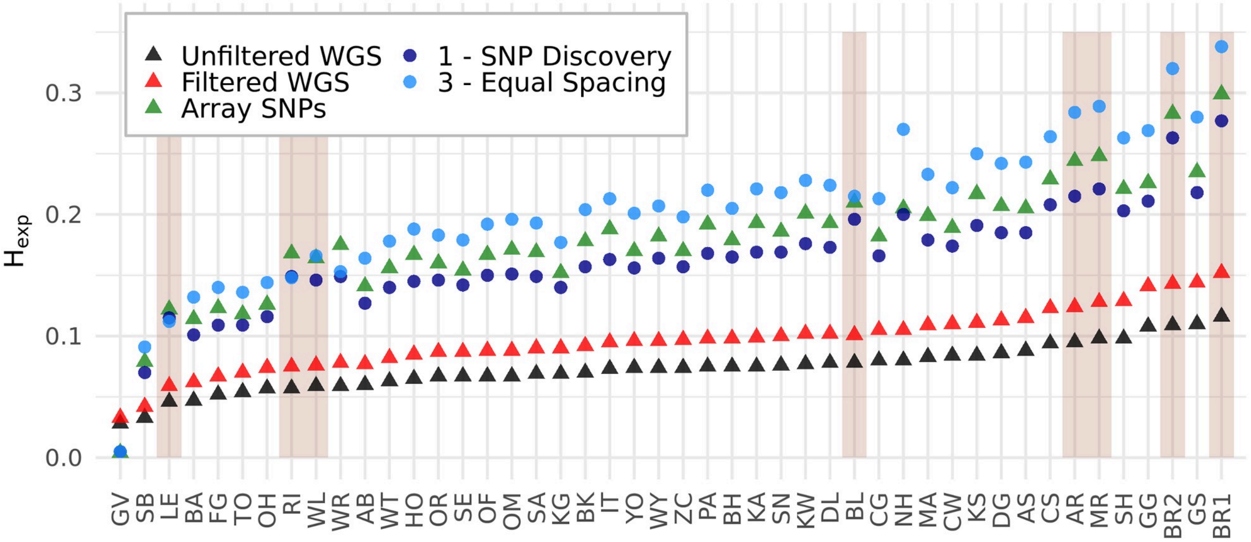

**Fig 5. Expected Heterozygosity ($H_{exp}$) by population and SNP set.** Populations are ordered by the $H_{exp}$ of the unfiltered WGS SNP set. Only the reference sets and relevant steps of the array design are shown. Discovery populations are shaded with a darker background.

https://doi.org/10.1371/journal.pone.0245178.g005

from the original array SNPs (1.41; Fig 5; Table 1) laid in the range covered by the remodeling steps.

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 (Fig 5; Table 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

**Table 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 $H_{exp}$.

https://doi.org/10.1371/journal.pone.0245178.t001

Table 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.

https://doi.org/10.1371/journal.pone.0245178.t002

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%.

These findings were supported by the 50 random groupings (S8 Fig). The LSMEANS (Table 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 $H_{exp}$

Fig 6A 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 (Fig 6B)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 (S9D 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).

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 (Fig 7A). 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 (Fig 7B).

## 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 [13]. Parts of our set were also involved in the development process of the Axiom™

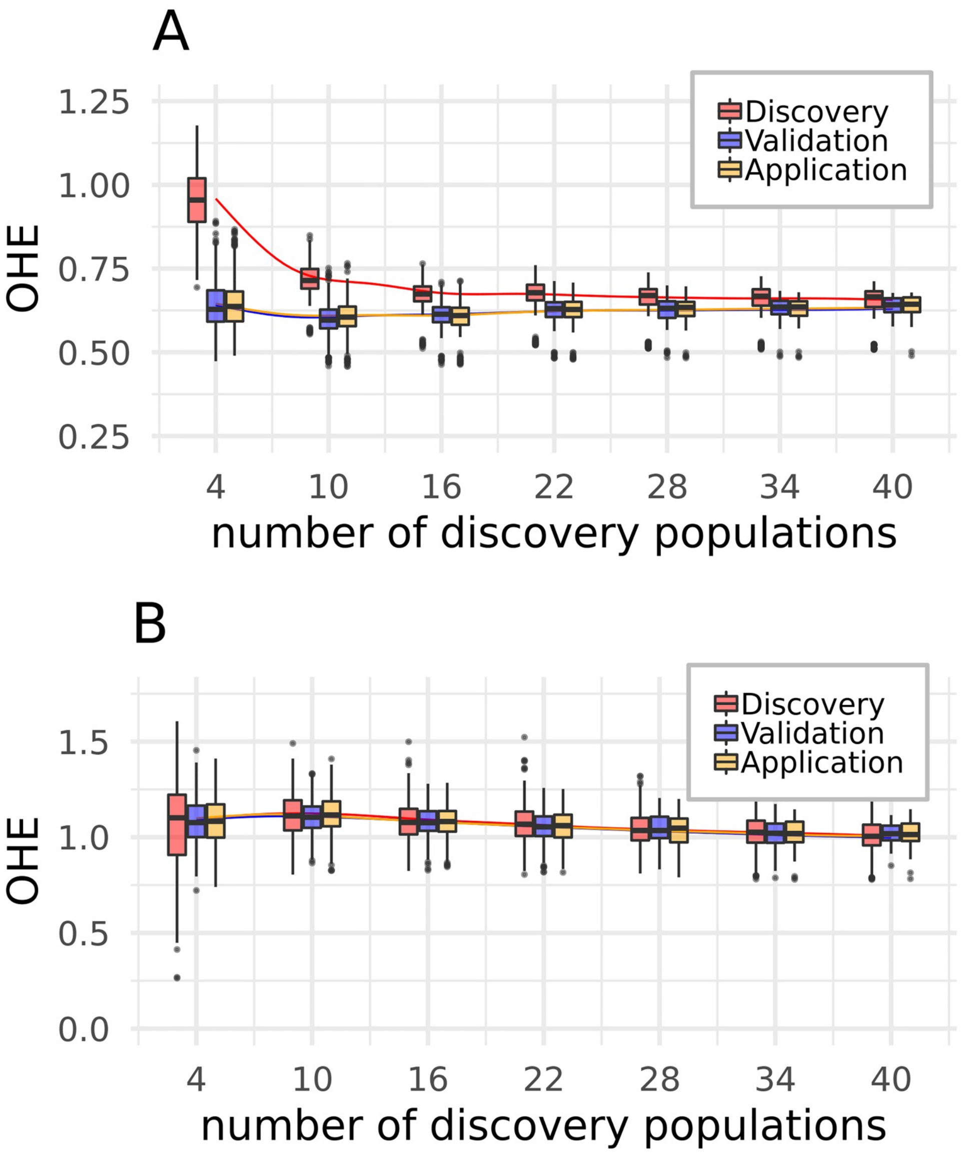

**Fig 6. Relation of the OHE as a function of the number of discovery populations. A**—discovery, **B**—equal spacing. While the number of discovery populations was varied from 4 to 40 by increments of one, the Boxplots are only shown for a subset of the number of discovery populations to avoid a crowded figure. The smoothing lines, which show the trend, are calculated from all observations. Plots for all five steps can be found in S9 Fig.

https://doi.org/10.1371/journal.pone.0245178.g006

Genome-Wide 580 k Chicken Array [21]. 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 [13]. 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* [61] accompanied by limiting the discovery set to one of the two clades, should increase the impact of population structure dependent ascertainment bias.

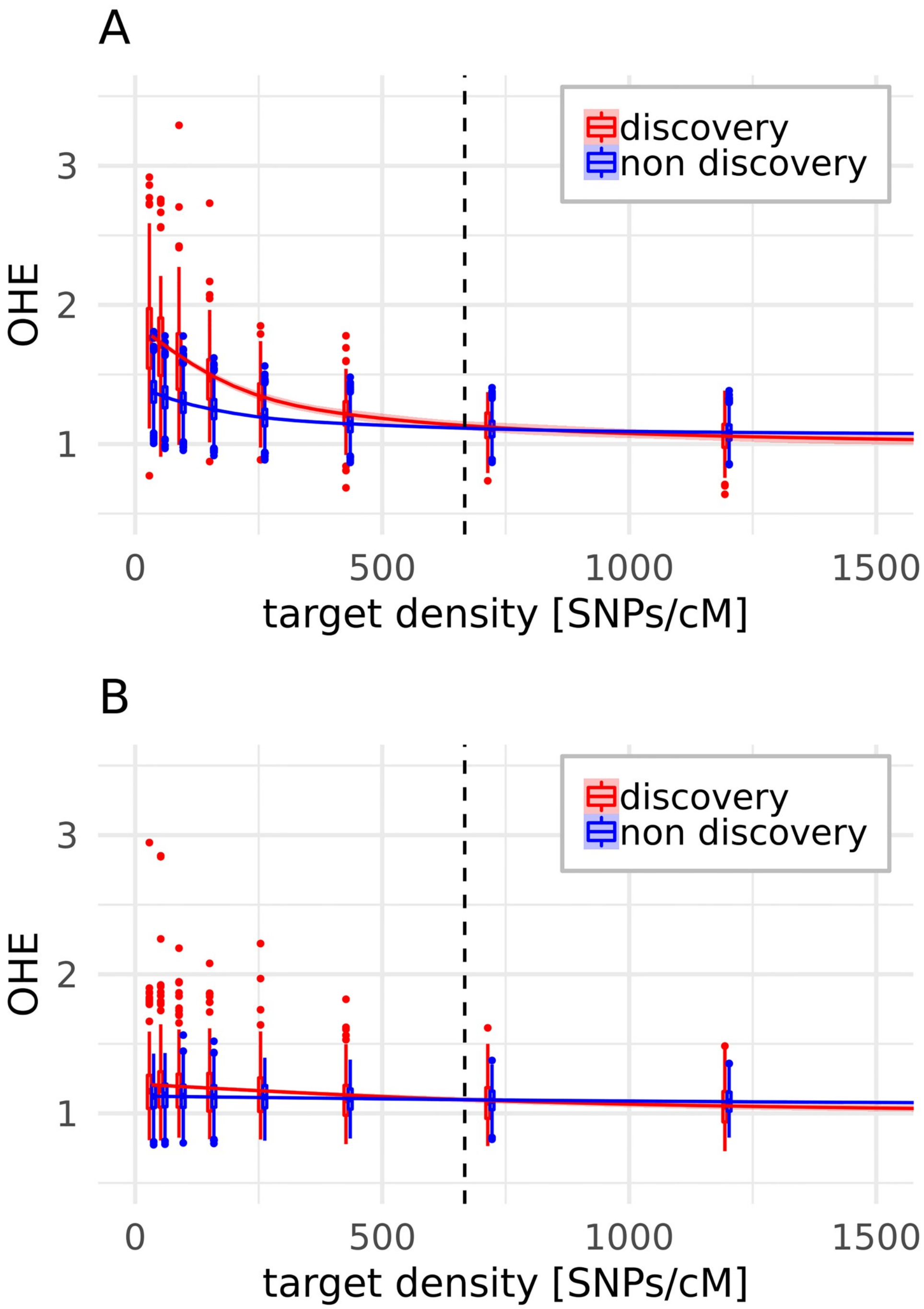

**Fig 7. OHE after equal spacing (step 3) by target density in SNPs/cM and population group.** The smoothing lines show the trend and the dashed lines the target density of 667 SNPs/cM, used for the remodeling according to the original array [21]. The algorithm was run including the initial backbone SNPs (**A**) or not including them (**B**). *Gallus varius* is not included, as it is constantly underestimated.

https://doi.org/10.1371/journal.pone.0245178.g007

## 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 [62], 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 [63–66]. 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 [25,26,30,34]. 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 (Fig 6A). Similarly, the influence of the validation step on validation populations was strongest in case of a small number of validation populations (S9D 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 [25] 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.* [30].

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.* [67] 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 [29] 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 (Fig 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.* [68] 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 (Fig 7A). 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* discovery panels and, vice versa, variants from *Gallus* were not variable in *Gallus varius* (Fig 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 [69] or genome wide association studies [70]) and/or individually sequenced or genotyped individuals (e.g. linkage disequilibrium decay [71] or runs of homozygosity analyses [72]) 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 [73] and therefore some SNP ascertainment schemes intentionally bias the used SNPs towards high MAF within the desired populations [17]. 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 [73–75]. However, the increase in efficiency by higher numbers of SNPs levels off when going towards WGS data [76]. 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 [73] 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. 25,30,31,35,71).

## 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 [62,77], even though such methods introduce other problems related to the use of restriction enzymes which are reviewed by Andrews *et al*. [78].

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.

## Supporting information

**S1 File. Abbreviations for breeds and accession numbers for the sequencing data.**
(XLSX)

**S2 File. Supplementary methods.**
(DOCX)

**S3 File. Pairwise Nei's standard genetic distances.**
(CSV)

**S4 File. Pairwise $F_{ST}$ values.**
(CSV)

**S1 Fig. Pooled vs. ploidy two calling.** Expected Heterozygosity ($H_{exp}$) from the calling with assuming the correct ploidy vs. assuming ploidy two for all samples (A) and alternative allele frequency distributions of called alleles for individually sequenced (B) and pooled sequenced (C) populations.
(TIF)

**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.**
(TIF)

**S3 Fig. UPGMA tree based on pairwise Nei's standard genetic distances.** The tree was calculated from the filtered WGS SNPs. Populations defined as layers or broilers, which form in total the discovery set for the array design close to the original array, are highlighted. The plot was produced using the R package ape [55]. Note that the plot is only supposed to reveal close clustering chicken populations and cannot be interpreted in depth as chickens show a rich history of hybridization events. The interested reader can find all underlying pairwise Nei's standard genetic distances [53] in S3 File and additionally pairwise FST values [54] in S4 File.
(TIF)

**S4 Fig. Cumulative number of SNPs in million retained during the equal spacing step.** The red line and points represent the first remodeling according to the original array [21], while the dashed lines and the boxplots represent the 50 random population groupings and the black line the according median values. The algorithm starts with a very basic initial backbone and then adds SNPs to the backbone which are variable in either all layer lines or all broiler lines. Separated by a vertical line, the second part of the algorithm successively fills up potential gaps to achieve an equidistant coverage of 667 segregating SNPs/cM for each discovery population.
(TIF)

**S5 Fig. SNPs retained by number of populations in the discovery set.**
(TIF)

**S6 Fig. Number of SNPs retained after the equal spacing step by target density.**
(TIF)

**S7 Fig. Relative amount of SNPs shared by a specific number of runs from 50 random population assignments.**
(TIF)

**S8 Fig. OHE for discovery validation and application populations after the five steps of array design.** Discovery populations are chosen to represent populations which are comparable to the original array (blue) or 50 times random sampled (grey).
(TIF)

**S9 Fig. Relation of the OHE as a function of the number of discovery populations.** A—discovery, B–cluster removal, C–equal spacing, D–validation, E–downsampling. The Boxplots are only shown for a subset of the number of discovery populations, while the smoothing lines, which show the trend, are calculated from all observations.
(TIF)

**S1 Table. Number of SNPs from the remodeling processes.**
(DOCX)

**S2 Table. Pearson correlations between the $H_{exp}$ of the different SNP sets.**
(DOCX)

## Acknowledgments

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

## Author Contributions

**Conceptualization:** Johannes Geibel, Christian Reimer, Henner Simianer.

**Data curation:** Johannes Geibel, Annett Weigend.

**Formal analysis:** Johannes Geibel.

**Funding acquisition:** Steffen Weigend, Henner Simianer.

**Investigation:** Johannes Geibel, Christian Reimer.

**Methodology:** Johannes Geibel, Torsten Pook.

**Project administration:** Henner Simianer.

**Software:** Johannes Geibel, Torsten Pook.

**Supervision:** Steffen Weigend, Henner Simianer.

**Visualization:** Johannes Geibel.

**Writing – original draft:** Johannes Geibel.

**Writing – review & editing:** Johannes Geibel, Christian Reimer, Steffen Weigend, Torsten Pook, Henner Simianer.

## References

1. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008; 456:98. https://doi.org/10.1038/nature07331 PMID: 18758442
2. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192:1065–93. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212.
3. Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, et al. Linkage Disequilibrium in Wild Mice. PLoS Genet. 2007; 3:e144. https://doi.org/10.1371/journal.pgen.0030144 PMID: 17722986

4. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The Scale of Population Structure in Arabidopsis thaliana. PLoS Genet. 2010; 6:e1000843. https://doi.org/10.1371/journal.pgen.1000843 PMID: 20169178

5. Travis AJ, Norton GJ, Datta S, Sarma R, Dasgupta T, Savio FL, et al. Assessing the genetic diversity of rice originating from Bangladesh, Assam and West Bengal. Rice. 2015; 8:35. https://doi.org/10.1186/s12284-015-0068-z PMID: 26626493

6. Mayer M, Unterseer S, Bauer E, de Leon N, Ordas B, Schön C-C. Is there an optimum level of diversity in utilization of genetic resources. Theor Appl Genet. 2017; 130:2283–95. https://doi.org/10.1007/s00122-017-2959-4 PMID: 28780586

7. Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, et al. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc Natl Acad Sci. 2008; 105:17312–7. https://doi.org/10.1073/pnas.0806569105 PMID: 18981413

8. Gibbs RA, Taylor JF, van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009; 324:528–32. https://doi.org/10.1126/science.1167936 PMID: 19390050.

9. Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, et al. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One. 2009; 4:e4668. https://doi.org/10.1371/journal.pone.0004668 PMID: 19270757

10. Gautier M, Laloe D, Moazami-Goudarzi K. Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. PLoS One. 2010; 5. https://doi.org/10.1371/journal.pone.0013038 PMID: 20927341.

11. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, et al. A genome-wide scan for signatures of recent selection in Holstein cattle. Animal genetics. 2010; 41:377–89. https://doi.org/10.1111/j.1365-2052.2009.02016.x PMID: 20096028

12. McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM. New World cattle show ancestry from multiple independent domestication events. Proc Natl Acad Sci. 2013; 110:E1398–E1406. https://doi.org/10.1073/pnas.1303367110 PMID: 23530234

13. Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high genomic resolution. BMC Genomics. 2019; 20:345. https://doi.org/10.1186/s12864-019-5727-9 PMID: 31064348.

14. Perkel J. SNP genotyping. Six technologies that keyed a revolution. Nature Methods. 2008; 5:447. https://doi.org/10.1038/nmeth0508-447

15. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics. 2014; 15:823. https://doi.org/10.1186/1471-2164-15-823 PMID: 25266061

16. Singh N, Jayaswal PK, Panda K, Mandal P, Kumar V, Singh B, et al. Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. Sci Rep. 2015; 5:11600. https://doi.org/10.1038/srep11600 PMID: 26111882

17. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009; 4:e5350. https://doi.org/10.1371/journal.pone.0005350 PMID: 19390634

18. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009; 4:e6524. https://doi.org/10.1371/journal.pone.0006524 PMID: 19654876

19. Groenen MAM, Megens H-J, Zare Y, Warren WC, Hillier LW, Crooijmans RPMA, et al. The development and characterization of a 60K SNP chip for chicken. BMC Genomics. 2011; 12:274. https://doi.org/10.1186/1471-2164-12-274 PMID: 21627800

20. Boichard DA, Chung H, Dassonneville R, David X, Eggen A, Fritz S, et al. Design of a bovine low-density SNP array optimized for imputation. PLoS One. 2012; 7:e34130. https://doi.org/10.1371/journal.pone.0034130 PMID: 22470530

21. Kranis A, Gheyas AA, Boschiero C, Turner F, Le Yu, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013; 14:59. https://doi.org/10.1186/1471-2164-14-59 PMID: 23356797

22. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans RPMA, Dong Y, et al. Design and characterization of a 52K SNP chip for goats. PLoS One. 2014; 9:e86227. https://doi.org/10.1371/journal.pone.0086227 PMID: 24465974

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

24. Fan B, Du Z-Q, Gorbach DM, Rothschild MF. Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Australas J Anim Sci. 2010; 23:833–47.

25. Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004; 1:1. https://doi.org/10.1186/1479-7364-1-3-218 PMID: 15588481

26. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15:1496–502. https://doi.org/10.1101/gr.4107905 PMID: 16251459

27. Eller E. Effects of Ascertainment Bias on Recovering Human Demographic History. Human Biology. 2001; 73:411–27. https://doi.org/10.1353/hub.2001.0034 PMID: 11459422

28. Nielsen R, Signorovitch J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol. 2003; 63:245–55. https://doi.org/10.1016/s0040-5809(03)00005-4 PMID: 12689795

29. McTavish EJ, Hillis DM. How do SNP ascertainment schemes and population demographics affect inferences about population history. BMC Genomics. 2015; 16:1.

30. Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics. 2018; 19:22. https://doi.org/10.1186/s12864-017-4416-9 PMID: 29304727

31. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays. 2013; 35:780–6. https://doi.org/10.1002/bies.201300014 PMID: 23836388

32. Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004; 168:2373–82. https://doi.org/10.1534/genetics.104.031039 PMID: 15371362

33. Guillot G, Foll M. Correcting for ascertainment bias in the inference of population structure. Bioinformatics. 2009; 25:552–4. https://doi.org/10.1093/bioinformatics/btn665 PMID: 19136550.

34. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010; 27:2534–2547. https://doi.org/10.1093/molbev/msq148 PMID: 20558595

35. Quinto-Cortés CD, Woerner AE, Watkins JC, Hammer MF. Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. Sci Rep. 2018; 8:10209. https://doi.org/10.1038/s41598-018-28539-y PMID: 29977040

36. West B, Zhou B-X. Did chickens go north? New evidence for domestication. Journal of archaeological science. 1988; 15:515–33.

37. Lawal RA, Martin SH, Vanmechelen K, Vereijken A, Silva P, Al-Atiyat RM, et al. The wild species genome ancestry of domestic chickens. BMC Biology. 2020; 18:13. https://doi.org/10.1186/s12915-020-0738-1 PMID: 32050971

38. Tixier-Boichard M, Bed'hom B, Rognon X. Chicken domestication. From archeology to genomics. Comptes rendus biologies. 2011; 334:197–204. https://doi.org/10.1016/j.crvi.2010.12.012 PMID: 21377614

39. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al. Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. PLoS Genet. 2008; 4: e1000010. https://doi.org/10.1371/journal.pgen.1000010 PMID: 18454198

40. Crawford RD. Poultry genetic resources. evolution, diversity and conservation. In: Crawford RD, editor. Poultry breeding and genetics. 2nd ed. Amsterdam: Elsevier; 1993.

41. Qanbari S, Rubin C-J, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of adaptation in modern chicken. PLoS Genet. 2019; 15:e1007989. https://doi.org/10.1371/journal.pgen.1007989 PMID: 31034467

42. Reference Genome Gallus gallus 5.0. UCSC 2016 [cited 25 Oct 2016]. http://hgdownload.soe.ucsc.edu/goldenPath/galGal5/bigZips/galGal5.fa.gz.

43. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves TA, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3. 2017; 7:109–17. https://doi.org/10.1534/g3.116.035923 PMID: 27852011

44. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491. https://doi.org/10.1038/ng.806 PMID: 21478889

**45.** van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls. The Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43 PMID: 25431634.

**46.** Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [updated 16 Mar 2013]. http://arxiv.org/pdf/1303.3997v2.

**47.** Picard Tools 2.0.1. Broad Institute 2015. https://broadinstitute.github.io/picard/.

**48.** McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit. A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

**49.** ENSEMBL. Chicken Germline SNP and INDELS. 2016 [cited 6 Jan 2017]. http://e87.ensembl.org/Gallus_gallus.

**50.** Broad Institute. GATK User Guide. 2018 [cited 20 Mar 2018]. https://software.broadinstitute.org/gatk/documentation/.

**51.** Rocha D, Billerey C, Samson F, Boichard D, Boussaha M. Identification of the putative ancestral allele of bovine single-nucleotide polymorphisms. J Anim Breed Genet. 2014; 131:483–6. https://doi.org/10.1111/jbg.12095 PMID: 24862839

**52.** Kimura M. The neutral theory of molecular evolution. A review of recent evidence. Jpn J Genet. 1991; 66:367–86. https://doi.org/10.1266/jjg.66.367 PMID: 1954033

**53.** Nei M. Genetic Distance between Populations. The American Naturalist. 1972; 106:283–92.

**54.** Wright S. The genetical structure of populations. Ann Eugen. 1949; 15:323–54. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x PMID: 24540312

**55.** Paradis E, Claude J, Strimmer K. APE. Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004; 20:289–90. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327

**56.** Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, et al. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 2009; 19:510–9. https://doi.org/10.1101/gr.086538.108 PMID: 19088305

**57.** McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17:122. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

**58.** Lenth R. emmeans: Estimated Marginal Means, aka Least-Squares Means.; 2019.

**59.** R Core Team. R. A Language and Environment for Statistical Computing. Vienna, Austria; 2017.

**60.** Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015; 67:1–48. https://doi.org/10.18637/jss.v067.i01

**61.** Hiendleder S, Lewalski H, Janke A. Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. Cytogenet Genome Res. 2008; 120:150–6. https://doi.org/10.1159/000118756 PMID: 18467841

**62.** Heslot N, Rutkoski J, Poland JA, Jannink J-L, Sorrells ME. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One. 2013; 8:e74612. https://doi.org/10.1371/journal.pone.0074612 PMID: 24040295

**63.** Futschik A, Schlötterer C. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. Genetics. 2010; 186:207–18. https://doi.org/10.1534/genetics.110.114397 PMID: 20457880

**64.** Chen X, Listman JB, Slack FJ, Gelernter J, Zhao H. Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples. Genet Epidemiol. 2012; 36:549–60. https://doi.org/10.1002/gepi.21648 PMID: 22674656

**65.** Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals [mdash] mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014; 15:749–63. https://doi.org/10.1038/nrg3803 PMID: 25246196

**66.** Wang J, Skoog T, Einarsdottir E, Kaartokallio T, Laivuori H, Grauers A, et al. Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples. Sci Rep. 2016; 6:33256. https://doi.org/10.1038/srep33256 PMID: 27633116

**67.** Bradbury IR, Hubert S, Higgins B, Bowman S, Paterson IG, Snelgrove PVR, et al. Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, Gadus morhua. Mol Ecol Res. 2011; 11:218–25. https://doi.org/10.1111/j.1755-0998.2010.02949.x PMID: 21429176

**68.** Frascaroli E, Schrag TA, Melchinger AE. 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. Theoretical and Applied Genetics. 2013; 126:133–41. https://doi.org/10.1007/s00122-012-1968-6 PMID: 22945268

**69.** Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 2001; 157:1819–29. Available from: https://www.genetics.org/content/157/4/1819. PMID: 11290733

**70.** Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009; 10:381–91. https://doi.org/10.1038/nrg2575 PMID: 19448663

**71.** Qanbari S, Pausch H, Jansen S, Somel M, Strom T-M, Fries R, et al. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014; 10:e1004148. https://doi.org/10.1371/journal.pgen.1004148 PMID: 24586189.

**72.** Peripolli E, Munari DP, Silva M. V. G. B., Lima ALF, Irgang R, Baldi F. Runs of homozygosity: current knowledge and applications in livestock. Anim Genet. 2017; 48:255–71. https://doi.org/10.1111/age.12526 PMID: 27910110

**73.** Perez-Enciso M, Rincon JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol. 2015; 47:43. https://doi.org/10.1186/s12711-015-0117-5 PMID: 25956961.

**74.** Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity. 2014; 112:39–47. https://doi.org/10.1038/hdy.2013.13 PMID: 23549338.

**75.** Wainschtein P, Jain DP, Yengo L, Zheng Z, Cupples LA, Shadyab AH, et al. Recovery of trait heritability from whole genome sequence data. bioRxiv. 2019. https://doi.org/10.1101/588020

**76.** Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 2012; 8: e1002685. https://doi.org/10.1371/journal.pgen.1002685 PMID: 22570636.

**77.** Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011; 6:e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248

**78.** Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016; 17:81. https://doi.org/10.1038/nrg.2015.28 PMID: 26729255