High imputation accuracy from informative low-to-medium density single nucleotide polymorphism genotypes is achievable in sheep

Aine C. O'Brien, Michelle M. Judge, Seán Fair, Donagh P. Berry

High imputation accuracy from informative low-density to medium-density single nucleotide polymorphism genotypes is achievable in sheep[1]

Aine C. O'Brien*†, Michelle M. Judge*, Sean Fair†, Donagh P. Berry*[2]

*Animal and Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy P61 P302, Co. Cork, Ireland

† Laboratory of Animal Reproduction, Department of Biological Sciences, Faculty of Science and Engineering, University of Limerick, Limerick, V94 T9PX, Ireland

[1] This research was part of the OviGen project (14/S/849) which is funded by the Department of Agriculture, Food and Marine, Ireland.

[2] Corresponding Author donagh.berry@teagasc.ie

Abstract

The objective of the present study was to quantify the accuracy of imputing medium-density single nucleotide polymorphism (SNP) genotypes from lower-density panels (384 to 12,000 SNPs) derived using alternative selection methods to select the most informative SNPs. Four different selection methods were used to select SNPs based on genomic characteristics (i.e. minor allele frequency (MAF) and linkage disequilibrium (LD)) within five sheep breeds (642 Belclare; 645 Charollais; 715 Suffolk; 440 Texel; and 620 Vendeen) separately. Selection methods evaluated included 1) random, 2) splitting the genome into blocks of equal length and selecting SNPs within block based on MAF and LD patterns, 3) equidistant location while optimising MAF, 4) a combination of MAF, distance from already selected SNPs, and weak LD with the SNP(s) already selected. All animals were genotyped on the Illumina OvineSNP50 Beadchip containing 51,135 SNPs of which 44,040 remained after edits. Within each breed separately, the youngest 100 animals were assumed to represent the validation population; the remaining animals represented reference population. Imputation was undertaken under three different conditions; 1) SNPs were selected within a given breed and imputed for all breeds individually; 2) all breeds were collectively used to select SNPs and were included as the reference population; and 3) the SNPs were selected for each breed separately and imputation was undertaken for all breeds but excluding from the reference population, the breed from which the SNPs were selected. Regardless of SNP selection method, mean animal allele concordance rate improved at a diminishing rate while the variability in mean animal allele concordance rate reduced as the panel density increased. The SNP selection method impacted the accuracy of imputation although the effect reduced as the density of the panel increased. Overall, the most accurate SNP selection method for panels with <9,000 SNPs was that based on MAF and LD pattern within genomic blocks. The mean animal allele concordance rate varied from 0.89 in Texel to 0.97 in Vendeen. Greater

imputation accuracy was achieved when SNPs were selected and imputed within each breed individually compared to when SNPs were selected across all breeds and imputed using a multi-breed reference population. In all, results indicate that accurate genotype imputation to medium-density is achievable with low-density genotype panels with at least 6,000 SNPs.

Key words linkage disequilibrium, minor allele frequency, multi-breed, single nucleotide polymorphism selection

Introduction

The magnitude of return-on-investment is a major factor affecting the uptake of any technology and this can be improved by either increasing the return or by reducing the investment requirement. While the logistics underpinning the procurement of a biological sample contributes to the overall cost of acquiring a genotype on an individual, the cost of the genotype panel, as well as the cost of the genotyping service itself, whatever the chosen technology, also impacts the overall cost of generating a genotype. Therefore, any strategy to reduce the cost of individual components contributing to the overall cost of genotyping warrants investigation. One such strategy could be to reduce the number of necessary single nucleotide polymorphisms (SNPs) for genotyping without compromising the downstream analyses. The high uptake in the use of genomic technologies in ruminants is attributable to the desire for accurately identifying genetically elite animals in a process now termed genomic selection (Meuwissen et al., 2001). Genomic evaluations in both cattle and sheep are mostly based on traditional best linear unbiased prediction (BLUP) approaches but where the numerator relationship matrix, traditionally generated from solely pedigree information, is

replaced by a relationship matrix derived from genotype information. The genomic relationship matrices in farmed species are usually developed using 38,000 to 50,000 SNPs scattered across the genome (cattle: Berry and Kearney, 2011; sheep: Aurvay et al., 2014; goats: Mucha et al., 2015; and pigs: Wellman et al., 2013); hence, any lower-density genotype panel develop should ideally be imputable to higher-density. The use of cheaper lower-density panels would be especially useful to increase the uptake of genotyping in low value animals such as sheep.

The objective of the present study was to quantify the accuracy of imputing medium-density SNP genotypes from lower-density genotyping panels derived using alternative approaches to select the most informative SNPs. As many sheep breeding programmes comprise more than a single breed, greater uptake of genomic technologies may materialise if the lower-density panels were applicable across multiple breeds (and populations) including those not represented in the development of the panel. This was also investigated in the present study.

Materials and Methods

Genotype data

A total of 51,135 biallelic SNPs were available on 3,241 animals genotyped using the Illumina OvineSNP50 Beadchip. The animals all originated from five flockbook recorded sheep populations namely the Belclare (n=650), Charollais (n=674), Suffolk (n=783), Texel (n=494), and Vendeen (n=640); animals were retained if they had a call rate of ≥0.95. These animals originated from 20, 105, 68, 79, and 32 individual seedstock breeders for the Belclare, Charollais, Suffolk, Texel, and Vendeen, respectively. Only autosomal SNPs with a

known genomic position, a call rate ≥0.95, and an Illumina GenCall (GC) score ≥0.55 (http://www.illumina.com/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf) were retained. Parentage analysis was undertaken using the edited SNP dataset based on the proportion of autosomal SNPs in each putative parent-offspring pair that did not adhere to expected Mendelian inheritance patterns; where the extent of Mendelian inconsistencies were >2%, the parent of the individual was set to missing for the subsequent analyses. Inconsistency in the Mendelian inheritance pattern of each SNP was subsequently determined based on the proportion of genotypes per SNP that were opposing homozygotes in a validated parent-offspring pair; a total of 986 parent-offspring pairs existed among the 3,062 genotyped animals. A total of 321 of the remaining autosomal SNPs were discarded where >2% of the parent-offspring autosomal genotypes did not conform to normal Mendelian inheritance. Finally, the extent to which each SNP genotype deviated from Hardy-Weinberg equilibrium was calculated within each of the five breeds separately; SNPs that deviated from Hardy-Weinberg equilibrium ($P < 0.01 \times 10^{-7}$) in any one of the five breeds were not considered further. Following edits, 44,040 autosomal SNPs from 3,062 animals remained across the five breeds (Belclare n=642; Charollais n=645; Suffolk n=715; Texel n=440; and Vendeen n=620). Within the population there were 101, 155, 177, 90, and 106 paternal half-sibs families in the Belclare, Charollais, Suffolk, Texel and Vendeen, respectively; the respective mean size of the paternal half-sib families (range in parenthesis) was 5.960 (2 to 26), 3.129 (2 to 8), 3.248 (2 to 9), 3.044 (2 to 8), and 5.462 (2 to 23).

To quantify the accuracy of imputation, animals were partitioned into either a reference or a validation imputation population based on their date of birth. Within each breed separately, the youngest 100 animals were assumed to represent the validation population; the remaining animals were assumed to be part of the reference population. The average (pedigree-based) relationship between the reference and validation per population

was 0.048, 0.011, 0.034, 0.018, and 0.029 for the Belclare, Charollais, Suffolk, Texel, and Vendeen, respectively.

Development of low-density SNP panels

SNP selection methods

Low-density genotype panels were developed for each of the five breeds individually to imitate seven different panel densities namely; 384 SNPs, 1,000 SNPs, 2,000 SNPs, 3,000 SNPs, 6,000 SNPs, 9,000 SNPs, and 12,000 SNPs. Four different methods were used to select the most informative SNP, within each breed separately, primarily based on the approaches used by Judge et al. (2016) for cattle. The number of SNPs selected per chromosome in the present study differed per panel density and was a function of the length of the chromosome; within panel size, the number of SNPs chosen per chromosome was the same for all four selection methods tested. The number of SNPs selected per chromosome for each low-density panel is in Supplementary Table 1. The four SNP selection methods were;

1) *Random SNP selection method:* Single nucleotide polymorphisms were randomly selected within each chromosome until the pre-defined number of SNPs per chromosome was reached for the respective panel density.
2) *Block SNP selection method:* Each chromosome was divided into blocks of equal length. Chromosome length was defined as the distance from the genomic position of the first SNP to the genomic position of the last SNP. The number of blocks per chromosome was equal to the predefined number of SNPs for that chromosome, less two, so that an extra SNP could be chosen in the blocks at the start and the end of each chromosome (heron referred to as the periphery blocks of the chromosome). All SNPs were ranked on an

index consisting of the minor allele frequency (MAF) of the SNP plus the mean linkage disequilibrium (LD) between that SNP and all other candidate SNPs within that block; an equal weighting was placed on both average LD and MAF, and the highest ranking SNP was then chosen within each block. A second informative SNP was then selected from the periphery blocks of each chromosome. The partial correlation of each candidate SNP in the block with all other candidate SNPs in the block after adjustment for the correlation with the already chosen SNP was calculated as:

$$r\left(SNP_i, SNP_j \middle| SNP_{sel}\right) = \left[r\left(SNP_i, SNP_j\right) - r(SNP_i, SNP_{sel})r(SNP_j, SNP_{sel})\right] / \{[1 - r^2(SNP_i, SNP_{sel})]^{\frac{1}{2}}[1 - r^2(SNP_j, SNP_{sel})]^{\frac{1}{2}}\}$$

where $r\left(SNP_i, SNP_j \middle| SNP_{sel}\right)$ is the correlation of two candidate SNPs $\left(SNP_i \; and \; SNP_j\right)$ after adjusting for the relationship of these SNPs with the already selected SNP (SNP_{sel}). The highest ranked SNP on an index of MAF and the mean partial correlations between the SNP and all other remaining SNPs in that block (standardized to have equal variances) was selected as the second most informative SNP (Judge et al., 2016).

3) *EquiMAF SNP selection method:* Each chromosome was divided equally in length depending on the predefined number of SNPs required per chromosome to identify an ideal distance between SNPs on the lower-density panel. Chromosome length was defined as above. Each SNP was assigned a number corresponding to the order of that SNP by its position within the chromosome. The ideal distance was then calculated by $\left(\frac{length}{count}\right) * SNP \; number$ where length is the total length of the chromosome and count is the number of predefined SNPs per chromosome desired. An index was then created using the following equation:

$$(positionSNP_i - ideal \; distance) * (0.5 - MAF)$$

where MAF is the minor allele frequency. The highest ranked SNP on an index of ideal distance and MAF was then selected (adapted from Corbin et al., 2014).

4) *Wellman SNP selection method (Wellman et al., 2013):* SNPs selected were favoured for high MAF, distance from already selected SNPs, and weak LD with the SNP(s) already selected as described in Wellman et al. (2013). Two SNPs were first chosen at each periphery (<0.5 Mb) of each chromosome as described previously. After the SNPs on the peripheries were selected, additional SNPs were selected using three steps. The purpose of the first step was to define a distance measure d between two SNPs m_i and m_j on the same chromosome i.e., $Chr_{m_i} = Chr_{m_j}$; this distance was calculated as:

$$d(m_i, m_j) = \lambda\left|loc_i - loc_j\right| + (1-\lambda)\mathrm{K}(1 - 0.99 * \left|r\left(G_{m_i}, G_{m_j}\right)\right|)$$

where $\lambda = \min(1, \frac{|loc_i - loc_j|}{K})$ with K = 5 Mb; loc_{m_i} and loc_{m_j} represented the genomic location in megabases of the SNPs m_i and m_j, respectively and $r\left(G_{m_i}, G_{m_j}\right)$ was the correlation between genotypes for SNPS m_i and m_j. For loci that are in close proximity (λ<1), the correlation between the genotypes contributed to the distance measure to enable two markers at similar genomic positions to be included in the low-density panel if the SNPs were not in LD. In the second step, a score was calculated for each SNP i based on its minor allele frequency (MAF_i):

$$Score_m = MAF_m u_m$$

where $u_m = 1$ where the position of m_i is known. In the final step, SNPs were selected based on their scores and the distance measure d. If n markers had already been selected, then marker m_{n+1} was chosen such that $MAF_{m_{n+1}} * \min(d(m_{n+1}, m_k) : k = 1, \ldots, n$ was maximised. This method prioritizes SNPs of both higher MAF and at a large distance

(both standardized to have equal weighting) from SNPs already chosen (Wellman et al., 2013).

Imputation

All imputation was undertaken by chromosome across the entire genome simultaneously using FImpute version 2.2 (Sargolzaei et al., 2014) exploiting both family- and population-based imputation. Pedigree information was supplied to FImpute version 2.2. Imputation from the low density panels was initially carried out for each breed separately. The reference population for imputation consisted solely of the breed for which the genotype panel was developed and, for validation; only the masked genotypes of animals of this breed were imputed. This method was repeated for each of the 5 breeds to determine the most accurate SNP selection method.

Once the most accurate SNP selection method was identified, two further scenarios were investigated; these involved modifications to the method of selecting SNPs and reference population structure used for imputation:

S*cenario 1:* Single nucleotide polymorphisms were selected in a multi-breed population containing all five breeds. When selecting SNPs for inclusion on the lower-density panels, the MAF used was the minimum MAF of that SNP in any of the five breeds while the mean LD between all candidate SNPs used was the maximum of the within-breed mean LD estimated in any of the five breeds. The masked genotypes of the validation animals in all breeds were then imputed using a reference population that included all five breeds together.

S*cenario 2:* Single nucleotide polymorphisms were selected based on the genomic characteristics of each breed separately as described previously. The masked genotypes of validation animals from all breeds were then simultaneously imputed using a single-breed reference population. Every other breed was separately included in the reference population, except the breed from which the SNPs were selected. The whole process was repeated so that SNPs were selected based on the genomic characteristics of each breed individually with every other breed individually included in the reference population. For example, if the genomic characteristics of the Belclare breed were used to selected SNPs, masked genotypes of all validation animals would be imputed where just the Charollais (or the Suffolk, Texel or Vendeen) were individually included in the reference population. A summary of the imputation scenarios is in Appendix 1.

Imputation accuracy statistics

Measures of imputation accuracy were undertaken within each breed separately. Imputation accuracy was carried out on all SNPs on the higher-density genotype panel (i.e., 44,040 SNPs). Five imputation accuracy statistics were estimated:

1) The genotype concordance rate, defined as the mean proportion of correctly imputed genotypes within animal (Berry et al., 2014).
2) The allele concordance rate defined as the mean proportion of correctly imputed alleles within animal; where a genotype was imputed to be heterozygote but was truly homozygote then it was assumed to have been imputed with an accuracy of 0.5 (Berry et al., 2014).

3) The (raw) correlation between true and imputed genotypes; genotypes were denoted as 0, 1 or 2 to represent homozygous, heterozygous, and opposite homozygous, respectively.
4) The adjusted genotype correlation between the actual and imputed genotype per animal. This measure adjusted each genotype for the respective SNP allele frequency to account for differing allelic frequency per SNP as previously described by Mulder et al. (2012) and Berry et al. (2017). The adjusted genotype was achieved by subtracting twice the SNP allele frequency of the allele represented by the homozygous 2 genotype from both the actual and imputed genotype of that SNP; the adjusted correlation was subsequently estimated. The allele frequency per SNP was estimated solely on the reference population used for imputation. All accuracy statistics were undertaken within each breed separately.
5) The rare allele concordance rate for SNPs with a MAF >0 but ≤0.05 defined as the mean proportion of correctly imputed rare alleles per animal. Where a genotype was imputed to heterozygote but was truly homozygote for the minor allele (and vice versa) then it was assumed to have been imputed with an accuracy of 0.5. Genotypes that were truly homozygote for the major allele were not considered in the estimation of rare allele concordance rate.

Results

Regardless of SNP selection method or imputation scenario, the mean allele concordance rate was always greater than genotype concordance rate (Supplementary Figure 1) while the adjusted genotype correlation between true and imputed genotypes was consistently weaker than the raw genotype correlation (Supplementary Figure 2). The allele concordance rate and adjusted genotype correlation between true and imputed genotypes for different SNP

selection methods are presented in Figure 1 and Figure 2, respectively. Hereafter, the only imputation statistics reported and discussed are allele concordance rate and the adjusted genotype correlation.

Single nucleotide polymorphism selection method

The impact of SNP selection method was only evaluated where the breed-specific genomic statistics (i.e., MAF and LD) used to select the SNPs that were from the same breed used in the reference and the validation population. Selection method impacted the accuracy of imputation achievable although the impact reduced as the density of the low-density panel increased. Imputation accuracy from SNP panels composed of SNPs selected using the block selection method outperformed all other SNP selection method across all breeds and panel densities up to 1,000 SNPs with, on average, a superior allele concordance rate compared to the next best method (i.e., Wellman method) across the five breeds of 0.008 and 0.004 for 384 SNPs and 1,000 SNPs, respectively. For the Belclare, Charollais, Suffolk, and Texel breeds, the block SNP selection method also outperformed all other methods for the panel densities up to 6,000 SNPs in terms of both better allele concordance rate (an average better allele concordance rate of 0.003, 0.002, and 0.001 for 2,000, 3,000, and 6,000 SNPs, respectively compared to the next best method (i.e., Wellman)) and adjusted genotype correlation (on average 0.007, 0.005, and 0.002 superior for 2,000, 3,000, and 6,000 SNPs, respectively compared to the next best method). However, the Wellman SNP selection method proved to be slightly superior in the Suffolk breed for genotype panels containing 2,000 and 3,000 SNPs with better allele concordance rate of 0.002 and 0.001, respectively compared to the block method. The differences in imputation accuracy between the block and

274 the Wellman SNP selection methods were negligible on the 9,000 and 12,000 genotype
275 panels, although both were consistently superior to both the random and equiMAF methods.

276 Imputation accuracy from SNPs selected using the equiMAF method consistently
277 resulted in the poorest imputation accuracy (both in terms of allele concordance rate and
278 adjusted genotype correlation between true and imputed genotypes) across all breeds and
279 panel densities with the exception of the 384 SNP panel in each of the five breeds (Figure 1;
280 Figure 2). Compared with the next poorest method (i.e., the random method), the equiMAF
281 method resulted in a poorer allele concordance rate of 0.033, 0.073, 0.092, 0.103, 0.095, and
282 0.085, for genotype panels containing 1,000, 2,000, 3,000, 6,000, 9,000, and 12,000 SNPs,
283 respectively. The difference in both allele concordance rate and adjusted genotype correlation
284 between the equiMAF SNP selection method and the random SNP selection method was
285 negligible when the genotype panel contained just 384 SNPs (Figure 1; Figure 2).

286 The variability in both the allele concordance rate per animal and the adjusted
287 genotype correlation between the true and imputed genotypes per animal was also affected by
288 SNP selection method (Figure 1; Figure 2). The SNP selection methods that achieved the
289 greatest mean imputation accuracy (i.e., block and Wellman methods) across all panels and
290 breeds were also characterised by the least variability in both the mean allele concordance
291 rate and adjusted genotype correlation between true and imputed genotypes per individual
292 compared to the poorer methods (i.e., EquiMAF and random methods). When SNPs were
293 selected for the 6,000 SNP panel, for example, in the Vendeen breed using the block SNP
294 selection method, the mean allele concordance rate per animal for the 6,000 SNP panel was
295 0.9651 with a standard deviation (SD) of 0.05 while for the same density panel, SNPs
296 selected using the random method had a mean allele concordance rate of 0.9529 but with a
297 SD of 0.06.

The relationship between minor allele frequency (MAF) and imputation accuracy (i.e., allele concordance rate and adjusted genotype correlation between true and imputed genotypes) of masked genotypes in the 6,000 SNP panel when the block SNP selection method was used in all breeds is presented in Table 1. Both the allele concordance rate (when the minor allele frequency was ≤0.45) and the adjusted genotype correlation worsened as the MAF (bin) increased. For the Charollais, Suffolk, Texel and Vendeen breeds, allele concordance rate was better when MAF was between 0.45 and 0.50 than when MAF was between 0.40 and 0.45; this was primarily due to fewer SNPs in having a MAF between 0.45 and 0.50 compared to other MAF bins.

The rare allele concordance rate (Figure 5) was only undertaken for the block SNP selection method where SNPs were selected within a single breed and only that breed was included in both the reference and validation populations. The number of SNPs with a MAF >0 but ≤ 0.05 present in the Belclare, Charollais, Suffolk, Texel, and Vendeen were 3,562, 3,437, 6,802, 4,658, and 5,214, respectively. As the panel density increased, the rare allele concordance rate also increased albeit at a diminishing rate; however, large variability in rare allele concordance rate per animal existed across breeds and SNP panel density. The imputation of rare alleles in the Vendeen and Texel were consistently better and worse, respectively than other breeds for all panel densities. For the 6,000 SNP panel, the allele concordance for rare alleles was 0.487 and 0.374 for the Vendeen and Texel, respectively.

Single nucleotide polymorphism panel density

Regardless of the SNP selection method used, the mean animal allele concordance rate improved at a diminishing rate as the panel density increased (Figure 1). When SNPs were selected using the block method, and imputation was undertaken solely within the same

breed, the mean animal allele concordance rate was better, on average, by 4.88 percentage units across all breeds (maximum of 6.95 percentage units and minimum of 3.01 percentage units in the Charollais and Vendeen, respectively) when the panel density doubled from 1,000 to 2,000 SNPs. Subsequently, when the SNP panel density doubled from 3,000 to 6,000, the mean animal allele concordance rate improved by, on average, by 2.21%; a maximum difference of 4.00% and a minimum difference of 0.84% were observed in the Texel and Vendeen breeds, respectively.

The variability in the mean allele concordance rate per animal also reduced as the SNP panel density increased, independent of the SNP selection method used. When SNPs were selected within each breed individually using the block method, and imputation was undertaken within a single breed, the mean allele concordance rate per animal across each of the five breeds was 0.818 for 384 SNPs selected (average minimum 0.740 and average maximum 0.923), whereas for 9,000 selected SNPs, the mean allele concordance rate per animal across each of the five breeds was 0.98 (average minimum 0.899 and average maximum 0.998).

Imputation scenario

The block method was the most accurate SNP selection method and was therefore the only SNP selection method used in the remaining imputation scenarios. The imputation accuracy of the two scenarios where the breed composition of the reference and validation population differed as well as the genomic characteristics used to generate the panels are summarised in Figure 3 and Figure 4. The accuracy of imputation was affected by the composition of the reference population (i.e., whether the reference population contained only a single breed or all five breeds simultaneously). For all breeds, greater imputation accuracy was observed for

genotype panels containing <3,000 SNPs when just one breed was used to develop the genotype panel and imputation undertaken with only that breed included in the reference and validation populations compared to when the genomic characteristics of all breeds were used both to develop the genotype panels and included in the reference population. On average, across all five breeds, a better allele concordance rate of 0.075, 0.072, 0.045, and 0.029, was observed for SNP panels containing 384, 1,000, 2,000, and 3,000 SNPs, respectively when imputation was undertaken with breeds individually included in the reference population compared with when all five breeds were simultaneously included in the reference population. Within the Vendeen breed, the effect of the composition of the reference population was negligible when the low-density panels contained ≥ 6,000 SNPs. The allele concordance rate for the remaining four breeds increased by, on average, 0.138 for the 6,000 SNP genotype panel when the reference population for imputation contained only the breed in which the low-density panels were developed (Figure 3). For the Belclare and Suffolk breeds, differences between the accuracy of imputation when all five breeds were simultaneously included in the reference population compared to when just the Belclare and Suffolk breed were, respectively included in the reference population for the 9,000 SNP panel were negligible (<0.001 for allele concordance rate). For all breeds, with the exception of the Texel breed, the effect of the composition of the reference population was negligible for the 12,000 SNP panel; a stronger adjusted genotype correlation between true and imputed genotypes of 0.012 was observed for the Texel breed when imputation was undertaken with just the Texel breed included in the reference population.

When the SNP genotype panels were developed within an individual breed and a single breed was used as the reference population, the impact on accuracy of imputation of which breed was actually included in the reference population was large (Figure 3 and 4). Where a 6,000 SNP panel was built in the Belclare using the block method, and only the

Belclare animals were included in the reference population, an allele concordance rate of 0.988 was achieved for the Belclare breed compared to allele concordance rates of 0.762, 0.743, and 0.752 when the reference population was composed solely of Charollais, Suffolk or Vendeen, respectively. When the genotype panels were built within the Belclare breed, and the Texel breed was the only breed included in the reference population, a stronger correlation between the true and imputed genotypes of 0.831 was observed for the 6,000 SNP panel compared to the average adjusted genotype correlation of the Charollais, Suffolk or Vendeen. Better imputation accuracy was also observed across all panel densities when the genotype panels were built in the Texel population and Texel reference population was imputed using a reference population that only included the Belclare. Similar improvements in imputation accuracy were observed for the Charollais when the Vendeen breed was solely included in the reference population and vice versa (Figure 3; Figure 4).

Discussion

While many studies have quantified the accuracy of imputation from lower-density genotype panels to higher-density genotype panels in cattle (Zhang and Druet, 2010; Berry and Kearney, 2011; Judge et al., 2016), fewer such studies exist in sheep (Hayes et al., 2012; Bolormaa et al., 2015; Moghaddar et al., 2015). Previous imputation-based studies in sheep have mainly been confined to wool and meat sheep breeds in Australia and New Zealand (Hayes et al., 2012; Bolormaa et al., 2015; Ventura et al., 2016). Furthermore, studies to date on the use of lower-density genotype panels in sheep have focused primarily on factors affecting imputation accuracy; these factors include the multi- or single-breed structure of the reference and validation populations (Bolormaa et al., 2015; Ventura et al., 2016), the degree of relatedness among and between animals in the reference and validation populations

(Bolormaa et al., 2015; Moghaddar et al., 2015), and the size of the reference population (Moghaddar et al., 2015; Ventura et al., 2016). Imputation studies quantifying the impact of alternative approaches to selecting the SNPs for genotyping panels differing in SNP densities in sheep do not exist. Several alternative approaches to select such SNPs in cattle have been evaluated including random selection (Szyda et al., 2013), a combination of equidistant physical location and high MAF (Boichard et al., 2012) as well as dividing each chromosome into equally sized segments and selecting SNPs within the segment with the greatest MAF (Mulder et al., 2012). Further SNP selection methods have been evaluated in other species; one such method in pigs involved selecting SNPs based on high MAF, relatively equally spaced and weak correlations with the SNPs already selected (Wellman et al., 2013). With the exception of the equiMAF method (adapted from Corbin et al., 2014) evaluated in the present study, all other strategies to SNP selection evaluated in the present study have been documented in dairy and beef cattle (Judge et al., 2016).

When SNPs were selected in the present study using a combination of the LD and MAF of a single breed and that breed was itself solely included in the reference population, greater imputation accuracy for the 12,000 SNP genotype panel was achieved compared to that reported by Bolormaa et al. (2015) when a 11,267 SNP panel was imputed to 48,599 SNPs. Bolormaa et al. (2015) reported a range in mean animal allele concordance rate of 0.88 to 0.94 in multiple Australian sheep breeds (i.e., Border Leicester, Poll Dorset, White Suffolk, Merino and crossbreds) with a weighted mean allele concordance rate across breeds of 0.89. The range in allele concordance rate in the present study (with SNPs selected using the block method) was 0.983 to 0.996 with a weighted average of 0.992. However, with the exception of the Merino breed, the size of the reference populations of the individual breeds reported by Bolormaa et al. (2015) was smaller (157 to 341 animals) than those in the present study (with the exception of the Texel breed). Furthermore, a range in allele concordance rate

of 0.90 to 0.94 was reported by Bolormaa et al. (2015) when imputation was carried out using a multi-breed population on a 11,267 SNP panel was undertaken; the range in allele concordance rate for imputation where SNPs were selected using the block method within all breeds and all breeds were included in the reference population was still greater in the present study (0.976 to 0.996). In both the present study and that of Bolormaa et al. (2015), the lower-density panel was developed within the same multi-breed population as that included in the reference population.

The SNPs included on the low-density sheep panels proposed by Hayes et al. (2012) were chosen from 48,640 SNPs as every nth marker by chromosome position. Hayes et al. (2012) used the fastPHASE imputation method (Scheet and Stephens, 2006), and achieved imputation accuracy comparable to that achieved in the present study from the genotype panels developed using randomly selected SNPs. Hayes et al. (2012) reported the genotype concordance rate for all breeds was <0.80 (i.e., Border Leicester, Merino, and Poll Dorset and White Suffolk combined) when 5,000 SNPs were imputed to 48,640 SNPs using a single breed reference population. While the 5,000 SNP panel used by Hayes et al. (2012) is closer in density to the 6,000 SNP panel in the present study, the genotype concordance rate obtained by Hayes et al. (2012) is more similar to the average genotype concordance rate (0.884) reported for genotype panel containing 3,000 randomly selected SNPs in the present study.

The trend observed for a declining allele concordance rate as MAF increased corroborates other studies in sheep (Bolormaa et al., 2015) and cattle (Berry and Kearney, 2011; Judge et al., 2016). While Bolormaa et al. (2015) observed an increase in raw genotype correlation between true and imputed genotypes as MAF increased, this trend was not observed in the present study.

444 Where possible, the approaches taken in the present study aimed to produce results
445 that reflect real-life. As it tends to be the younger animals that are genotyped on a lower-
446 density panel, the youngest 100 animals per breed were chosen to be the validation
447 population. Furthermore, the inclusion of the unmasked genotypes (i.e., 100% concordance
448 with the real genotypes) in the estimation of imputation accuracy was also to simulate a real-
449 life scenario. The allele concordance rate (ACR) of the unmasked genotypes for example of
450 the 6,000 SNP panel can be easily calculated using the formula $\frac{ACR(44{,}040)-1.0(6{,}000)}{44{,}040-6{,}000}$ where
451 44,040 is the number of SNPs on the higher-density panel and the assumed allele
452 concordance rate of the 6,000 unmasked SNPs was 1.0. Taking the Belclare breed as an
453 example, where SNPs were selected using the block method within the Belclare breed and
454 only the Belclare breed was included in the reference and validation population. The allele
455 concordance rate for all SNPs (masked and unmasked) imputed from the 6,000 SNP panel
456 was 0.988 while the allele concordance rate of the 38,040 masked SNPs only calculated using
457 the above formula was 0.986.

458

459 *Single nucleotide polymorphism selection method*

460 While the random method was expected to perform the poorest, the poor performance of the
461 equiMAF method is in direct contrast to the findings of Corbin et al. (2014) who also selected
462 SNPs based on equidistance, optimized for MAF in Thoroughbred horses. Corbin et al.
463 (2014) reported greater imputation accuracy for a 6,000 SNP panel (genotype concordance
464 rate of 0.98) compared to SNPs selected based solely on equidistance across the genome
465 (genotype concordance rate of 0.97) or selected based on a combination of LD pattern and
466 MAF (genotype concordance rate of 0.95). Carvalheiro et al. (2014) reported that SNPs
467 selected based on a combination of MAF and LD resulted in better imputation accuracy in

Nelore cattle when compared to SNPs selected using either MAF or LD. Of the SNP selection strategies evaluated in the present study, both the block method and the Wellman method placed equal emphasis on both high MAF and weak LD when selecting SNPs for inclusion on a lower-density genotype panel with the block method and the Wellman method outperforming all other methods in all breeds for all panel densities. The overall superiority of the block method was not entirely unexpected as is consistent with its superiority in selecting SNPs for imputation to higher-density in cattle (Judge et al., 2016). Judge et al. (2016) evaluated six alternative SNP selection approaches, three of which (i.e., random, Wellman and block) were common to those evaluated in the present study.

Where SNPs were selected using the block method based on LD and MAF statistics from one breed with that breed being included in the reference population, the imputation accuracies in the present study were poorer than those reported by Judge et al. (2016) for genotyping panels containing ≤3,000 SNPs. However, the size of the reference population (range of 340 to 615 animals) in the present study was much smaller than used by Judge et al. (2016; 1,484 animals). Nevertheless, the range in allele concordance rate for SNP panels developed using the block method in a single breed with that breed being solely included in the reference population in the present study containing 6,000 (0.963 to 0.993) and 12,000 (0.983 to 0.996) was similar to the allele concordance rate reported by Judge et al. (2016) for their 6,000 (0.988) and 12,000 (0.994) panels. The genotype concordance rate (Supplementary Figure 1) achieved for the 3,000 SNP panel in the Vendeen breed (0.966) when the Wellman SNP selection method was used is similar to that reported by Wellman et al. (2013; 0.96) for the same density using the same SNP selection method in German Piétrain boars. However, genotype allele concordance rates in the remaining breeds using the same SNP selection method were lower than that reported by Wellman et al. (2013) for the 3,000 panel. Judge et al. (2016) suggested that the reason for the superior performance of the

493 block method compared to the Wellman method may be due to the positioning of SNPs
494 across the genome. Single nucleotide polymorphisms selected using the block method were
495 forced to be more evenly distributed across the genome as only one SNP could be selected
496 per segment (or block). The Wellman method however enabled neighbouring SNPs to be
497 selected if the LD between them was low (and SNPs in other regions of the chromosomes had
498 already been selected; Judge et al., 2016). The mean SD in distance between neighbouring
499 SNPs on the 6,000 SNP panel was 177kb for the block SNP selection method (minimum SD
500 in the Texel 176kb; maximum SD in the Charollais 180kb) compared to 425kb for the
501 Wellman method (minimum SD in the Charollais 409kb; maximum SD in the Texel 444kb).
502 This therefore indicates that SNPs selected using the block method were more evenly spaced
503 across the genome.

504

505 *Single nucleotide polymorphism panel density*

506 The improvement in imputation accuracy (both allele concordance rate and adjusted genotype
507 correlation) with increasing panel density was expected and has previously been documented
508 in both sheep (Hayes et al., 2012) and cattle (Judge et al., 2016). While few sheep studies
509 have investigated the imputation accuracy of genotyping panels containing less than 6,000
510 SNPs, Hayes et al. (2012) reported a genotype concordance rate of >0.80 for a genotype
511 panel containing 5,000 SNPs. In a population of 6,369 dairy cattle, Judge et al. (2016)
512 reported allele concordance rates to from lower-density panels containing 384 (0.849), 1,000
513 (0.881), 2,000 (0.963), 3,000 (0.976), 6,000 (0.988), 12,000 (0.994) SNPs to higher-density
514 similar to those obtained (using panels containing SNPs selected using the block method) in
515 the present study; allele concordance rates averaged across breeds in the present study were

0.817, 0.889, 0.938, 0.959, 0.981, and 0.991 for panels containing 384, 1,000, 2,000, 3,000, 6,000, 12,000 SNPs, respectively.

When averaged across all breeds (when SNPs were selected using the block method within a single breed and that breed alone was included in the reference population), the adjusted genotype correlation between true and imputed genotypes improved by 0.130 when the density increased from 1,000 to 2,000 SNPs compared with an improvement of only 0.055 and 0.025 when density increased from 3,000 to 6,000 and from 6,000 to 12,000, respectively. The reduced improvement in imputation accuracy when panel density increased was primarily because allele concordance rate was already high across all breeds (>0.90) with the exception of the Texel breed (>0.85).

Reference population

The improved imputation accuracy when just the animals of the breed being imputed were included in the population compared to a multi-breed reference population was more evident in the lower-density panels (i.e., ≤3,000 SNPs); this corroborates results reported by Hayes et al. (2012) in sheep. However, the impact of the breed representation in the reference population was negligible once the lower-density panel contained at least 9,000 SNPs (with the exception of the Texel breed). The improved imputation accuracy for genotype panels <9,000 SNPs when just a single-breed reference population was used for imputation in the present study is, nonetheless, in contrast with those of Bolormaa et al. (2015) who reported a marked increase in imputation accuracy of Australian sheep when a multi-breed reference population was used. The increase in imputation accuracy using a multi-breed reference population compared to a single-breed reference population in Bolormaa et al. (2015) may be due to the smaller reference population size for the individual single-breed reference

population. Nonetheless, Bolormaa et al. (2015) documented that failure to include the breed being imputed in the reference population contributed to substantial erosion in imputation accuracy (e.g., the mean genotype correlation between true and imputed genotypes per animal of the Merino breed reduced from 0.91 to 0.80 when 11,267 SNPs were imputed to 48,599 SNPs); a conclusion also deduced from the present study.

The purpose of imputation scenario two was to the investigate the effects of 1) the application of a panel built in one breed and applied to another, and 2) the composition of the reference population. While there was a marked reduction in imputation accuracy when the breed being imputed was not included in the reference population, there were exceptions. The reduction in imputation accuracy was not as severe when the Texel breed was included in the reference population with the Belclare breed being the breed imputed, or vice versa. This is most likely be due to the Texel being one of the breeds included in the Belclare composite breed. When the Belclare breed was initially formed, the Galway, Finnish Landrace, and Lleyn breeds served as the founder breeds while the Texel was later introduced. Similarly, the reduction in imputation accuracy from not having the breed to be imputed also included in the reference population was not as severe when the Charollais was included in the reference population and the Vendeen was being imputed and vice versa. As both the Charollais and the Vendeen are French breeds, they may have a closer genetic relationship that any of the other breeds.

The association between a larger imputation reference population and greater imputation accuracy has been well documented in both sheep (Bolormaa et al., 2012; Ventura et al., 2016) and cattle (Hozé et al., 2013). The smaller reference population of the Texel breed in the present study (n=340) may help explain the poorer imputation accuracy across all imputation scenarios in this breed relative to the other breeds. Within the Texel breed, 16 validation animals had a parent in the reference population; no Texel animals in the

565 validation population had both parents in the reference population. The mean allele
566 concordance rate of the 16 Texel animals with a parent in the reference population for the
567 6,000 SNP panel (selected using the block method) was superior (0.983) compared to those
568 that did not have any parent in the reference population (0.958). While the Vendeen did not
569 have the largest reference population, the greatest imputation accuracy was observed in the
570 Vendeen. This is most likely due to a large number of Vendeen animals in the validation
571 population (n=91) that had at least one parental genotype in the reference population. Where
572 Vendeen animals had at least one parent in the reference population, the allele concordance
573 rate for the 6,000 SNP panel (selected using the block method) was 0.993 compared to an
574 allele concordance rate of 0.987 where the validation animals did not have either parent in the
575 reference population. While the Suffolk had both the greatest number of animals in the
576 reference population and the strongest mean linkage disequilibrium (LD) between adjacent
577 SNPs of 0.377 (compared to the next breed which was the Texel with an LD of 0.356) on the
578 medium density panel (i.e., 44,040 SNPs), it did not result in the greatest imputation
579 accuracy. Other studies have reported that stronger LD between SNPs on the higher-density
580 panels lends itself to greater imputation accuracy (Hickey et al., 2012; Pimentel et al., 2013;
581 Corbin et al., 2014). Linkage disequilibrium was calculated between the SNPs included in the
582 6,000 genotype panel for all breeds and SNP selection method. Where 6,000 SNPs were
583 selected in Suffolks using the block method, a stronger LD (0.406) between SNPs existed
584 compared to other breeds (closest breed was Vendeen with weighted mean LD between
585 neighbouring SNPs of 0.331 between SNPs). The mean MAF of SNPs on the higher-density
586 panel (i.e., 44,040 SNPs) in the Suffolk breed was lower (0.239) compared to the all other
587 breeds; the Vendeen (i.e., the next closest breed) had a mean MAF of 0.254. The SNPs
588 selected to be on the lower-density panels should, in theory, be in weaker LD with the other
589 selected SNPs. Therefore, because the mean MAF of the candidate SNPs for the Suffolk

590 6,000 SNP panel was greater than other breeds, the block selection method was forced to
591 choose SNPs with a greater MAF thereby reducing the weight on LD.

592

593 **Conclusion**

594 Accurate genotype imputation is achievable using genotype panels as low as 6,000 SNPs.
595 Careful SNP selection for inclusion on the lower-density panel is, however, paramount to
596 achieve accurate imputation. The block SNP selection method which combines relatively
597 equally positioning of SNPs, MAF and LD outperformed the other methods in imputation
598 accuracy. However, the Wellman method which combines MAF, distance from already
599 selected SNPs, and weak LD with the SNP(s) already selected would be a suitable alternative.
600 With 6,000 SNPs chosen using the block method, the mean allele concordance rate per
601 animal per breed varied from 0.975 to 0.992; the minimum mean allele concordance rate of
602 any animal achieved with a 6,000 panel was 0.847. Imputation accuracy could possibly be
603 improved with larger reference populations. If possible, the breed included in the target
604 imputation population should always be included in the reference population to mitigate any
605 erosion in imputation accuracy.

606

607 **Literature cited**

608 Auvray, B., J.C. McEwan, S-A.N. Newman, M. Lee, K.G. Dodds. 2014. Genomic prediction
609 of breeding values in the New Zealand sheep industry using a 50K SNP chip. J. Anim. Sci.
610 2014.92:4375 – 4389. doi:10.2527/jas.2014-7801.

611 Berry, D.P., and J.F. Kearney. 2011. Imputation of genotypes from low- to high-density
612 genotyping platforms and implications for genomic selection. Animal (2011), 5:8, pp 1162 –
613 1169. doi:10.1017/S1751731111000309.

614 Berry, D.P., M.C. McClure, M.P. Mullen. 2014. Within- and across-breed imputation of
615 high-density genotypes in dairy and beef cattle from medium- and low-density genotypes.
616 Anim Breed. Genet 131 (2014) 165 – 172. doi:10.1111/jbg.12067.

617 Berry D.P., N. McHugh, S. Randles, E. Wall, K. McDermott, M. Sargolzaei, A.C. O'Brien.
618 2017. Imputation of non-genotyped sheep from genotypes of their mates and resulting
619 progeny. Animal (2017) 12:2, pp 191 – 198. doi:10.1017/S1751731117001653.

620 Boichard D., H. Chung, R. Dassonnville, X. David, A. Effen, S. Fritz, K.J. Gietzen, B.J.
621 Hayes, C.T. Lawley, T.S. Sonstegard, C.P. Van Tassell, P.M. VanRaden, K.A. Viaud-
622 Martinez, G.R. Wiggans, for the Bovine LD Consortium. 2012. Design of a Bovine Low-
623 Density SNP Array Optimized for Imputation. PLos ONE 7:3.
624 doi:10.1371/journal.pone.0034130.

625 Bolormaa S., K. Gore, J.H.J. van der Werf, B.J. Hayes, H.D. Daetwyler. 2015. Design of a
626 low-density SNP chip for the main Australian sheep breeds and its effect on imputation and
627 genomic prediction accuracy. Animal Genetics, 46, 544 – 556. doi:10.1111/age.12340.

628 Carvalheiro R., S.A. Boison, H.H.R. Neves, M. Sargolzaei, F.S. Schenkel, Y.T. Utsunomiya,
629 A.M. Pérez O' Brien, J. Sölkner, J.C. McEwan, C.P. Van Tassell, T.S. Sonstegard, J.
630 Fernando Garcia. 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol.
631 2014, 46:69. doi:10.1186/s12711-014-0069-1.

Corbin L.J., A. Kranis, S.C. Blott, J.E. Swinburne, M. Vaudin, S.C. Bishop, J.A. Woolliams. 2014. The utility of low-density genotyping for imputation in the Thoroughbred horse. Genet. Sel. Evol. 2014, 46:9. doi:10.1186/1297-9686-46-9.

Habier D., R.L. Fernando, J.C. Dekkers. 2009. Genomic Selection using low-density marker panels. Genetics 182(1):343 – 353. doi:10.1534/genetics.108.100289.

Hayes B.J., P.J. Bowman, H.D. Daetwyler, J.W. Kijas, J.H.J. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Animal Genetics, 43, 72 – 80. doi:10.1111/j.1365-2052.2011.02208.x.

Hickey J.M., J. Crossa, R. Babu, G. de los Campos. 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Science (2012) 52:654 – 663. doi:10.2135/cropsci2011.07.0358.

Hozé C., M-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 2013, 45:33. doi:10.1186/1297-9686-45-33.

Judge M.M., J.F. Kearney, M.C. McClure, R.D. Sleator, D.P. Berry. 2016. Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations. J. Anim. Sci. 2016. 94:949 – 962. doi:10.2527/jas.2015-0044.

Khatkar M.S., G. Moser, B.J. Hayes, H.W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 2012, 13:538. doi:10.1186/1471-2164-13-538.

Meuwissen T.H.E., B.J. Hayes, M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819 – 1829.

654 Moghaddar N., K.P. Gore, H.D. Daetwyler, B.J. Hayes, J.H.J. van der Werf. 2015. Accuracy
655 of genotype imputation based on random and selected reference sets in purebred and
656 crossbred sheep populations and its effect on accuracy of genomic prediction. Genet Sel Evol
657 (2015) 47:97. doi:10.1186/s12711-015-0175-8.

658 Mucha S., R. Mrode, I. MacLaren-Lee, M. Coffey, J. Conington. 2015. Estimation of
659 genomic breeding values for milk yield in UK dairy goats. J. Dairy Sci. 98:8201 – 8208.
660 doi:10.3168/jds.2015-9682.

661 Mulder, H.A., M.P.L. Calus, T. Druet, C. Schrooten. 2012. Imputation of genotypes with
662 low-density chips and its effect on reliability of direct genomic values in Dutch Holstein
663 cattle. J. Dairy. Sci. 95:876 - 889. doi:10.3168/jds.2011-4490.

664 Pimental, E.C.G., M. Wensch-Dorendorf, S. König, H.H. Swalve. 2013. Enlarging a training
665 set for genomic selection by imputation of un-genotyped animals in population of varying
666 genetic architecture. Genet. Sel. Evol. 2013, 45:12. doi:10.1186/1297-9686-45-12.

667 Sargolzaei, M., J.P. Chesnais, F.S. Schenkel. 2014. A new approach for efficient genotype
668 imputation using information from relatives. BMC Genomics 2014, 15:478.
669 doi:10.1186/1471-2164-15-478.

670 Scheet, P.and M. Stephens. 2006. A fast and flexible statistical model for large-scale
671 population genotype data: applications to inferring missing genotypes and haplotypic phase.
672 The American Journal of Human Genetics, 78, 629-644. doi: 10.1086/502802.

673 Szyda, J., K. Żukowski, S. Kaminski, and A. Żarnecki. 2013. Testing different single
674 nucleotide polymorphism selection for prediction of genomic breeding values in dairy cattle
675 based on low density panels. Czech J. Anim. Sci. 58(3):136–145.

676 Weigel, K.A., C.P. Van Tassell, J.R. O' Connell, P.M. VanRaden, G.R. Wiggans. 2010.
677 Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using
678 reference panels and population-based imputation algorithms. J. Dairy. Sci. 93:2229 – 2238.
679 doi:10.3168/jds.2009-2849.

680 Wellman R., S. Preuß, E. Tholen, J. Heinkel, K. Wimmers, J. Bennewitz. 2013. Genomic
681 selection using low density marker panels with application to a sire line in pigs. Genetics
682 Selection Evolution 2013, 45:28. doi:10.1186/1297-9686-45-28.

683 Ventura, R.V., S.P. Miller, K.G. Dodds, B. Aurvay, M. Lee, M. Bixley, S.M. Clarke, J.C.
684 McEwan. 2016. Assessing accuracy of imputation using different SNP panel densities in a
685 multi-breed sheep population. Genet. Sel. Evol. (2016) 48:71. doi:10.1186/s12711-016-0244-
686 7.

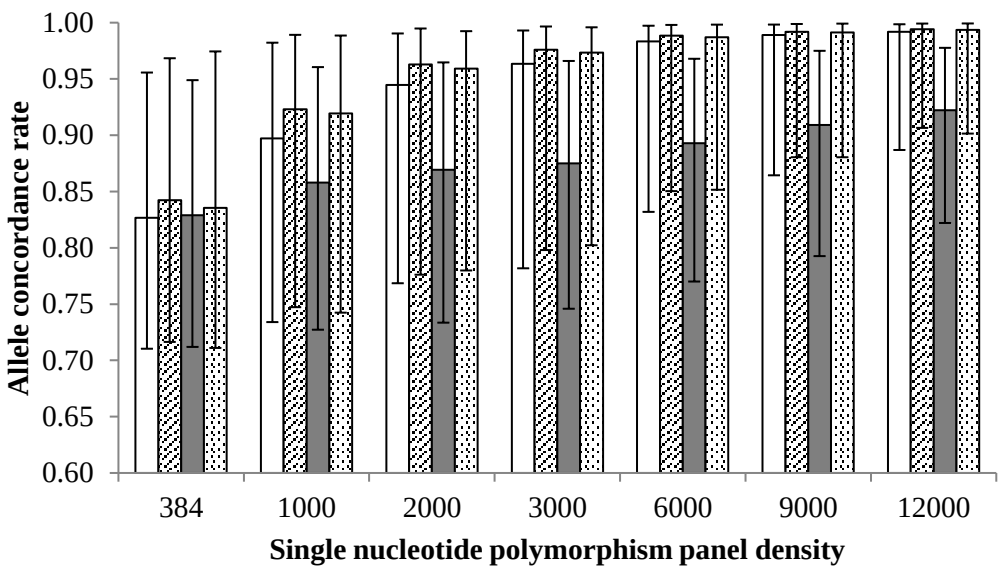
687 Zhang, Z., and T, Druet. 2010. Marker imputation with low-density marker panels in Dutch
688 Holstein cattle. J. Dairy Sci. 93:5487 – 5494. doi:10.3168/jds.2010-3501.

689 **Table 1.** Mean allele concordance rate (ACR) and adjusted genotype correlation (r) per masked single nucleotide polymorphism (SNP) for
690 different minor allele frequency (MAF) bins in each of five sheep breeds on the 6,000 SNP panel where SNPs were selected using the block
691 method.

	Belclare		Charollais		Suffolk		Texel		Vendeen	
MAF bin	ACR	r	ACR	r	ACR	r	ACR	r	ACR	r
>0.00: ≤0.05	0.993	0.993	0.993	0.993	0.997	0.995	0.985	0.990	0.998	0.996
>0.05: ≤0.10	0.990	0.989	0.984	0.986	0.992	0.991	0.971	0.978	0.995	0.994
>0.10: ≤0.15	0.989	0.984	0.979	0.978	0.989	0.986	0.963	0.968	0.993	0.990
>0.15: ≤0.20	0.988	0.982	0.974	0.970	0.987	0.982	0.956	0.954	0.992	0.987
>0.20: ≤0.25	0.988	0.977	0.970	0.959	0.985	0.976	0.952	0.940	0.991	0.982
>0.25: ≤0.30	0.987	0.970	0.969	0.947	0.985	0.970	0.950	0.921	0.990	0.977
>0.30: ≤0.35	0.987	0.965	0.969	0.934	0.985	0.963	0.952	0.907	0.991	0.971
>0.35: ≤0.40	0.987	0.958	0.970	0.921	0.986	0.955	0.954	0.885	0.991	0.963
>0.40: ≤0.45	0.987	0.952	0.975	0.909	0.990	0.951	0.962	0.872	0.992	0.959
>0.45: ≤0.50	0.987	0.951	0.981	0.901	0.993	0.946	0.972	0.862	0.994	0.958

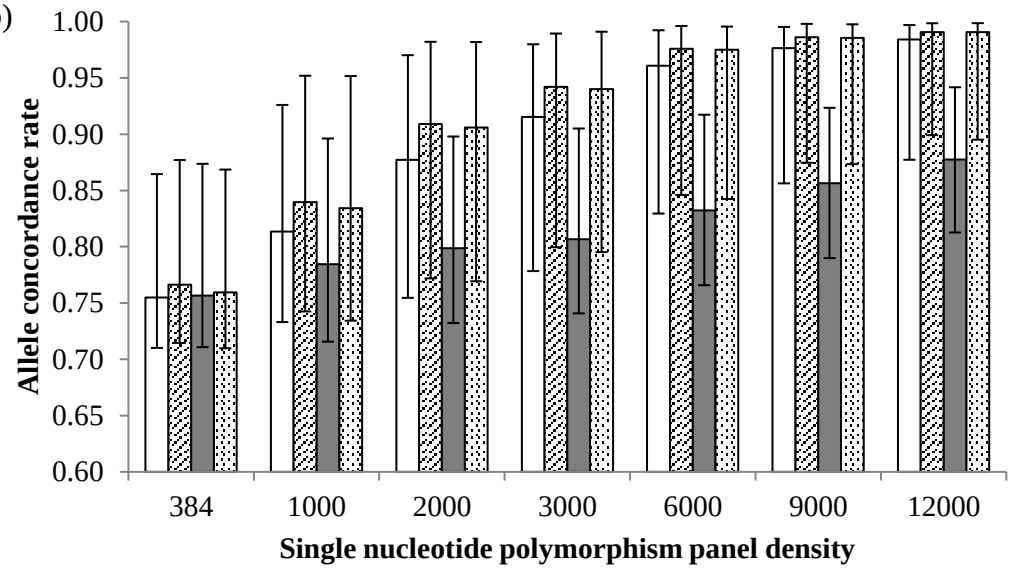
692

a)



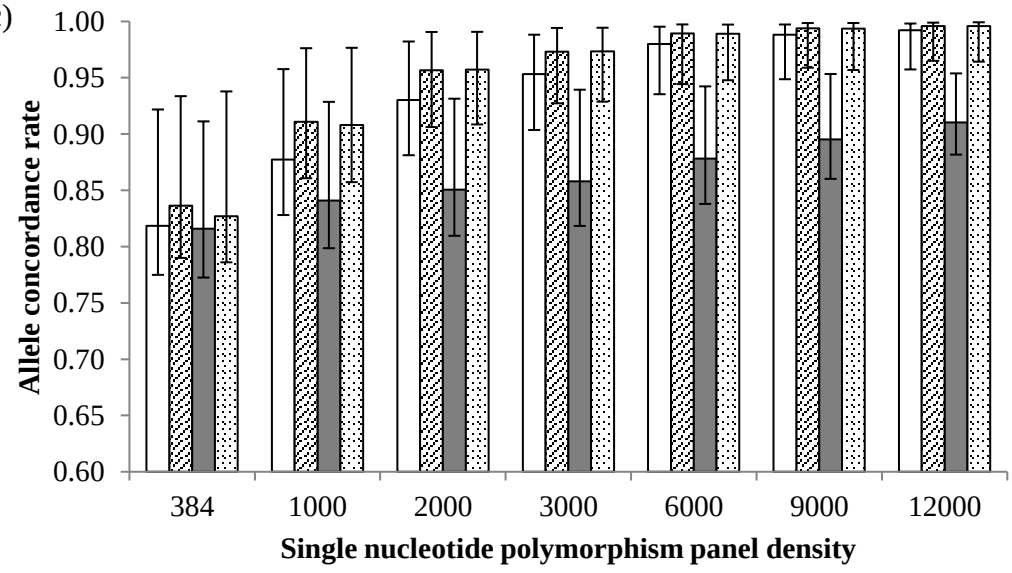
693

b)

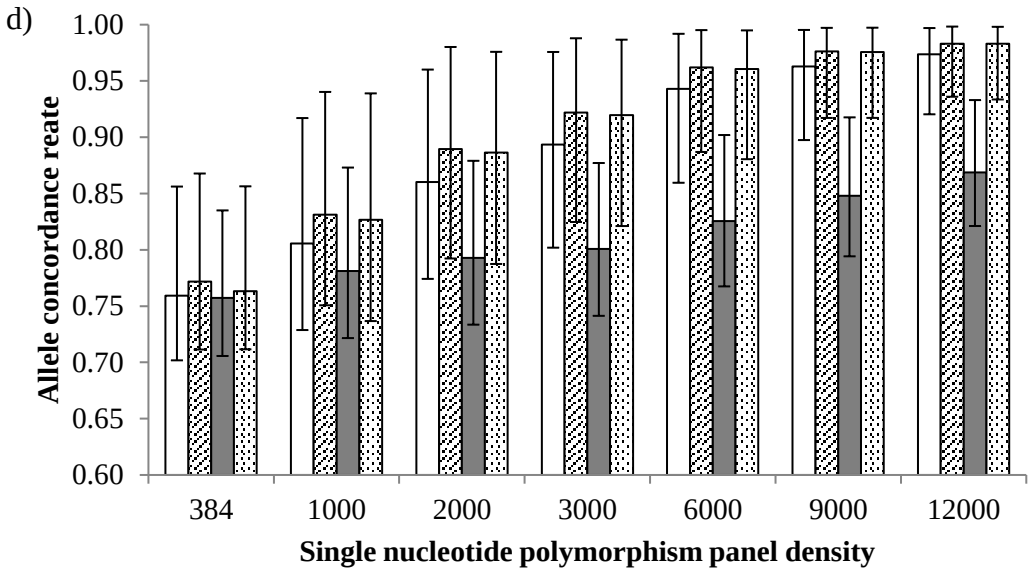


694

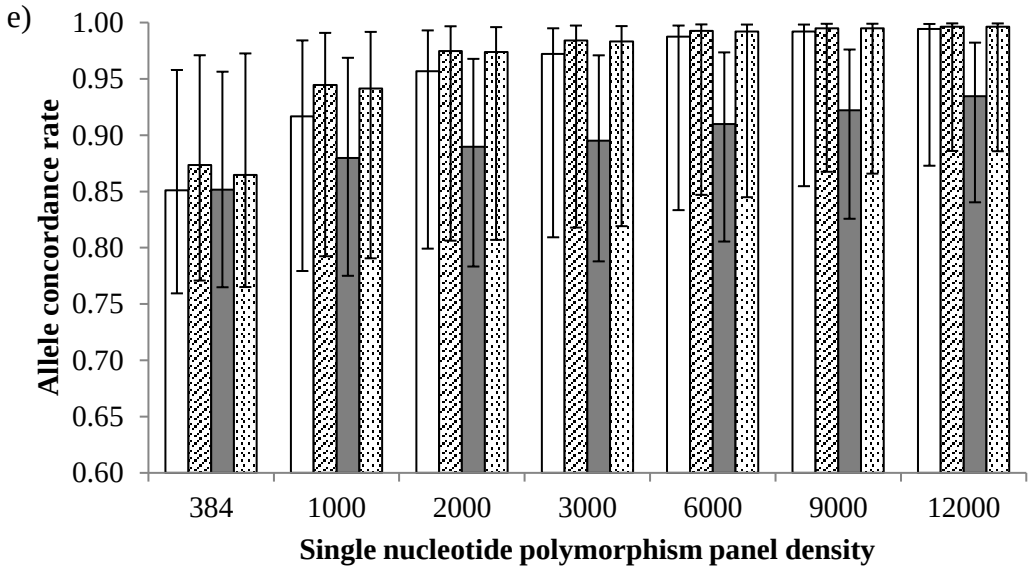
c)



695



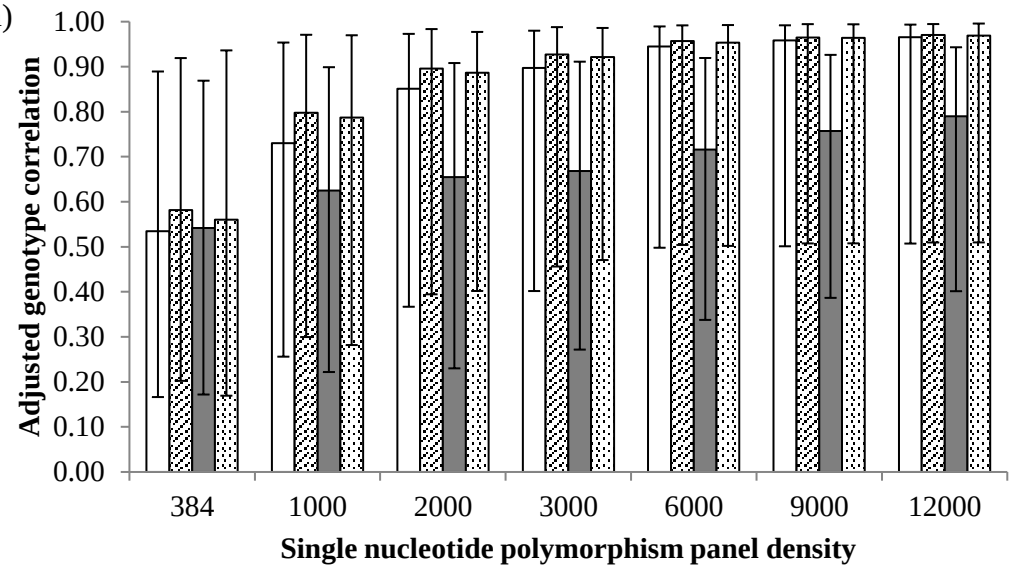
696



697

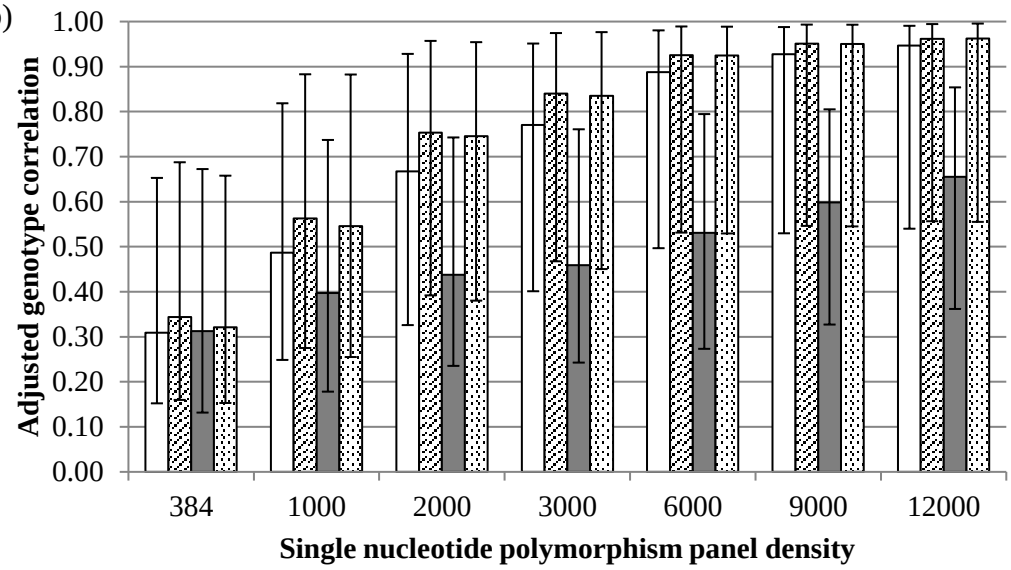
698 **Figure 1.** Mean allele concordance rate per animal across multiple single nucleotide
699 polymorphism (SNP) panels for a) Belclare, b) Charollais, c) Suffolk, d) Texel and e)
700 Vendeen. SNPs were selected randomly (white bar), using the block method (striped bar),
701 using the EquiMAF method (dark grey bar) or using the Wellman method (spotted bar). SNP
702 panels were created within each breed separately and the reference and validation populations
703 were composed only of animals of that breed (Scenario 1). The error bars represent the best
704 and worst mean allele concordance rate per animal.

a)



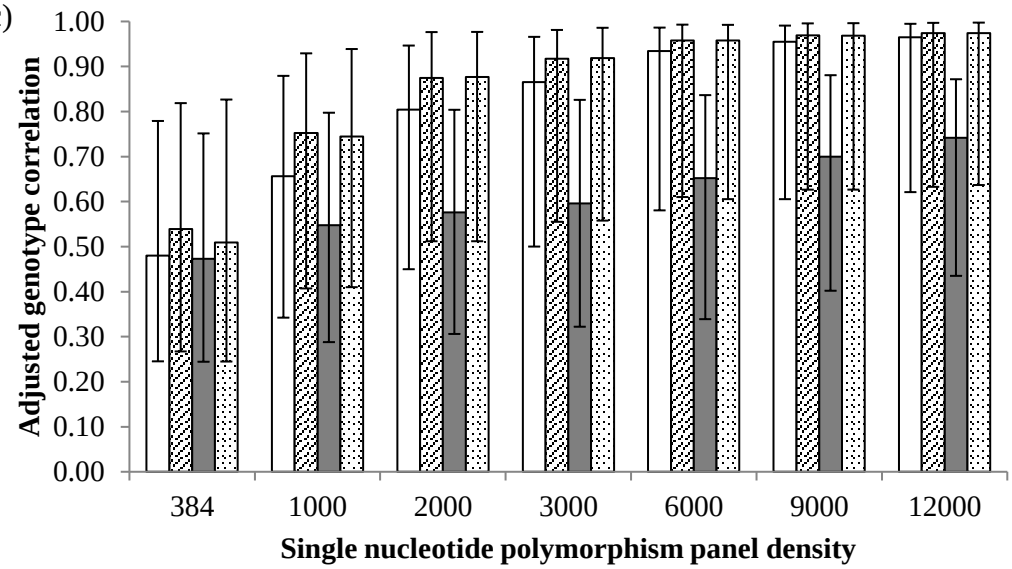
705

b)

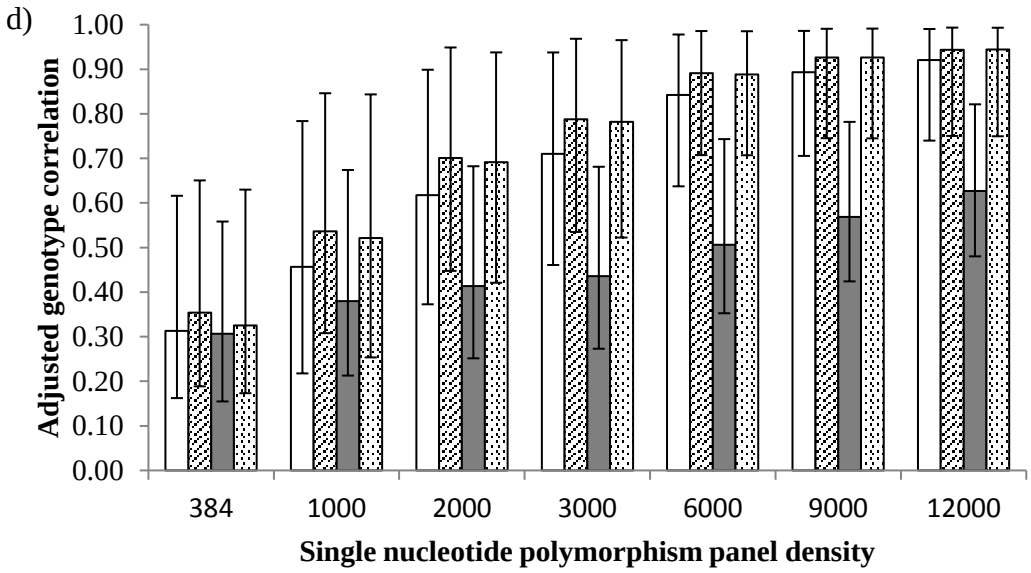


706

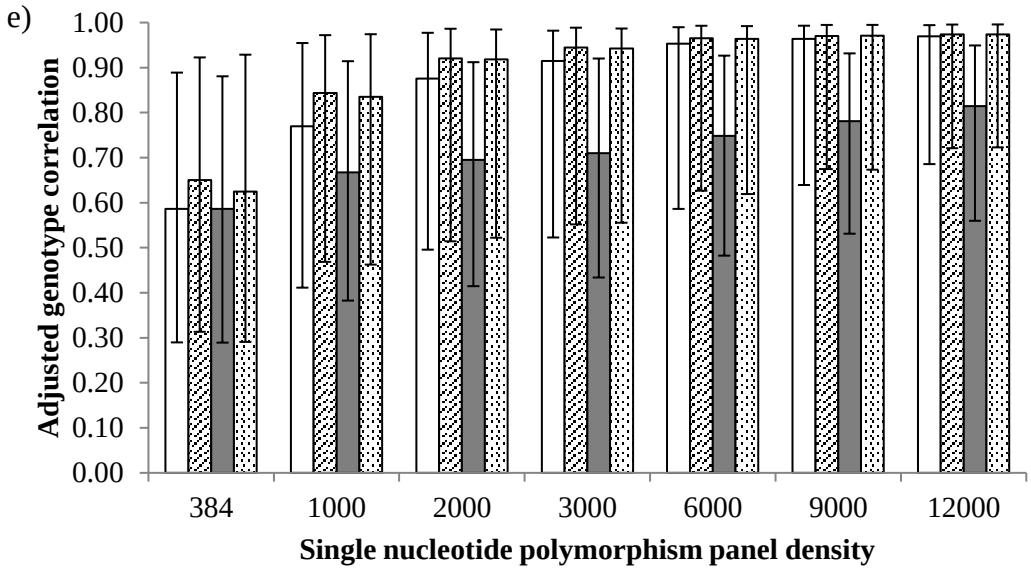
c)



707



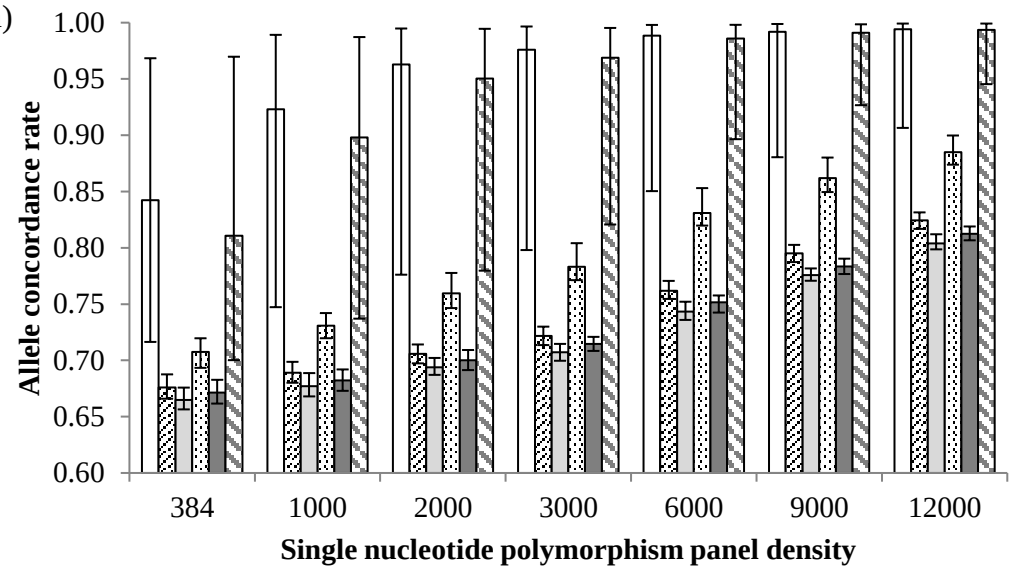
708



709

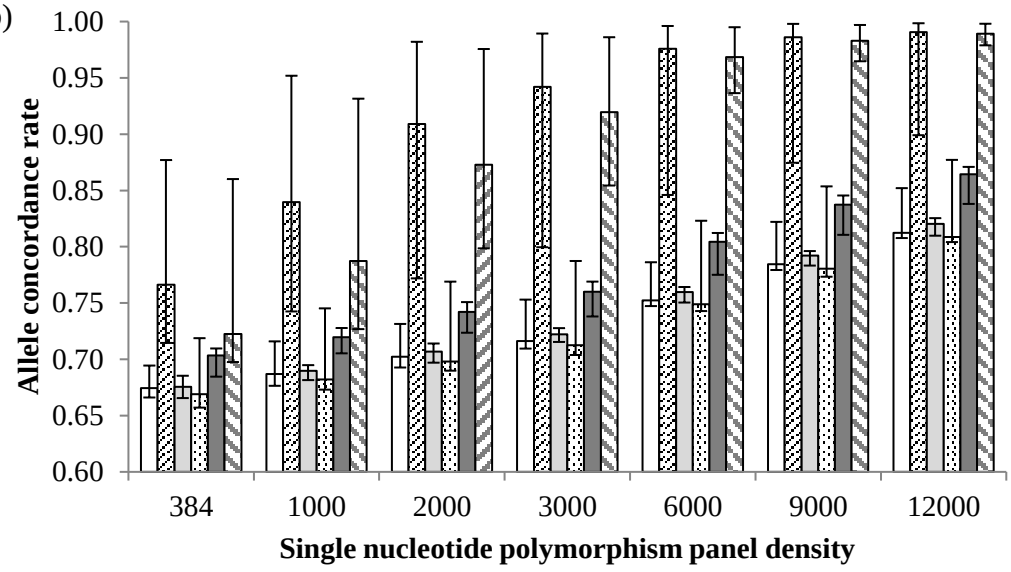
710 **Figure 2.** Mean adjusted genotype correlation per animal across multiple single nucleotide
711 polymorphism (SNP) panels for a) Belclare, b) Charollais, c) Suffolk, d) Texel and e)
712 Vendeen. SNPs were selected randomly (white bar), using the block method (striped bar),
713 using the EquiMAF method (dark grey bar) or using the Wellman method (spotted bar). SNP
714 panels were created within each breed separately and the reference and validation populations
715 were composed only of animals of that breed (Scenario 1). The error bars represent the best
716 and worst mean allele concordance rate per animal.

a)



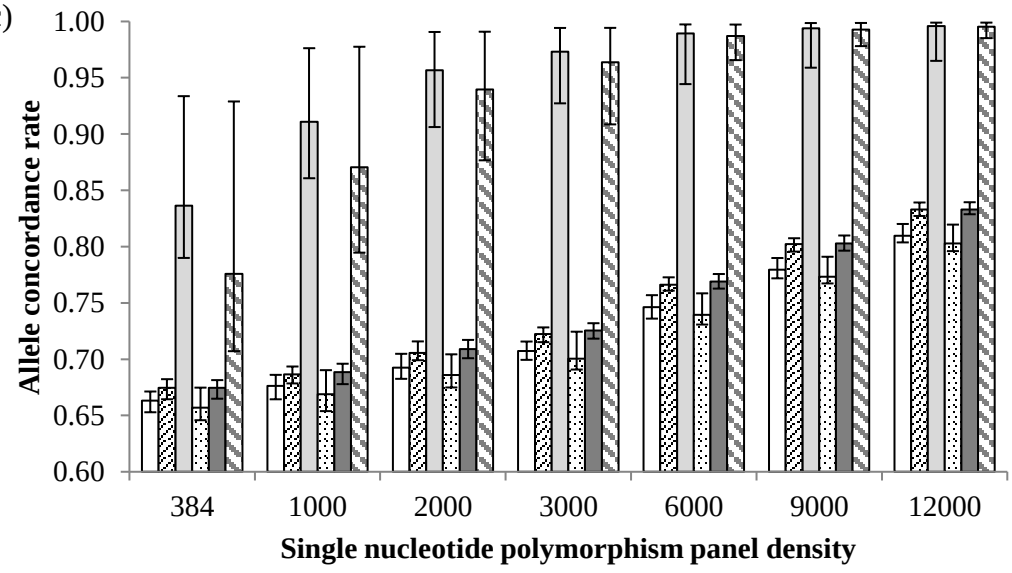
717

b)

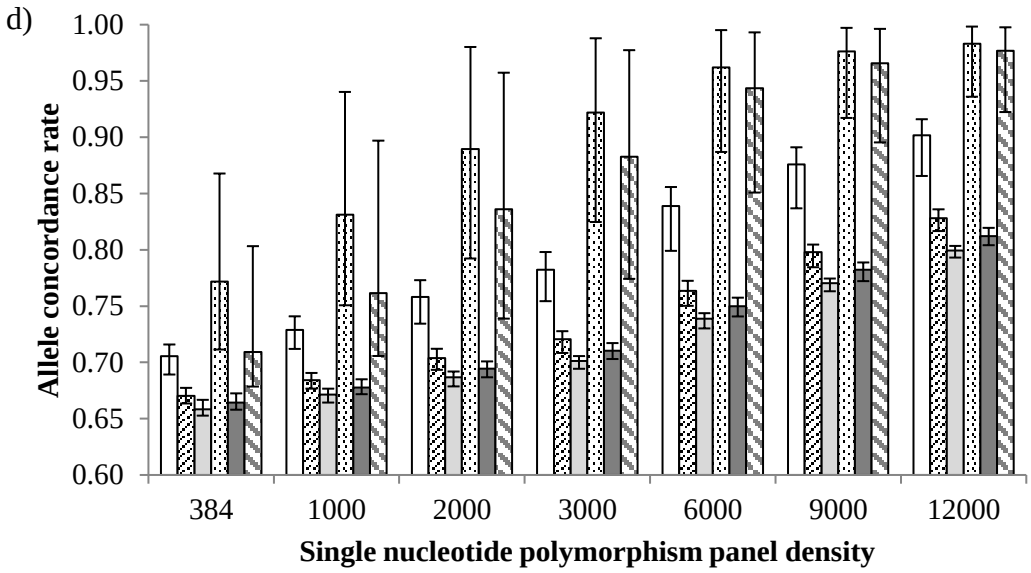


718

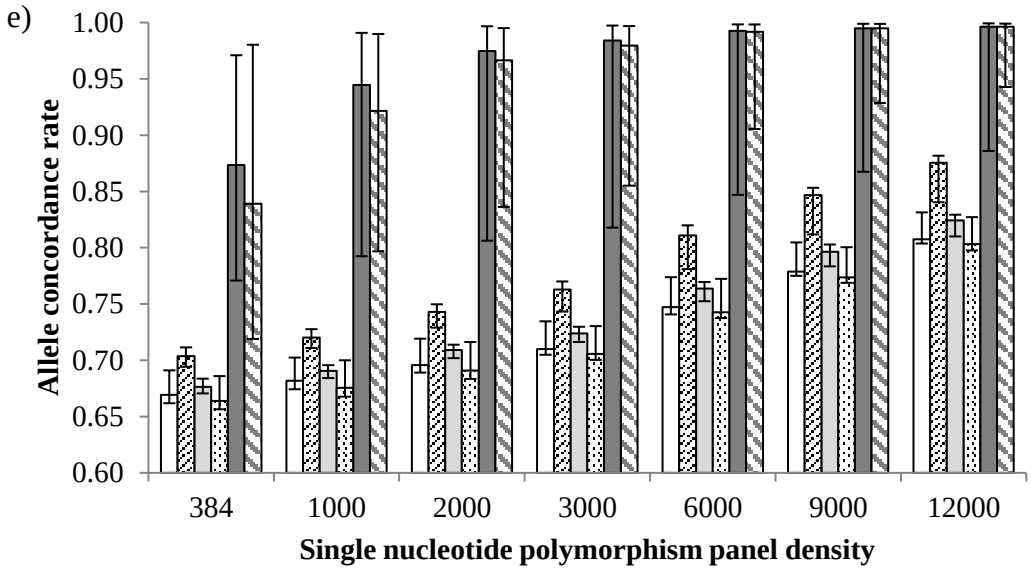
c)



719



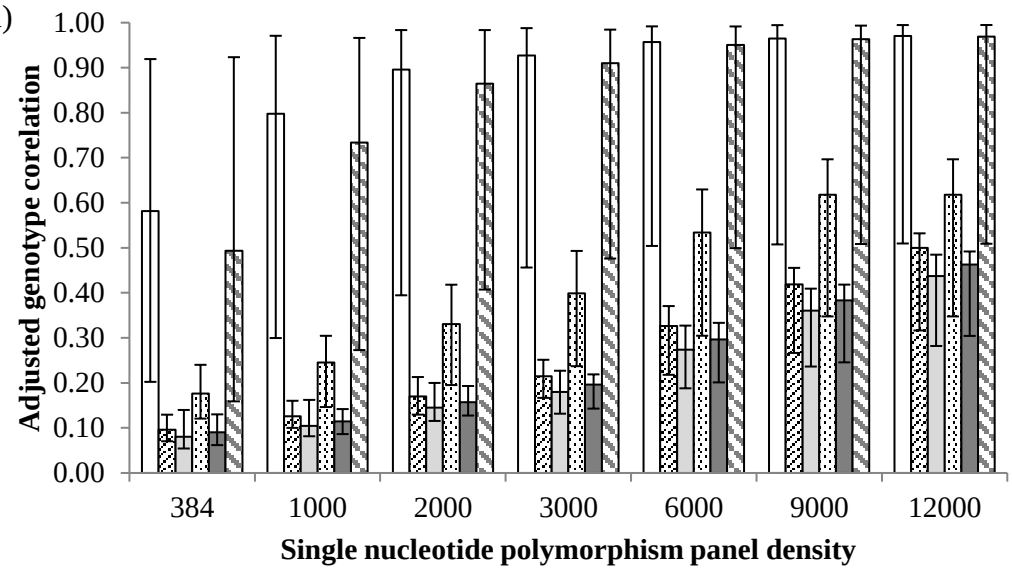
720



721

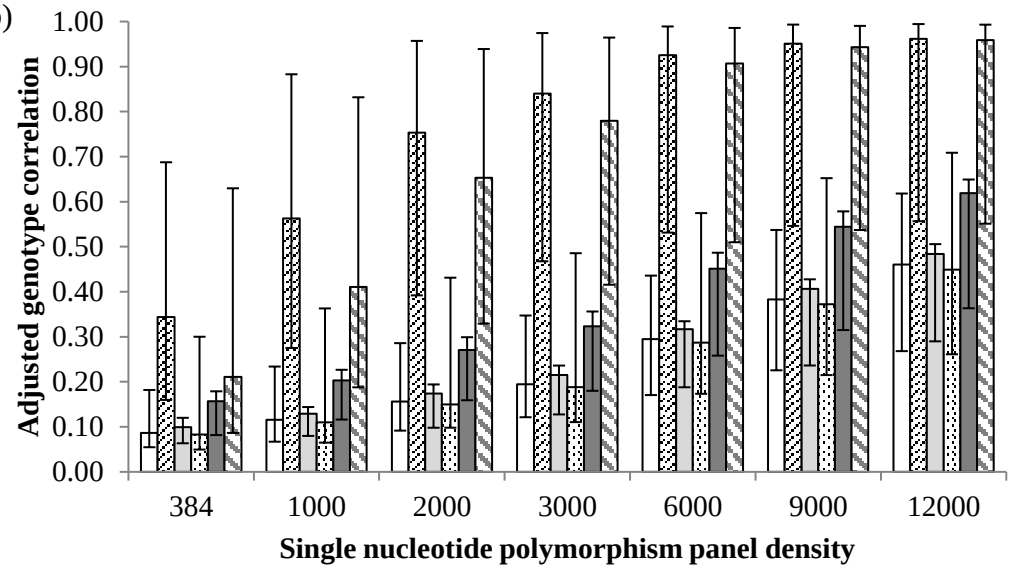
722 **Figure 3.** Mean allele concordance rate per animal across multiple single nucleotide
723 polymorphism (SNP) panels for a) Belclare, b) Charollais, c) Suffolk, d) Texel and e)
724 Vendeen. SNPs were selected using the block method while the reference population used to
725 impute the genotypes are denoted as Belclare (white bar), Charollais (black thin striped bar),
726 Suffolk (light grey bar), Texel (spotted bar), Vendeen (dark grey bar). Multi-breed imputation
727 was undertaken including all breeds in the SNP selection and the reference population (grey
728 thick striped bar). The error bars represent the best and worst mean allele concordance rate
729 per animal.

a)



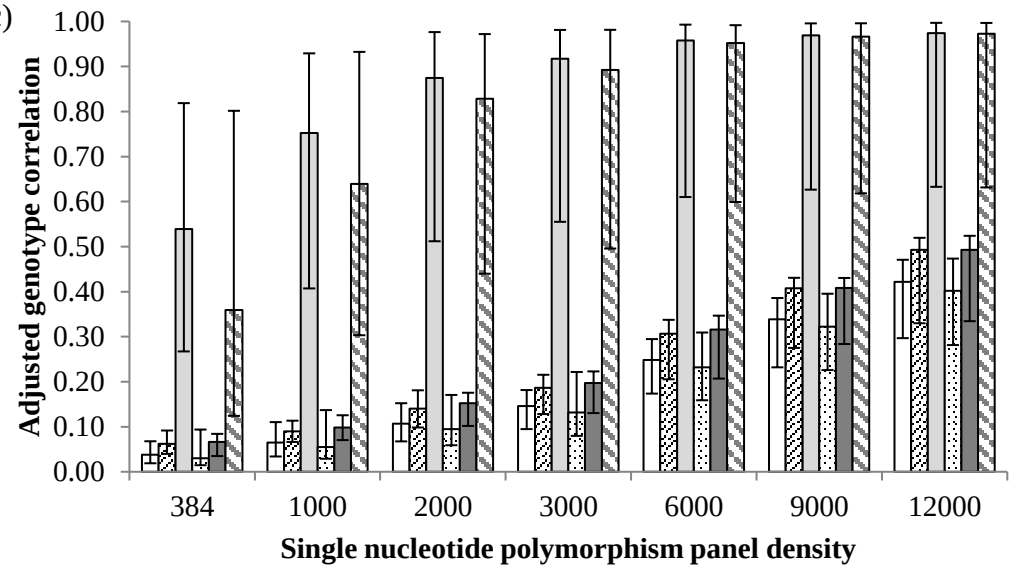
730

b)

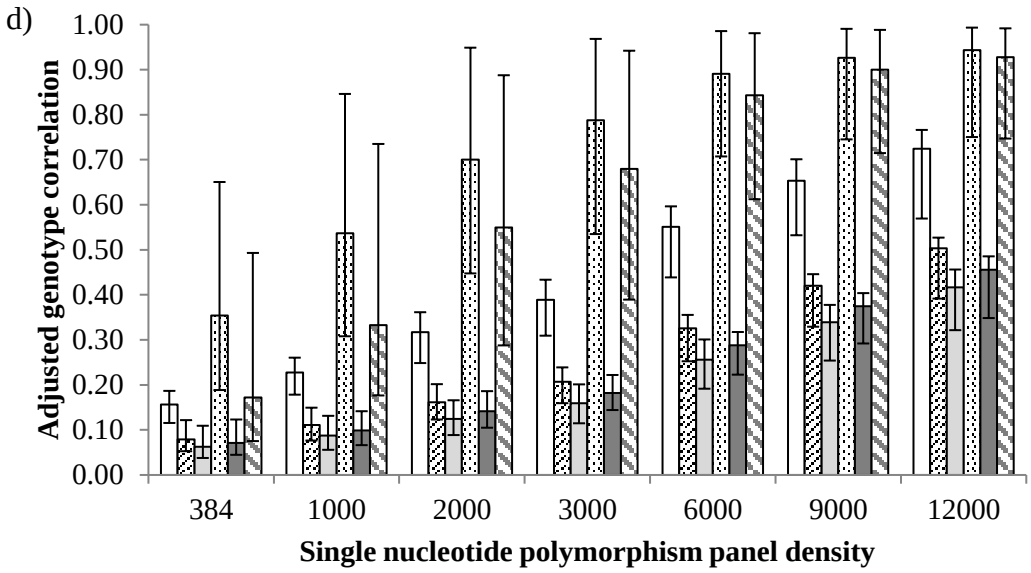


731

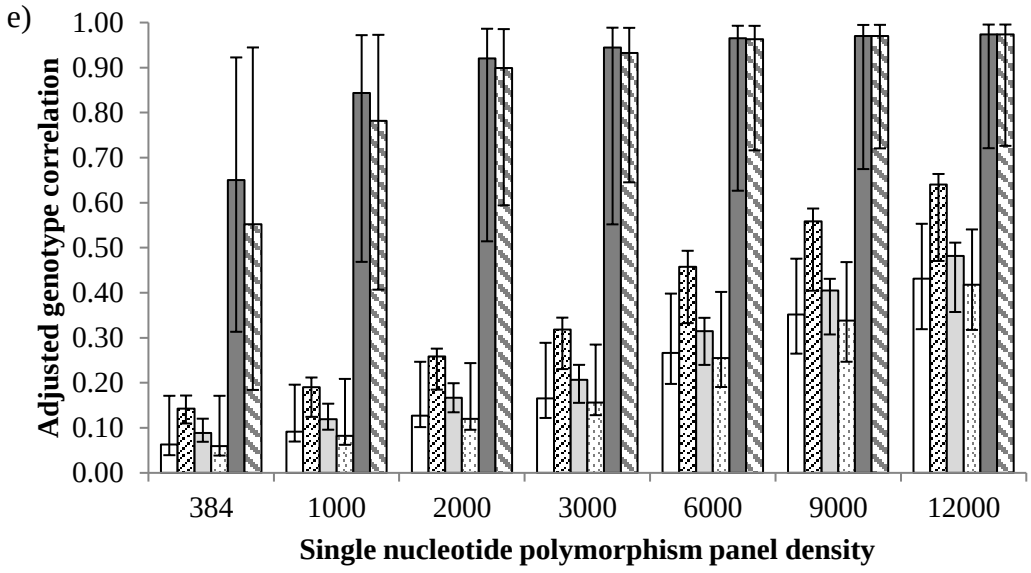
c)



732

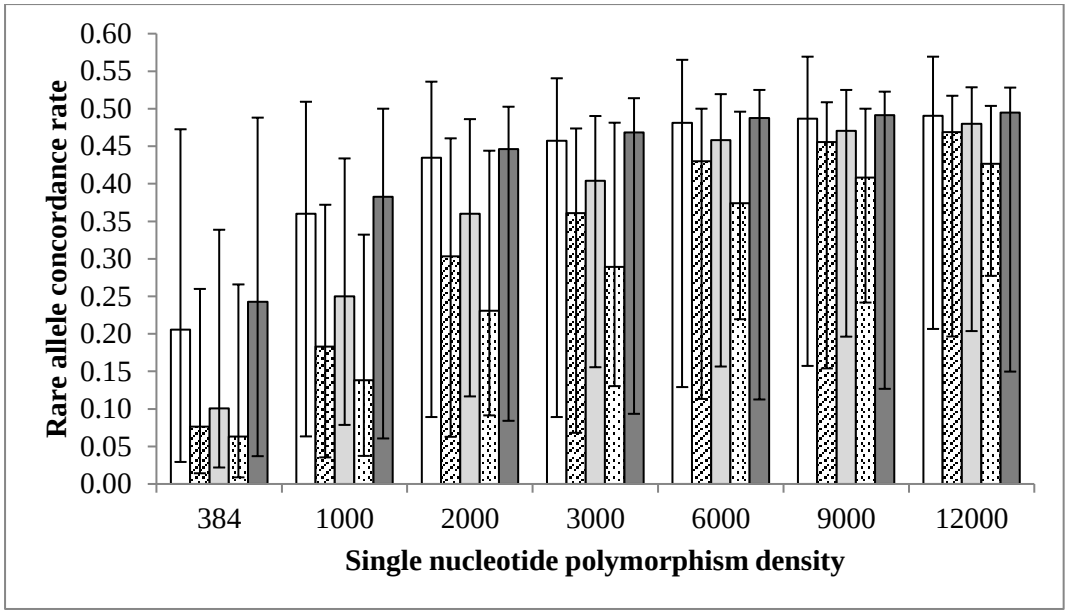


733



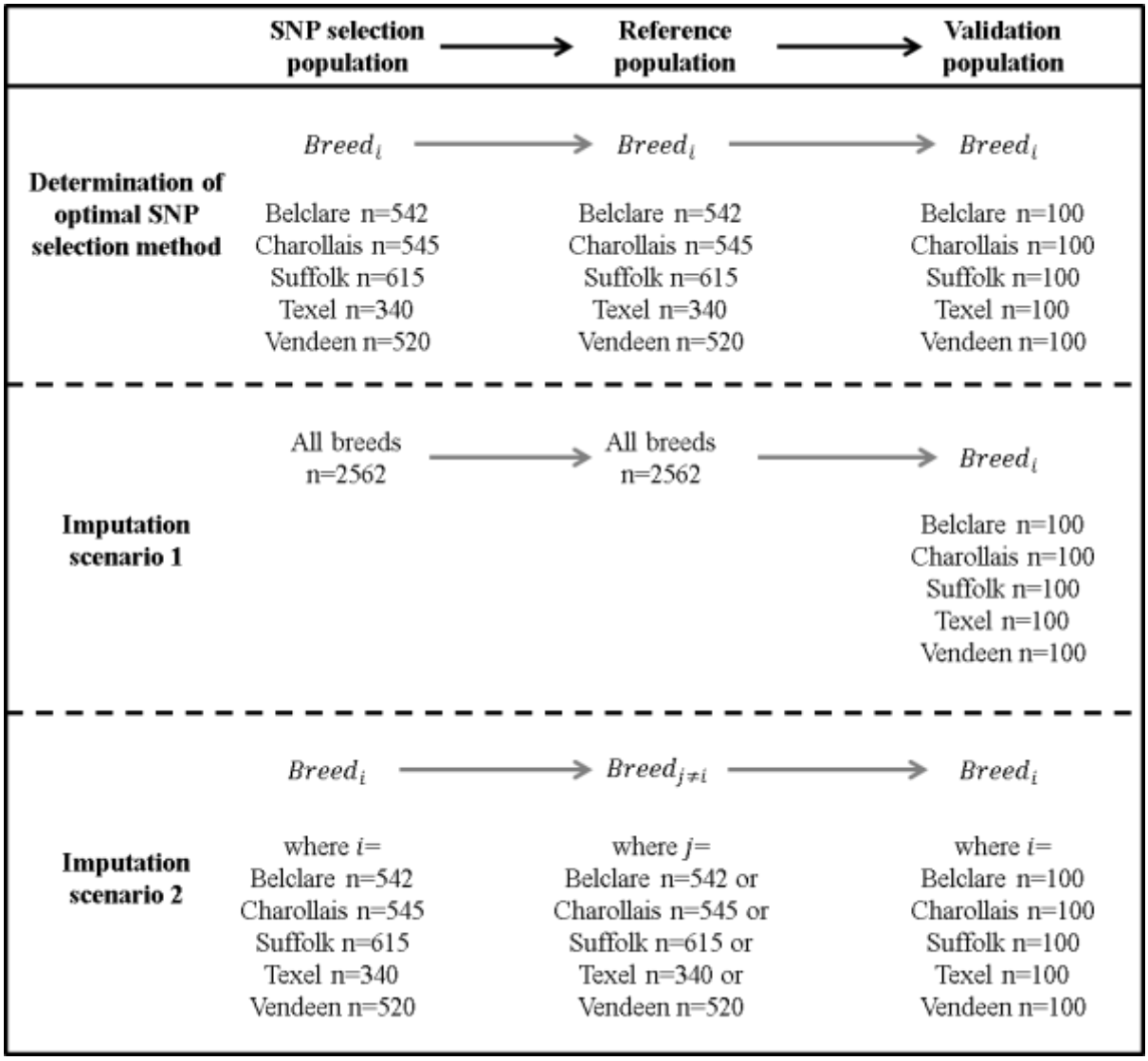
734

735 **Figure 4.** Mean adjusted genotype correlation between true and imputed genotype per animal
736 across multiple single nucleotide polymorphism (SNP) panels densities for a) Belclare, b)
737 Charollais, c) Suffolk, d) Texel and e) Vendeen. Single nucleotide polymorphisms were
738 selected using the block method while the reference population used to impute the genotypes
739 are denoted as Belclare (white bar), Charollais (black thin striped bar), Suffolk (light grey
740 bar), Texel (spotted bar), Vendeen (dark grey bar). Multi-breed imputation was undertaken
741 including all breeds in the SNP selection and the reference population (grey thick striped
742 bar). The error bars represent the best and worst mean allele concordance rate per animal.



743

744 **Figure 5.** The mean rare allele concordance rate per animal across genotype panels of
745 multiple densities selected using the block method within each breed individually; only the
746 breed within which the single nucleotide polymorphisms (SNPs) were selected were included
747 solely in the reference and validation population. The number of SNPs with a MAF ≤ 0.05
748 differed between breeds; Belclare (3,526 SNPs; white bars), Charollais (3,437 SNPs; striped
749 bar), Suffolk (6,802 SNPs; light grey bar), Texel (4,652 SNPs; spotted bar), and Vendeen
750 (5,214 SNPs; dark grey bar). The error bars represent the best and worst mean rare allele
751 concordance rate per animal.



Appendix 1. Summary of the imputation scenarios used in the present study.