How Array Design affects 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

* Corresponding author

E-Mail: johannes.geibel@uni-goettingen.de (JG)

Abstract

Single nucleotide polymorphisms (SNPs), genotyped with SNP 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 with main influencing factors being 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 effectively genotyping high numbers of Single Nucleotide Polymorphisms (SNP) for many individuals in parallel via SNP arrays led to an increased usage of them 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 a range from 40 to 60 k [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 [20] 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 SNP 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].

63 The first factor is a relatively small panel of individuals being used for discovery of SNPs,
64 leading to a large proportion of globally rare variants not being selected, since they appear
65 monomorphic in the discovery panel [26,28]. The second factor is the across population use
66 of arrays. Arrays are developed based on the variation within the discovery panel, thus miss-
67 ing variation present in distantly related individuals or populations [25,27]. This second
68 source of bias was shown to be of relatively high importance for livestock studies, where
69 arrays are usually developed for large commercial breeds and later used to genotype diverse
70 sets of local breeds all over the world [29,30].

71 Besides different strategies to minimize the impact of ascertainment bias [30,31], there
72 are some attempts to correct the allele frequency spectrum via Bayesian methods
73 [25,28,32]. However, those corrections highly rely on detailed statistical assumptions of the
74 ascertainment process [33,34] and are currently only tested for corrections of the first
75 source of ascertainment bias, the small discovery panel [25,28,32]. In practice, detailed in-
76 formation on the design process is limited [34] and the complexity of the processes makes
77 statistical models for the corrections inaccurate.

78 Agricultural species such as chickens often show a complex domestication history, and
79 therefore allow for few prior assumptions on ascertainment bias. Domestic chickens are as-
80 sumed to originate from red jungle fowl (*Gallus gallus*) ancestors in Southeast Asia [35,36],
81 represented by the five subspecies *G. g. gallus*, *G. g. spadiceus*, *G. g. murghi*, *G. g. bankiva*
82 and *G. g. jabouillei* [37]. Additionally, some hybridization events with other *Gallus* species
83 (e.g. grey jungle fowl; *Gallus sonneratii*) have been suggested [36,38]. The diversity of to-
84 day's local breeds of chickens in Europe originates from chickens that reached the continent
85 about 3000 years ago via a northern and a southern route, followed by selection and cross-

ing with Asian chicken breeds introduced in the 19th century [37]. 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 crossbreeding of paternal lines (e.g. White Cornish) with maternal lines which descend from a comparable basis as brown layers (e.g. White Plymouth Rock) [39]. For more detailed information on chicken ancestry we refer to Qanbari *et al.* [36] 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 ascertainment bias, 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

Populations and sequencing

The analysis is based on WGS data of a diverse set of 46 commercial, non-commercial and wild chicken populations, sampled within the framework of the projects AVIANDIV (www.aviandiv.fli.de) and SYNBREED (www.synbreed.tum.de). Commercial brown (BL) and

107 white layer (WL) populations consist of 25 individually re-sequenced animals each, while the
108 two commercial broiler lines (BR1 and BR2) include 20 individually sequenced animals each.
109 For 41 populations, pooled DNA from 9 - 11 animals per population was sequenced, while
110 *Gallus varius* (green jungle fowl; GV) samples of only two animals were sequenced as a pool.
111 More detailed information about the samples can be found in **S1 File** and two previously
112 published papers, from Malomane *et al.* [30] and Qanbari *et al.* [40]. Coverage was between
113 7X and 10X for the individual sequences, while DNA pools were sequenced with 15X to 70X
114 coverage. Sequencing was conducted on Illumina HiSeq machines at the Helmholtz Zentrum,
115 German Research Center for Environmental Health in Munich, Germany.

116 Raw data preparation and SNP calling

117 Sequences were aligned to the reference genome Gallus gallus 5.0 [41,42] and the SNP
118 calling was conducted according to GATK Best Practices guidelines [43,44]. BWA-MEM 0.7.12
119 [45] was used for the alignment step, duplicates were marked using Picard Tools 2.0.1 [46]
120 MarkDuplicatesWithMateCigar and base qualities were recalibrated with GATK 3.7 [47]
121 BaseQualityRecalibrator. The set of known SNPs, necessary for base quality score recalibra-
122 tion, was downloaded from ENSEMBL release 87 [48]. SNPs were called for all samples sepa-
123 rately using the GATK 3.7 HaplotypeCaller and later on simultaneously genotyped across
124 samples with GATK 3.7 GenotypeGVCFs. Due to computational limitations, a ploidy of two
125 instead of the higher true ploidy of the pooled sequences had to be assumed, which led to
126 calling slightly less rare alleles. However, effects of this limitation are negligible (**S2 File; S1**
127 **Fig**).

128 SNP filtering was conducted using GATK 3.7 VariantRecalibrator, which filtered the called
129 SNPs by a machine learning approach (use of a Gaussian mixture model), which uses both, a
130 set of previous known (low confidence needed) and a set of highly reliable (assumed to be
131 true) variants as training resources [49]. The source for known SNPs (prior 2) provided to
132 VariantRecalibrator was again ENSEMBL (release 87) and the SNPs of the Axiom™ Genome-
133 Wide Chicken Array were defined as true training set (prior 15). The algorithm was trained
134 on the quality parameters DP, QD, FS, SOR, MQ and MQRankSum. Filters were set to recover
135 99 % of the training SNPs in the filtered set, which resulted in a Transition/Transversion ratio
136 of 2.52 for known SNPs, and a Transition/Transversion ratio of 2.26 for novel SNPs. Only
137 biallelic autosomal SNPs were used in all further analyses.

138 Identification of the ancestral allele

139 Ancestral alleles were defined using allele frequency information from the three wild
140 populations *Gallus gallus gallus* (GG), *Gallus gallus spadiceus* (GS) and *Gallus varius* (GV) by
141 an approach comparable to 50 [50]. It was assumed that the *Gallus gallus* and *Galllus varius*
142 species emerged from a common ancestor and *Gallus gallus* later split into *Gallus gallus gal-*
143 *lus* and *Gallus gallus spadiceus* subspecies. Additionally, assuming neutral molecular evolu-
144 tion [51], the ancestral allele was most likely the major allele within those three populations,
145 when weighting the allele frequency of *Gallus varius* twice. This procedure assigned the an-
146 cestral status to the reference allele for 86 % of the SNPs and to the alternative allele for 14
147 % of the SNPs. The change in the allele frequency spectrum was only relevant for the interval
148 from 0.95 – 1.00, which was reduced by 111'851 SNPs (0.39 % of all SNPs) when switching
149 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 [47] 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 populations were divided into four groups for the array design:

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

For SNP discovery firstly the 4 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 [52] As the two broiler populations were closest related (**Fig 1**), the next two closest populations were chosen. This resulted in the inclusion of White Leghorn (LE), Rhode Island Red (RI), Marans (MR) and Rumpless Araucana (AR). Note that these populations are closely related to the populations used as discovery populations for the development of the Axiom™ Genome-Wide Chicken Array [21] with exception to some inbred lines from the Roslin Institute

172 in Edinburgh we do not know the genetic origin of. The discovery set used for the original
173 array [21] additionally consisted of more animals than ours. Further, the discovery set had to
174 be split into broilers (BR1, BR2, MR, AR) and layers (WL, LE, BL, RI) for the equal spacing step.
175 From the remaining populations, 19 were randomly chosen for SNP validation (validation
176 populations), 18 populations (which were not included in the array development) were used
177 as showcase for an application of the array (application populations), and *Gallus varius* as a
178 different species was defined as outgroup. The interested reader can find all underlying
179 pairwise Nei's standard genetic distances [52] in **S3 File** and additionally pairwise F_{ST} values
180 [53] in **S4 File**.

181

182 **Fig 1. UPGMA tree based on pairwise Nei's standard genetic distances.** The tree was calculated from the fil-
183 tered WGS SNPs. Populations defined as layers or broilers, which form in total the discovery set for the array
184 design close to the original array, are highlighted. The plot was produced using the R package ape [54]. Note
185 that the plot is only supposed to reveal close clustering chicken populations and cannot be interpreted in depth
186 as chickens show a rich history of hybridization events. The interested reader can find all underlying pairwise
187 Nei's standard genetic distances [52] in **S3 File** and additionally pairwise FST values [53] in **S4 File**.

188

189 Based on the unfiltered SNP set, the sampling of the SNPs for an approximately 600 k
190 sized array was remodeled *in silico* in five consecutive steps according to the design process
191 of the original array which was described by Kranis *et al.* [21], starting from the unfiltered
192 SNP set:

193 1) **SNP discovery → 10.9 M SNPs**

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

196 2) **Cluster removal → 8.8 M SNPs**

197 SNP clusters with less than 4 bp invariant sites at one side of a SNP (no matter of

198 whether 3' or 5' direction) and less than 10 bp invariant sites at the other side of
199 the SNP within the discovery populations were removed. This step is justified ra-
200 ther technically to enable probe binding, but could also lead to an overrepresenta-
201 tion of conserved regions compared to highly variable regions of the genome.

202 3) **Equal spacing → 2.1 M SNPs**

203 Reduction of SNPs to achieve approximately equidistant spacing between variable
204 SNPs within discovery populations based on genetic distances. The used algorithm
205 was modeled according to Kranis *et al.* [21] and followed a two-step procedure.
206 The first step was setting up an initial backbone of common SNPs (three sub
207 steps). It started with selecting SNPs which segregated in all discovery populations
208 (MAF within each population > 0) while requiring a minimal distance of 2 kb, re-
209 sulting in about 8 k SNPs. This was complemented by a backbone of SNPs which
210 segregated in all layer populations and another one of SNPs which segregated in
211 all broiler populations. Note that Kranis *et al.* [21] additionally constructed a back-
212 bone from a group of inbred lines for which no comparable samples were availa-
213 ble for this study. In the second step, the algorithm iterated over all single popula-
214 tions and filled in potential gaps between backbone SNPs which are variable with-
215 in the according population. This was done by choosing the SNPs closest to equi-
216 distant positions within the gap while aiming for a predefined local target density
217 of 667 segregating SNPs/cM (linkage map taken from [55]). See **S3 Fig** for the de-
218 tailed contribution of additional SNPs from each sub step of the algorithm.

219 4) **SNP validation → 1.7 M SNPs**

220 Removing SNPs (~ 20 %) which were not variable in at least 8 of the 19 validation

221 populations. This step would in reality be done by genotyping with preliminary
222 test arrays and therefore allows the use of a broader set of populations than the
223 discovery step.

224 5) **Downsampling → 580 k SNPs**

225 Downsampling of SNPs comparable to step 3, but without adding the broiler/ layer
226 specific backbones and instead keeping all exonic SNPs (annotation using Ensembl
227 VEP 89.7; 56). Additionally, the target density in broiler lines was set as three
228 times the target density of the layer lines. The increased target density in broilers
229 is intended to account for lower levels of linkage disequilibrium in these lines.

230 Variation of the design process

231 The whole design process was repeated 50 times with populations being randomly as-
232 signed to be discovery, validation or application populations, while the *Gallus varius* popula-
233 tion was always kept as the outgroup. In this process, the number of populations per group
234 was the same as in the previous scenario.

235 To assess the impact of the number of discovery populations on the design process, the
236 number of discovery populations was varied in additional runs from 4 to 40 randomly chosen
237 populations (while assigning the remaining populations, except *Gallus varius,* to validation
238 and application groups of equal size) with 20 random replicates for each number of discov-
239 ery populations. In a last scenario, equal spacing was varied with respect to the target densi-
240 ty (33 – 3333 SNPs/cM) with 20 independent population groupings for each target density,
241 with or without the initial backbone. As the number of SNPs from the backbone was con-
242 stant, the increase of the target density led to a higher number of SNPs chosen by the algo-

243 rithm due to the equal spacing itself and hence the relative influence of the fixed number of
244 common backbone SNPs decreased.

245 Analyses of the results

246 Per-locus-allele frequencies for individually sequenced populations were estimated from
247 genotypes, whereas the estimation for the sequenced DNA-pools was based on the allelic
248 depth. Influences on the allele frequency spectra were examined by comparing density esti-
249 mates of derived allele frequency spectra (unfolded frequency spectrum). Further H_{exp}, the
250 expected heterozygosity assuming Hardy Weinberg frequencies of the genotypes, for the
251 different populations were used as summary statistics of the within population allele fre-
252 quency spectra and calculated as in equation (1), where $p_{ref;l}$ denotes the frequency of the
253 reference allele at locus l and L the total number of loci.

254 $$H_{\exp} = \frac{\sum_{l} 2 p_{ref;l}(1 - p_{ref;l})}{L} \quad (1)$$

255 Deviations in the estimation of H_{exp} from the various SNP sets were quantified as differ-
256 ences between the H_{exp} calculated from the respective SNP set and the H_{exp} calculated from
257 the filtered WGS SNPs relative to the H_{exp} from the filtered WGS SNPs, further called overes-
258 timation of H_{exp} (OHE; equation (2)). An OHE of zero means that the estimates are equal,
259 while an OHE of one describes doubling of the unbiased estimate.

260 $$OHE = \frac{H_{\text{exp; SNP set}} - H_{\text{exp; filtered WGS SNPs}}}{H_{\text{exp; filtered WGS SNPs}}} \quad (2)$$

261 The effects of the population group assignments on the OHE of the random population
262 assignments were evaluated by pairwise comparisons of least square means (LSMEANS; cal-

263 culated with the R package emmeans [57,58] by using Tukey correction for multiple pairwise
264 contrasts) of the population groups. An underlying mixed linear model for the estimation of
265 LSMEANS was fitted using the R package lme4 [59] as shown in equation (3), where the OHE
266 depended on an overall mean μ, the fixed effect of the population group $popG_i$ (i can be
267 discovery-, validation-, application- or outgroup), a random effect for the jth repetition of
268 random population grouping ($rep_j \sim N(0, I\sigma^2_{rep})$) and a random error $e_{ijk} \sim N(0, I\sigma^2_e)$. The
269 procedure is comparable to simple pairwise comparisons of group means, the correction by
270 the repetition only reduces the error variance and thus decreases the confidence intervals.

271 $$OHE_{ijk} = \mu + popG_i + rep_j + e_{ijk} \quad (3)$$

272

273 Results

274 Numbers of SNPs

275 The SNP calling identified 28.5 M biallelic autosomal SNPs from which 20.9 M SNPs
276 passed GATK's filtering procedure. 540 k SNPs from the unfiltered WGS SNP set are also
277 mapped on the original Axiom™ Genome-Wide 580 k Chicken Array. The remodeling of the
278 array according to the design process of the original array returned 10.9 M SNPs from the
279 discovery step, which were reduced to approximately 580 k in steps as described. Numbers
280 of identified SNPs for the additional runs differed depending on the populations and settings
281 used and are listed in **S1 Table**. It has to be noted that the different sub steps of the equal
282 spacing algorithm contributed with different amounts of SNPs (**S3 Fig**). Especially the much

283 higher contribution of SNPs which were segregating in all broiler populations compared to
284 SNPs segregating in all layer populations in the first remodeling, which is due to closer rela-
285 tionships between the broiler populations and their generally higher heterozygosity, was
286 remarkable. Additional information about the identified number of SNPs depending on the
287 number of discovery populations and target density as well as information about the share
288 of SNPs of different random runs can be found in **S4 – S6 Figs**.

289 Underrepresentation of rare SNPs

290 A clear underrepresentation of rare SNPs in all ascertained SNP sets compared to WGS is
291 evident from the allele frequency spectra (**Fig 2**). Major changes in the allele frequency spec-
292 tra during the array development process were observed after the SNP discovery step and
293 the equal spacing step. The SNP discovery led to an underrepresentation of rare SNPs com-
294 pared to sequence data, which was intensified by the equal spacing step (**Fig 2**) and resulted
295 finally in a spectrum close to the spectrum of the original array. Randomly choosing popula-
296 tions as discovery populations confirmed the shape of the first remodeling, where the popu-
297 lation groups were chosen according to the original array [21]. As major changes in the spec-
298 tra mainly occurred after the SNP discovery and equal spacing, further results will concen-
299 trate on those steps.

300

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

303

304 The allele frequency spectra (**Fig 3**) within discovery populations, compared to the spectra
305 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.

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 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.

Fig 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. 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.

330 A very low target density, indicating that SNPs were mostly called due to being common
331 backbone SNPs, resulted in an allele frequency spectrum with the majority of alleles having a
332 MAF of around 0.5 (**Fig 5**). Increasing the target density for the equal spacing and thus reduc-
333 ing the influence of the initial backbone of common SNPs shifted the peak of the allele fre-
334 quency spectrum left towards a higher proportion of alleles with small derived allele fre-
335 quencies. Using only the backbone SNPs common over all discovery populations and thus
336 calling SNPs mostly by the equal spacing procedure resulted, independently from the target
337 density, in a spectrum similar to the one obtained with a high target density with backbone
338 (**Fig 5**).
339

340 **Fig 5. Derived allele frequency spectra for different target densities [SNPs/cM] after the equal spacing step.**
341 Blue spectra show the runs with the initial backbone included, red ones without backbone. Different target
342 densities are indicated by a log scaled increase in color intensity. Only one representative run per target density
343 is shown.
344

345 Overestimation of H_{exp}

346 **Fig 6** shows the H_{exp} of different SNP sets by population. The H_{exp} obtained from the fil-
347 tered WGS SNPs were slightly higher than from the unfiltered WGS SNPs. H_{exp} obtained from
348 the ascertained SNP sets showed an even more pronounced overestimation together with
349 an increase during the design steps. In general, the correlations between the H_{exp} obtained
350 in the different SNP sets were relatively high (≥ 0.95; **S2 Table**). Especially the H_{exp} of the two
351 WGS SNP sets showed a nearly perfect correlation of > 0.99, which led to an almost constant
352 OHE of -0.23 (**Table 1**) for the unfiltered WGS SNPs. As already recognizable from the H_{exp}
353 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; **Fig 6**; **Table 1**) laid in the range covered by the remodeling steps.

Fig 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.

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}.

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

376 showed an underestimation of the H_{exp}, resulting in an OHE of < -0.84 for all ascertained SNP
377 sets (**Fig 6**; **Table 1**).
378 A closer look on the contribution of the sub-steps during the equal spacing step revealed
379 that 62 % of the SNPs which were preserved during equal spacing were variable in all of the
380 four closely related broiler populations (BR1, BR2, MR, AR; maximum pairwise Nei's distance
381 of 0.06 and FST of 0.17 in the filtered SNP set), while only 3 % of the SNPs were retained due
382 to being variable in all of the four less closely related layer populations (WL, LE, BL, RI; max-
383 imum pairwise Nei's distance of 0.15 and FST of 0.48 in the filtered SNP set). The first popu-
384 lation used to fill in the gaps in the backbone (WL) contributed 17 % of the SNPs, while the
385 other populations contributed < 8 %.
386
387 **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	$-0.85_{\pm 0.008}^{c}$	$-0.86_{\pm 0.008}^{c}$	$-0.85_{\pm 0.015}^{c}$	$-0.84_{\pm 0.017}^{c}$	$-0.84_{\pm 0.019}^{c}$

388 LSMEANS for OHE ± standard error.
389 An OHE of zero means no bias and an OHE of 1 means doubling the H_{exp}.
390 Different lowercase letters within columns indicate significant differences to the 5 % level.
391
392 These findings were supported by the 50 random groupings (**Fig 7**). The LSMEANS (**Table**
393 **2**) of the population groups revealed 24 % larger OHE for discovery populations than for val-
394 idation and application populations after discovery and cluster removal step, which was de-
395 creased to a numerically insignificant difference after the equal spacing step. Interestingly,
396 and in contrast to the findings from the first remodeling, SNP validation led to a significantly

397 higher OHE (5 % larger) for application populations than discovery and validation popula-
398 tions.
399
400 **Fig 7. OHE for discovery validation and application populations after the five steps of array design.** Discovery
401 populations are chosen to represent populations which are comparable to the original array (blue) or 50 times
402 random sampled (grey).
403

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

406 **Fig 8 A** shows that increasing the number of discovery populations reduces the mean OHE
407 of discovery populations after SNP discovery while not affecting the OHE of validation and
408 application populations. After equal spacing (**Fig 8 B**), no dependency of OHE on the number
409 of discovery populations remained. As, due to the limited number of populations in the
410 complete set, the number of validation populations decreased with more populations in the
411 discovery set, the individual validation populations gained more impact on the ascertain-
412 ment. This led to a higher OHE of validation populations with a high number of discovery
413 populations (**S7 D Fig**), comparable to the higher OHE of discovery populations for a small
414 number of discovery populations.
415
416 **Fig 8. Relation of the OHE as a function of the number of discovery populations. A** - discovery, **B** - equal spac-
417 ing. The Boxplots are only shown for a subset of the number of discovery populations, while the smoothing
418 lines, which show the trend, are calculated from all observations. Plots for all five steps can be found in **S7 Fig**.
419
420 In the equal spacing step, using only backbone SNPs resulted in a higher OHE for discov-
421 ery than for non- discovery populations. Increasing the target density and thus increasing the

422 proportion of SNPs due to the equal spacing part of the algorithm reduced the difference in
423 OHE between the population groups (**Fig 9 A**). If the SNPs from the initial backbone were not
424 used, no difference of OHE between discovery and non- discovery populations was present,
425 regardless of the target density (**Fig 9 B**).
426

427 **Fig 9. OHE after equal spacing (step 3) by target density in SNPs/cM and population group.** The smoothing
428 lines show the trend and the dashed lines the target density of 667 SNPs/cM, used for the remodeling accord-
429 ing to the original array [21]. The algorithm was run including the initial backbone SNPs (**A**) or not including
430 them (**B**). *Gallus varius* is not included, as it is constantly underestimated.
431

432 Discussion

433 In this study we used a uniquely diverse collection of sequenced wild, commercial and
434 non-commercial chicken populations, mainly based on samples of the Synbreed Chicken Di-
435 versity Panel [13]. Parts of our set were also involved in the development process of the Axi-
436 om™ Genome-Wide 580 k Chicken Array [21]. This offered an excellent possibility for as-
437 sessing the impact of ascertainment bias on real data in a complex scenario. In general, re-
438 sults derived in this study should therefore be comparable to other species. However, do-
439 mestic chickens show a rich history of hybridization and crossbreeding events [13]. The ef-
440 fects of using a discovery set closely related to the commercial populations and distributing
441 the discovery set randomly across the spectrum of populations were therefore comparably
442 small in this study. Special patterns of population structure e.g. the stronger differentiation
443 in cattle due to the two subspecies *Bos taurus* and *Bos indicus* [60] accompanied by limiting
444 the discovery set to one of the two clades, should increase the impact of population struc-
445 ture dependent ascertainment bias.

Potential impacts of the SNP calling pipeline

As the state of the art pipeline of GATK relies on a supervised machine learning approach for filtering the SNP calls, which needs a highly reliable set of known SNPs, we started with examining potential impacts of the filtering procedure on ascertainment bias. The number of rare variants was slightly reduced by the filtering procedure and thus increased estimates of H_{exp} were obtained in the filtered WGS set. As rare variants have a higher risk to be discarded as sequencing errors [61], 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**).

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

468 MAF alleles (MAF filter for quality control in discovery step and use of common alleles for
469 the backbone in the equal spacing step) and/ or was applied to a small number of popula-
470 tions (small discovery set vs. small validation set). Thereby, the strongest shifts of the allele
471 frequency spectra and increases of H_{exp} are observed after SNP discovery and equal spacing.
472 Both, cluster removal and second downsampling had almost no effect on the allele frequen-
473 cy spectra and H_{exp}, while the validation step slightly decreased the share of rare SNPs.

474 The discovery step had the strongest impact on discovery populations, when a small set
475 of discovery populations was used (**Fig 8 A**). Similarly, the influence of the validation step on
476 validation populations was strongest in case of a small number of validation populations
477 (**S7D Fig**). A balancing of these two groups of samples is therefore necessary, if the number
478 of available DNA samples for array development is limited. We rather suggest the use of an
479 extended discovery set for validation as best practice instead of using completely different
480 populations.

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

485 Differences between groups

486 If allele frequency spectra are changed in the same way for all populations and are there-
487 fore biasing heterozygosity estimates to the same extent, findings for between population
488 comparisons will be little affected. Ascertainment bias then is only of importance if one
489 compares populations based on different arrays, and corrections of the allele frequency

490 spectrum as reviewed by Nielsen [25] should be possible. As correlations between H_{exp} of
491 ascertained SNP sets and unfiltered/ filtered WGS SNP sets were consistently high (> 0.94),
492 arrays designed in the way as performed in this paper should mostly be suitable for robust
493 and cost efficient analyses. Biasedness of estimates could be reduced even more by consid-
494 ering filter strategies according to Malomane *et al.* [30].

495 However, we could show that the bias acts with different extent on different population
496 groups (population structure dependent bias) and therefore changes ranking of populations
497 and can affect conclusions. This population structure dependent bias was already shown to
498 have severe impact on findings from SNP arrays. For example, Bradbury *et al.* [62] found a
499 demographic decline up to an approximately 30 % lower H_{exp} for Atlantic cod based on the
500 distance to the sampling location of the discovery panel and McTravish and Hillis [29]
501 showed strong deviations between simulated and observed polymorphisms for different
502 combinations of migration- and ascertainment scenarios on simulated cattle populations. In
503 concordance with this, populations which are closely related to the discovery populations of
504 the original array in our study on average showed a 24 % higher OHE than validation and
505 application populations for the original array.

506 This population structure dependent bias was mainly introduced by the initial discovery
507 step. It was also observed for the random population groupings, but to a slightly different
508 extent. The difference in overestimation decreased with an increase in the number of dis-
509 covery populations (**Fig 8**) and was smallest if the discovery populations showed minimum
510 distance to the application and validation populations (results not shown). Comparable ob-
511 servations were already made by Frascaroli *et al.* [63] which found very small ascertainment
512 bias for European elite maize lines when using a SNP panel discovered in a combination of a

513 maize diversity set and inbred lines, but strong ascertainment bias when using SNPs which
514 were discovered in American elite lines. Therefore, we suggest to ideally choose an array
515 where the discovery panel does span the scope of populations it will be applied to, and by
516 this covers the existing variation in a most representative way, or to design such an array for
517 oneself if it does not exist.

518 The equal spacing step lowered the difference in mean OHE between population groups
519 in most of our remodeling scenarios, but obviously not in case of the original array. In the
520 remodeling, we saw this difference only with a low target density and thus calling SNPs in
521 the equal spacing step mainly due to being common over many populations (**Fig 9 A**). How-
522 ever, the equal spacing step also increased the variance of OHE in the discovery panel,
523 meaning that the OHE was increased more for some of the discovery populations than for
524 others, thus causing more uncertainty for resulting effects. This effect is driven by the une-
525 qual contribution of variable SNPs to the chosen SNP set by the different populations during
526 the equal spacing step (**S3 Fig**). The equal spacing step increases the OHE for some of the
527 discovery populations, while it decreases it for others, and hence it does not remove the
528 population structure dependent bias. This means that the knowledge of which discovery
529 populations were used is not sufficient to draw conclusions regarding a possible ascertain-
530 ment bias, since their relative contribution varies through the described pipeline.

531 Outgroup

532 *Gallus varius* as an outgroup showed a different behavior than all other populations. It al-
533 ready exhibited the lowest H_{exp} in the unfiltered WGS SNP set, which was most likely driven
534 by the small number of only two samples in the pool, and showed less upward bias of H_{exp} in

535 the filtered WGS SNP set than all other populations. The *Gallus varius* sequence reads on
536 average showed weak Phred-scaled mapping quality scores of 19 (1.3 % probability of misa-
537 lignment), while the mean quality scores of the other populations ranged from 25 (0.3 %) to
538 28 (0.1 %) – meaning that variation, only present in *Gallus varius*, will be more likely missed
539 due to misplacement of the reads or discarded as possible sequencing errors. Additionally,
540 every ascertained SNP set showed an OHE for *Gallus varius* of < -0.84, as variation being pre-
541 sent only in *Gallus varius* was not found in *Gallus gallus* discovery panels and contrary vari-
542 ants from *Gallus gallus* were not variable in *Gallus varius* (**Fig 3**). This demonstrates that ar-
543 rays should not be used if different species (even closely related ones) are included in the
544 research project. Even sequence based estimates can be slightly biased, if the reference ge-
545 nome does not fit properly.

546 Potential impact on other breeding applications

547 In general, we cannot infer the impact on breeding applications which require phenotypic
548 data (e.g. genomic selection [64] or genome wide association studies [65]) and/or individual-
549 ly sequenced or genotyped individuals (e.g. linkage disequilibrium decay [66] or runs of
550 homozygosity analyses [67]) from this study. However, literature highlights the increased
551 power of high MAF SNPs to capture/ detect effects which are caused by common variants
552 due to stronger linkage disequilibrium and higher levels of variance explained. Therefore,
553 increasing MAF in a first instance increases prediction accuracy when the number of SNPs is
554 limited [68] and therefore some SNP ascertainment schemes intentionally bias the used
555 SNPs towards high MAF within the desired populations [17]. The switch to WGS data, and
556 therefore the additional inclusion of rare alleles, is then expected to increase the possibility

557 of capturing the effects of rare alleles [68–70]. However, the increase in efficiency by higher
558 numbers of SNPs levels off when going towards WGS data [71]. Nevertheless, we would ex-
559 pect negative impacts of ascertainment bias due to the across population use of the arrays.
560 When biasing the genotyped variation towards the discovery population, the variability in
561 populations, which are less related to the discovery populations, is less increased or even
562 reduced, and arrays therefore become less valuable in non-target populations. Slight effects
563 of this were demonstrated by simulation [68] and we can clearly support these findings by
564 the levels of differences in the genotyped heterozygosity which we observed in this study.
565 For the effect of ascertainment bias on a broader set of applications, we further refer the
566 interested reader to studies which specifically address those issues (e.g. 30,31,66).

567 Further recommendations for future studies

568 We showed that existing arrays come with a large, and due to a diverse set of promoting
569 factors barely predictable, potential for ascertainment bias. Strongly decreasing costs for
570 WGS and increasing availability of powerful computing resources therefore promote an in-
571 tensified use of WGS for population genetic analyses, especially when diverse populations
572 are included in the studies. However, costs and computational effort will still be substantial
573 for large scale projects. Possible cost effective alternatives could be reduced library sequenc-
574 ing approaches like Genotyping-by-Sequencing [61,72], even though such methods intro-
575 duce other problems related to the use of restriction enzymes which are reviewed by An-
576 drews *et al.* [73].

577 For the purpose of monitoring genetic diversity in a large set of small non-commercial
578 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, it is 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.

However, when considering the findings of this study, choosing an array for which the SNP ascertainment process is well documented and a discovery panel that matches the populations to be studied, should enable researchers to derive appropriate conclusions when robust estimators are used.

Acknowledgments

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

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. doi: 10.1038/nature07331.

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. doi: 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. doi: 10.1371/journal.pgen.0030144.

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. doi: 10.1371/journal.pgen.1000843.

604 **5**. Travis AJ, Norton GJ, Datta S, Sarma R, Dasgupta T, Savio FL, et al. Assessing the genetic
605 diversity of rice originating from Bangladesh, Assam and West Bengal. Rice. 2015; 8:35.
606 doi: 10.1186/s12284-015-0068-z.
607 **6**. Mayer M, Unterseer S, Bauer E, Leon N de, Ordas B, Schön C-C. Is there an optimum
608 level of diversity in utilization of genetic resources. Theor Appl Genet. 2017; 130:2283–
609 95. doi: 10.1007/s00122-017-2959-4.
610 **7**. Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, et al. Ge-
611 nome-wide assessment of worldwide chicken SNP genetic diversity indicates significant
612 absence of rare alleles in commercial breeds. Proc Natl Acad Sci. 2008; 105:17312–7.
613 doi: 10.1073/pnas.0806569105.
614 **8**. Gibbs RA, Taylor JF, van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-
615 wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science.
616 2009; 324:528–32. doi: 10.1126/science.1167936 PMID: 19390050.
617 **9**. Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, et al. A genome
618 wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One.
619 2009; 4:e4668.
620 **10**. Gautier M, Laloe D, Moazami-Goudarzi K. Insights into the genetic history of French
621 cattle from dense SNP data on 47 worldwide breeds. PLoS One. 2010; 5.
622 doi: 10.1371/journal.pone.0013038 PMID: 20927341.
623 **11**. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, et al. A ge-
624 nome-wide scan for signatures of recent selection in Holstein cattle. Animal genetics.
625 2010; 41:377–89.
626 **12**. McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM. New World cattle show an-
627 cestry from multiple independent domestication events. Proc Natl Acad Sci. 2013;
628 110:E1398-E1406.
629 **13**. Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The
630 SYNBREED chicken diversity panel. A global resource to assess chicken diversity at high
631 genomic resolution. BMC Genomics. 2019; 20:345. doi: 10.1186/s12864-019-5727-9
632 PMID: 31064348.
633 **14**. Perkel J. SNP genotyping. Six technologies that keyed a revolution. Nature Methods.
634 2008; 5:447. doi: 10.1038/nmeth0508-447.
635 **15**. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool
636 for genome analysis in maize: development and evaluation of the high density 600 k
637 SNP genotyping array. BMC Genomics. 2014; 15:823.
638 **16**. Singh N, Jayaswal PK, Panda K, Mandal P, Kumar V, Singh B, et al. Single-copy gene
639 based 50 K SNP chip for genetic studies and molecular breeding in rice. Sci Rep. 2015;
640 5:11600.
641 **17**. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Devel-
642 opment and characterization of a high density SNP genotyping assay for cattle. PLoS
643 One. 2009; 4:e5350.
644 **18**. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. De-
645 sign of a high density SNP genotyping assay in the pig using SNPs identified and charac-
646 terized by next generation sequencing technology. PLoS One. 2009; 4:e6524.

647 **19**. Groenen MAM, Megens H-J, Zare Y, Warren WC, Hillier LW, Crooijmans RPMA, et al.
648 The development and characterization of a 60K SNP chip for chicken. BMC Genomics.
649 2011; 12:274.
650 **20**. Boichard DA, Chung H, Dassonneville R, David X, Eggen A, Fritz S, et al. Design of a bo-
651 vine low-density SNP array optimized for imputation. PLoS One. 2012; 7:e34130.
652 **21**. Kranis A, Gheyas AA, Boschiero C, Turner F, Le Yu, Smith S, et al. Development of a high
653 density 600K SNP genotyping array for chicken. BMC Genomics. 2013; 14:59.
654 doi: 10.1186/1471-2164-14-59.
655 **22**. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans RPMA, Dong Y, et al. Design
656 and characterization of a 52K SNP chip for goats. PLoS One. 2014; 9:e86227.
657 **23**. Sandenbergh L, Cloete SWP, Roodt-Wilding R, Snyman MA, Bester-van der Merwe AE.
658 Evaluation of the OvineSNP50 chip for use in four South African sheep breeds. S Afr J
659 Anim Sci. 2016; 46:89–93.
660 **24**. Fan B, Du Z-Q, Gorbach DM, Rothschild MF. Development and application of high-
661 density SNP arrays in genomic studies of domestic animals. Asian-Australas J Anim Sci.
662 2010; 23:833–47.
663 **25**. Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004;
664 1:1.
665 **26**. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in
666 studies of human genome-wide polymorphism. Genome Res. 2005; 15:1496–502.
667 **27**. Eller E. Effects of Ascertainment Bias on Recovering Human Demographic History. Hu-
668 man Biology. 2001; 73:411–27. doi: 10.1353/hub.2001.0034.
669 **28**. Nielsen R, Signorovitch J. Correcting for ascertainment biases when analyzing SNP data:
670 applications to the estimation of linkage disequilibrium. Theor Popul Biol. 2003;
671 63:245–55.
672 **29**. McTavish EJ, Hillis DM. How do SNP ascertainment schemes and population de-
673 mographics affect inferences about population history. BMC Genomics. 2015; 16:1.
674 **30**. Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. Efficiency of
675 different strategies to mitigate ascertainment bias when using SNP panels in diversity
676 studies. BMC Genomics. 2018; 19:22. doi: 10.1186/s12864-017-4416-9.
677 **31**. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is
678 important, and how to correct it. Bioessays. 2013; 35:780–6.
679 **32**. Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained
680 single-nucleotide polymorphism data. Genetics. 2004; 168:2373–82.
681 **33**. Guillot G, Foll M. Correcting for ascertainment bias in the inference of population struc-
682 ture. Bioinformatics. 2009; 25:552–4. doi: 10.1093/bioinformatics/btn665 PMID:
683 19136550.
684 **34**. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect
685 measures of population divergence. Mol Biol Evol. 2010; 27:2534–2547.
686 **35**. West B, Zhou B-X. Did chickens go north? New evidence for domestication. Journal of
687 archaeological science. 1988; 15:515–33.
688 **36**. Lawal RA, Martin SH, Vanmechelen K, Vereijken A, Silva P, Al-Atiyat RM, et al. The wild
689 species genome ancestry of domestic chickens. BMC Biology. 2020; 18:13.
690 doi: 10.1186/s12915-020-0738-1.

691 **37**. Tixier-Boichard M, Bed'hom B, Rognon X. Chicken domestication. From archeology to
692 genomics. Comptes rendus biologies. 2011; 334:197–204.
693 **38**. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al.
694 Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken.
695 PLoS Genet. 2008; 4:e1000010. doi: 10.1371/journal.pgen.1000010.
696 **39**. Crawford RD. Poultry genetic resources. evolution, diversity and conservation. In: Craw-
697 ford RD, editor. Poultry breeding and genetics. 2nd ed. Amsterdam: Elsevier; 1993.
698 **40**. Qanbari S, Rubin C-J, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of
699 adaptation in modern chicken. PLoS Genet. 2019; 15:e1007989.
700 doi: 10.1371/journal.pgen.1007989.
701 **41**. Reference Genome Gallus gallus 5.0. UCSC 2016 [cited 25 Oct 2016]. Available from:
702 http://hgdownload.soe.ucsc.edu/goldenPath/galGal5/bigZips/galGal5.fa.gz.
703 **42**. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves TA, et al. A New
704 Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3. 2017;
705 7:109–17. doi: 10.1534/g3.116.035923.
706 **43**. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework
707 for variation discovery and genotyping using next-generation DNA sequencing data. Nat
708 Genet. 2011; 43:491. doi: 10.1038/ng.806.
709 **44**. van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et
710 al. From FastQ data to high confidence variant calls. The Genome Analysis Toolkit best
711 practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1-11.10.33.
712 doi: 10.1002/0471250953.bi1110s43 PMID: 25431634.
713 **45**. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
714 2013 [updated 26 May 2013; cited 21 Mar 2019]. Available from:
715 https://arxiv.org/pdf/1303.3997.
716 **46**. Picard Tools 2.0.1. Broad Institute 2015. Available from:
717 https://broadinstitute.github.io/picard/.
718 **47**. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Ge-
719 nome Analysis Toolkit. A MapReduce framework for analyzing next-generation DNA se-
720 quencing data. Genome Res. 2010; 20:1297–303. doi: 10.1101/gr.107524.110.
721 **48**. ENSEMBL. Chicken Germline SNP and INDELS. 2016 [cited 6 Jan 2017]. Available from:
722 http://e87.ensembl.org/Gallus_gallus.
723 **49**. Broad Institute. GATK User Guide. 2018 [cited 20 Mar 2018]. Available from:
724 https://software.broadinstitute.org/gatk/documentation/.
725 **50**. Rocha D, Billerey C, Samson F, Boichard D, Boussaha M. Identification of the putative
726 ancestral allele of bovine single-nucleotide polymorphisms. J Anim Breed Genet. 2014;
727 131:483–6. doi: 10.1111/jbg.12095.
728 **51**. Kimura M. The neutral theory of molecular evolution. A review of recent evidence. Jpn J
729 Genet. 1991; 66:367–86.
730 **52**. Nei M. Genetic Distance between Populations. The American Naturalist. 1972;
731 106:283–92.
732 **53**. Wright S. The genetical structure of populations. Ann Eugen. 1949; 15:323–54.
733 doi: 10.1111/j.1469-1809.1949.tb02451.x.
734 **54**. Paradis E, Claude J, Strimmer K. APE. Analyses of phylogenetics and evolution in R lan-
735 guage. Bioinformatics. 2004; 20:289–90.

736 **55**. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, et al.
737 A high-density SNP-based linkage map of the chicken genome reveals sequence fea-
738 tures correlated with recombination rate. Genome Res. 2009; 19:510–9.
739 **56**. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant
740 Effect Predictor. Genome Biol. 2016; 17:122. doi: 10.1186/s13059-016-0974-4.
741 **57**. Lenth R. emmeans: Estimated Marginal Means, aka Least-Squares Means. ; 2019.
742 **58**. R Core Team. R. A Language and Environment for Statistical Computing. Vienna, Aus-
743 tria; 2017.
744 **59**. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using
745 lme4. Journal of Statistical Software. 2015; 67:1–48. doi: 10.18637/jss.v067.i01.
746 **60**. Hiendleder S, Lewalski H, Janke A. Complete mitochondrial genomes of *Bos taurus* and
747 *Bos indicus* provide new insights into intra-species variation, taxonomy and domestica-
748 tion. Cytogenet Genome Res. 2008; 120:150–6. doi: 10.1159/000118756.
749 **61**. Heslot N, Rutkoski J, Poland JA, Jannink J-L, Sorrells ME. Impact of marker ascertain-
750 ment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One.
751 2013; 8:e74612.
752 **62**. Bradbury IR, Hubert S, Higgins B, Bowman S, Paterson IG, Snelgrove PVR, et al. Evalu-
753 ating SNP ascertainment bias and its impact on population assignment in Atlantic cod,
754 Gadus morhua. Mol Ecol Res. 2011; 11:218–25.
755 **63**. Frascaroli E, Schrag TA, Melchinger AE. Genetic diversity analysis of elite European
756 maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertain-
757 ment bias for a subset of SNPs. Theoretical and Applied Genetics. 2013; 126:133–41.
758 doi: 10.1007/s00122-012-1968-6.
759 **64**. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total Genetic Value Using Ge-
760 nome-Wide Dense Marker Maps. Genetics. 2001; 157:1819–29. Available from:
761 https://www.genetics.org/content/157/4/1819.
762 **65**. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their
763 use in breeding programmes. Nat Rev Genet. 2009; 10:381–91. doi: 10.1038/nrg2575.
764 **66**. Qanbari S, Pausch H, Jansen S, Somel M, Strom T-M, Fries R, et al. Classic selective
765 sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014; 10:e1004148.
766 doi: 10.1371/journal.pgen.1004148 PMID: 24586189.
767 **67**. Peripolli E, Munari DP, Silva, M. V. G. B., Lima ALF, Irgang R, Baldi F. Runs of
768 homozygosity: current knowledge and applications in livestock. Anim Genet. 2017;
769 48:255–71. doi: 10.1111/age.12526.
770 **68**. Perez-Enciso M, Rincon JC, Legarra A. Sequence- vs. chip-assisted genomic selection:
771 accurate biological information is advised. Genet Sel Evol. 2015; 47:43.
772 doi: 10.1186/s12711-015-0117-5 PMID: 25956961.
773 **69**. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome se-
774 quence data: impact of sequencing design on genotype imputation and accuracy of
775 predictions. Heredity. 2014; 112:39–47. doi: 10.1038/hdy.2013.13 PMID: 23549338.
776 **70**. Wainschtein P, Jain DP, Yengo L, Zheng Z, Cupples LA, Shadyab AH, et al. Recovery of
777 trait heritability from whole genome sequence data. bioRxiv. 2019.
778 doi: 10.1101/588020.

71. 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. doi: 10.1371/journal.pgen.1002685 PMID: 22570636.
72. 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.
73. 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.

Supporting information

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

S2 File. Supplementary methods.

S3 File. Pairwise Nei's standard genetic distances.

S4 File. Pairwise F_{ST} values.

S1 Fig. Pooled vs. ploidy two calling. 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.

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

S3 Fig. 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

805 up potential gaps to achieve an equidistant coverage of 667 segregating SNPs/cM for each
806 discovery population.

807 **S4 Fig. SNPs retained by number of populations in the discovery set.**

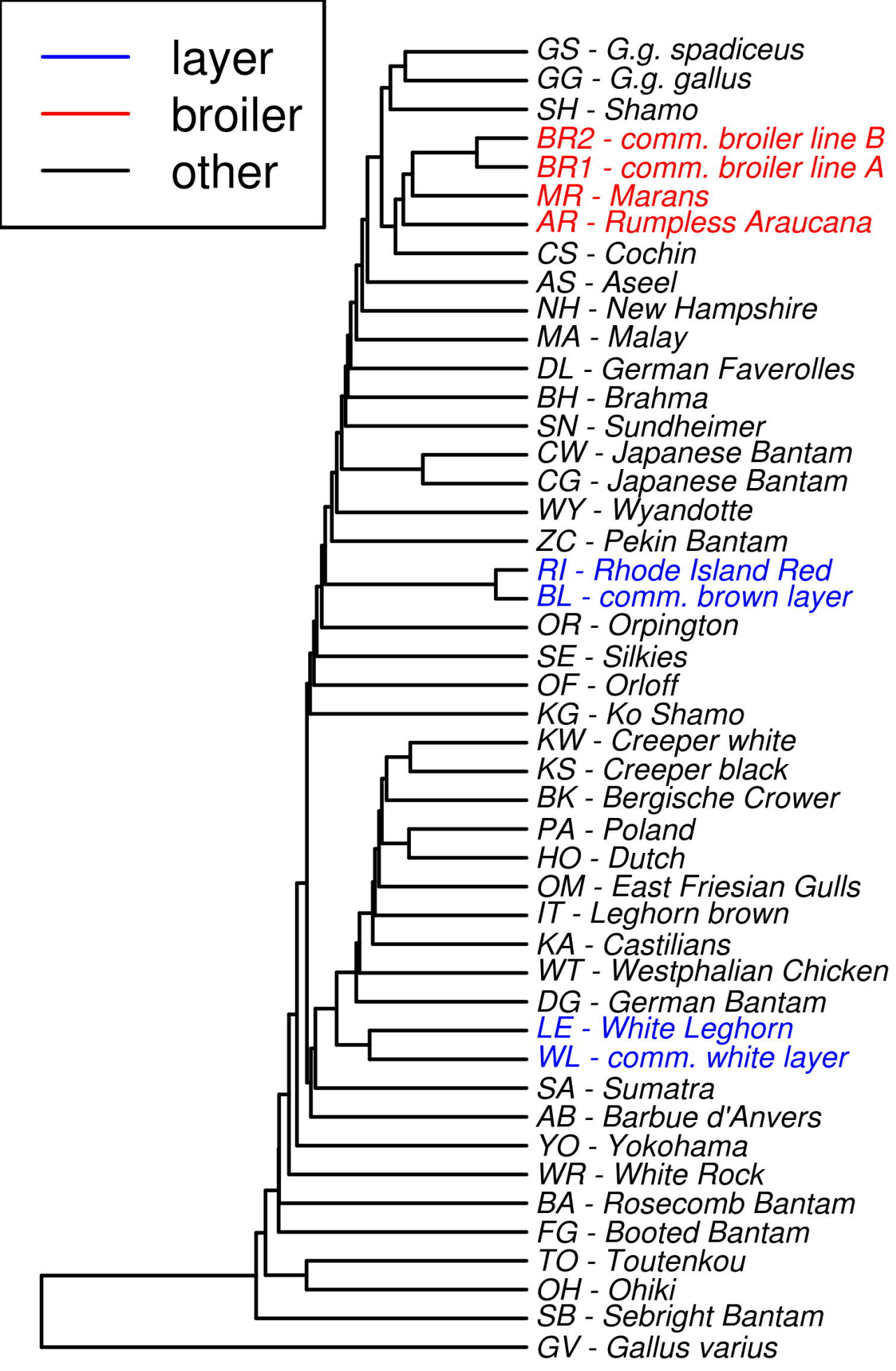
808 **S5 Fig. Number of SNPs retained after the equal spacing step by target density.**

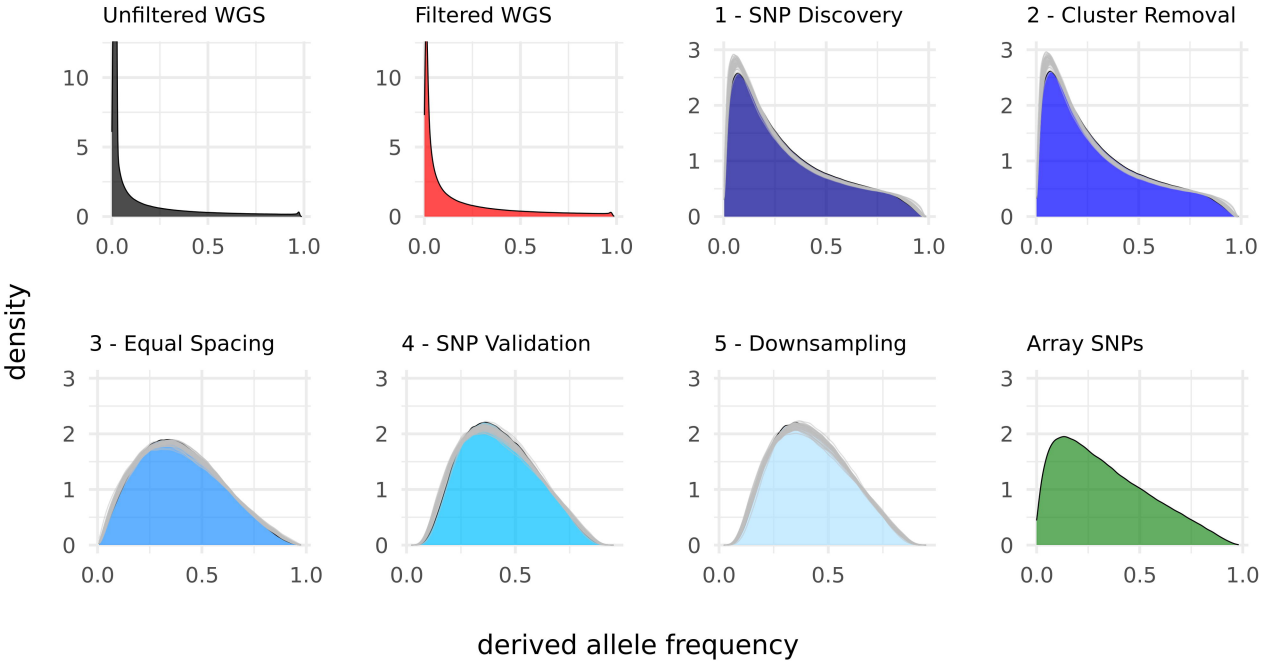
809 **S6 Fig. Relative amount of SNPs shared by a specific number of runs from 50 random popu-**
810 **lation assignments.**

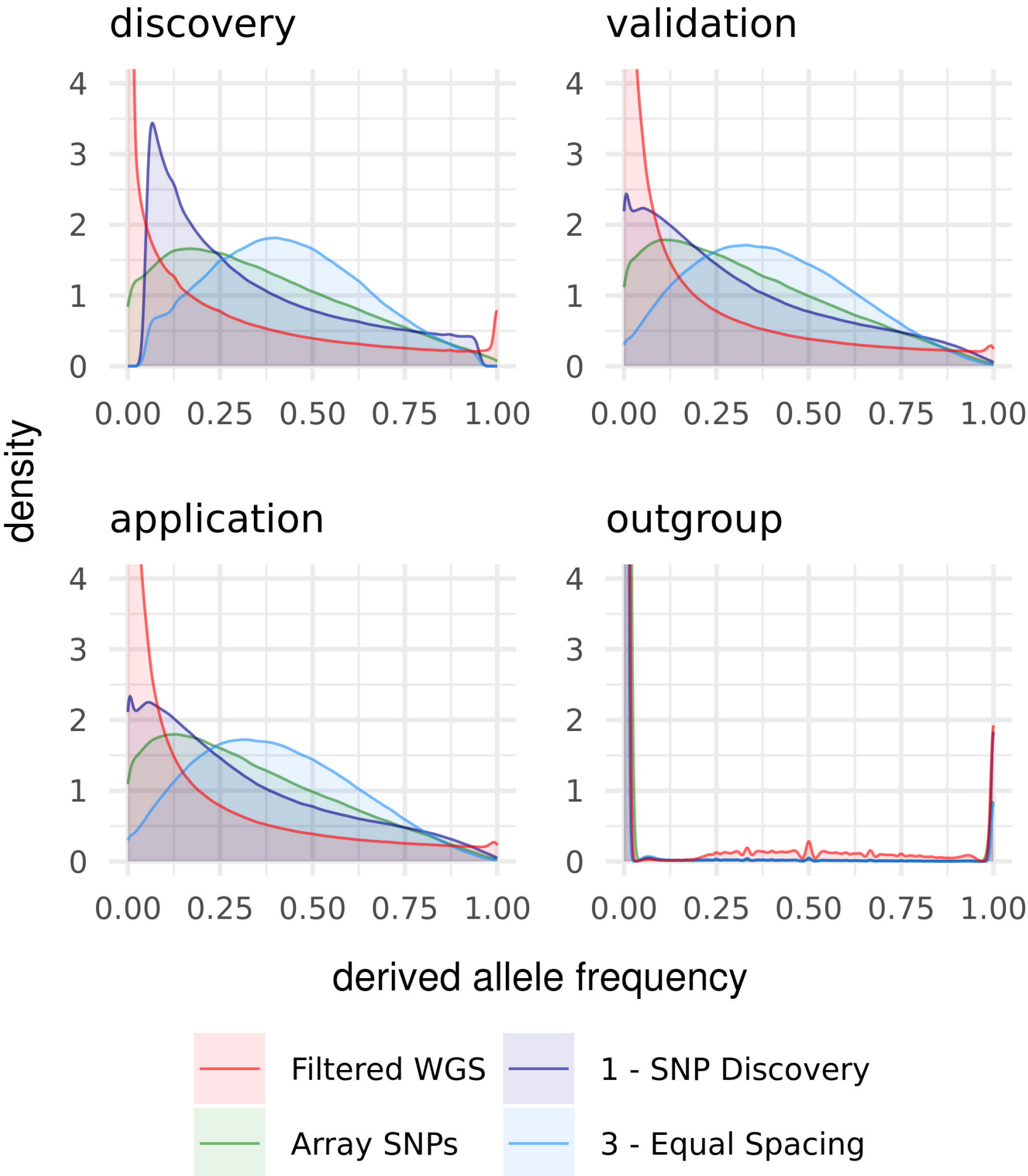
811 **S7 Fig. Relation of the OHE as a function of the number of discovery populations.**
812 **A** - discovery, **B –** cluster removal, **C –** equal spacing, **D –** validation, **E –** downsampling. The
813 Boxplots are only shown for a subset of the number of discovery populations, while the
814 smoothing lines, which show the trend, are calculated from all observations.

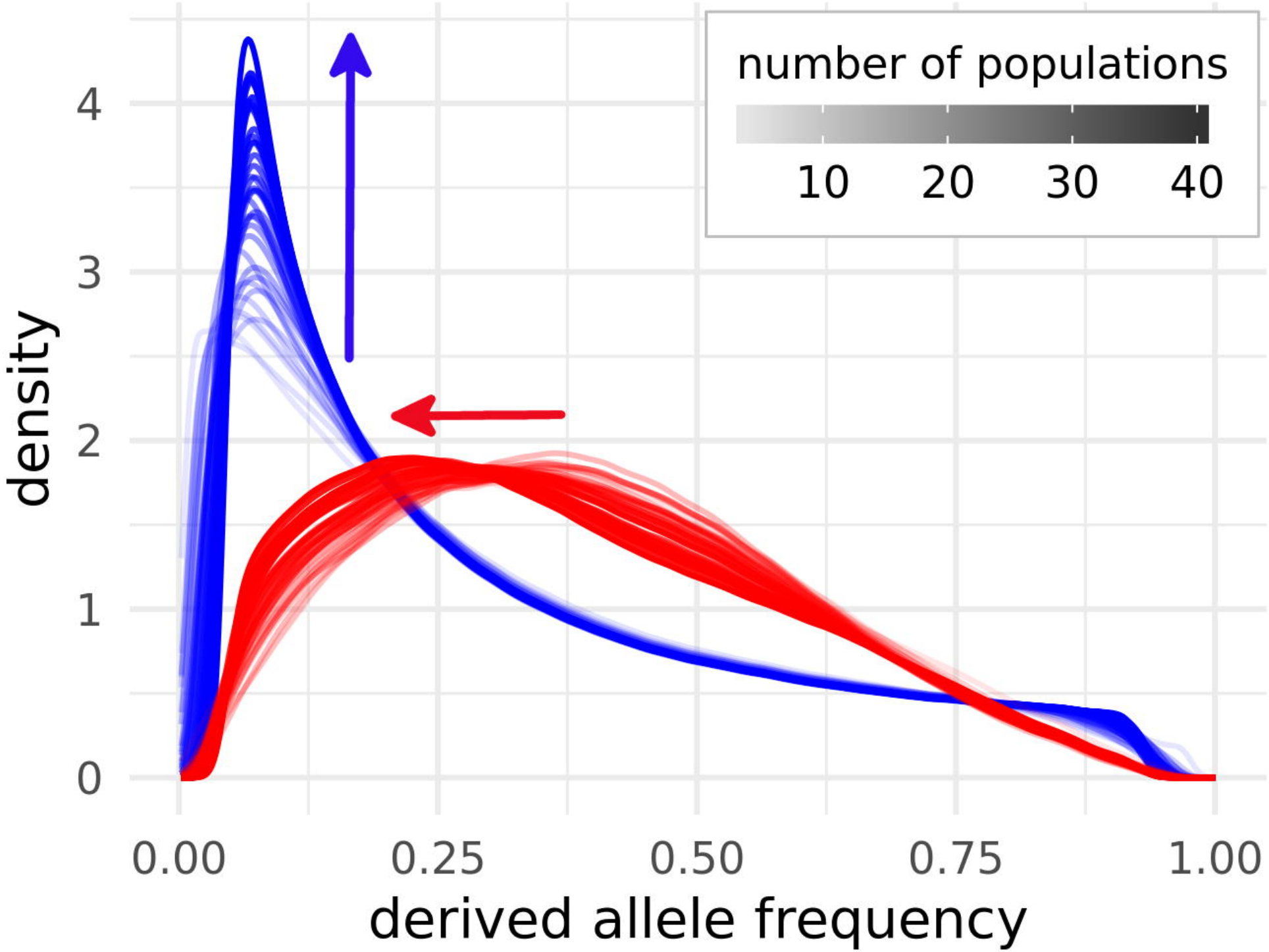
815 **S1 Table. Number of SNPs from the remodeling processes**

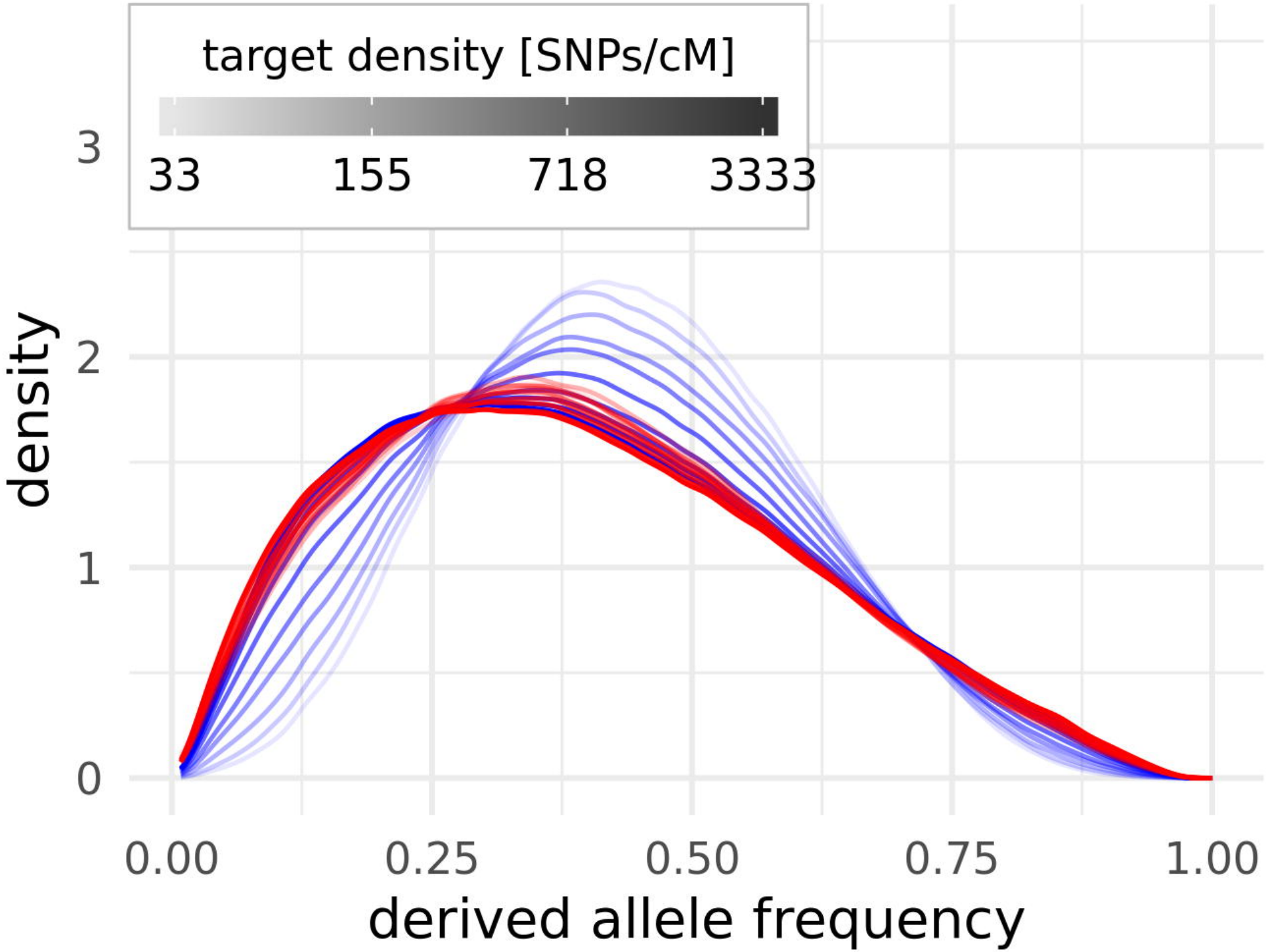
816 **S2 Table. Pearson correlations between the H_{exp} of the different SNP sets**

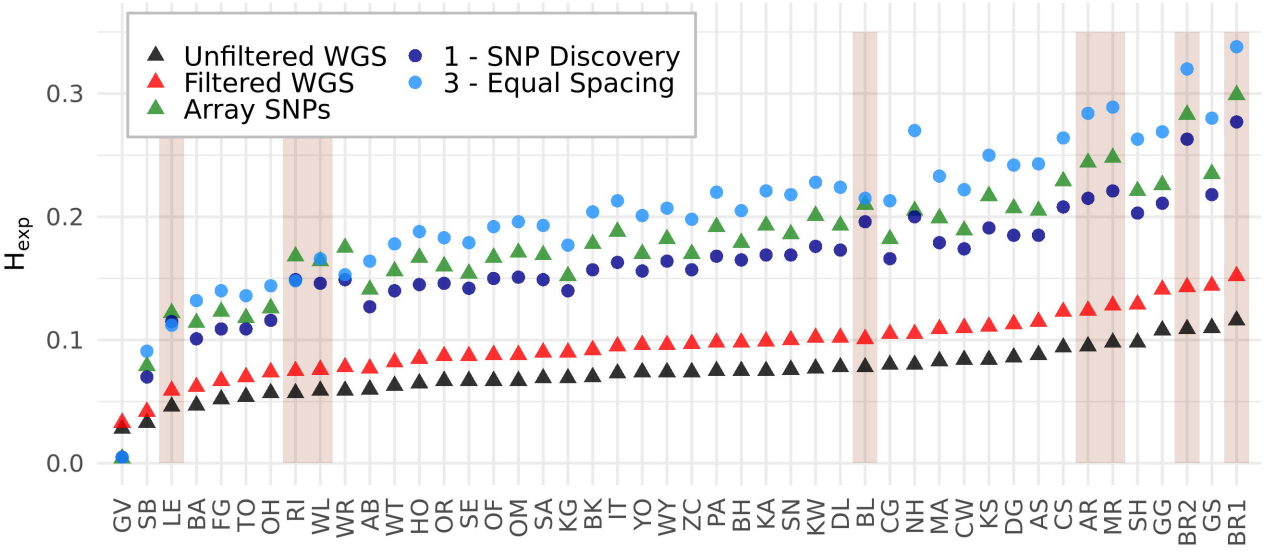

layer
broiler
other
GS - G.g. spadiceus
GG - G.g. gallus
SH - Shamo
BR2 - comm. broiler line B
BR1 - comm. broiler line A
MR - Marans
AR - Rumpless Araucana
CS - Cochin
AS - Aseel
NH - New Hampshire
MA - Malay
DL - German Faverolles
BH - Brahma
SN - Sundheimer
CW - Japanese Bantam
CG - Japanese Bantam
WY - Wyandotte
ZC - Pekin Bantam
RI - Rhode Island Red
BL - comm. brown layer
OR - Orpington
SE - Silkies
OF - Orloff
KG - Ko Shamo
KW - Creeper white
KS - Creeper black
BK - Bergische Crower
PA - Poland
HO - Dutch
OM - East Friesian Gulls
IT - Leghorn brown
KA - Castilians
WT - Westphalian Chicken
DG - German Bantam
LE - White Leghorn
WL - comm. white layer
SA - Sumatra
AB - Barbue d'Anvers
YO - Yokohama
WR - White Rock
BA - Rosecomb Bantam
FG - Booted Bantam
TO - Toutenkou
OH - Ohiki
SB - Sebright Bantam
GV - Gallus varius

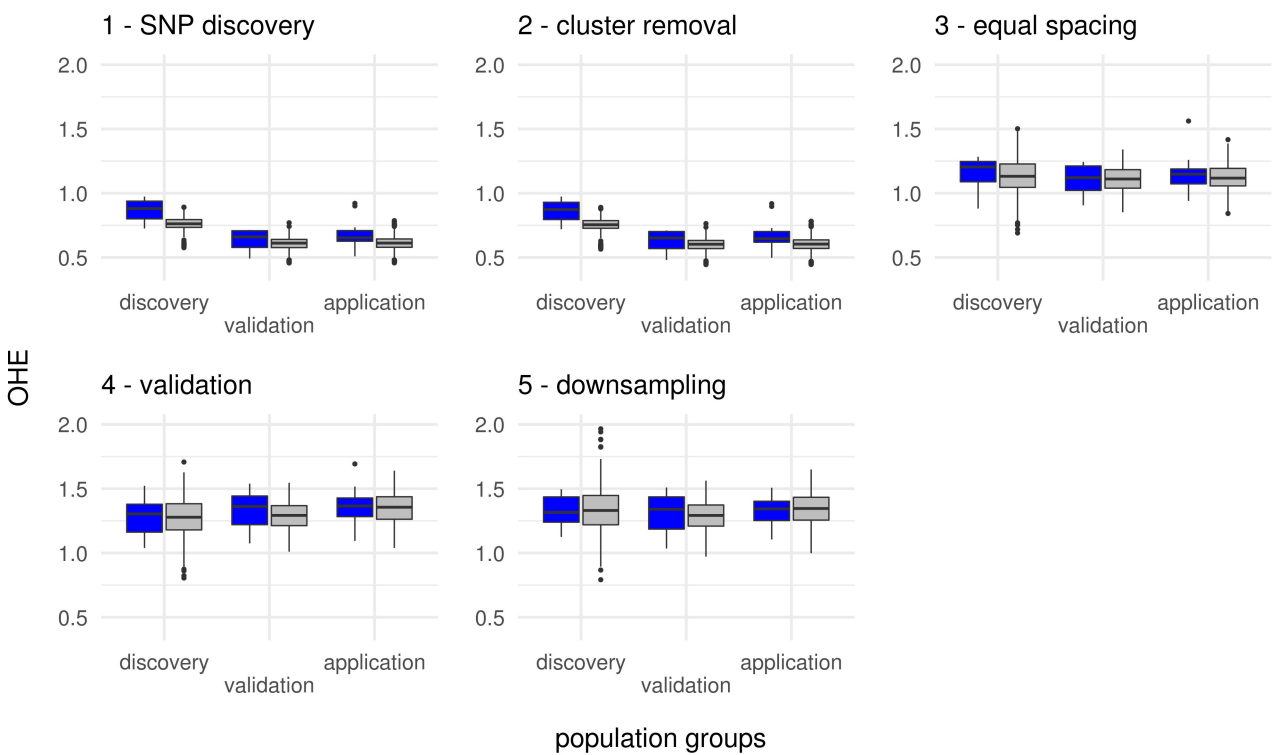

Unfiltered WGS
Filtered WGS
1 - SNP Discovery
2 - Cluster Removal
3 - Equal Spacing
4 - SNP Validation
5 - Downsampling
Array SNPs
density
derived allele frequency

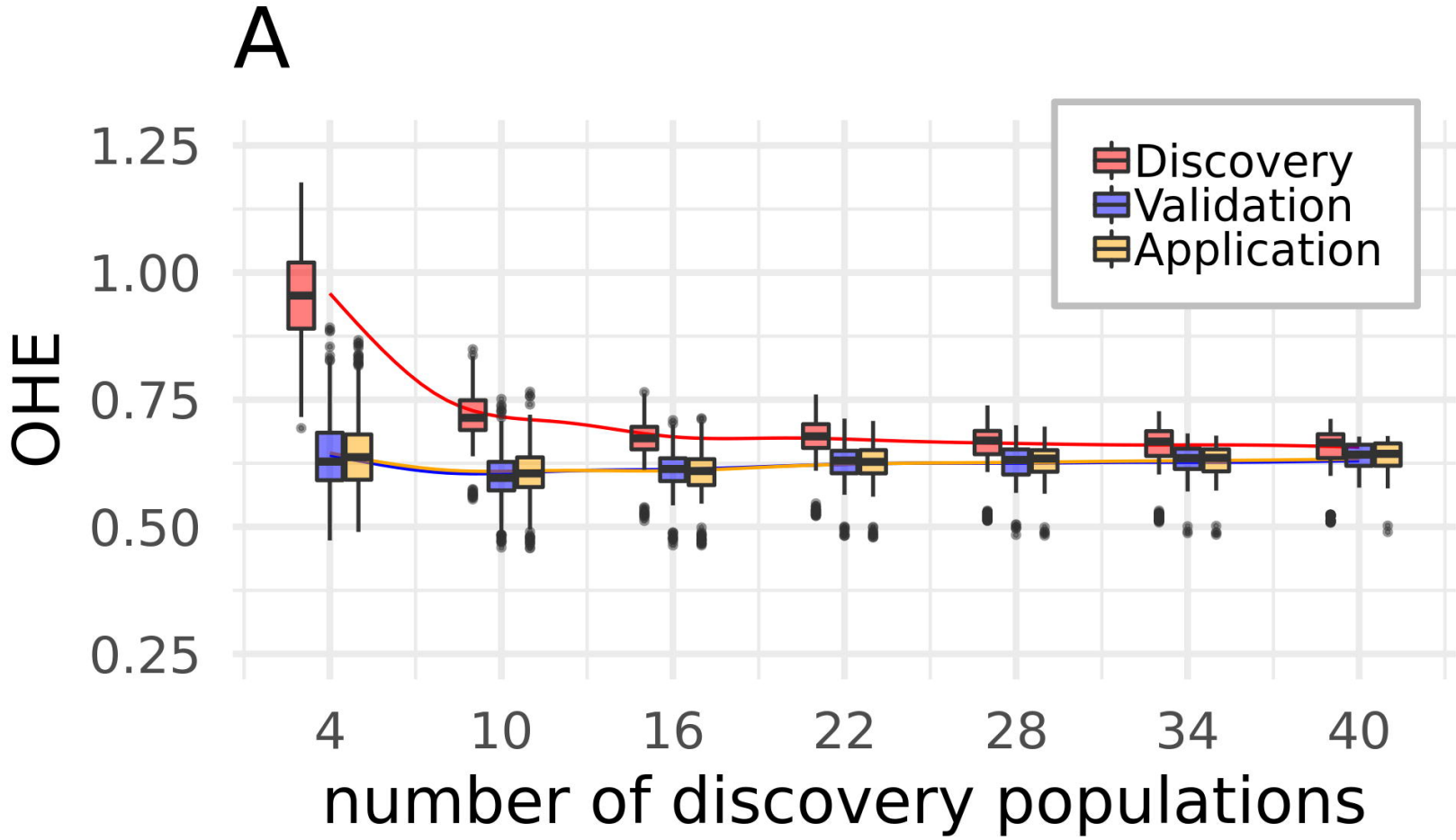

discovery
validation
application
outgroup
density
derived allele frequency
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
Filtered WGS
1 - SNP Discovery
Array SNPs
3 - Equal Spacing

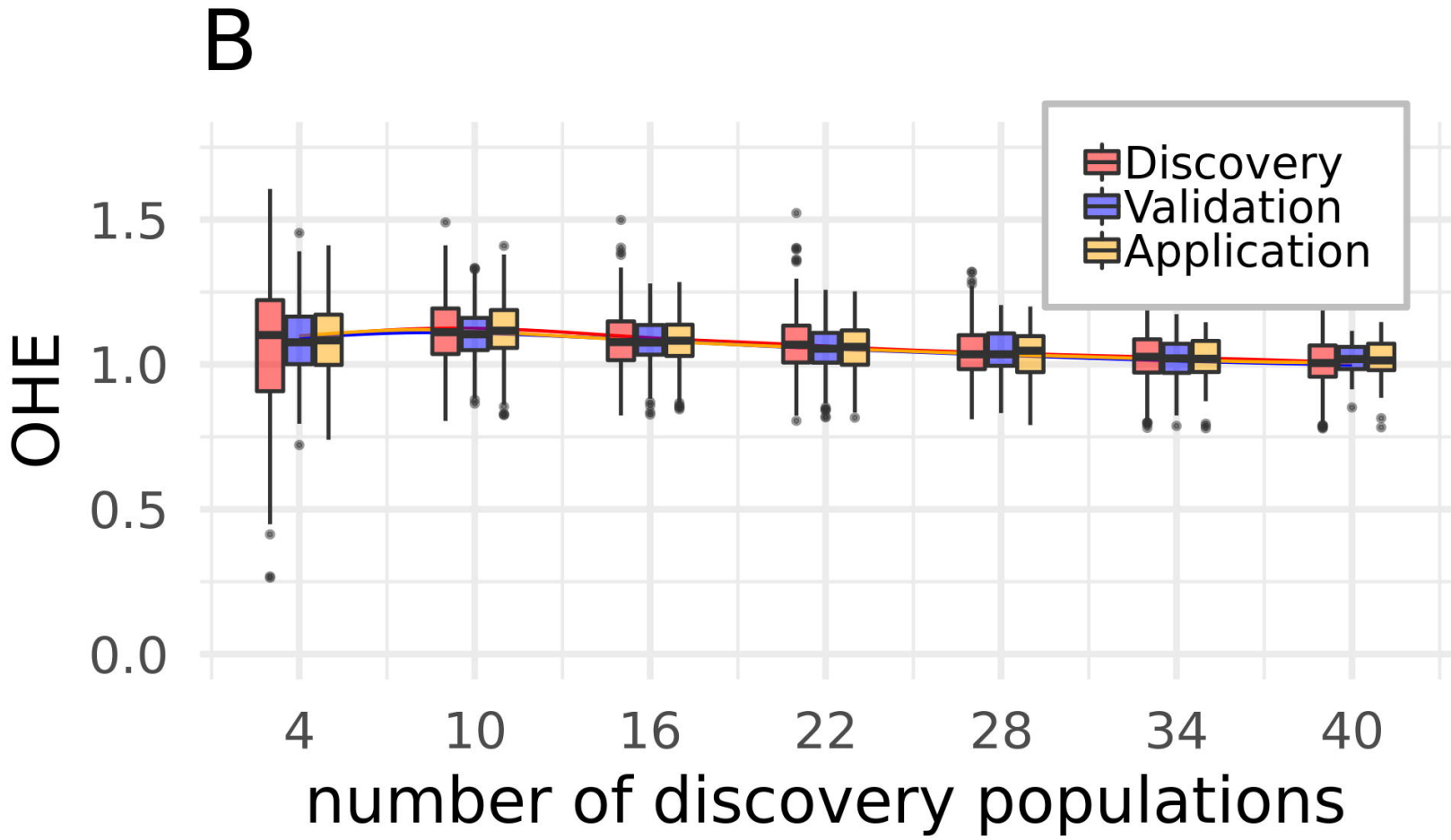

number of populations
10
20
30
40
density
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
derived allele frequency

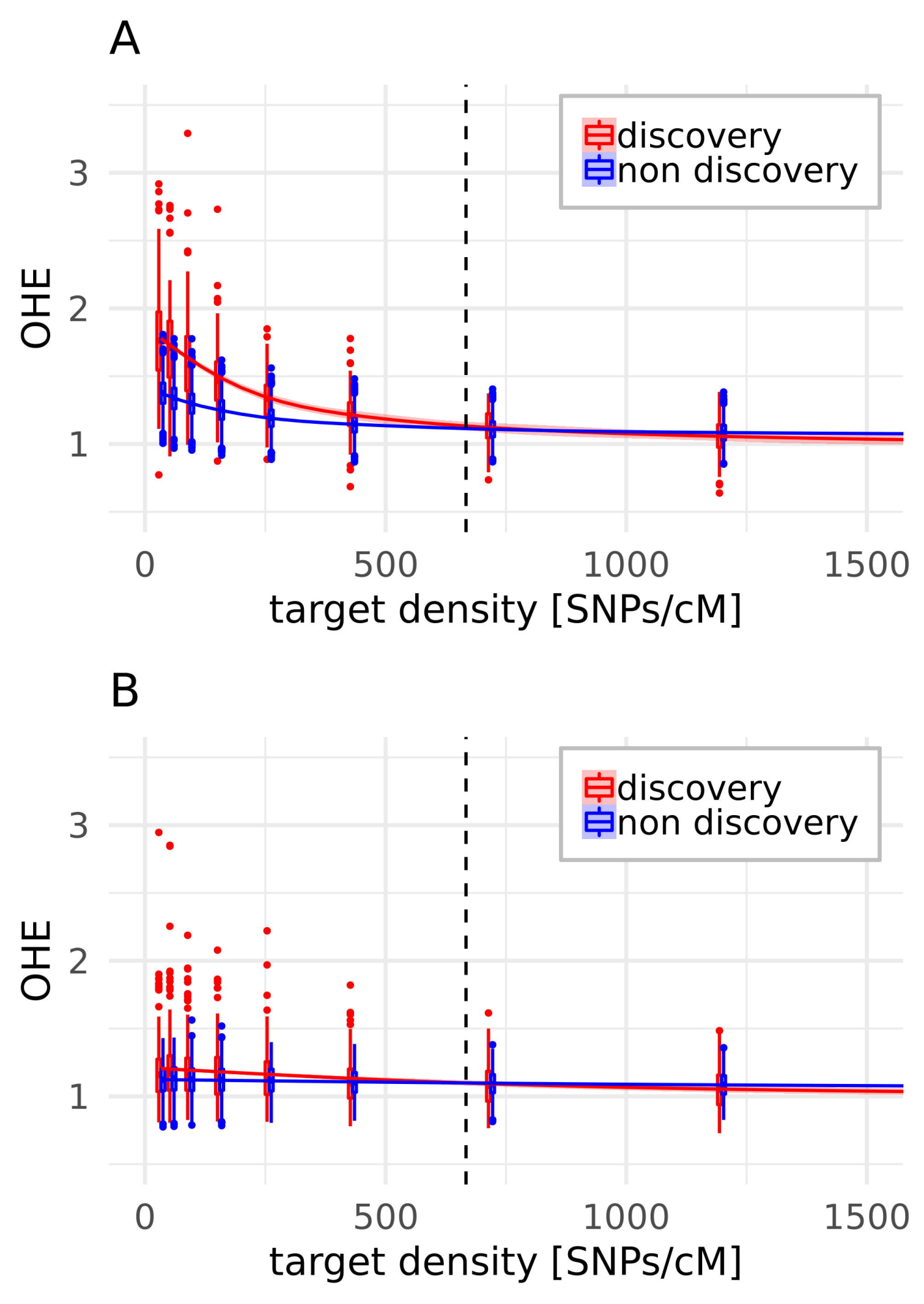

target density [SNPs/cM]
33
155
718
3333
density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
derived allele frequency


Unfiltered WGS
Filtered WGS
Array SNPs
1 - SNP Discovery
3 - Equal Spacing
H_{exp}
0.0
0.1
0.2
0.3
GV
SB
LE
BA
FG
TO
OH
RI
WL
WR
AB
WT
HO
OR
SE
OF
OM
SA
KG
BK
IT
YO
WY
ZC
PA
BH
KA
SN
KW
DL
BL
CG
NH
MA
CW
KS
DG
AS
CS
AR
MR
SH
GG
BR2
GS
BR1


1 - SNP discovery
2 - cluster removal
3 - equal spacing
4 - validation
5 - downsampling
OHE
2.0
1.5
1.0
0.5
discovery
validation
application
population groups


A
Discovery
Validation
Application
OHE
1.25
1.00
0.75
0.50
0.25
4
10
16
22
28
34
40
number of discovery populations


B
Discovery
Validation
Application
OHE
1.5
1.0
0.5
0.0
4
10
16
22
28
34
40
number of discovery populations


A
discovery
non discovery
OHE
3
2
1
0
500
1000
1500
target density [SNPs/cM]
B
discovery
non discovery
OHE
3
2
1
0
500
1000
1500
target density [SNPs/cM]