How Imputation Can Mitigate Ascertainment Bias

Johannes Geibel (johannes.geibel@uni-goettingen.de)
University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics 6 Group, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

Christian Reimer
University of Göttingen

Torsten Pook
University of Göttingen

Steffen Weigend
Friedrich-Loeffler-Institut

Annett Weigend
Friedrich-Loeffler-Institut

Henner Simianer
University of Göttingen

Research Article

Keywords: SNP ascertainment bias, imputation, chickens, population genetics

Posted Date: January 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-150008/v1

How Imputation Can Mitigate Ascertainment Bias

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

[1]University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany; johannes.geibel@uni-goettingen.de; creimer@gwdg.de; torsten.pook@uni-goettingen.de; hsimian@gwdg.de

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

[3]Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, Höltystrasse 10, 31535 Neustadt-Mariensee, Germany; steffen.weigend@fli.de; annett.weigend@fli.de

[*] johannes.geibel@uni-goettingen.de (Corresponding Author)

Abstract

Background

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

Results

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

and to 1.44 when using the larger but unbalanced reference panel. This generally supported the results from simulation but was less favorable, advocating for a larger reference panel when imputing to WGS.

Conclusions

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

Keywords

SNP ascertainment bias; imputation; chickens; population genetics

Background

To realize cost- and computational efficiency, many of the population genetic studies of the last 10 years for humans [1, 2], as well as for model- [3, 4] and agricultural species [5–8] were based on single nucleotide polymorphisms (SNP), which were genotyped by commercially available SNP arrays. Those arrays are based on a non-random selection (ascertainment) of SNPs, mostly performed in more intensively researched populations (e.g. commercially used livestock breeds). This results in a shift of the allele frequency spectrum towards mean allele frequencies and thereby an overestimation of heterozygosity compared to whole genome re-sequencing (WGS) data. As this overestimation is stronger for populations involved in the array design (discovery populations) than for those not involved [9], follow-up analyses can be biased. This effect is widely known as SNP Ascertainment Bias [10–12].

Different population genetic estimators are affected by SNP ascertainment bias to a different extent [11] and implementing bias-reduced estimators requires strong assumptions on the

60 design process of the used SNP array [13] which is often not public knowledge or too compli-
61 cated to be remodeled [14, 15]. Malomane *et al.* [16] therefore screened different raw data
62 filtering strategies on mitigation of ascertainment bias in SNP data and identified linkage prun-
63 ing to result in slightly decreasing ascertainment bias. Due to strongly decreasing sequencing
64 costs and the complexity of the ascertainment bias correction strategies, more and more stud-
65 ies started using WGS data for population genetic analysis during the last years [17–21]. How-
66 ever, costs for broad WGS based studies are still rather high, resulting in large-scale collabo-
67 rations such as the 1000 Genomes Project [22], the 1000 Bull Genomes Project [23], or the
68 1001 Arabidopsis Genomes Project [24].

69 A commonly used method to *in silico* increase the resolution of SNP data sets is imputation
70 [25]. Over the years a variety of imputation approaches [26–32] have been proposed that uti-
71 lize linkage, pedigree, and haplotype information. To increase the marker density, an addi-
72 tional reference panel of individuals that were genotyped/sequenced by the intended resolu-
73 tion is required to additionally infer information from SNPs missing on the respective lower
74 density study set.

75 Imputation-based studies mostly either used a reference panel of the same population as the
76 study set itself [33–35] or utilized large global reference panels as of the 1000 Genomes [22,
77 36, 37] or 1000 Bull genomes [23, 38] projects. Especially for admixed or small endangered
78 populations, the use of additional distantly related populations in the reference panel was
79 investigated. On one hand, Brøndum *et al.* [39], Ye *et al.* [40] and Rowan *et al.* [41] identified
80 multi-breed reference panels to increase imputation accuracy especially in admixed breeds
81 and for low frequent alleles when imputing from high-density genotypes to sequence data.
82 On the other hand, Berry *et al.* [42] observed that smaller within breed reference panels

(140 - 688 reference cattle individuals per breed) performed always superior compared to the combined across breed reference panel when imputing from low density to high-density array genotypes. Korkuć *et al.* [43] showed that adding 100 to 500 Holstein cattle sequences to a reference panel of 30 German Black Pied cattle significantly decreased the imputation accuracy in comparison to the pure panel when imputing from array to sequence data. Adding the same numbers of a multi-breed reference panel only outperformed the pure panel when at least 300 reference animals were added. Pook *et al.* [44] investigated the inclusion of chicken populations to the reference set which were differently distantly related to the study set. While error rates generally decreased for rare alleles, the inclusion of distantly related populations slightly increased error rates for previously good imputed SNPs. Overall, the ideal setup of a reference panel seems to be highly dependent on the application with positive effects for some, but also potential harm in other cases.

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

Material and Methods

Data

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

1) Individual sequence data of 68 chickens from 68 different populations, sequenced within the scope of the EU project Innovative Management of Animal Genetic Resources (IMAGE; www.imageh2020.eu) [45]. They were complemented by 25 sequences (17 + 8) from two commercial white layer lines, 25 sequences (19 + 6) from two commercial brown layer lines, and 40 sequences (20 each) from two commercial broiler lines [20]. In total 158 sequences from 74 populations.
2) Pooled sequence data from 37 populations (9-11 chickens per population) [16]. All except 4 chickens from two populations were part of set 3.
3) Genotypes of 1,566 chickens from 74 populations, either genotyped (sub-set of the Synbreed Chicken Diversity Panel; SCDP) [46] with the Affymetrix Axiom™ 580k Genome-Wide Chicken Array [47], or complemented from set 1.

The intersection of the used data sets is shown in Figure 1 and accession information of the raw data per sample can be found in Supplementary File 1. All three data sets came with their own characteristics. While individual sequences are considered to be the gold standard throughout this study, genotypes of the Affymetrix Axiom™ 580k Genome-Wide Chicken Array [47] are biased towards variation which is common in the commercial chicken lines [9] and pooled sequences only allow for an estimate of population allele frequencies and show a slight bias due to sample size and coverage (Supplementary File 2) [48, 49].

Figure 1: UpSet plot showing the distinct intersections of chickens between the used sequencing/ genotyping technologies. The left bar plot contains the total number of individuals that were genotyped (array), individually sequenced (indSeq), or pooled sequenced (poolSeq). The upper bar plot contains the number of individuals within each distinct intersection, indicated by the connected points below.

Calling of WGS SNPs and generation of genotype set

Alignment of the raw sequencing reads against the latest chicken reference genome GRCg6a [50] and SNP calling was conducted for individual and pooled sequenced data following GATK best practices [51, 52]. As the Affymetrix Axiom™ 580k Genome-Wide Chicken Array [47] does not contain enough SNPs on chromosomes 30 – 33 for imputation (and chromosome 29 is not annotated in the reference genome), only up to chromosome 28 was used. This resulted in 20,829,081 biallelic SNPs on chromosomes 1 - 28 which were used in further analyses. Additionally, all individual sequences were genotyped for the positions of the Affymetrix Axiom™ 580k Genome-Wide Chicken Array [47].

To ensure compatibility between Array- and WGS data, the genotypes of the Synbreed Chicken Diversity panel were lifted over from galGal5 to galGal6 and corrected for switches of reference and alternate alleles. Only SNPs with known autosomal position, call rates > 0.95 and genotype recall rates > 0.95 were further considered. MAF filters were later used when subsampling the different sets and thus not considered in this step. Further, missing genotypes were imputed using Beagle 5.0 [32] with ne=1000 [44] and the genetic map taken from Groenen *et al.* [53]. This resulted in a final set of 1,566 animals from 74 populations (18 - 37 animals per population) and 462,549 autosomal SNPs, further referred to as the **genotype set**.

As Malomane *et al.* [16] described LD-based pruning as an effective filtering strategy to minimize the impact of ascertainment bias in SNP array data, the genotype set was additionally LD pruned using plink 1.9 [54] with --indep 50 5 2 flag. This reduced the genotype set to 136,755 SNPs (30 %) and will be referred to as **pruned genotype set**.

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

Analyses based on simulation of ascertainment bias within the genotype set

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

1) allPop_74_740: Equally distributed across all populations by sampling one to ten chickens per population (74 - 740 reference samples).
2) randSamp_5_50: 5, 10, …, 50 randomly sampled chickens (5-50 reference samples).
3) randPop_5_50: Five chickens from each of one to ten randomly sampled populations (5 - 50 reference samples).
4) minPop_5_50: Five chickens from each of one to ten populations which were closest related to the discovery population, based on Nei's Distance ([55]; 5 - 50 reference samples).
5) maxPop_5_50: Five chickens from each of one to ten populations which were most distantly related to the discovery population, based on Nei's Distance ([55]; 5 - 50 reference samples).

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

above-mentioned strategies. Imputation of the *in silico* arrays to the reference set was performed by running Beagle 5.0 [32] with ne=1000 [44], the genetic distances taken from Groenen *et al.* [53] and the according reference set. The schematic workflow can be found in Figure 2.

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

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

Imputation of genotype set to sequence level

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

158_all), which resulted in a strong imbalance towards the two broiler lines (20 reference samples per broiler line).

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

Assessment of imputation accuracy

Assessment of imputation accuracy was done by using Pearson correlation (r) between true and imputed genotypes [42, 56] for the *in silico* designed arrays. Pearson correlation puts a higher relative weight on imputation errors in rare alleles than plain comparison of allele- or genotype concordance rates [56]. In case of the imputation to sequence level, no 'true' set was available for the evaluation of imputation accuracy. Instead, we used the internal Beagle quality measure, the dosage r-squared (DR2) [57]. This, however, has the drawback that it only shows the theoretical imputation accuracy, cannot capture biases due to biased reference sets and also does not allow for a per population evaluation of imputation accuracy.

Comparison of population genetic estimators

Ascertainment bias is most crucial when comparing populations with varying relatedness to the populations used for the SNP ascertainment [9]. To investigate these effects, we concentrated on two heterozygosity estimates: expected (H_E) and observed (H_O) heterozygosity; and two distance measurements: Wright's fixation index (F_{ST}) [58] and Nei's distance (D) [55].

H_O, as the proportion of heterozygous genotypes in a population, could only be calculated when the genotypic status of a population was known (individual sequences or genotypes). In

contrast, H_E could also be calculated from pooled sequences which allow the estimation of allele frequencies (p). Thereby, H_O and H_E (equation (1)) are calculated as average over all loci (l = 1, …, L).

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

As pooled sequence data comes with a slight but systematic underestimation of H_E ([48]; Supplementary File 2), H_E for pooled sequences was multiplied with the correction factor $\frac{n}{n-1}$, introduced by Futschik and Schlötterer [48], where n is the number of haplotypes in the pool. This partially corrected the H_E estimates for the bias introduced by pooled sequencing (Supplementary File 2).

D was calculated as given by equation (2), where D_{xy} accounts for the genetic distance between populations X and Y, while x_{il} and y_{il} represent the frequency of the i^{th} allele at the l^{th} locus in population X and Y, respectively.

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

Pairwise F_{ST} values between populations X and Y were estimated using equation (3), where HT_l accounts for the H_E within the total population at locus l and $\overline{HS}_l$ for the mean H_E within the two subpopulations at locus l [58].

$$F_{ST} = \frac{\sum_l (HT_l - \overline{HS}_l)}{\sum_l HT_l} \quad (3)$$

D and F_{ST} both show a downward bias that is comparable to HE when estimated from pooled data (Supplementary File 2). The effect of ascertainment bias is much larger than the effect of pooling for D. In contrast, F_{ST} is generally robust against the effects of ascertainment bias when a sufficiently large discovery panel was used for array development [11]. Therefore, it shows underestimation when calculated from pooled sequence data which is larger than the effect of ascertainment bias (Supplementary File 2). We therefore could not dissect the effects of the two biases in the comparisons on sequence level and did not include F_{ST} there.

Having no ascertainment bias would mean that estimates of a respective set would lie on the line of identity (diagonal) when regressing the set against the true values. The magnitude of the bias can therefore be defined as the distance of the estimates to that line. We therefore regressed the estimates from biased data (y_{ij}) on the unbiased ones (x_{ij}) while fitting group specific intercepts ($group_i$) as well as group-specific slopes ($group_i \times \beta_i$) and a random error ($\grave{o}_{ij}, \grave{o} \sim N(0, \mathbf{I}\sigma_e^2)$) as in equation (4).

$$y_{ij} = group_i + group_i \times \beta_i x_{ij} + \grave{o}_{ij} \quad (4)$$

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

distance estimates from biased and true set within groups were used as indicators for the magnitude of bias and random estimation error.

To get a measure for a fixed estimation error, we also calculated the mean overestimation across populations (j = 1 ... J) as in equation (5).

$$mean\ \text{overestimation} = \frac{\sum_j \frac{biased\ estimate_j - true\ estimate_j}{true\ estimate_j}}{J} \quad (5)$$

Note, that we did not have a population-wide estimate of allele frequency on sequence level for each population we did have single individual sequences for. Comparisons of population estimates on sequence level are therefore only based on 45 populations instead of all 74 populations which were used as study and reference set for the imputation process.

Results

In silico array to genotype

For the *in silico* array to genotype set imputation, median per-animal imputation accuracies (r) were already high with 0.94 when using one reference individual per population (allPop_74). Increasing the number of reference individuals subsequently increased the accuracy up to 0.99 for 10 reference individuals per population (allPop_740). The accuracy was consistently higher for individuals which were part of the discovery population (Figure 3).

Figure 3: Development of the per-animal imputation accuracy for the *in silico* array to genotype set imputation with an increasing number of reference animals per population. Individuals are grouped on whether they belong to the population used for SNP discovery or not and reference individuals were chosen as in scenario allPop_74_740. The lines show the trend of

283 the median and outliers are not shown in the plot as they do not add valuable information
284 due to the high number of repetitions.

285

286 Accuracies were lower for the other strategies, mainly due to fewer reference individuals
287 (maximum 50). However, Figure S 2 indicates that the best imputation accuracies were
288 achieved for populations which contained at least one reference individual. The discovery
289 populations showed slightly to strongly increased accuracies compared to the application pop-
290 ulations for all scenarios besides the scenario where the reference populations were chosen
291 to be maximum distant to the discovery population (scenario maxPop_5_50; Figure S 2). Ran-
292 domly choosing reference individuals across the complete population set (scenario rand-
293 Samp_5_50; Figure S 2) resulted in the overall best imputation accuracies. When assigning the
294 reference samples to randomly chosen populations in blocks of five samples per population
295 (scenario randPop_5_50, Figure S 2), the accuracies for the reference populations were only
296 slightly increased, while accuracies for discovery- and application populations dropped mas-
297 sively. Interestingly, the accuracies for application populations were higher in scenario
298 minPop_5_50 (Figure S 2), in which the reference populations were chosen according to their
299 minimum distance to the discovery population, than with the random reference population
300 selection from scenario randPop_5_50. A selection of reference populations to be maximum
301 distant to the discovery population (scenario maxPop_5_50, Figure S 2) performed overall
302 worst.

303 As expected, the *in silico* ascertained sets showed a strong overestimation of the H_E for nearly
304 all populations in all cases. The overestimation was much stronger for populations used for

SNP discovery (Figure 4 A). Imputation using an equal number of reference samples per population (scenario allPop_74_740) massively decreased this bias (Figure 4 B). The strength of the correction was increased by an increase in the number of reference populations.

Figure 4: True H_E vs. ascertained H_E (A) and imputed H_E (B) by population group. For the imputed case, the strategy of using the same number of reference samples per population (allPop_74_740) is shown, an increase in the number of reference samples per population (1-10) is marked by an increasing color gradient and the line of identity is marked by a solid black line.

To get an impression on the strength of the correction and the needed size of the reference panel, Figure 5 compares the correlation by population group, the slope for the within-group regression of the true H_E and H_O vs. the ascertained/ imputed cases and mean overestimation for strategy allPop_74_740. It shows that the effects of ascertainment bias were stronger for H_E than for H_O. Imputation when using the reference set with just one individual per population corrects the initially much lower correlation within population group to > 0.99. While slope and mean overestimation are also pushed nearly ultimately towards the intended values of one respectively zero for the non-discovery populations, there remains a small bias for the discovery populations, which decreases with an increasing number of reference samples.

Figure 5: Development of correlation within population group (A), slope (B) and mean overestimation (C) of the regression lines for the two heterozygosity estimates when distributing the reference samples equally across all populations (allPop_74_740). The intended value for unbiasedness and minimum variance is marked as dense black horizontal line. Note that the case without imputation is consistent with zero reference samples.

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

Population distances

The effects of ascertainment bias were also present but less pronounced for the distance measurements (D and F_{ST}; Figure S 4). The bias was thereby mainly present, when estimating the distances between populations which belong to differently strong biased population groups (Figure S 5). Note that F_{ST} was, all in all, less affected than D.

Genotype to sequence

For the imputation from array to sequence, the smallest reference panel (74_1perLine) showed a median DR2 of 0.93 with 5 % and 25 % quantiles being 0.73 and 0.88. Increasing the number of reference samples from the commercial populations to 158_all increased the 5% quantile of DR2 by 0.05 while increasing all higher quantiles only by 0.01 (Table S 1). Note that chromosomes 16, 22 and 25 clearly showed a higher proportion of badly imputed SNPs than the other chromosomes (Figure S 6; Figure S 7).

The effect of imputation to WGS on ascertainment bias of HE is shown in Figure 6. Given the situation that we cannot completely exclude pooling bias for the pooled sequenced samples (Supplementary File 2), only the effect on the individually sequenced samples can be discussed

with adequate reliability. While the array-based estimates showed a slope of 1.94, the linkage pruning slightly reduced this slope to 1.71. The clearly best result was achieved with imputation to WGS (slope = 1.26; 74_1perLine; Figure 6 A). However, the effect was also observed for all samples and they additionally showed a slight increase of correlation (0.97 to 0.98), which however is also influenced by pooling bias. Slightly increasing the reference panel (Figure 6 B) up to five samples per commercial line (98_5perLine) does not show any effect, while using all commercial samples in the reference panel (158_all) and thereby clearly biasing the reference panel towards the broiler samples increases HE again for all samples (slope = 1.44).

Figure 6: Effect of different correction strategies on ascertainment bias for expected heterozygosity (HE). A – uncorrected array, linkage pruned array and imputed array (reference set 74_1perLine) based vs. sequence-based HE. B – array imputed with different reference sets vs. sequence-based HE. The solid black line represents the line of identity, the solid colored lines are regression lines within the individually sequenced populations (larger points) and the dashed lines regression lines within all populations which include individually and pooled (small points) sequenced populations.

The results for Nei's standard genetic distance (D; Figure 7) showed the same pattern as the results for HE. The slope for distances between individually sequenced populations decreased from 2.86 (array) and 1.77 (array_pruned) to 1.38 (imputed, 74_1perLine). The unbalanced reference panel 158_all then again increased the slope to 1.56. The correlation for all distances, besides being also influenced by pooling bias and therefore being a rough estimate, was increased from 0.93 (array) respectively 0.95 (array_pruned) to 0.98 (all reference sets).

Figure 7: Effect of different correction strategies on ascertainment bias for Nei's standard genetic distance (D). A – uncorrected array, linkage pruned array and imputed array (reference set 74_1perLine) based vs. sequence-based D. B – array imputed with different reference sets

vs. sequence-based HE. The solid black line represents the line of identity, the solid colored lines are regression lines within distances between individually sequenced populations (larger points) and the dashed lines regression lines within distances between all populations which include individually and pooled (small points) sequenced populations.

Discussion

Overall performance of the correction method

Imputation of SNP data sets from lower to higher density is a commonly used technique to either increase the resolution of data sets [33, 34, 38] or make them comparable across different platforms [59, 60]. The according studies mostly use a relatively homogeneous study set and a closely related and large reference set [33, 34]. However, studies exist which investigate the effect of increasing the reference set to a multi-population reference set to make use of the increased number of reference haplotypes [39–43]. To our knowledge, we here present the first study which investigates the use of a relatively small and diverse reference set on a large and diverse study set to correct for a genotyping platform-specific bias, the SNP ascertainment bias.

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

Especially the *in silico* ascertained SNP arrays showed that even a very small reference panel consisting of one individual of each population showed very good results for all investigated estimators (Figure 5, Figure S 4) and became better with an increasing number of reference populations. The results were less beneficial for the real WGS data, but also showed a strong

405 decrease of the slope towards one. From the imputed *in silico* arrays, we could additionally
406 realize a fast closing of the gap of the stronger overestimation of heterozygosity within dis-
407 covery populations and the less severe overestimation in non-discovery populations. This also
408 seemed to be the case when imputing to WGS level where we observed that the slope within
409 the commercial populations (closely related to discovery populations of the real array) de-
410 creased more than the slope within all populations due to imputation. However, this observa-
411 tion in the WGS data has to be regarded with caution, as we additionally identified a non-
412 negligible bias due to pooled sequencing which interfered with the assessment of ascertain-
413 ment bias and which was, in our study, confounded with the difference between commercial
414 populations (sequenced individually) and non-commercial populations (sequenced as pools).

415 The use of WGS information via imputation also consistently showed better results in regard
416 of reduction of ascertainment bias than using linkage pruned array SNPs which was reported
417 to be an effective filtering strategy for ascertainment bias mitigation by Malomane *et al.* [16].

418 Generally, the effect of imputation on the investigated estimators was shown to be compara-
419 ble across estimators, regardless of their initial reaction to ascertainment bias. An interesting
420 side observation was that F_{ST} did not show any ascertainment bias on the real array data (Fig-
421 ure S 10) when calculated in the form of summing the numerator across SNPs and dividing by
422 the sum of the denominator as calculated in this study. F_{ST} was only affected when used to
423 estimate differentiation between the, in regard of heterozygosity differently strong affected,
424 discovery- and non-discovery populations in the simulated array data. This strongly supports
425 the findings of Albrechtsen *et al.* [11], who showed F_{ST} to be relatively robust against the ef-
426 fects of ascertainment bias.

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

Results from literature suggest that multi-breed reference panels generally increase imputation accuracy especially for rare variants and within admixed populations [39–41]. Additionally, Rowan *et al.* [41] argue that they do not seem to introduce variation at a relevant scale for markers for which the breeds are actually fixed. However, some studies also showed that strongly unbalanced reference sets can reduce imputation accuracy [42, 43]. In this study, including additional reference samples in a biased way when going from reference set 74_1perLine to 158_all increased the effects of ascertainment bias. However, theoretical imputation accuracies rather increased than decreased (Figure S 6; Table S 1) for previously poorly imputed SNPs. On one hand, this effect supports the findings of Brøndum [39], Rowan *et al.* [41] and Ye *et al.* [40] and on the other hand, nicely highlights the main reason for ascertainment bias. One can only identify variation which is present in the investigated sample. When developing an array, it is the variation in the discovery set, while in our case it is the reference set used for imputation. Therefore, it is crucial to use a reference set for imputation which covers the intended range of variation.

Besides the previously described effects of imputation on ascertainment bias, we also identified an effect of array design on imputation accuracy. Discovery populations show higher imputation accuracies than non-discovery populations (Figure 3). As markers on arrays are more representative for discovery populations than non-discovery populations, relatively more of the genetic variability in discovery populations is explained by the array and imputation is more accurate on average.

Conclusion

Imputation is generally able to mitigate ascertainment bias. The effect is already present when using a very small reference set of only one sequenced individual per population. It also performs better than simple filtering strategies based on the array data alone. However, care has to be taken in designing an evenly spaced reference panel to not introduce a new bias towards variation present in the reference panel. We also suggest using a larger reference panel than the one which was available for this study to achieve better results. Additionally, we observed an effect of array design on imputation accuracy as discovery populations showed a higher imputation accuracy than non-discovery populations. This should be taken into account when designing studies based on imputed SNPs by choosing an appropriate genotyping array for the intended study populations.

Abbreviations

D – Nei's Standard Genetic Distance

F_{ST} – Wright's Fixation Index

H_E – expected heterozygosity

H_O – observed heterozygosity

MAF – minor allele frequency

r – Pearson Correlation

SNP – single nucleotide polymorphism

WGS – whole genome sequencing

Declarations

Ethics approval and consent to participate

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

Consent for publication

Not applicable

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests

Funding

The positions of JG and TP were partly funded by the project IMAGE - Innovative Management of Animal Genetic Resources (www.imageh2020.eu, funded by the EU Horizon 2020 research and innovation program No 677353). The sampling, genotyping and sequencing of the chicken populations involved funding by the EC project AVIANDIV (EC Contract No. BIO4-CT98_0342; 1998 – 2000; www.aviandiv.fli.de), the German Federal Ministry of Education and Research (BMBF) via the SYNBREED project (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de) and the EU project IMAGE - Innovative Management of Animal Genetic Resources (www.imageh2020.eu, funded by the EU Horizon 2020 research and innovation program No 677353). Further, the publication fees were covered by the Open Access Publication Funds of the University of Goettingen.

Authors' contributions

JG conceptualized and designed the study, analyzed and interpreted the data, wrote the initial draft and revised the manuscript. CR and TP substantially contributed to design of the study, interpretation of the data and revision of the manuscript. SW substantially contributed to ac-

quisition and interpretation of the data and revised the manuscript. AW substantially contributed to acquisition and curation of the data. HS substantially contributed to conception and design of the study, interpretation of the data and revision of the manuscript. All authors have red and agreed on the final draft.

Acknowledgement

Not applicable

Supplementary

Supplementary File 1: Accession Information of raw data per sample

Supplementary File 2: Supplementary Methods

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

Figure S 2: Development of the per-animal imputation accuracy with an increasing number of reference animals per population. A – scenario randSamp_5_50; B – scenario randPop_5_50; C – scenario minPop_5_50; D – scenario maxPop_5_50. Individuals are grouped on whether they belong to the population which contains reference individuals, was used as for SNP discovery or none of them (application). the lines show the trend of the median.

Figure S 3: Development of correlations within population group (r), slope and mean overestimation of the regression lines for H_E and H_O estimates and different reference panel strategies. The intended value for unbiasedness and minimum variance is marked as dense black horizontal line. Note that the case without imputation is consistent with zero reference samples.

Figure S 4: Development of correlation within population group (A), slope (B) and intercept (C) of the regression lines for D and F_{ST} when distributing the reference samples equally over all populations (allPop_74_740). The intended value for unbiasedness and minimum variance is marked as dense black horizontal line. Note that the case without imputation is consistent with zero reference samples.

Figure S 5: Development of correlation within population group (r), slope and mean overestimation of the regression lines for Nei's Distance (D) and F_{ST} estimates and different reference panel strategies. The intended value for unbiasedness and minimum variance is marked as dense black horizontal line. Note that the case without imputation is consistent with zero reference samples.

Figure S 6: Distribution of DR2 values by chromosome and reference set. Note that outliers are not shown due to a large number of underlying values.

Figure S 7: Two-dimensional distributions of DR2 values vs. MAF by chromosome when imputed with the reference set 74_1perLine. The red line represents the median within 0.05 MAF bins.

Figure S 8: Effect of pooled sequencing and the correction factor of Futschik and Schlötterer [48] on expected heterozygosity (HE) and ascertainment bias. A – HE estimated from array positions of the sequencing data vs. HE directly estimated from array data. The color indicates the state before and after correcting the pooled sequence estimates and the accordingly colored solid lines the group specific regression lines while the black solid line indicates the line of identity in all three plots. The plot therefore shows the magnitude of the bias introduced by pooled sequencing and the according effect of the correction factor. B – HE estimated from the array data vs. HE estimated from the complete sequence data. The color again shows the values before and after implementing the correction of the pooled sequence estimates. While the solid regression line and dense circles indicate the individually sequenced samples, the dashed regression lines and triangles indicate pooled sequenced samples. The plot therefore shows the combined effect of ascertainment bias and pooled sequencing bias. C – HE estimated from array positions of the sequencing data vs. HE estimated from all positions of the sequencing data. The plot therefore shows the pure ascertainment bias.

Figure S 9: Effect of pooled sequencing on the expression of the ascertainment bias in Nei's standard genetic distance (D). The biased D was either estimated directly from the array genotypes (D.arr, pooled bias + ascertainment bias) or from the array positions of the sequencing data (D.arr.seq, pure ascertainment bias), while the estimates from the complete sequence were assumed to be the true estimates. The black solid line represents the line of identity, solid colored regression lines and dense points represent estimates between individually sequenced populations and dashed lines and triangles represent estimates between two populations of which at least one was pooled sequenced.

Figure S 10: Effect of pooled sequencing on the expression of the ascertainment bias in Wright's fixation index (F_{ST}). The biased F_{ST} was either estimated directly from the array genotypes (FST.arr, pooled bias + ascertainment bias) or from the array positions of the sequencing data (FST.arr.seq, pure ascertainment bias), while the estimates from the complete sequence were assumed to be the true estimates. The black solid line represents the line of identity, solid colored regression lines and dense points represent estimates between individually sequenced populations and dashed lines and triangles represent estimates between two populations of which at least one was pooled sequenced.

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

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.
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.
5. Travis AJ, Norton GJ, Datta S, Sarma R, Dasgupta T, Savio FL, et al. Assessing the genetic diversity of rice originating from Bangladesh, Assam and West Bengal. Rice. 2015;8:35. doi:10.1186/s12284-015-0068-z.
6. Mayer M, Unterseer S, Bauer E, Leon N de, Ordas B, Schön C-C. Is there an optimum level of diversity in utilization of genetic resources? Theor Appl Genet. 2017;130:2283–95. doi:10.1007/s00122-017-2959-4.
7. Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, et al. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc. Natl. Acad. Sci. 2008;105:17312–7. doi:10.1073/pnas.0806569105.

8. Gibbs RA, Taylor JF, van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324:528–32. doi:10.1126/science.1167936.
9. Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. How Array Design Affects SNP Ascertainment Bias. bioRxiv. 2019:833541. doi:10.1101/833541.
10. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005;15:1496–502.
11. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol. Biol. Evol. 2010;27:2534–2547.
12. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays. 2013;35:780–6.
13. Nielsen R. Population genetic analysis of ascertained SNP data. Hum. Genomics. 2004;1:1.
14. Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004;168:2373–82.
15. Quinto-Cortés CD, Woerner AE, Watkins JC, Hammer MF. Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. Sci. Rep. 2018;8:10209. doi:10.1038/s41598-018-28539-y.
16. Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics. 2018;19:22. doi:10.1186/s12864-017-4416-9.
17. Qanbari S, Pausch H, Jansen S, Somel M, Strom T-M, Fries R, et al. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014;10:e1004148. doi:10.1371/journal.pgen.1004148.

18. Qanbari S, Seidel M, Strom T-M, Mayer KFX, Preisinger R, Simianer H. Parallel Selection Revealed by Population Sequencing in Chicken. Genome Biol. Evol. 2015;7:3299–306. doi:10.1093/gbe/evv222.
19. Lawal RA, Al-Atiyat RM, Aljumaah RS, Silva P, Mwacharo JM, Hanotte O. Whole-Genome Resequencing of Red Junglefowl and Indigenous Village Chicken Reveal New Insights on the Genome Dynamics of the Species. Front Genet. 2018;9:264. doi:10.3389/fgene.2018.00264.
20. Qanbari S, Rubin C-J, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of adaptation in modern chicken. PLoS Genet. 2019;15:e1007989. doi:10.1371/journal.pgen.1007989.
21. Peripolli E, Reimer C, Ha N-T, Geibel J, Machado MA, Panetto, João Cláudio do Carmo, et al. Genome-wide detection of signatures of selection in indicine and Brazilian locally adapted taurine cattle breeds using whole-genome re-sequencing data. BMC Genomics. 2020;21:624. doi:10.1186/s12864-020-07035-6.
22. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. doi:10.1038/nature15393.
23. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annual Review of Animal Biosciences. 2019;7:89–102. doi:10.1146/annurev-animal-020518-115024.
24. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166:481–91. doi:10.1016/j.cell.2016.05.063.

649 25. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat.
650 Rev. Genet. 2010;11:499 EP -. doi:10.1038/nrg2796.
651 26. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination
652 hotspots using single-nucleotide polymorphism data. Genetics. 2003;165:2213–33.
653 27. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for ge-
654 nome-wide association studies by imputation of genotypes. Nat. Genet. 2007;39:906–
655 13.
656 28. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method
657 for the Next Generation of Genome-Wide Association Studies. PLoS Genet. 2009;5:1–15.
658 doi:10.1371/journal.pgen.1000529.
659 29. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of
660 genomes. Nature Methods. 2012;9:179–81.
661 30. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputa-
662 tion using information from relatives. BMC Genomics. 2014;15:478. doi:10.1186/1471-
663 2164-15-478.
664 31. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. LinkImpute: Fast
665 and Accurate Genotype Imputation for Nonmodel Organisms. G3. 2015;5:2383.
666 doi:10.1534/g3.115.021667.
667 32. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Genera-
668 tion Reference Panels. Am. J. Hum. Genet. 2018;103:338–48.
669 doi:10.1016/j.ajhg.2018.07.015.
670 33. Pausch H, Aigner B, Emmerling R, Edel C, Götz K-U, Fries R. Imputation of high-density
671 genotypes in the Fleckvieh cattle population. Genet Sel Evol. 2013;45:3.
672 doi:10.1186/1297-9686-45-3.

673 34. Heidaritabar M, Calus MPL, Megens H-J, Vereijken A, Groenen MAM, Bastiaansen JWM.
674 Accuracy of genomic prediction using imputed whole-genome sequence data in white
675 layers. J Anim Breed Genet. 2016;133:167–79. doi:10.1111/jbg.12199.
676 35. van den Berg S, Vandenplas J, van Eeuwijk FA, Bouwman AC, Lopes MS, Veerkamp RF.
677 Imputation to whole-genome sequence using multiple pig populations and its use in ge-
678 nome-wide association studies. Genetics Selection Evolution. 2019;51:2.
679 doi:10.1186/s12711-019-0445-y.
680 36. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identi-
681 fies novel and refined associations for the Wellcome Trust Case Control Consortium
682 phase 1 Data. European Journal Of Human Genetics. 2012;20:801–5.
683 doi:10.1038/ejhg.2012.3.
684 37. Artigas MS, Wain LV, Miller S, Kheirallah AK, Huffman JE, Ntalla I, et al. Sixteen new lung
685 function signals identified through 1000 Genomes Project reference panel imputation.
686 Nat. Commun. 2015;6:8658. doi:10.1038/ncomms9658.
687 38. Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. Utility of
688 whole-genome sequence data for across-breed genomic prediction. Genetics Selection
689 Evolution. 2018;50:27. doi:10.1186/s12711-018-0396-8.
690 39. Brøndum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to
691 whole genome sequence using a single or multi-breed reference population in cattle.
692 BMC Genomics. 2014;15:728. doi:10.1186/1471-2164-15-728.
693 40. Ye S, Yuan X, Huang S, Zhang H, Chen Z, Li J, et al. Comparison of genotype imputation
694 strategies using a combined reference panel for chicken population. Animal.
695 2019;13:1119–26. doi:10.1017/S1751731118002860.

41. Rowan TN, Hoff JL, Crum TE, Taylor JF, Schnabel RD, Decker JE. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. Genetics Selection Evolution. 2019;51:77. doi:10.1186/s12711-019-0519-x.
42. Berry DP, McClure MC, Mullen MP. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J Anim Breed Genet. 2014;131:165–72. doi:10.1111/jbg.12067.
43. Korkuć P, Arends D, Brockmann GA. Finding the Optimal Imputation Strategy for Small Cattle Populations. Front Genet. 2019;10:52. doi:10.3389/fgene.2019.00052.
44. Pook T, Mayer M, Geibel J, Weigend S, Cavero D, Schoen CC, Simianer H. Improving Imputation Quality in BEAGLE for Crop and Livestock Data. G3. 2019:g3.400798.2019. doi:10.1534/g3.119.400798.
45. Bortoluzzi C, Megens H-J, Bosse M, Derks MFL, Dibbits B, Laport K, et al. Parallel Genetic Origin of Foot Feathering in Birds. Mol. Biol. Evol. 2020;37:2465–76. doi:10.1093/molbev/msaa092.
46. Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The SYNBREED chicken diversity panel: A global resource to assess chicken diversity at high genomic resolution. BMC Genomics. 2019;20:345. doi:10.1186/s12864-019-5727-9.
47. Kranis A, Gheyas AA, Boschiero C, Turner F, Le Yu, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59. doi:10.1186/1471-2164-14-59.
48. Futschik A, Schlötterer C. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. Genetics. 2010;186:207–18. doi:10.1534/genetics.110.114397.

719 49. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals [mdash] mining
720 genome-wide polymorphism data without big funding. Nat. Rev. Genet. 2014;15:749–
721 63.
722 50. Genome Reference Consortium GRCg6a. GRCg6a chicken reference genome. 2018.
723 http://hgdownload.soe.ucsc.edu/goldenPath/galGal6/bigZips/galGal6.fa.gz. Accessed 9
724 Apr 2019.
725 51. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for
726 variation discovery and genotyping using next-generation DNA sequencing data. Nat.
727 Genet. 2011;43:491. doi:10.1038/ng.806.
728 52. van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et
729 al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best
730 practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1-11.10.33.
731 doi:10.1002/0471250953.bi1110s43.
732 53. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, et al. A
733 high-density SNP-based linkage map of the chicken genome reveals sequence features
734 correlated with recombination rate. Genome Res. 2009;19:510–9.
735 54. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK:
736 Rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
737 doi:10.1186/s13742-015-0047-8.
738 55. Nei M. Genetic Distance between Populations. The American Naturalist. 1972;106:283–
739 92.
740 56. Hickey JM, Crossa J, Babu R, los Campos G de. Factors Affecting the Accuracy of Geno-
741 type Imputation in Populations from Several Maize Breeding Programs. Crop Science.
742 2012;52:654. doi:10.2135/cropsci2011.07.0358.

57. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 2009;84:210–23. doi:10.1016/j.ajhg.2009.01.005.
58. Wright S. The genetical structure of populations. Ann. Eugen. 1949;15:323–54. doi:10.1111/j.1469-1809.1949.tb02451.x.
59. Al-Tassan NA, Whiffin N, Hosking FJ, Palles C, Farrington SM, Dobbins SE, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci. Rep. 2015;5:10442. doi:10.1038/srep10442.
60. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat. Genet. 2018;50:362–7. doi:10.1038/s41588-018-0056-5.

Figures

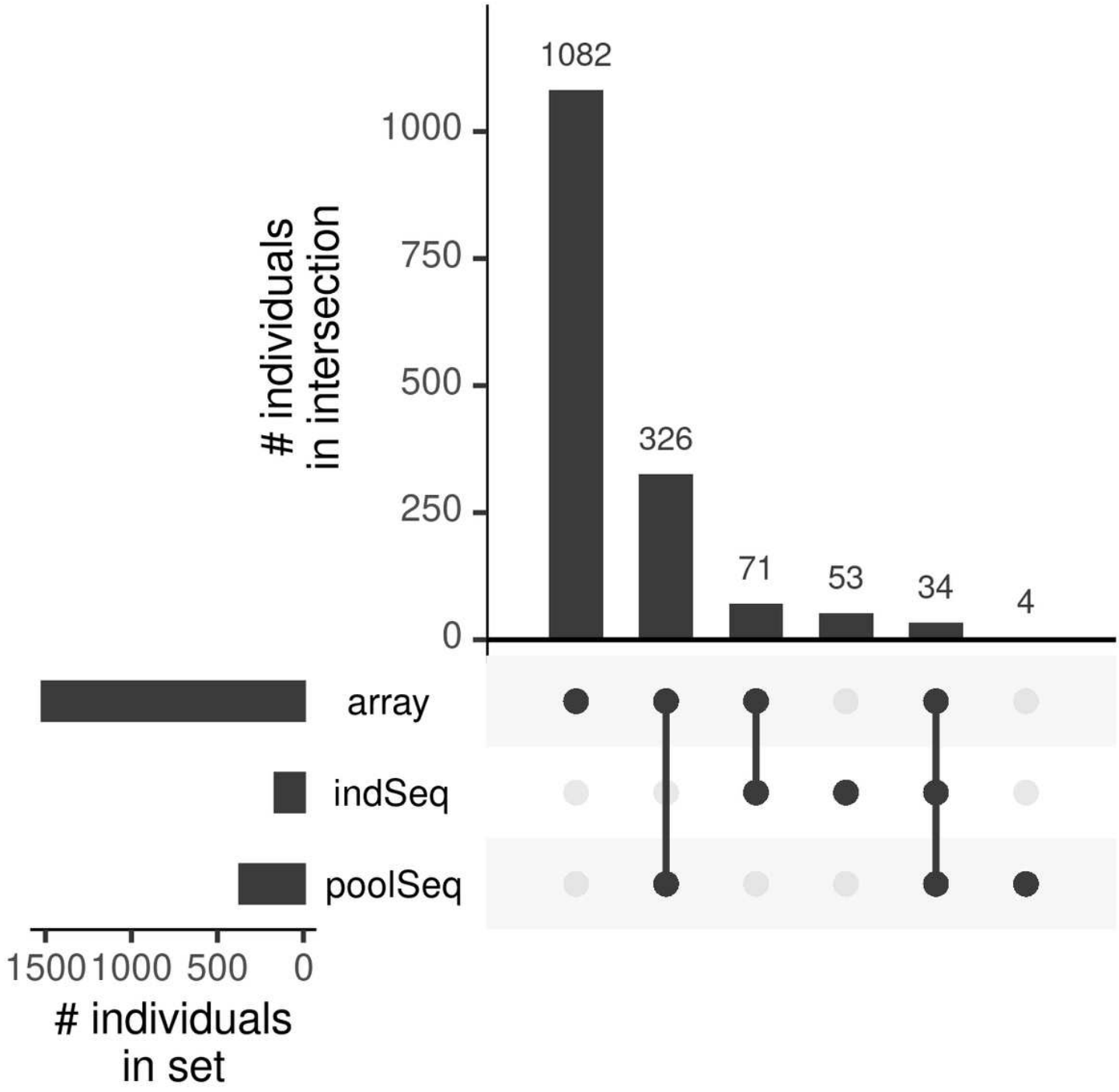

Figure 1

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

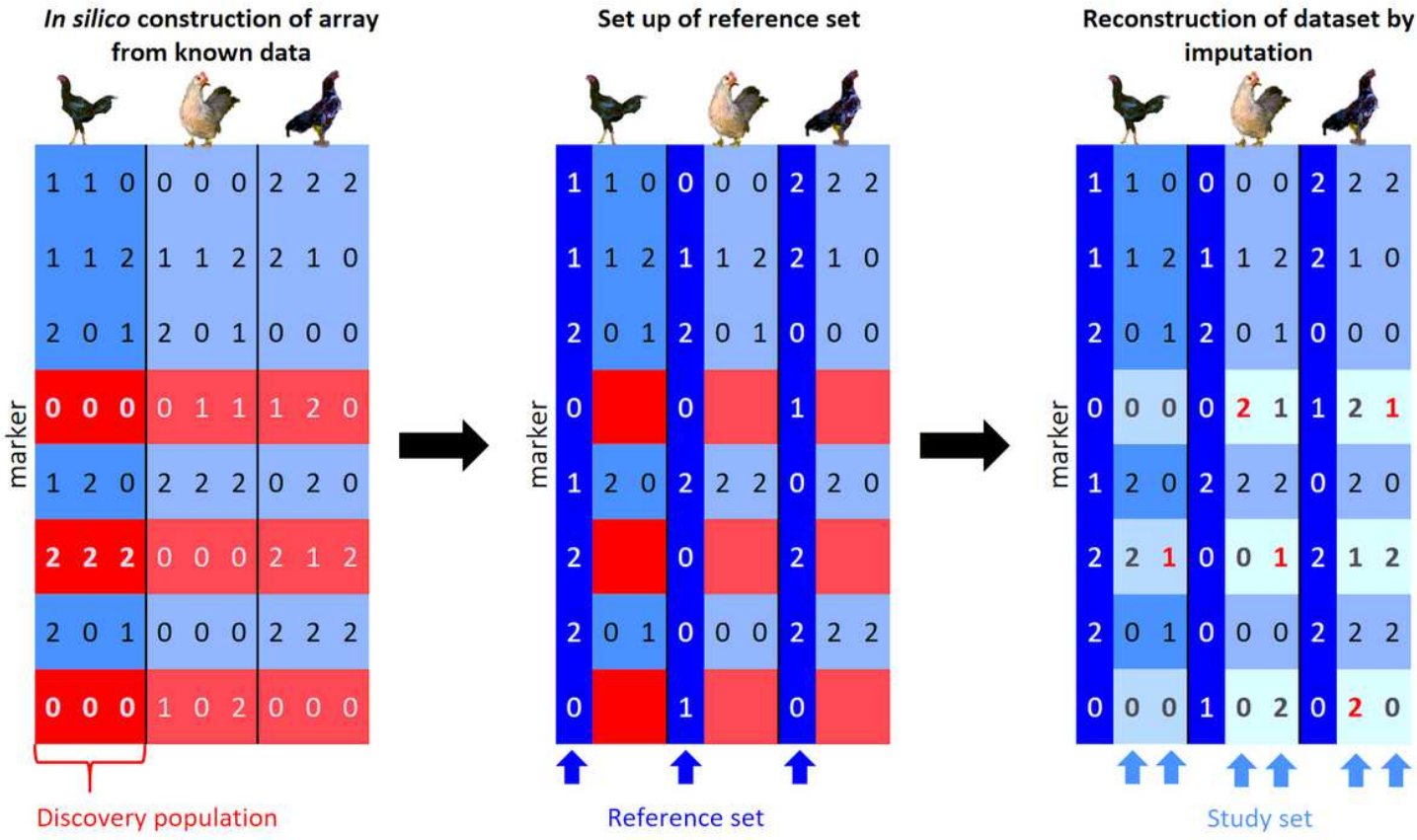

Figure 2

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

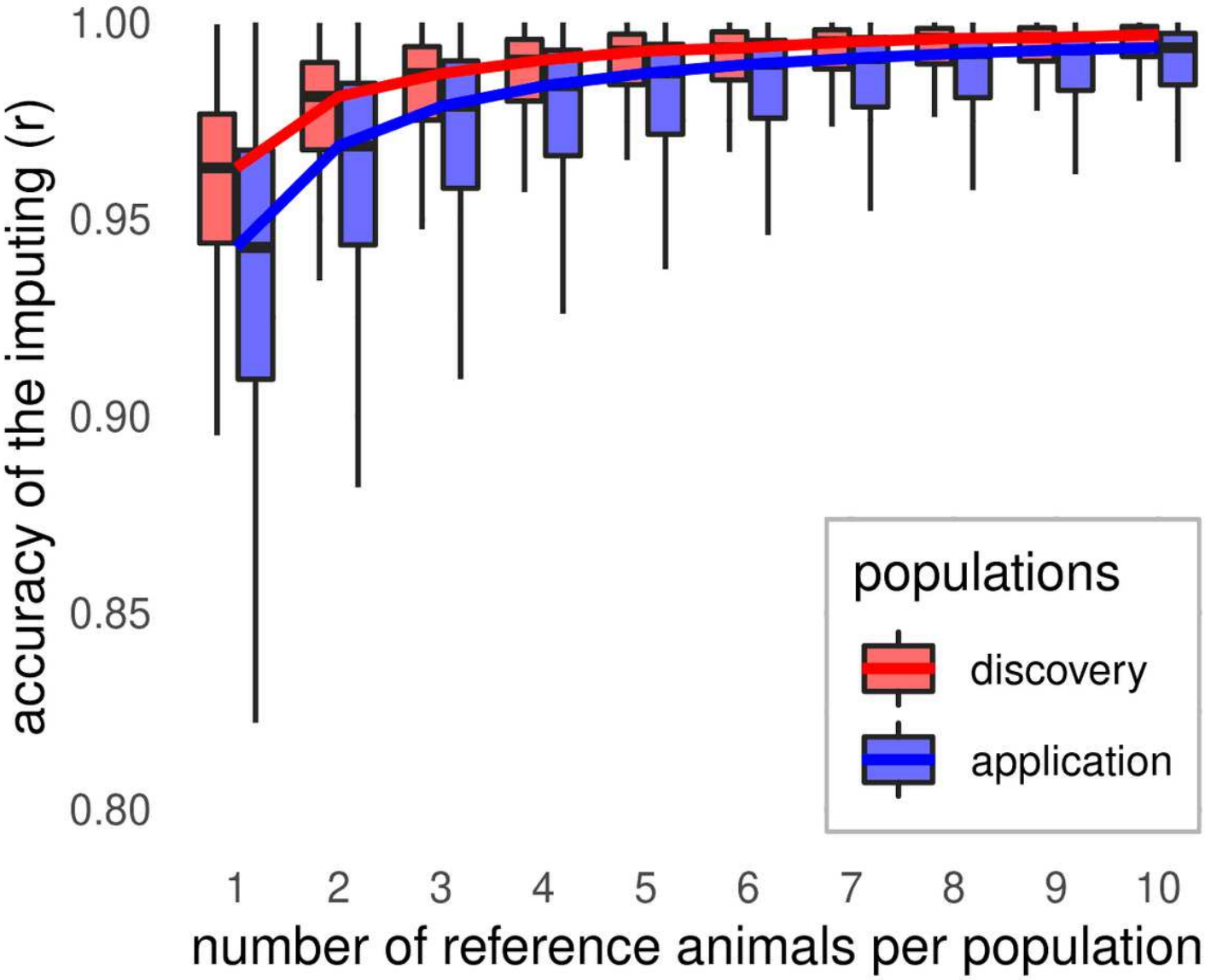

Figure 3

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

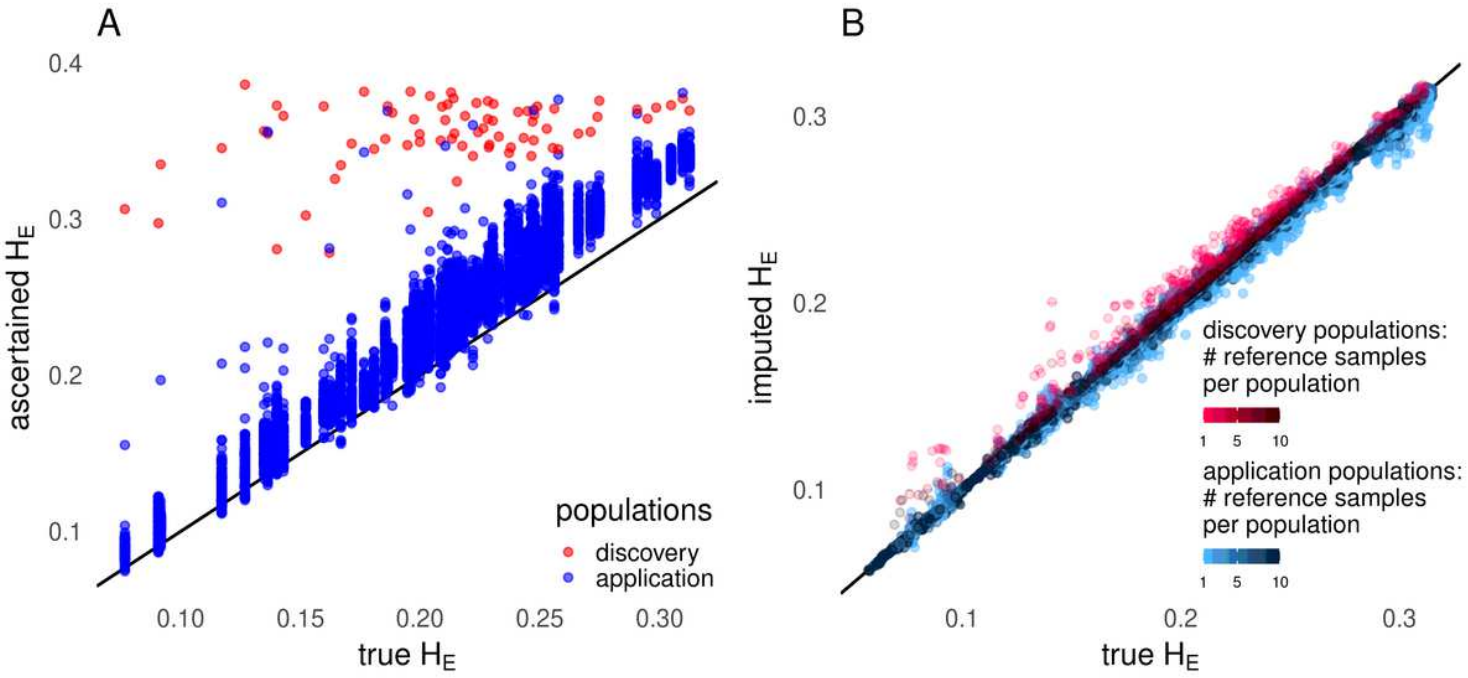

Figure 4

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

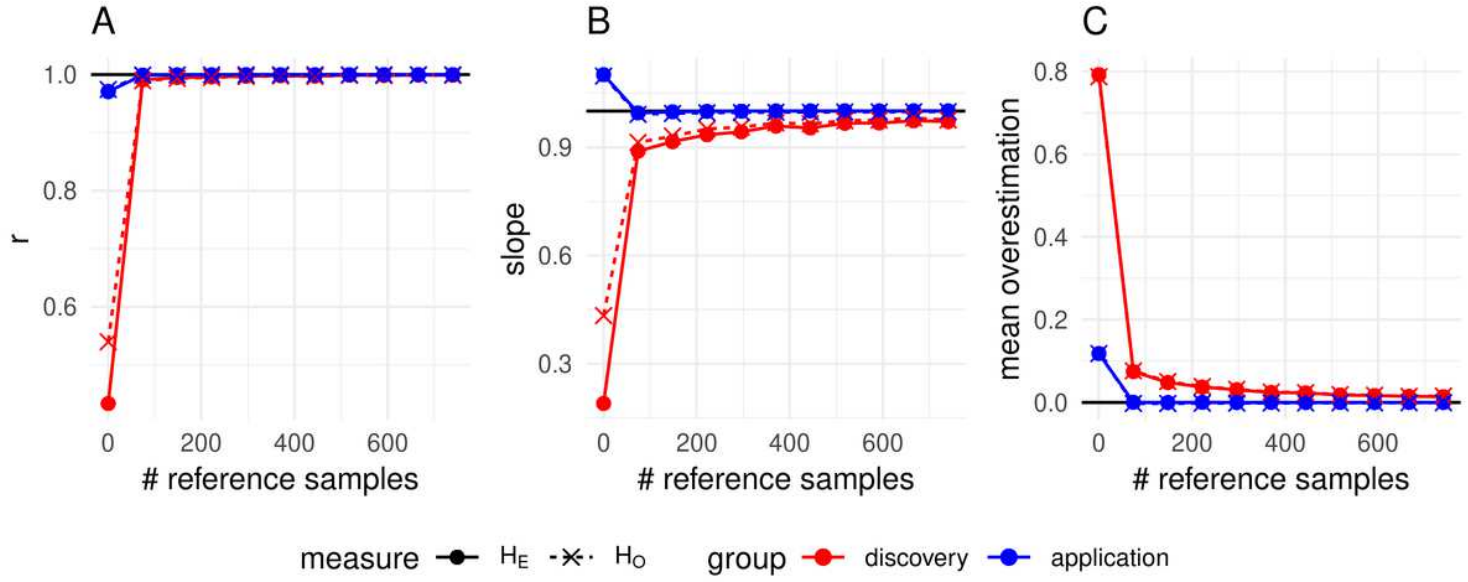

Figure 5

Development of correlation within population group (A), slope (B) and mean overestimation (C) of the regression lines for the two heterozygosity estimates when distributing the refer-ence samples equally across all populations (allPop_74_740). The intended value for unbi-asedness and minimum variance is

marked as dense black horizontal line. Note that the case without imputation is consistent with zero reference samples.

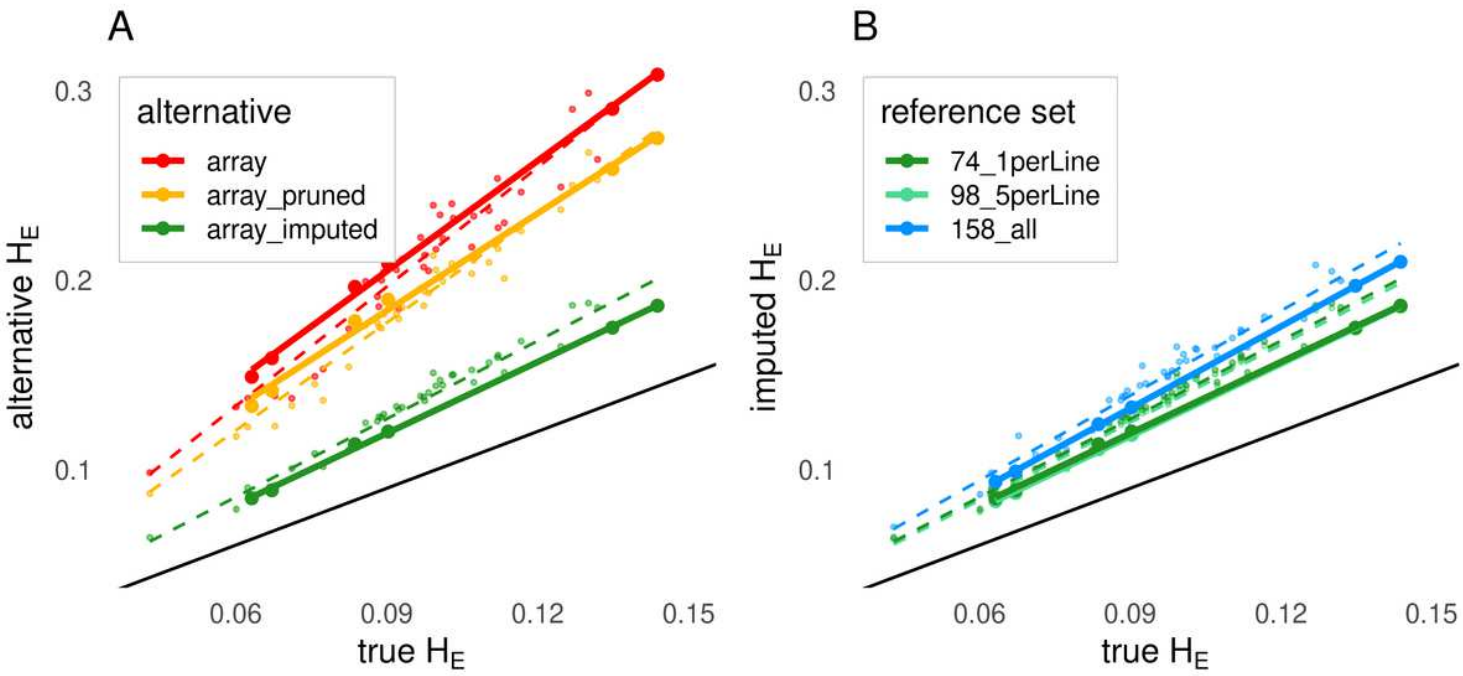

Figure 6

Effect of different correction strategies on ascertainment bias for expected heterozygosity (HE). A – uncorrected array, linkage pruned array and imputed array (reference set 74_1perLine) based vs. sequence-based HE. B – array imputed with different reference sets vs. sequence-based HE. The solid black line represents the line of identity, the solid colored lines are regression lines within the individually sequenced populations (larger points) and the dashed lines regression lines within all populations which include individually and pooled (small points) sequenced populations.

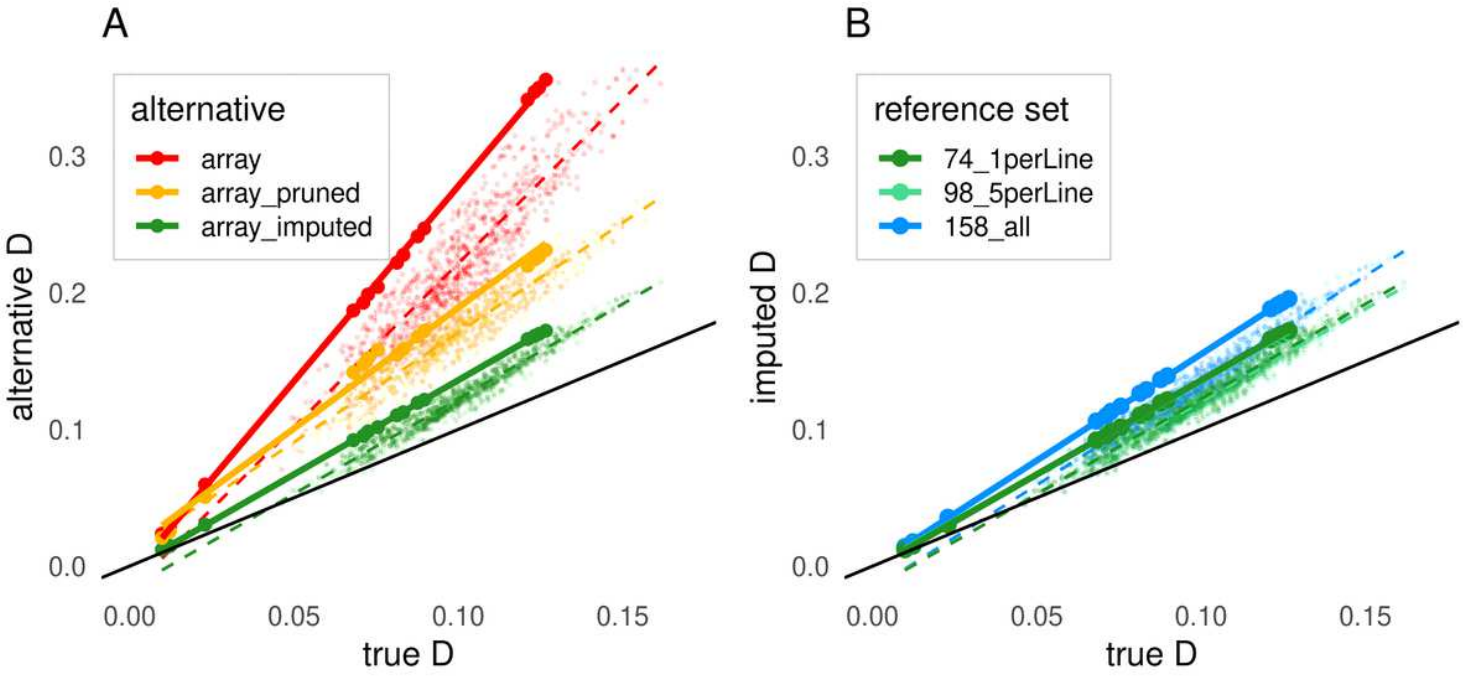

Figure 7

Effect of different correction strategies on ascertainment bias for Nei's standard genetic distance (D). A – uncorrected array, linkage pruned array and imputed array (reference set 74_1perLine) based vs. sequence-based D. B – array imputed with different reference sets vs. sequence-based HE. The solid black line represents the line of identity, the solid colored lines are regression lines within distances between individually sequenced populations (larg-er points) and the dashed lines regression lines within distances between all populations which include individually and pooled (small points) sequenced populations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.tiff
- FigureS2.tiff
- FigureS3.tiff
- FigureS4.tiff
- FigureS5.tiff
- FigureS6.tiff
- FigureS7.tiff
- FigureS8.tiff
- FigureS9.tiff
- FigureS10.tiff
- SupplementaryFile1.csv
- SupplementaryFile2.pdf
- TableS1.docx