How Array Design Affects SNP Ascertainment Bias

Johannes Geibel[*†], Christian Reimer[*†], Steffen Weigend[†‡], Annett Weigend[‡], Torsten Pook[*†] & Henner Simianer[*†]

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

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

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

Data Availability: Additional to the sequences already published by Malomane *et al.* (2018) or Qanbari *et al.* (2019), sequencing reads of *Gallus varius* (GV), Rhode Island Red (RI) and White Rock (WR) pools were uploaded to European Nucleotide Sequence Database (ENA; https://www.ebi.ac.uk/ena/data/view/PRJEB31780) and will made publically available upon acceptance of the paper. The study accession numbers of former and newly submitted reads can be found in supplementary file 1. Note that Malomane *et al.* (2018) already published a subset of the individually sequenced white layers, published in total by Qanbari *et al.* (2019). Supplementary material can be found on figshare.

Array Design and SNP Ascertainment Bias

KEYWORDS

ascertainment bias; array design; whole genome re-sequencing; chicken

CORRESPONDING AUTHOR

Johannes Geibel
Albrecht-Thaer-Weg 3
37075 Göttingen, Germany
+49 551 3925609
johannes.geibel@uni-goettingen.de

ABSTRACT

Single nucleotide polymorphisms (SNPs), genotyped with SNP arrays, have become the most widely used marker types 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 with main influencing factors being the identification of SNPs in a limited set of discovery populations and a within-line selection of 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 28 % higher for populations involved in the discovery process than not involved populations for the original array. The same was observed after the SNP discovery step in the redesign. However, an unequal contribution of populations in the equal spacing process 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.

54 **INTRODUCTION**

55 Starting in the first decade of this century, the possibility of cost effectively genotyping
56 high numbers of Single Nucleotide Polymorphisms (SNP) for many individuals in parallel via
57 SNP arrays led to an increased usage of them for population genetic analyses in humans
58 (Novembre *et al.* 2008; Patterson *et al.* 2012), model species (Laurie *et al.* 2007; Platt *et al.*
59 2010), plants (Travis *et al.* 2015; Mayer *et al.* 2017) and livestock (Muir *et al.* 2008; Gibbs *et*
60 *al.* 2009; Kijas *et al.* 2009; Gautier *et al.* 2010; Qanbari *et al.* 2010; McTavish *et al.* 2013;
61 Malomane *et al.* 2019).

62 Various SNP arrays exist for humans (Perkel 2008), plants (Unterseer *et al.* 2014; Singh *et*
63 *al.* 2015) and all major livestock species (Matukumalli *et al.* 2009; Ramos *et al.* 2009;
64 Groenen *et al.* 2011; Boichard *et al.* 2012; Kranis *et al.* 2013; Tosser-Klopp *et al.* 2014; San-
65 denbergh *et al.* 2016). SNP numbers within these arrays range from 10 k SNPs (Boichard *et*
66 *al.* 2012) over approximately 50 k (Groenen *et al.* 2011; Matukumalli *et al.* 2009; Tosser-
67 Klopp *et al.* 2014; Singh *et al.* 2015) up to 600 k (Kranis *et al.* 2013; Matukumalli *et al.* 2009;
68 Unterseer *et al.* 2014). The design process of every array has an initial step of SNP discovery
69 in common, where SNPs are identified from existing databases and/or from a small set of
70 sequenced individuals. SNPs are then selected based on different quality criteria like minor
71 allele frequency (MAF) thresholds and platform specific design scores (Fan *et al.* 2010). Addi-
72 tional criteria like equidistant spacing over the genome (Kranis *et al.* 2013), overrepresenta-
73 tion of some areas (Boichard *et al.* 2012) or increased overrepresentation of high MAF SNPs
74 (Matukumalli *et al.* 2009) are applied dependent on the design intentions. In the end, draft
75 arrays are validated either on the discovery set itself (Ramos *et al.* 2009) and/or on a broad
76 set of individuals from different populations (Kranis *et al.* 2013; Fan *et al.* 2010).

77 In contrast to whole genome re-sequencing (WGS) data, SNP arrays often show a clear
78 underrepresentation of SNP with extreme allele frequencies (Nielsen 2004). As population
79 genetic statistics are mostly based on estimates of allele frequencies, this context leads to
80 biased population genetic estimators (Nielsen 2004) and is known as SNP ascertainment bi-
81 as.

82 The absence of rare alleles is mainly driven by two factors in the array design process
83 where SNPs are selected (ascertained) based on different requirements and decisions (Eller
84 2001). The first factor is a relatively small panel of individuals used for discovery of SNPs,
85 leading to a large proportion of globally rare variants not being selected, since they appear
86 monomorphic in the discovery panel (Nielsen and Signorovitch 2003; Clark *et al.* 2005). The
87 second factor is the across population use of arrays. Arrays are developed based on the vari-
88 ation within the discovery panel, thus missing variation present in distantly related individu-
89 als or populations (Nielsen 2004; Eller 2001). This second source of bias was shown to be of
90 relatively high importance for livestock studies, where arrays are usually developed for large
91 commercial breeds and later used to genotype diverse sets of local breeds all over the world
92 (McTavish and Hillis 2015; Malomane *et al.* 2018).

93 Besides different strategies to minimize the impact of ascertainment bias (Malomane *et*
94 *al.* 2018; Lachance and Tishkoff 2013), there are some attempts to correct the allele fre-
95 quency spectrum via Bayesian methods (Nielsen and Signorovitch 2003; Nielsen *et al.* 2004;
96 Nielsen 2004). However, those corrections highly rely on detailed statistical assumptions of
97 the ascertainment process (Guillot and Foll 2009; Albrechtsen *et al.* 2010) and are currently
98 only tested for corrections of the first source of ascertainment bias, the small discovery pan-
99 el (Nielsen 2004; Nielsen *et al.* 2004; Nielsen and Signorovitch 2003). In practice, detailed

100 information on the design process is limited (Albrechtsen *et al.* 2010) and the complexity of
101 the processes makes statistical models for the corrections inaccurate.

102 Given the complexity of modern array design processes, this study aims at highlighting
103 the mechanisms which promote the bias by illustrating the effects of the different steps of
104 the array design process on ascertainment bias, using real data in a typical setting from live-
105 stock sciences. For this purpose, the design process of the Axiom™ Genome-Wide Chicken
106 Array (www.affymetrix.com) was remodeled in a set of diverse chicken WGS data. Allele fre-
107 quency spectra as well as expected heterozygosity (H_{exp}) were compared to the WGS data
108 and the SNPs of the Axiom™ Genome-Wide Chicken Array. Finally, some recommendations
109 are made on how to design an array for monitoring genetic diversity.

110 **MATERIAL AND METHODS**

111 **Populations and sequencing:** The analysis is based on WGS data of a diverse set of 46 com-
112 mercial, non-commercial and wild chicken populations, sampled within the framework of the
113 projects AVIANDIV (www.aviandiv.fli.de) and SYNBREED (www.synbreed.tum.de). Commer-
114 cial brown (BL) and white layer (WL) populations consist of 25 individually re-sequenced an-
115 imals each, while the two commercial broiler lines (BR1 and BR2) include 20 individually se-
116 quenced animals each. For 41 populations, pooled DNA from 9 - 11 animals per population
117 was sequenced, while *Gallus varius* (GV) samples of only two animals were sequenced as a
118 pool. More detailed information about the samples can be found in supplementary file 1 and
119 two previously published papers, Malomane *et al.* (2018) and Qanbari *et al.* (2019). Cover-
120 age ranged between 7X and 10X for the individual sequences, while DNA pools were se-
121 quenced with 15X to 70X coverage. Paired-end sequencing was conducted on Illumina HiSeq

122 machines at the Helmholtz Zentrum, German Research Center for Environmental Health in
123 Munich, Germany.

124 **Raw data preparation and SNP calling:** Sequences were aligned to the reference genome
125 Gallus gallus 5.0 (UCSC 2016; Warren *et al.* 2017) and the SNP calling was conducted accord-
126 ing to GATK Best Practices guidelines (DePristo *et al.* 2011; van der Auwera *et al.* 2013). BWA
127 0.7.12 (Li 2013) was used for the alignment step, duplicates were marked using Picard
128 Tools 2.0.1 (Broad Institute 2015) MarkDuplicatesWithMateCigar and base qualities were
129 recalibrated with GATK 3.7 (McKenna *et al.* 2010) BaseQualityRecalibrator. The set of known
130 SNPs, necessary for base quality score recalibration, was downloaded from ENSEMBL release
131 87 (ENSEMBL 2016). SNPs were called for all samples separately using the GATK 3.7 Haplo-
132 typeCaller and later on simultaneously genotyped across samples with GATK 3.7 Geno-
133 typeGVCFs. Due to computational limitations, a ploidy of two instead of the higher true
134 ploidy of the pooled sequences had to be assumed, which led to calling slightly less rare al-
135 leles. However, effects of this limitation are negligible (supplementary file 2; Figure S 1).

136 SNP filtering was conducted using GATK 3.7 VariantRecalibrator, which filtered the called
137 SNPs by a machine learning approach (use of a Gaussian mixture model), which uses both, a
138 set of previous known (low confidence needed) and a set of highly reliable (assumed to be
139 true) variants as training resources (Broad Institute 2018). The source for known SNPs (pri-
140 or 2) provided to VariantRecalibrator was again ENSEMBL (release 87) and the SNPs of the
141 Axiom™ Genome-Wide Chicken Array were defined as true training set (prior 15). The algo-
142 rithm was trained on the quality parameters DP, QD, FS, SOR, MQ and MQRankSum. Filters
143 were set to recover 99 % of the training SNPs in the filtered set, which resulted in an Transi-

144 tion/Transversion ratio of 2.52 for known SNPs, and an Transition/Transversion ratio of 2.26
145 for novel SNPs. Only biallelic autosomal SNPs were used in all further analyses.

146 **Identification of the ancestral allele:** Ancestral alleles were defined using allele frequency
147 information from the three wild populations *Gallus gallus gallus* (GG), *Gallus gallus spadi-*
148 *ceus* (GS) and *Gallus varius* (GV) by an approach comparable to Rocha *et al.* (2014). It was
149 assumed that the *Gallus gallus* and *Galllus varius* species emerged from a common ancestor
150 and *Gallus gallus* later split into *Gallus gallus gallus* and *Gallus gallus spadiceus* subspecies.
151 Additionally, assuming neutral molecular evolution (Kimura 1991), the ancestral allele was
152 most likely the major allele within those three populations, when weighting the allele fre-
153 quency of *Gallus varius* twice. This procedure assigned the ancestral status to the reference
154 allele for 86 % of the SNPs and to the alternative allele for 14 % of the SNPs, causing close to
155 no change in the allele frequency spectrum (supplementary file 2; Figure S 2).

156 **Reference Sets:** Three different reference sets were defined as follows: the **unfiltered WGS**
157 SNPs (28.5 M SNPs), SNPs filtered using GATK 3.7 (McKenna *et al.* 2010) VariantRecalibrator
158 (20.9 M SNPs; **filtered WGS**) and **array SNPs** (540 k SNPs), which are the intersection of the
159 unfiltered SNPs and the SNPs of the Axiom™ Genome-Wide Chicken Array. The separate use
160 of unfiltered and filtered WGS SNPs was done to assess the effect of filtering (especially the
161 use of an ascertained SNP set as the true set) on ascertainment bias.

162 **Redesigning the SNP Array:** The populations were divided into four groups for the array de-
163 sign. Eight populations were chosen for SNP discovery (Figure 1): two non-commercial popu-
164 lations (LE, NH), one commercial white (WL) and three commercial brown layer lines (BL, RI,
165 WR), and two commercial broiler lines (BR1 and BR2). Further, the discovery set had to be

split into broilers (BR1, BR2, NH) and layers (WL, LE, BL, RI, WR) for the equal spacing step. Even though the New Hampshire (NH) population is a dual purpose breed, it was chosen as broiler as it is assumed that the breed has contributed to founders of contemporary commercial broiler lines and shows the smallest genetic distances (Nei 1972) to the two broiler populations in this study. From the remaining populations, 19 were randomly chosen for SNP validation, 18 populations were used for an application of the array, and *Gallus varius* as a different species was defined as outgroup.

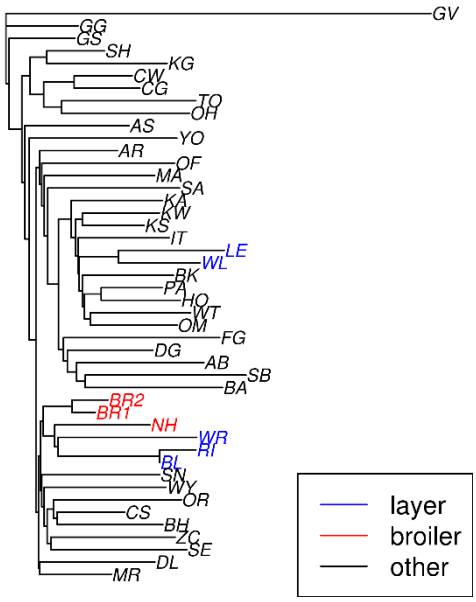


Figure 1: Neighbor Joining phylogenetic tree based on pairwise Nei's standard genetic distances (Nei 1972). The tree was calculated from the filtered WGS SNPs and rooted by using Gallus varius (GV) as outgroup. Layer and broiler populations which form in total the discovery set for the array design according to Kranis *et al.* (2013) are highlighted. The plot was produced using the R package ape (Paradis *et al.* 2004)and the abbreviations for the populations can be found in supplementary file 1.

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

185 1) **SNP discovery → 10.8 M SNPs**
186 Discovery of SNPs fulfilling basic criteria (quality ≥ 60; MAF ≥ 0.05; cover-
187 age ≤ mean + three standard deviations) within the discovery populations.

188 2) **Cluster removal → 8.8 M SNPs**
189 SNP clusters with less than 4 bp invariant sites at one side of a SNP (no matter of
190 whether 3' or 5' direction) and less than 10 bp fixation at the other side of the SNP
191 were removed.

192 3) **Equal spacing → 2.3 M SNPs**
193 Reduction of SNPs to achieve approximately equidistant spacing between SNPs
194 within discovery populations based on genetic distances. The used algorithm was
195 modeled according to Kranis *et al.* (2013) and initially sets up a backbone of SNPs
196 segregating either in all layer populations or in all broiler populations and then fills
197 up the gaps while aiming for a predefined local target density of 667 equidistant
198 segregating SNPs/cM (linkage map taken from Groenen *et al.* (2011)) within each
199 population.

200 4) **SNP validation → 1.9 M SNPs**
201 Removing SNPs (~ 20 %) which were not variable in at least 8 of the 19 validation
202 populations.

203 5) **Downsampling → 600 k SNPs**
204 Downsampling of SNPs comparable to step 3, but without any initial backbone.

205 Keeping all exonic SNPs (annotation using Ensembl VEP 89.7; McLaren *et al.* 2016)
206 and setting the target density in broiler lines as three times the target density of
207 the layer lines.

208 **Variation of the design process:** The whole design process was repeated 50 times with pop-
209 ulations being randomly assigned to be discovery, validation or application populations,
210 while the *Gallus varius* population was always kept as the outgroup.

211 To assess the impact of the number of discovery populations on the design process, the
212 number of discovery populations was varied in additional runs from 4 to 40 randomly chosen
213 populations (while assigning the remaining populations, except *Gallus varius,* to validation
214 and application groups of equal size) with 20 random replicates for each number of discov-
215 ery populations. In a last scenario, equal spacing was varied with respect to the target densi-
216 ty (33 – 3333 SNPs/cM) with 20 independent population groupings for each target density,
217 with or without the initial backbone. As the number of SNPs from the backbone was con-
218 stant, the increase of target density led to a higher number of SNPs chosen by the algorithm
219 due to the equal spacing itself and hence the relative influence of the fixed number of com-
220 mon backbone SNPs decreased.

221 **Analyses of the results:** Per-locus-allele frequencies for individually sequenced populations
222 were estimated from genotypes, whereas the estimation for the sequenced DNA-pools was
223 based on the allelic depth. Influences on the allele frequency spectra were examined by
224 comparing density estimates of derived allele frequency spectra (unfolded frequency spec-
225 trum). Further H_{exp}, the expected heterozygosity assuming Hardy Weinberg frequencies of
226 the genotypes, for the different populations were used as summary statistics of the within

227 population allele frequency spectra and calculated as in equation (1), where $p_{ref;l}$ denotes

228 the frequency of the reference allele at locus l and L the total number of loci.

229 $$H_{\exp} = \frac{\sum_{l} 2 p_{ref;l}(1 - p_{ref;l})}{L} \quad (1)$$

230 Deviations in the estimation of H_{exp} from the various SNP sets were quantified as differ-

231 ences between the H_{exp} calculated from the respective SNP set and the H_{exp} calculated from

232 the filtered WGS SNPs relative to the H_{exp} from the filtered WGS SNPs, further called overes-

233 timation of H_{exp} (OHE; equation (2)). An OHE of zero means that the estimates are equal,

234 while an OHE of one describes doubling of the unbiased estimate.

235 $$OHE = \frac{H_{\text{exp; SNP set}} - H_{\text{exp; filtered WGS SNPs}}}{H_{\text{exp; filtered WGS SNPs}}} \quad (2)$$

236 The effects of the population group assignments on the OHE of the random population

237 assignments were evaluated by pairwise comparisons of least square means (LSMEANS; cal-

238 culated with the R package emmeans (Lenth 2019; R Core Team 2017) by using Tukey cor-

239 rection for multiple pairwise contrasts) of the population groups. An underlying mixed linear

240 model for the estimation of LSMEANS was fitted using the R package lme4 (Bates *et al.* 2015)

241 as shown in equation (3), where the OHE depended on an overall mean μ, the fixed effect

242 of the population group $popG_i$ (i can be discovery-, validation-, application- or outgroup), a

243 random effect for the jth repetition of random population grouping ($rep_j \sim N(0, I\sigma^2_{rep})$) and

244 a random error $e_{ijk} \sim N(0, I\sigma^2_e)$. The procedure is comparable to simple pairwise compari-

245 sons of group means, the correction by the repetition only reduces the error variance and

246 thus decreases the confidence intervals.

$$OHE_{ijk} = \mu + popG_i + rep_j + e_{ijk} \quad (3)$$

Data Availability: Additional to the sequences already published by Malomane *et al.* (2018) or Qanbari *et al.* (2019), sequencing reads of *Gallus varius* (GV), Rhode Island Red (RI) and White Rock (WR) pools were uploaded to 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 supplementary file 1. Note that Malomane *et al.* (2018) already published a subset of the individually sequenced white layers, published in total by Qanbari *et al.* (2019). Supplementary material can be found on figshare.

257 **RESULTS**

258 **Numbers of SNPs:** The SNP calling identified 28.5 M biallelic autosomal SNPs from which
259 20.9 M SNPs passed GATK's filtering procedure. 540 k SNPs from the unfiltered WGS SNP set
260 are also mapped on the original Axiom™ Genome-Wide 580 k Chicken Array. The remodeling
261 of the array according to Kranis *et al.* (2013) returned 10.8 M SNPs from the discovery step,
262 which were reduced to approximately 600 k in steps as described. Numbers of identified
263 SNPs for the additional runs differed depending on the populations and settings used and
264 are listed in Table S 1. Additional information about the identified number of SNPs depend-
265 ing on the number of discovery populations and target density as well as information about
266 the share of SNPs of different random runs can be found from Figure S 3 – Figure S 5.

267 **Underrepresentation of rare SNPs:** A clear underrepresentation of rare SNPs in all ascer-
268 tained SNP sets compared to WGS is evident from the allele frequency spectra (Figure 2).
269 Major changes in the allele frequency spectra during the array development process were
270 especially observed after the SNP discovery step and the equal spacing step. The SNP discov-
271 ery led to an underrepresentation of rare SNPs compared to sequence data, which was in-
272 tensified by the equal spacing step (Figure 2) and resulted finally in a spectrum close to the
273 spectrum of the original array. Randomly choosing populations as discovery populations con-
274 firmed the shape of the first remodeling, where the population groups were chosen accord-
275 ing to Kranis *et al.* (2013). As major changes in the spectra mainly occurred after the SNP
276 discovery and equal spacing, further results will concentrate on those steps.

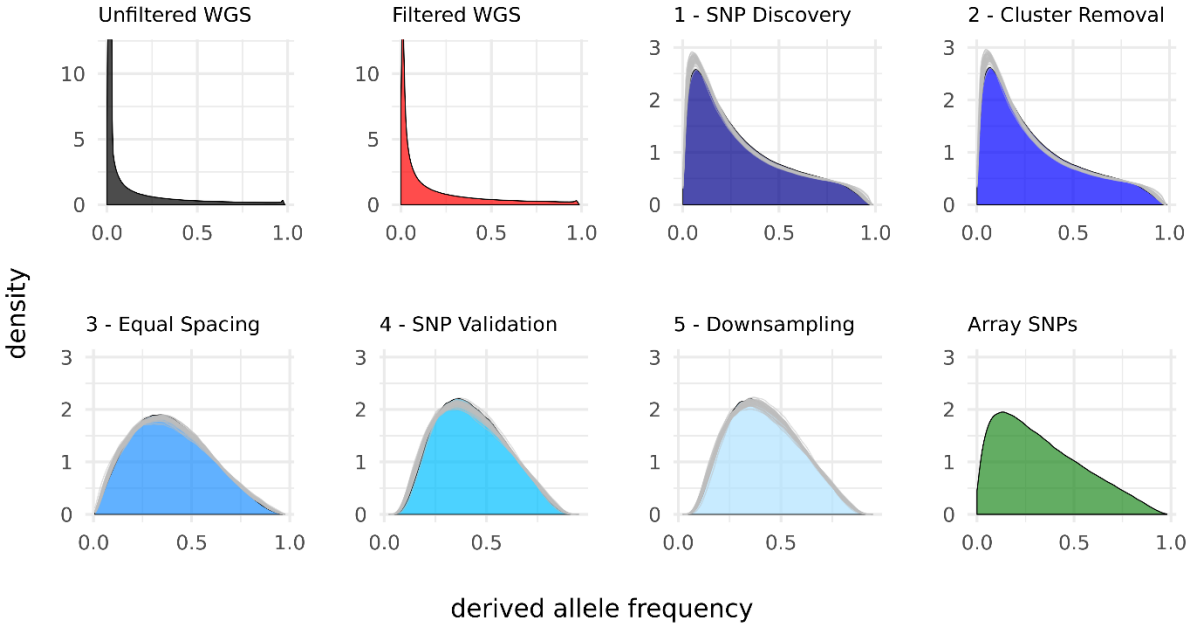


Figure 2: Derived allele frequency spectra for the different SNP sets. For the remodeled sets, areas show the modelling according to Kranis *et al.* (2013) while grey lines represent the 50 random population groupings.

The allele frequency spectra (Figure 3) within discovery populations, compared to the spectra over all populations, clearly showed the cutoff from the MAF 0.05 filter. After the equal spacing, they additionally revealed a higher share of common SNP than the overall spectra. 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.

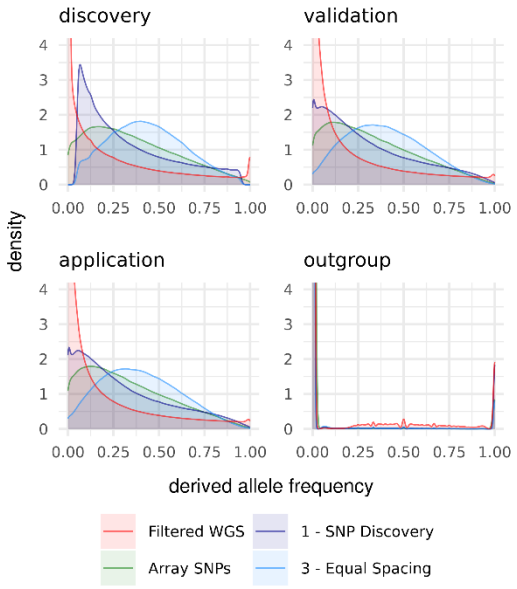


Figure 3: Derived allele frequency spectra within the population groups.

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 4). 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.

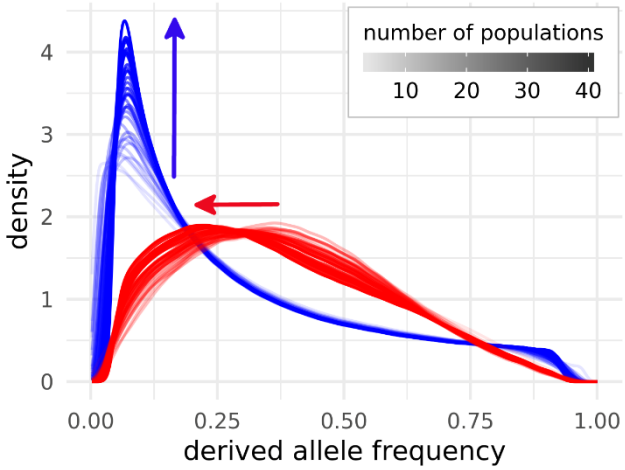


Figure 4: Derived allele frequency spectra after SNP discovery (blue) and equal spacing (red) for varying numbers of populations in the discovery set. Different numbers of populations in the discovery set (4 to 40) are indicated by an intensifying color gradient and only one randomly picked run per population number is shown. 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 5). 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 5).

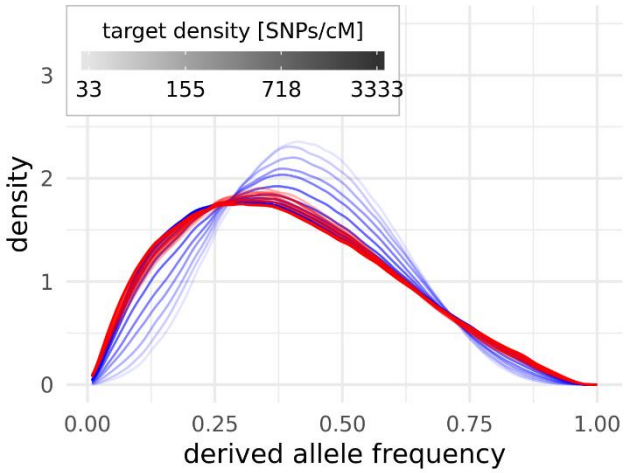


Figure 5: Derived allele frequency spectra for different target densities [SNPs/cM] after the equal spacing step. Blue spectra show the runs with the initial backbone included, red ones without backbone. Different target densities are indicated by a log scaled increase in color intensity. Only one representative run per target density is shown.

Overestimation of H_{exp}: Figure 6 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 over-estimation 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 (≥ 0.94; Table S 2). 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 from the original array SNPs (1.41; Figure 6; Table 1) lay in the range covered by the remodeling steps.

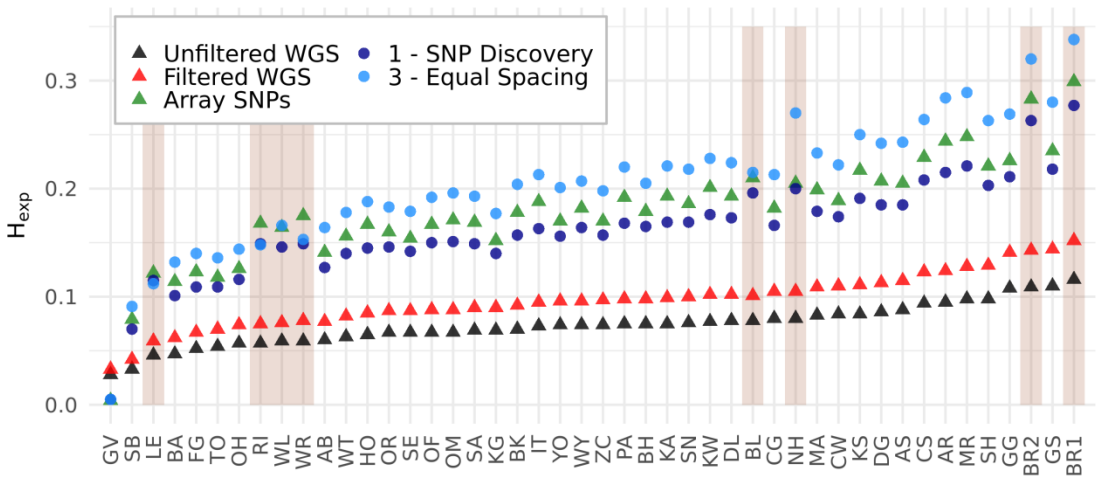


Figure 6: Expected Heterozygosity (H_{exp})by population and SNP set. Populations are ordered by the H_{exp} of the unfiltered WGS SNP set. Only the reference sets and relevant steps of the array design are shown. Discovery populations are shaded with a darker background.

343

344 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 – down-sampling
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	1.08 ± 0.12	0.91 ± 0.05	0.90 ± 0.05	1.15 ± 0.22	1.26 ± 0.21	1.28 ± 0.13
Validation	-0.23 ± 0.00	0.84 ± 0.11	0.65 ± 0.07	0.65 ± 0.07	1.13 ± 0.11	1.38 ± 0.13	1.35 ± 0.14
Application	-0.23 ± 0.00	0.83 ± 0.11	0.65 ± 0.07	0.65 ± 0.07	1.13 ± 0.10	1.32 ± 0.12	1.30 ± 0.13
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}

345

346 Averaging the OHE within the population groups revealed a 40 % higher OHE of the dis-
347 covery populations compared to validation and application populations after the discovery
348 step. The equal spacing step reduced this difference to an only 1 % larger OHE for discovery
349 populations, while it came with a substantial increase of the variance of OHE, which was
350 larger for the discovery populations than validation and application populations. The valida-
351 tion step then increased the OHE of the validation populations more than the OHE of discov-
352 ery and application populations. This stronger OHE of discovery populations was also appar-
353 ent within the array SNPs (28 % higher). In contrast to the other populations, the outgroup
354 showed an underestimation of the H_{exp}, resulting in an OHE of < -0.84 for all ascertained SNP
355 sets (Figure 6; Table 1).

356 A closer look on the contribution of the sub-steps during the equal spacing step revealed
357 that 65 % of the SNPs which were preserved during equal spacing were variable in all of the
358 three closely related broiler populations (BR1, BR2, NH; maximum pairwise Nei's distances
359 (Nei 1972) of 0.07 in the unfiltered SNP set), while only 1.2 % of the SNPs were retained due
360 to being variable in all of the five less closely related layer populations (WL, LE, WR, BL, RI;
361 maximum pairwise Nei's distances of 0.15 in the unfiltered SNP set). The first population

362 used to fill in the gaps in the backbone (WL) contributed 17 % of the SNPs, while the other

363 populations contributed < 5 %.

364

365 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

366

367 These findings were supported by the 50 random groupings (Figure 7). The LSMEANS (Ta-

368 ble 2) of the population groups revealed 24 % larger OHE for discovery populations than for

369 validation and application populations after discovery and cluster removal step, which was

370 decreased to a numerically insignificant difference after the equal spacing step. Interestingly,

371 and in contrast to the findings from the first remodeling, SNP validation led to a significantly

372 higher OHE (5 % larger) for application populations than discovery and validation popula-

373 tions.

374

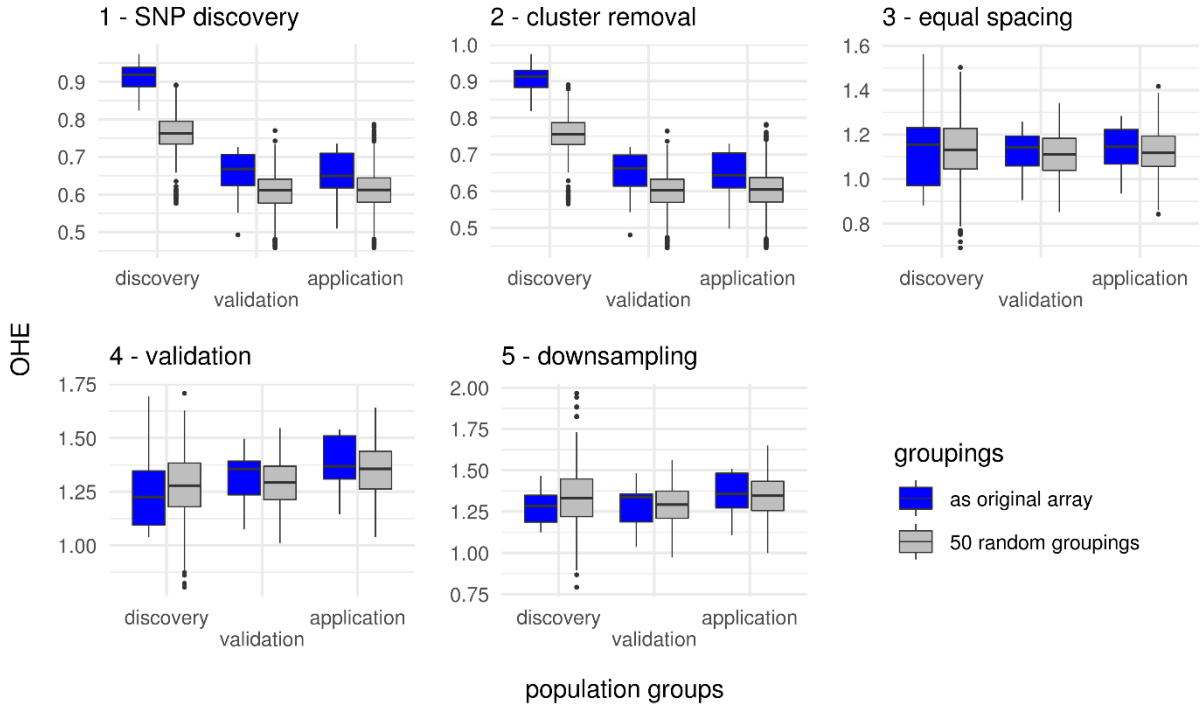
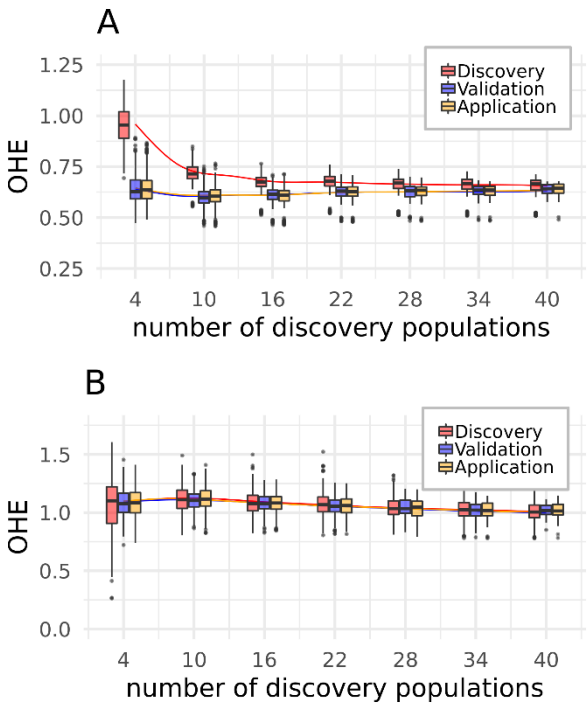


Figure 7: OHE for discovery validation and application populations after the five steps of array design. Discovery populations are chosen to represent populations which are comparable to the original array (blue) or 50 times randomly sampled (grey.

380 **Influence of number of discovery populations and target density on H_{exp}:** Figure 8 A shows
381 that increasing the number of discovery populations reduces the mean OHE of discovery
382 populations after SNP discovery while not affecting the OHE of validation and application
383 populations. After equal spacing (Figure 8 B), no dependency of OHE on the number of dis-
384 covery populations remained. As, due to the limited number of populations in the complete
385 set, the number of validation populations decreased with more populations in the discovery
386 set, the individual validation populations gained more impact on the ascertainment. This led
387 to a higher OHE of validation populations with a high number of discovery populations (Fig-
388 ure S 6 D), comparable to the higher OHE of discovery populations for a small number of
389 discovery populations.



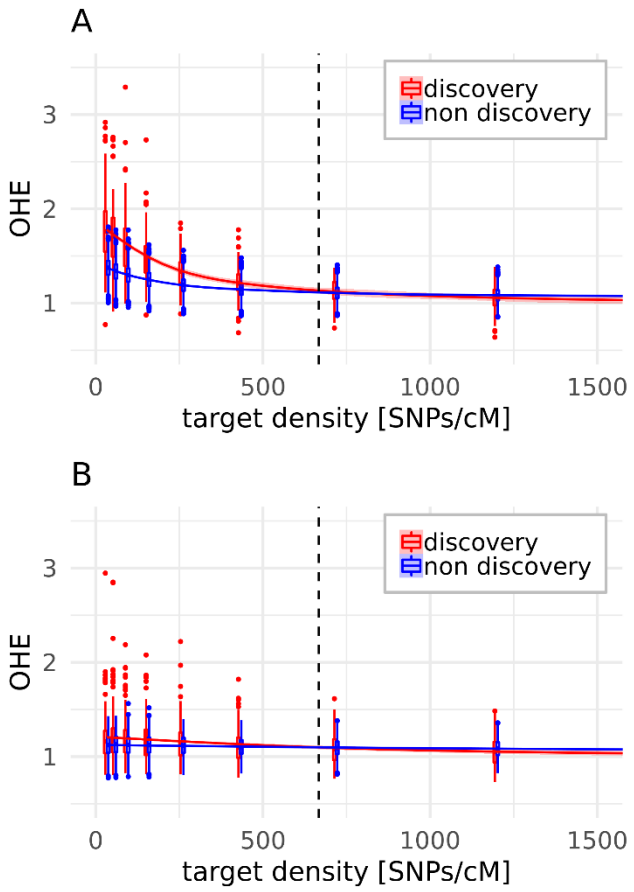
390

391 Figure 8: Relation of the OHE as a function of the number of discovery populations.
392 **A** - discovery, **B**– equal spacing. The boxplots are only shown for a subset of the number of
393 discovery populations, while the smoothing lines, which show the trend, are calculated from
394 all observations. Plots for all five steps can be found in Figure S 6.

395

396 In the equal spacing step, using only backbone SNPs resulted in a higher OHE for discov-
397 ery than for non- discovery populations. Increasing the target density and thus increasing the
398 proportion of SNPs due to the equal spacing part of the algorithm reduced the difference in
399 OHE between the population groups (Figure 9 A). If the SNPs from the initial backbone were
400 not used, no difference of OHE between discovery and non- discovery populations was pre-
401 sent, regardless of the target density (Figure 9 B).

402



403

404 Figure 9: OHE after equal spacing (step 3) by target density in SNPs/cM and population
405 group. The smoothing lines show the trend and the dashed lines the target density of 667
406 SNPs/cM, used for the remodeling according to Kranis *et al.* (2013). The algorithm was run
407 including the initial backbone SNPs (**A**) or not including them (**B**). Gallus varius is not includ-
408 ed, as it is constantly underestimated.

409 **DISCUSSION**

410 In this study we used a uniquely diverse collection of sequenced wild, commercial and
411 non-commercial chicken populations, mainly based on samples of the Synbreed Chicken Di-
412 versity Panel (Malomane *et al.* 2019). Parts of our set were also involved in the development
413 process of the Axiom™ Genome-Wide 580 k Chicken Array (Kranis *et al.* 2013). This offered
414 an excellent possibility for assessing the impact of ascertainment bias on real data in a com-
415 plex scenario.

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

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

431 Additionally, due to computational limitations, we had to assume a ploidy of two for
432 pools during the SNP calling process, which resulted in a minimal reduction of rare alleles.
433 However, this effect was shown to have a very minor impact on the findings of this study
434 (supplementary file 2).

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

445 The discovery step had the strongest impact on discovery populations, when a small set
446 of discovery populations was used (Figure 8 A). Similarly, the influence of the validation step
447 on validation populations was strongest in case of a small number of validation populations
448 (Figure S 6D). A balancing of these two groups of samples is therefore necessary, if the num-
449 ber of available DNA samples for array development is limited. We rather suggest the use of
450 an extended discovery set for validation as best practice instead of using completely differ-
451 ent populations.

452 If the equal spacing step contains a preselection of SNPs based on their variability within
453 population groups, the bias is stronger towards high MAF SNPs and thus yields a higher OHE.

454 This effect was reduced by increasing the target density and thus selecting relatively more
455 SNPs due to the equal spacing instead of common occurrence.

456 **Differences between groups:** If allele frequency spectra are changed in the same way for all
457 populations and are therefore biasing heterozygosity estimates to the same amount, find-
458 ings for between population comparisons will be little affected. Ascertainment bias then is
459 only of importance if one compares populations based on different arrays, and corrections
460 of the allele frequency spectrum as reviewed by Nielsen (2004) should be possible. As corre-
461 lations between H_{exp} of ascertained SNP sets and unfiltered/ filtered WGS SNP sets were
462 consistently high (> 0.94), arrays designed in the way as performed in this paper should
463 mostly be suitable for robust and cost efficient analyses. Unbiasedness of estimates could be
464 increased even more by considering filter strategies according to Malomane *et al.* (2018).

465 However, we could show that a structural bias due to the second source of ascertainment
466 bias, the across population use of arrays, exists. This structural bias was already shown to
467 have severe impact on findings from SNP arrays. For example Bradbury *et al.* (2011) found a
468 demographic decline up to an approximately 30 % lower H_{exp} for Atlantic cod based on the
469 distance to the sampling location of the discovery panel and McTavish and Hillis (2015)
470 showed strong deviations between simulated and observed polymorphisms for different
471 combinations of migration- and ascertainment scenarios on simulated cattle populations. In
472 our study, populations which are closely related to the discovery populations of the original
473 array on average showed a 28 % higher OHE than validation and application populations for
474 the original array, which supports the mentioned findings from literature.

475 This structural bias was mainly introduced by the initial discovery step. It was also ob-
476 served for the random population groupings, but to a slightly smaller extent. The difference

477 in overestimation decreased with an increase in the number of discovery populations (Figure
478 8) and was smallest if the discovery populations showed minimum distance to the applica-
479 tion and validation populations (results not shown). Comparable observations were already
480 made by Frascaroli *et al.* (2013) who found very small ascertainment bias for European elite
481 maize lines when using a SNP panel discovered in a very diverse set of maize and inbred
482 lines, but strong ascertainment bias when using SNP discovered in American elite lines.
483 Therefore, we suggest to ideally choose an array where the discovery panel does span the
484 scope of populations it will be applied to, and by this covers the existing variation in a most
485 representative way, or to design such an array for oneself if it does not exist.

486 The equal spacing step removed the structural bias in most of our remodeling scenarios,
487 but obviously not in case of the original array. In the remodeling, we saw this structural bias
488 only with a low target density and thus calling SNPs in the equal spacing step mainly due to
489 being common over many populations (Figure 9 A). However, the equal spacing step also
490 increased the variance of OHE in the discovery panel, meaning that the OHE was increased
491 more for some of the discovery populations than for others, causing more uncertainty for
492 resulting effects. This effect is driven by the unequal contribution of populations during the
493 equal spacing step. It increases the bias for some of the discovery populations, while it de-
494 creases it for others, and hence it does not remove the structural bias. This means that the
495 knowledge of which discovery populations were used alone is not sufficient to draw conclu-
496 sions regarding a possible ascertainment bias, since their relative contribution varies through
497 the described pipeline.

498 **Outgroup:** *Gallus varius* as an outgroup showed a different behavior than all other popula-
499 tions. It already exhibited the lowest H_{exp} in the unfiltered WGS SNP set, which was most

500 likely driven by the small number of only two samples in the pool, and showed less upward
501 bias of H_{exp} in the filtered WGS SNP set than all other populations. The *Gallus varius* se-
502 quence reads on average showed weak Phred-scaled mapping quality scores of 19 (1.3 %
503 probability of misalignment), while the mean quality scores of the other populations ranged
504 from 25 (0.3 %) to 28 (0.1 %) – meaning that variation, only present in *Gallus varius*, will be
505 more likely missed due to misplacement of the reads or discarded as possible sequencing
506 errors. Additionally, every ascertained SNP set showed an OHE for *Gallus varius* of < -0.84, as
507 variation being present only in *Gallus varius* was not found in *Gallus gallus* discovery panels
508 and contrary variants from *Gallus gallus* were not variable in *Gallus varius* (Figure 3). This
509 demonstrates that arrays should not be used if different species (even closely related ones)
510 are included in the research project. Even sequence based estimates can be slightly biased, if
511 the reference genome does not fit properly.

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

522 For the purpose of monitoring genetic diversity in a large set of small non-commercial
523 populations, the development of a specialized new array for high throughput cost effective
524 genotyping could be still a good option. For the design of such an array, it is crucial to extend
525 the discovery set in a way which represents the total variability over all populations as bal-
526 anced as possible. The use of publicly available sequences can be helpful to reach this goal.
527 The ascertainment of the SNPs should then be done preferably over all available populations
528 instead of biasing the process towards common alleles by performing within population as-
529 certainment.
530 However, when considering the findings of this study, choosing an array for which the
531 SNP ascertainment process is well documented and a discovery panel that matches the pop-
532 ulations to be studied, should give sufficient results when robust estimators are used.
533

534 **ABBREVIATIONS**

535 H_{exp} – expected heterozygosity

536 MAF – minor allele frequency

537 OHE – overestimation of expected heterozygosity

538 SNP – single nucleotide polymorphism

539 WGS – whole genome re-sequencing

540 **DECLARATIONS**

541 **Ethics approval and consent to participate**

542 Sampling was approved by the animal welfare committees of the involved institutions
543 and was done in strict accordance to the German Animal Welfare regulations and with writ-
544 ten consent of the animal owners. In case of the samples which were taken during the SYN-
545 BREED project, according to a change in German regulations, notice was additionally given to
546 the Lower Saxony State Office for Consumer Protection and Food Safety (33.9-42502-05-
547 10A064).

548 **Funding**

549 The sampling and sequencing of the chicken populations was funded by the EC project
550 AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; www.aviandiv.fli.de) and the
551 German Federal Ministry of Education and Research (BMBF) via the SYNBREED project (FKZ
552 0315528E; 2009 – 2014; www.synbreed.tum.de).

553 **Authors' contributions**

554 JG designed the study, analyzed and interpreted the data, wrote the final draft and pre-
555 pared the manuscript. HS supervised the study, substantially contributed to design and con-
556 ception of the study as well as revision and editing of the manuscript. CR substantially con-
557 tributed to design and conception of the study, participated in interpreting the data and was
558 involved in revision and editing of the manuscript. SW contributed to design and conception
559 of the study, participated in the provision of the data as well as in revision and editing of the
560 manuscript. AW participated in provision and preparation of the data. TP wrote a part of the
561 equal spacing algorithm and participated in revision and editing of the manuscript. All au-
562 thors have read and approved the final manuscript.

563 **Acknowledgments**

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

565 **REFERENCES**

566 Albrechtsen, A., F. C. Nielsen, and R. Nielsen, 2010 Ascertainment biases in SNP chips affect
567 measures of population divergence. Mol Biol Evol 27 (11): 2534–2547.
568 Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe, 2016 Harnessing
569 the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17 (2): 81.
570 Bates, D., M. Mächler, B. Bolker, and S. Walker, 2015 Fitting Linear Mixed-Effects Models
571 Using lme4. Journal of Statistical Software 67 (1): 1–48.
572 https://doi.org/10.18637/jss.v067.i01
573 Boichard, D. A., H. Chung, R. Dassonneville, X. David, A. Eggen *et al.*, 2012 Design of a bo-
574 vine low-density SNP array optimized for imputation. PLoS One 7 (3): e34130.
575 Bradbury, I. R., S. Hubert, B. Higgins, S. Bowman, I. G. Paterson *et al.*, 2011 Evaluating SNP
576 ascertainment bias and its impact on population assignment in Atlantic cod, Gadus
577 morhua. Mol Ecol Res 11 (s1): 218–25.
578 Broad Institute, 2015 Picard Tools 2.0.1. https://broadinstitute.github.io/picard/.
579 Broad Institute, 2018 GATK User Guide. Accessed March 20, 2018.
580 https://software.broadinstitute.org/gatk/documentation/.

581 Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen,
582 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res
583 15 (11): 1496–1502.
584 DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework
585 for variation discovery and genotyping using next-generation DNA sequencing data. Nat
586 Genet 43: 491. https://doi.org/10.1038/ng.806
587 Eller, E., 2001 Effects of Ascertainment Bias on Recovering Human Demographic History.
588 Human Biology 73 (3): 411–27. https://doi.org/10.1353/hub.2001.0034
589 Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple
590 genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6 (5):
591 e19379.
592 ENSEMBL, 2016 Chicken Germline SNP and INDELS. Accessed January 06, 2017.
593 http://e87.ensembl.org/Gallus_gallus.
594 Fan, B., Z.-Q. Du, D. M. Gorbach, and M. F. Rothschild, 2010 Development and application
595 of high-density SNP arrays in genomic studies of domestic animals. Asian-Australas J Anim
596 Sci 23 (7): 833–47.
597 Frascaroli, E., T. A. Schrag, and A. E. Melchinger, 2013 Genetic diversity analysis of elite Eu-
598 ropean maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascer-
599 tainment bias for a subset of SNPs. Theoretical and Applied Genetics 126 (1): 133–41.
600 https://doi.org/10.1007/s00122-012-1968-6
601 Gautier, M., D. Laloe, and K. Moazami-Goudarzi, 2010 Insights into the genetic history of
602 French cattle from dense SNP data on 47 worldwide breeds. PLoS One 5 (9).
603 https://doi.org/10.1371/journal.pone.0013038
604 Gibbs, R. A., J. F. Taylor, C. P. van Tassell, W. Barendse, K. A. Eversole *et al.*, 2009 Genome-
605 wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324
606 (5926): 528–32. https://doi.org/10.1126/science.1167936
607 Groenen, M. A. M., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier *et al.*, 2011 The devel-
608 opment and characterization of a 60K SNP chip for chicken. BMC Genomics 12 (1): 274.
609 Guillot, G., and M. Foll, 2009 Correcting for ascertainment bias in the inference of popula-
610 tion structure. Bioinformatics 25 (4): 552–54.
611 https://doi.org/10.1093/bioinformatics/btn665
612 Heslot, N., J. Rutkoski, J. A. Poland, J.-L. Jannink, and M. E. Sorrells, 2013 Impact of marker
613 ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS
614 One 8 (9): e74612.
615 Kijas, J. W., D. Townley, B. P. Dalrymple, M. P. Heaton, J. F. Maddox *et al.*, 2009 A genome
616 wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One 4 (3):
617 e4668.
618 Kimura, M., 1991 The neutral theory of molecular evolution: A review of recent evidence.
619 Idengaku zasshi 66 (4): 367–86.
620 Kranis, A., A. A. Gheyas, C. Boschiero, F. Turner, Le Yu *et al.*, 2013 Development of a high
621 density 600K SNP genotyping array for chicken. BMC Genomics 14 (1): 59.
622 https://doi.org/10.1186/1471-2164-14-59
623 Lachance, J., and S. A. Tishkoff, 2013 SNP ascertainment bias in population genetic anal-
624 yses: why it is important, and how to correct it. Bioessays 35 (9): 780–86.

Laurie, C. C., D. A. Nickerson, A. D. Anderson, B. S. Weir, R. J. Livingston *et al.*, 2007 Linkage Disequilibrium in Wild Mice. PLoS Genet 3 (8): e144. https://doi.org/10.1371/journal.pgen.0030144

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

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Accessed March 21, 2019. https://arxiv.org/pdf/1303.3997.

Malomane, D. K., C. Reimer, S. Weigend, A. Weigend, A. R. Sharifi *et al.*, 2018 Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics 19 (1): 22. https://doi.org/10.1186/s12864-017-4416-9

Malomane, D. K., H. Simianer, A. Weigend, C. Reimer, A. O. Schmitt *et al.*, 2019 The SYNBREED chicken diversity panel: A global resource to assess chicken diversity at high genomic resolution. BMC genomics 20 (1): 345. https://doi.org/10.1186/s12864-019-5727-9

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan *et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4 (4): e5350.

Mayer, M., S. Unterseer, E. Bauer, N. de Leon, B. Ordas *et al.*, 2017 Is there an optimum level of diversity in utilization of genetic resources? Theor Appl Genet 130 (11): 2283–95. https://doi.org/10.1007/s00122-017-2959-4

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

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie *et al.*, 2016 The Ensembl Variant Effect Predictor. Genome Biol 17 (1): 122. https://doi.org/10.1186/s13059-016-0974-4

McTavish, E. J., J. E. Decker, R. D. Schnabel, J. F. Taylor, and D. M. Hillis, 2013 New World cattle show ancestry from multiple independent domestication events. Proc Natl Acad Sci 110 (15): E1398-E1406.

McTavish, E. J., and D. M. Hillis, 2015 How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics 16 (1): 1.

Muir, W. M., G. K.-S. Wong, Y. Zhang, J. Wang, M. A. M. Groenen *et al.*, 2008 Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc Natl Acad Sci 105 (45): 17312–17. https://doi.org/10.1073/pnas.0806569105

Nei, M., 1972 Genetic Distance between Populations. The American Naturalist 106 (949): 283–92.

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

Nielsen, R., M. J. Hubisz, and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168 (4): 2373–82.

Nielsen, R., and J. Signorovitch, 2003 Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol 63 (3): 245–55.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature 456: 98. https://doi.org/10.1038/nature07331

Paradis, E., J. Claude, and K. Strimmer, 2004 APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–90.

670 Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland *et al.*, 2012 Ancient admixture in
671 human history. Genetics 192 (3): 1065–93. https://doi.org/10.1534/genetics.112.145037
672 Perkel, J., 2008 SNP genotyping: Six technologies that keyed a revolution. Nature Methods
673 5: 447. https://doi.org/10.1038/nmeth0508-447
674 Platt, A., M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio *et al.*, 2010 The Scale of Population
675 Structure in Arabidopsis thaliana. PLoS Genet 6 (2): e1000843.
676 https://doi.org/10.1371/journal.pgen.1000843
677 Qanbari, S., E. C.G. Pimentel, J. Tetens, G. Thaller, P. Lichtner *et al.*, 2010 A genome-wide
678 scan for signatures of recent selection in Holstein cattle. Animal genetics 41 (4): 377–89.
679 Qanbari, S., C.-J. Rubin, K. Maqbool, S. Weigend, A. Weigend *et al.*, 2019 Massive whole
680 genome sequencing reveals strong purifying selection in commercial chicken: - accepted
681 but not yet published. PLoS Genet.
682 Qanbari, S., M. Seidel, T.-M. Strom, K. F. X. Mayer, R. Preisinger *et al.*, 2015 Parallel Selec-
683 tion Revealed by Population Sequencing in Chicken. Genome Biol Evol 7 (12): 3299–3306.
684 https://doi.org/10.1093/gbe/evv222
685 R Core Team, 2017 *R: A Language and Environment for Statistical Computing.* Vienna, Aus-
686 tria.
687 Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald *et al.*,
688 2009 Design of a high density SNP genotyping assay in the pig using SNPs identified and
689 characterized by next generation sequencing technology. PLoS One 4 (8): e6524.
690 Rocha, D., C. Billerey, F. Samson, D. Boichard, and M. Boussaha, 2014 Identification of the
691 putative ancestral allele of bovine single-nucleotide polymorphisms. J Anim Breed Genet
692 131 (6): 483–86. https://doi.org/10.1111/jbg.12095
693 Sandenbergh, L., S. W.P. Cloete, R. Roodt-Wilding, M. A. Snyman, and A. E. Bester-van der
694 Merwe, 2016 Evaluation of the OvineSNP50 chip for use in four South African sheep
695 breeds. S Afr J Anim Sci 46 (1): 89–93.
696 Singh, N., P. K. Jayaswal, K. Panda, P. Mandal, V. Kumar *et al.*, 2015 Single-copy gene based
697 50 K SNP chip for genetic studies and molecular breeding in rice. Sci Rep 5: 11600.
698 Tosser-Klopp, G., P. Bardou, O. Bouchez, C. Cabau, R. P. M. A. Crooijmans *et al.*, 2014 Design
699 and characterization of a 52K SNP chip for goats. PLoS One 9 (1): e86227.
700 Travis, A. J., G. J. Norton, S. Datta, R. Sarma, T. Dasgupta *et al.*, 2015 Assessing the genetic
701 diversity of rice originating from Bangladesh, Assam and West Bengal. Rice 8 (1): 35.
702 https://doi.org/10.1186/s12284-015-0068-z
703 UCSC, 2016 Reference Genome Gallus gallus 5.0. Accessed October 25, 2016.
704 http://hgdownload.soe.ucsc.edu/goldenPath/galGal5/bigZips/galGal5.fa.gz.
705 Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak *et al.*, 2014 A powerful tool for ge-
706 nome analysis in maize: development and evaluation of the high density 600 k SNP geno-
707 typing array. BMC Genomics 15 (1): 823.
708 van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel *et al.*, 2013 From
709 FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices
710 pipeline. Current protocols in bioinformatics 43 (1): 11.10.1-11.10.33.
711 https://doi.org/10.1002/0471250953.bi1110s43
712 Warren, W. C., L. W. Hillier, C. Tomlinson, P. Minx, M. Kremitzki *et al.*, 2017 A New Chicken
713 Genome Assembly Provides Insight into Avian Genome Structure. G3 7 (1): 109–17.
714 https://doi.org/10.1534/g3.116.035923

AUTHORS SUMMARY

Compared to whole genome sequencing data, SNP derived by genotyping arrays are biased towards common variants. This effect is widely known as SNP ascertainment bias. Especially the different strength of the bias for different populations, depending on whether they were involved in the development process or not, can strongly affect population genetic estimators. In this work Geibel et al. investigate the introduction of the bias in its magnitude during array development by redesigning a commercial chicken genotyping array under usage of a diverse set of sequenced chicken populations and give recommendations for developing bias minimized arrays.

728 **SUPPLEMENTARY FILE 2**

729

Correct ploidy calling vs. ploidy two calling: Due to the large diversity of the chicken populations and a series of bugs in GATK (up to version 4.2), it was only possible to call some intervals, spanning 28.2 % of the autosomal region, with the correct ploidy. Compared to the calling with the correct ploidy, assuming ploidy two for the pools came along with a slight reduction of SNPs with very rare alternative allele frequencies of ≤ 0.02 in the pools (Figure S 1 C). As those SNPs had to be invariable in the individually sequenced populations, a reduction of invariable alleles is present in the individually sequenced populations (Figure S 1 B). When comparing expected heterozygosities (H_{exp}) from the two callsets, the ploidy two calling comes along with a slight increase of H_{exp}, which is minimally larger for individually sequenced populations (Figure S 1 A), as the lost SNPs are invariable rather than rare in those populations. However, as the correlation between the two callsets is > 0.99 and the mean overestimation of H_{exp} (OHE) for the pools is 0.02 and 0.05 for the individual sequenced populations, the effect is negligible compared to the effects of ascertainment bias.

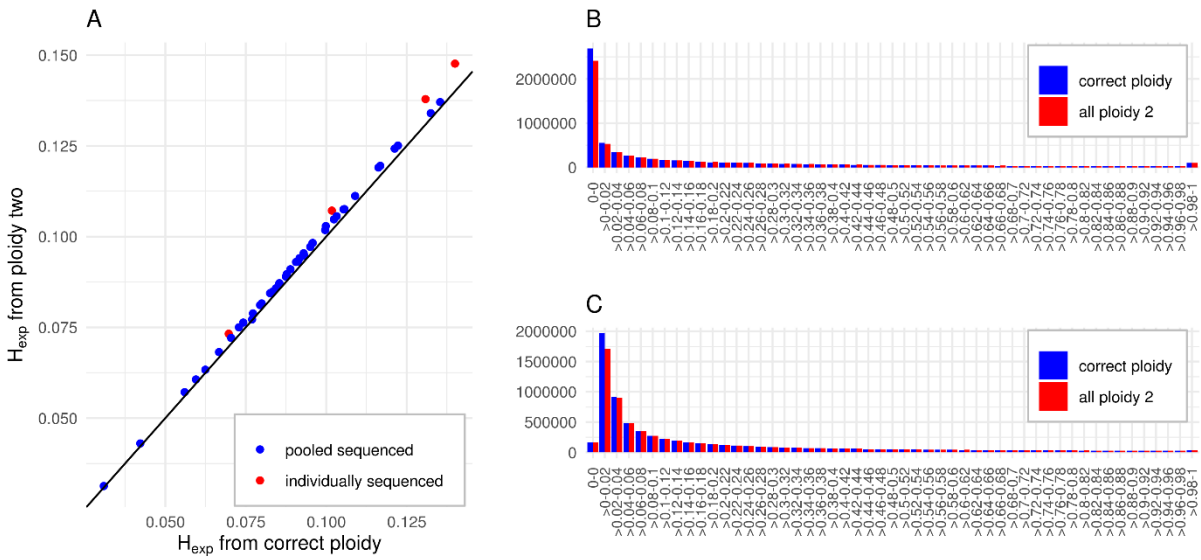


Figure S 1: Pooled vs. ploidy two calling. Expected Heterozygosity (H_{exp}) from the calling with assuming the correct ploidy vs. assuming ploidy two for all samples (A) and alternative allele

747 frequency distributions of called alleles for individually sequenced (B) and pooled sequenced
748 (C) populations.

749

750 **Alternative allele frequency spectrum vs. derived allele frequency spectrum:** As the chicken
751 reference genome is based on an inbred Red Jungle Fowl (*Gallus gallus gallus*; Warren *et al.*
752 2017), many publications handle the reference allele as ancestral allele (e.g. Qanbari *et al.*
753 2015). However, this assumption does not account for mutation, selection and inbreeding
754 happening in the Red Jungle Fowl after separation from the domesticated strains as well as
755 variation already present before separation. Therefore, ancestral alleles were defined by an
756 approach comparable to Rocha *et al.* (2014) using allele frequency information from the
757 three wild populations *Gallus gallus gallus* (GG), *Gallus gallus spadiceus* (GS) and *Gallus vari-*
758 *us* (GV). It was assumed that the *Gallus gallus* and *Galllus varius* species emerged from a
759 common ancestor and *Gallus gallus* later split up into *Gallus gallus gallus* and *Gallus gallus*
760 *spadiceus* subspecies. Additionally, assuming neutral molecular evolution (Kimura 1991), the
761 ancestral allele was most likely the major allele within those three populations, when
762 weighting the allele frequency of *Gallus varius* twice. This procedure assigned the ancestral
763 status to the reference allele for 86 % of the SNPs and to the alternative allele for 14 % of
764 the SNPs. SNPs where the ancestral state was assigned to the alternative allele were approx-
765 imately equal distributed over the range of allele frequencies (Figure S 2 A). Only alleles be-
766 ing close to fixation for the alternative allele were overrepresented, possibly being artefacts
767 of selection on de novo mutations in ancestors of the reference individual or sequencing
768 errors in the reference genome. Therefore, a comparison between alternative (Figure S 2 B)
769 and derived (Figure S 2 C) allele frequency spectra in the complete callset does only show
770 the reduction of a tiny peak being present at the end of the alternative allele frequency
771 spectrum.
772

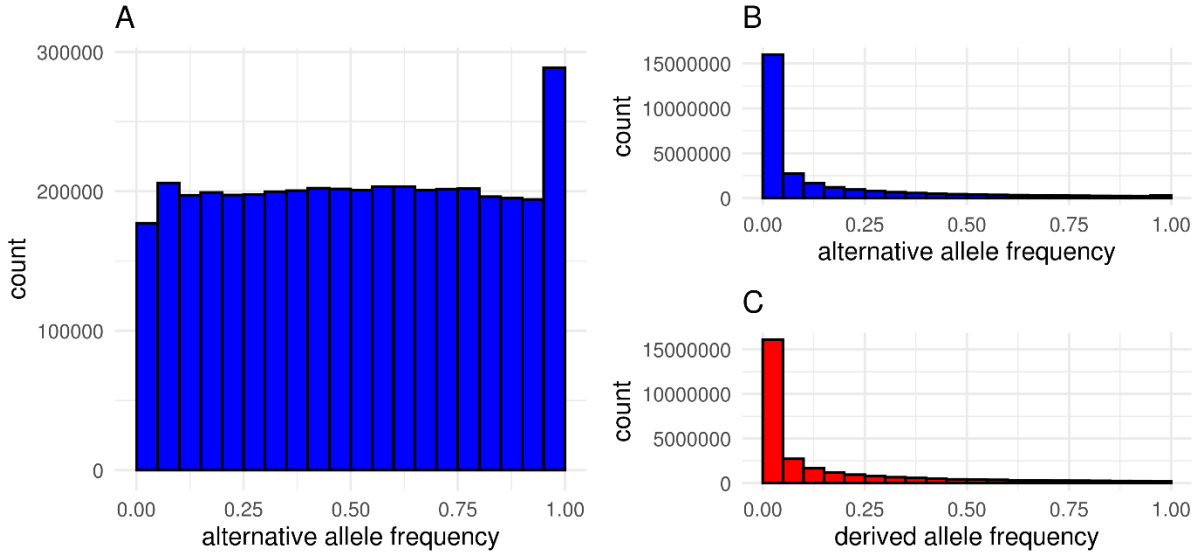


Figure S 2: Alternative allele frequency spectrum for SNPs where the reference allele is not the ancestral allele (A) and alternative (B) vs. derived (C) allele frequency spectrum of all SNPs.

Table S 1: Number of SNPs from the remodeling processes

	1 – SNP discovery	2 - cluster removal	3 – equal spacing	4 – validation	5 – down-sampling
Original Array (Kranis *et al.* 2013)	23 998	10 029	1 829	1 187	581
First Run	10 834	8 788	2 331	1 910	600
Repeated Remodeling	11 603 ± 226	9 328 ± 157	2 032 ± 147	1 698 ± 147	662 ± 33

Number of SNPs in 1,000
In case of repeated remodeling mean of the 50 runs ± standard deviation

Table S 2: Pearson correlations between the H_{exp} of the different SNP sets

	Filtered WGS	Array SNPs	1 – SNP discovery	2 - cluster removal	3 – equal spacing	4 – validation	5 – down-sampling
Unfiltered WGS	1.00	0.94	0.95	0.95	0.96	0.96	0.96
Filtered WGS		0.95	0.96	0.95	0.97	0.96	0.96
Array SNPs			1.00	1.00	0.98	0.97	0.98
1 – SNP discovery				1.00	0.98	0.97	0.97
2 - cluster					0.98	0.96	0.97

removal		
3 – equal spacing	1.00	0.99
4 – validation		1.00

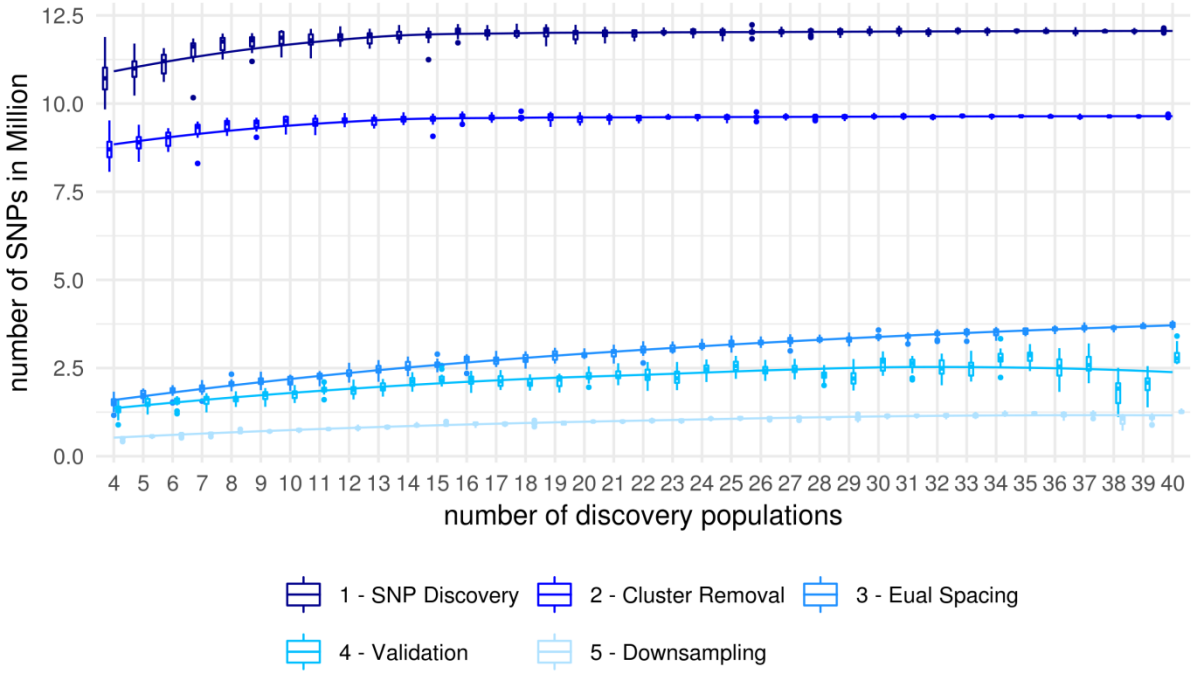
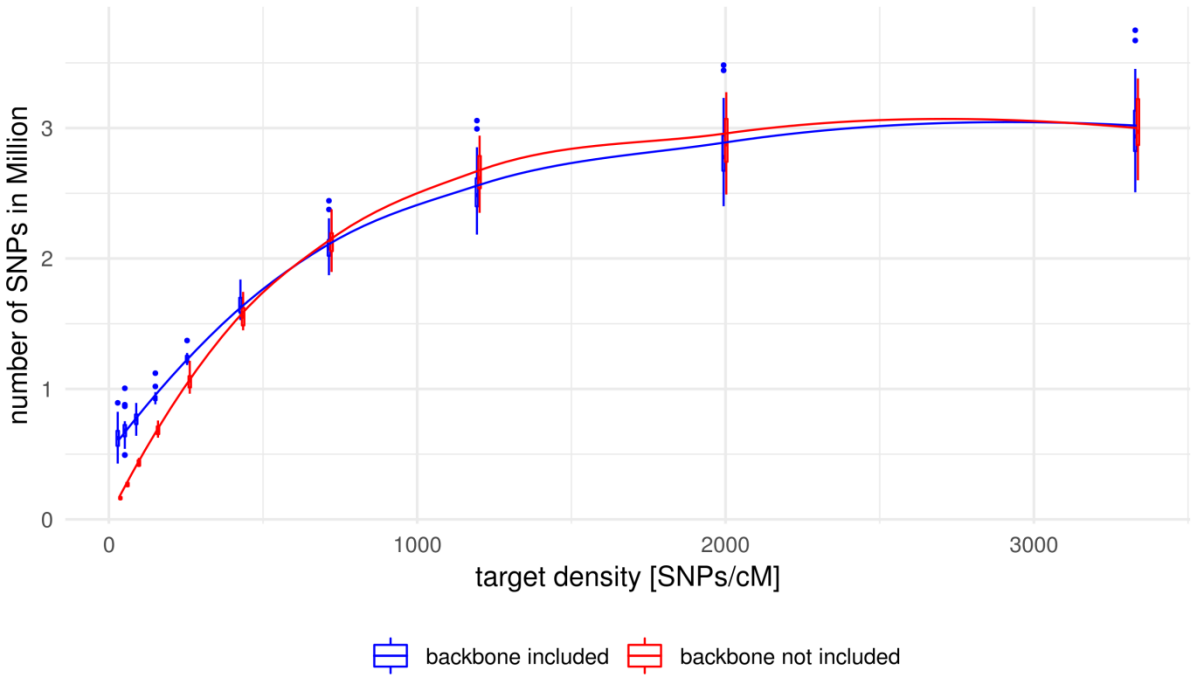


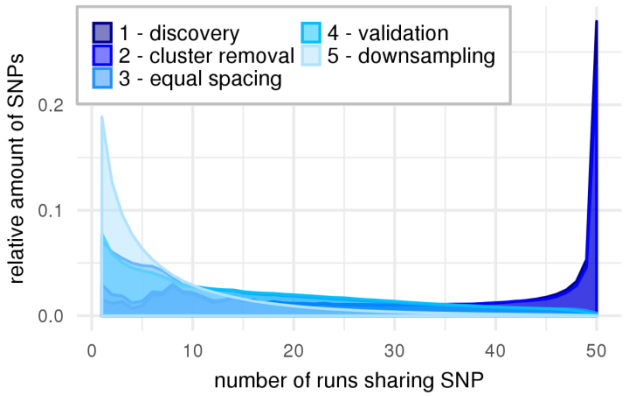
Figure S 3: SNPs retained by number of populations in the discovery set.



787

788 Figure S 4: Number of SNPs retained after the equal spacing step by target density.

789



790

791 Figure S 5: Relative amount of SNPs shared by a specific number of runs from 50 random
792 population assignments.

793

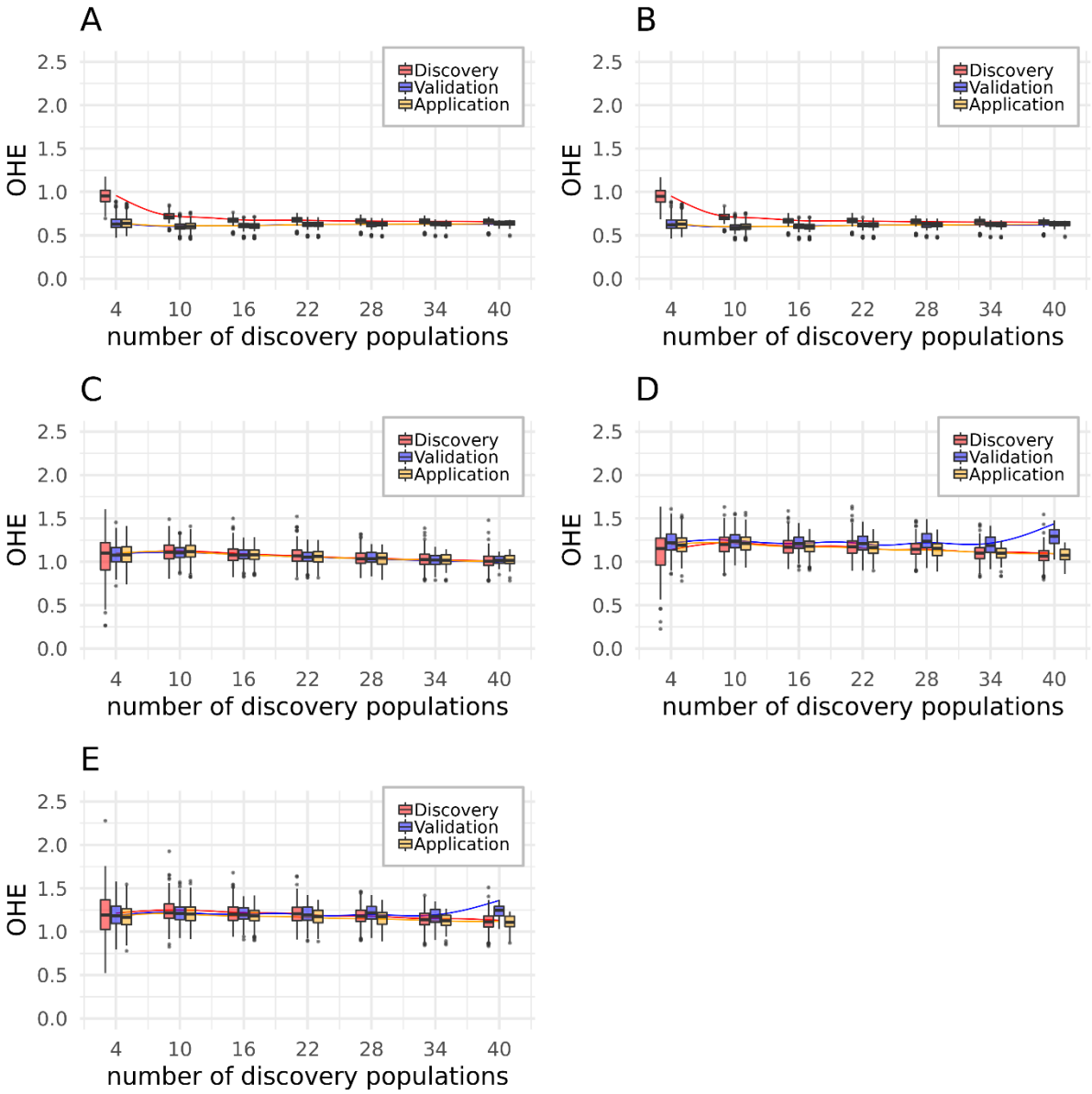


Figure S 6: Relation of the OHE as a function of the number of discovery populations. **A** - discovery, **B** – cluster removal, **C** – equal spacing, **D** – validation, **E** – downsampling. The Boxplots are only shown for a subset of the number of discovery populations, while the smoothing lines, which show the trend, are calculated from all observations.