



Genotype imputation accuracy in multiple equine breeds from medium- to high-density genotypes

Marjorie Chassier, Eric Barrey, Céline Robert, Arnaud Duluard, Sophie Danvy, Anne Ricard

► To cite this version:

Marjorie Chassier, Eric Barrey, Céline Robert, Arnaud Duluard, Sophie Danvy, et al.. Genotype imputation accuracy in multiple equine breeds from medium- to high-density genotypes. *Journal of Animal Breeding and Genetics*, 2018, 135 (6), pp.420-431. 10.1111/jbg.12358 . hal-02621670

HAL Id: hal-02621670

<https://hal.inrae.fr/hal-02621670>

Submitted on 26 May 2020


HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Genotype imputation accuracy in multiple equine breeds from medium- to high-density genotypes

Marjorie Chassier¹  | Eric Barrey¹ | Céline Robert^{1,2} | Arnaud Duluard³ |
Sophie Danvy⁴ | Anne Ricard^{1,4}

¹Unité Mixte de Recherche
1313 Génétique Animale et Biologie
Intégrative, Département Sciences du
Vivant, Institut National de la Recherche
Agronomique, AgroParisTech, Université
Paris Saclay, Jouy-en-Josas, France

²Ecole Nationale Vétérinaire d'Alfort,
Maisons Alfort, France

³Département élevage et santé animale,
Le Trot, Paris, France

⁴Institut Français du Cheval et de
l'Equitation, Pôle développement,
Innovation et Recherche, Exmes, France

Correspondence

Marjorie Chassier, Unité Mixte de
Recherche 1313 Génétique Animale et
Biologie Intégrative, Département
Sciences du Vivant, Institut National de la
Recherche Agronomique, AgroParisTech,
Université Paris Saclay, Jouy-en-Josas,
INRA de Jouy en Josas Domaine de
vilvert F-78352 JOUY EN JOSAS Cedex,
France.
Email: marjorie.chassier@inra.fr

Funding information

Institut Français du Cheval et de
l'Equitation; the Institut National de la
Recherche Agronomique (INRA); the
Fonds Eperon

Abstract

Genotype imputation is now a key component of genomic analyses as it increases the density of available genotypes within a population. However, many factors can influence imputation accuracy. The aim of this study was to assess and compare the accuracy of imputation of high-density genotypes (Affymetrix Axiom Equine genotyping array, 670,806 SNPs) from two moderate-density genotypes (Illumina Equine SNP50 BeadChip, 54,602 SNPs and Illumina Equine SNP70 BeadChip, 65,157 SNPs), using single-breed or multiple-breed reference sets. Genotypes were available from five groups of horse breeds: Arab (AR, 1,207 horses), Trotteur Français (TF, 979 horses), Selle Français (SF, 1,979 horses), Anglo-Arab (AA, 229 horses) and various foreign sport horses (FH, 209 horses). The proportions of horses genotyped with the high-density (HD) chip in each breed group were 10% in AA, 15% in AR and FH, 30% in TF and 57% in SF. A validation set consisting of one-third of the horses genotyped with the HD chip was formed and their genotypes deleted. Two imputation strategies were compared, one in which the reference population consisted only of horses from the same breed group as in the validation set, and another with horses from all breed groups. For the first strategy, concordance rates (CRs) ranged from 97.8% (AR) to 99.0% (TF) and correlations (r^2) from 0.94 (AR) to 0.99 (TF). For the second strategy, CR ranged from 97.4% (AR) to 98.9% (TF) and r^2 from 0.93 (AR) to 0.99 (TF). Overall, the results show a small advantage of within-breed imputation compared with multi-breed imputation. Adding horses from different breed groups to the reference population does not improve the accuracy of imputation. Imputation provides an accurate means of combining data sets from different genotyping platforms, now necessary with the increasing use of the recently developed Affymetrix Axiom Equine genotyping array.

KEYWORDS

breed group structure, genotype imputation, high- and medium-density SNP chips, horse, reference population

1 | INTRODUCTION

The first genome sequence of the domestic horse was published in November 2009 as the result of the collaborative effort of the worldwide equine research community (Wade et al., 2009). Shortly afterwards, a genotyping array (Illumina Equine SNP50 BeadChip, 54k) containing 54,602 single nucleotide polymorphism (SNP) markers was developed (McCue et al., 2012). In January 2011, a second-generation Illumina chip (Illumina Equine SNP70 BeadChip, 65k) was developed, containing 65,157 SNPs, of which 19,171 were new markers and 45,986 were already present in the Equine SNP50 BeadChip (Coleman et al., 2010). After these two medium-density (MD) chips, a high-density (HD) commercial chip, the Affymetrix Axiom Equine genotyping array (670k), was developed by the selection of 670,806 SNPs amongst the 2 million SNPs of an experimental chip designed from multiple breeds by the international consortium. This chip consisted of 626,710 new SNP markers and 44,096 SNPs already present on both MD chips (Schaefer et al., 2017). Currently, the Affymetrix Axiom Equine genotyping array and the Illumina Equine SNP70 BeadChip are the two commercial chips available. The Illumina Equine SNP50 BeadChip is no longer available.

These three chips have been used successively in equine studies, and therefore, the horses have been genotyped with three different SNP densities. However, combining the data sets generated which each chip results in a loss of information, especially in genomic studies that require standardized data because only SNPs present in all data sets can be used. Genotype imputation could help to overcome this problem by generating HD genotypes for animals genotyped with low- or MD chips. However, genotype imputation with different density chips may be a challenge in terms of imputation accuracy (Pereira et al., 2017).

In livestock, the accuracy of genotype imputation has mainly been investigated in cattle, for imputing MD (50k) to HD (777k) SNP panels (Pausch et al., 2013). Hozé et al. (2013) reported genotype imputation accuracies >97% in dairy cattle, suggesting that the imputation of HD genotypes was accurate. Genotype imputation has also been investigated in pig (Gualdrón Duarte et al., 2013) and sheep (Hayes, Bowman, Daetwyler, Kijas, & van der Werf, 2012; Moghaddar, Gore, Daetwyler, Hayes, & van der Werf, 2015), and these studies reported high imputation accuracy (>96%). Cross-breeding is accepted between most horse breeds leading to relatives open studbooks with breeds not as strictly defined as in other livestock. This difference can impact genotype imputation accuracy, relative to the other species. Few studies have investigated the accuracy of genotype imputation in the horse. McCoy and

McCue (2014) assessed imputation accuracy between the two generations of MD chips (54k and 65k) and demonstrated higher imputation accuracy with a breed-matched reference population than with a mixed-breed reference population. Corbin et al. (2014) assessed imputation accuracy from a very-low-density marker panel (1-6k), developed by them, to a MD chip (65k). Frischknecht et al. (2014) investigated the imputation to whole genome sequence from the first-generation Illumina chip (54k). And recently, Schaefer et al. (2017) assessed imputation accuracy from the HD chip (670k) to a set of 2 million SNPs selected by them. However, none of the studies investigated the accuracy of the genotype imputation from medium- to high-density chip genotypes.

Due to the current need to merge genotyping data produced with different marker densities in various equine breeds, the aim of this study was to assess the accuracy of the imputation of HD genotypes (670k) from MD genotypes (54k and 65k) and to compare “within-breed” imputation and “multi-breed” imputation.

2 | MATERIALS AND METHODS

2.1 | Animals

The data used in this study consisted of 4,603 genotyped horses with three different SNP density chips: the Illumina Equine SNP50 BeadChip (54k) that includes 54,602 SNPs, the Illumina Equine SNP70 BeadChip (65k) that includes 65,157 SNPs and the Affymetrix Axiom Equine genotyping array (670k) that includes 670,806 SNPs. The 54k and 65k chips are both referred to as MD chips, whilst the 670k is referred to as the HD chip.

Amongst the 4,603 horses genotyped, 37% were genotyped with the Illumina Equine SNP50 BeadChip, 26% with the Illumina Equine SNP70 BeadChip and 36% with the Affymetrix Axiom Equine genotyping array. The horses were born between 1967 and 2012, with 77.34% of them born between 2000 and 2012. The horses, 1,535 males, 1,689 females and 1,379 geldings, descended from 1,536 stallions and 4,148 mares, with an average of 2.99 offspring per stallion and 1.11 offspring per mare. Of these, 529 of the stallions and 41 of the mares were themselves genotyped with the MD chip. The genotyped horses belonged to 45 breeds which were grouped into five groups of horse breeds that, according to the rules of the genealogical books and past genealogical analyses (Leroy et al., 2009), are assumed to be relatively homogeneous: Arab (AR: 1,207 horses), Trotteur Français (TF: 979 horses), Selle Français (SF: 1,979 horses), Anglo-Arab (AA: 229 horses) and various foreign sport horses (FH: 209 horses). The proportion of horses genotyped with the various density arrays differed between the breed groups (Table 1):

26% of Arabs, 43% of Selle Français, 21% of Trotteur Français, 5% of Anglo-Arabs, and 5% of foreign sport horses.

2.2 | Genotypes and quality control

Genotyped animals with an average call rate <0.95 were not considered further for the analysis. Parentage tests were performed according to the French genomic evaluation procedure used in dairy cattle (Boichard et al., 2012), using the 54k, 65k and 670k data sets. This procedure uses 1,000 informative markers over the whole genome. For each marker, an incompatibility rate of Mendelian mismatch between parents and progeny is calculated. A parentage error was flagged if the incompatibility rate was $>5\%$. Progeny with inconsistent genotypes were removed, except if inconsistencies were found for at least two offspring of a given sire. In such cases, the sire was removed if his call rate was lower than 0.99.

Table 2 shows the number of SNPs retained and the number of overlapping SNPs for the three genotyping arrays considered. Quality control procedures, based on combined breeds and per chip, were applied to generate the list of SNPs used in the analysis. SNPs with an unknown chromosomal position and SNPs on the X and Y chromosomes were excluded, as were SNPs with identical chromosomal positions but different SNP-IDs (duplicates). Then, we retained SNPs that were genotyped in more than 90% of horses on at least one chip, had a minor allele frequency (MAF) $> 2\%$ in at least one breed group, showed non-significant ($p > 10^{-6}$) deviation from the Hardy–Weinberg equilibrium in at least one breed group and had a p -value for the difference of the within-group major allele frequency between chips $>10^{-5}$ in at least one breed group for each pair of chips. If a SNP was reported in the literature to have a significant effect on a trait (e.g., DMRT3 and MSTN), it was also retained for analysis. Finally, SNPs present on the 54k or 65k chip but not on the 670k chip were excluded because they were not relevant for measuring the quality of MD to HD genotype imputation.

TABLE 1 Distribution of genotyped horses by breed group

Chip	Density	AR	TF	SF	AA	FH	Total
Illumina 54K	MD	7	687	745	137	168	1,744
Illumina 65K	MD	1,021	0	100	70	9	1,200
Affymetrix 670K	HD	179	292	1,134	22	32	1,659
Total		1,207	979	1,979	229	209	4,603

Notes. AR, Arabs; AA, Anglo-Arabs; SF, Selle Français; TF, Trotteur Français; FH: foreign sport horses and type of chip (MD: medium density, HD: high density), after quality control.

TABLE 2 Distribution of the SNPs (0/1: missing/present) retained on the three available horse chips: the Illumina 54k and 65k (medium-density chips) and the Affymetrix 670k (high-density chip)

Illumina 54k	Illumina 65k	Affymetrix 670k	Number of SNPs
0	0	1	455,711
0	1	1	15,159
1	0	1	6,360
1	1	1	40,266

The final data set comprised 4,603 horses and 517,496 SNPs.

2.3 | Strategy for genotype imputation

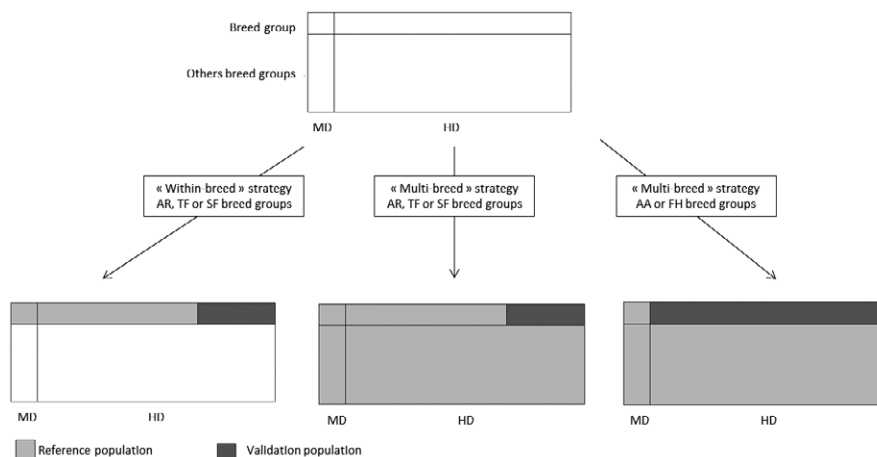
We applied a cross-validation scheme to assess imputation accuracy. The HD data set was divided into a reference and a validation population. Complete genotype information was retained for animals in the reference population. For animals in the validation population, markers that were only present on the HD chip were masked to mimic a target population genotyped with the MD chip. The MD data set was added to the reference population. These latter horses did not give information about SNP only present on HD chip but (a) will be in the future routine process of imputation and (b) were part of calculation of linkage disequilibrium between SNP present on both MD and HD chips. MD data were not used to calculate imputation accuracy.

For the data from the AR, TF and SF breed groups, two strategies were compared for the reference population: (a) only horses from the same breed group were included in the reference population (“within-breed”) or (b) all horses from the five breed groups were included in the reference population (“multi-breed”). The validation populations were the same for both strategies, comprised one-third of the horses with HD genotype data from the breed group studied, selected at random, and consequently, the reference population comprised the remaining two-third of horses with HD genotypes (Figure 1). The process was repeated three times with no replacements, leading to three validation populations and three reference populations for each of the AR, TF and SF breed groups, and the imputation results were pooled for analysis.

Due to the small number of individuals in AA and FH, only the second strategy was tested, and the validation population consisted of all the horses with HD genotypes for the breed studied.

The final aim of this study was to accurately impute the genotyped horses with MD genotypes to HD genotypes. FIMPUTE 2.2 software was used for the imputation. This

FIGURE 1 Composition of reference and validation populations for the Arab, Trotteur Français, Selle Français, Anglo-Arabs and foreign sport horses breed groups and for “within-breed” and “multi-breed” strategies. Breed group contains the horses genotyped with the medium-density chip (MD) and the high-density chip (HD) from the breed group studied. Other breed groups contains the horses from the four other breed groups, genotyped with the MD and HD



software uses linkage disequilibrium and pedigree information and provides similar or higher imputation accuracy than the alternative softwares Beagle and IMPUTE2, whilst being faster and easier to use with large data sets (Sargolzaei, Chesnais, & Schenkel, 2014). Imputation was performed for each chromosome separately.

2.4 | Evaluation of imputation accuracy

Imputation accuracy was assessed for the 455,711 imputed SNPs based on (a) the concordance rate (CR), determined as the proportion of the correctly imputed alleles out of all the alleles inferred by imputation, and (b) the correlation (r^2), determined as the squared Pearson's correlation between the true and imputed genotypic allele counts. These measures of accuracy were calculated per individual (over all SNPs) and per SNP (over all animals). Summary statistics were then calculated across all horses, per breed group, and across SNPs, per chromosome. SNPs that were monomorphic in a breed group were excluded from both sets of statistical analyses.

In order to explore the possible causes of differences in imputation accuracies, we considered the impact of distances between breed groups, allele frequencies, linkage disequilibrium and SNP density. Distances between breed groups were studied by cluster analysis performed on the genomic relationship matrix by principal component analysis using GenABEL packages (Aulchenko, Ripke, Isaacs, & van Duijn, 2007) and the genotypes of all horses genotyped with the HD chip ($n = 1,659$). The impact of allele frequency was characterized by the MAF. Linkage disequilibrium was measured by the squared correlation (r_{LD}^2) between genotypes (allele counts). SNP density was measured by the number of SNPs divided by then length of DNA segments in base pairs. Analyses were performed on different scales: (a) over the whole genome or breed group, (b) for each *Equus Caballus* chromosome (ECA) and (c) within individual chromosomes. At the genome level (a),

the following analyses were performed: for allele frequency analysis, SNPs were classified in 10 bins according to MAF: [0, 0.05], [0.05, 0.1], [0.1, 0.15], [0.15, 0.20], [0.20, 0.25], [0.25, 0.30], [0.30, 0.35], [0.35, 0.40], [0.40, 0.45], [0.45, 0.50]. The mean CR and r^2 were calculated for each bin. For linkage disequilibrium, pairs of SNPs were ranked according to distance by intervals of 1,000 bp from 0 to 100,000 bp, and the mean r_{LD}^2 was computed for each interval, and each breed group for the HD and MD chips. At the chromosome level (b), the mean values of MAF of all SNPs of the chromosome, r_{LD}^2 for all pairs of SNP distant from 3,001 bp to 4,000 bp, and SNP density (number of SNP on the chromosome divided by length of chromosome in bp) were plotted against the mean r^2 for each chromosome. At the intra-chromosomal level (c), the chromosome was divided in windows of 1 Mb. In each window, mean r^2 , mean MAF for all SNPs and r_{LD}^2 between all SNP pairs were calculated. Density was the number of SNP in the window divided by 10^6 . Multiple linear regression analysis was performed to estimate mean r^2 values in the 1 Mb window from MAF, r_{LD}^2 and SNP density within the five breed groups. The following model was used for the analysis of multiple linear regression models:

$$y = \alpha + \beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3$$

where y is the correlation (r^2), α is the constant, β_{1-3} are regression coefficients of MAF (w_1), r_{LD}^2 (w_2), and SNP density (w_3 in bp^{-1}) calculated for each 1 Mb window.

3 | RESULTS

3.1 | Breed group structure

Cluster analysis revealed three major clusters: AR, TF and a pool of SF, AA and FH (Figure 2). Moreover, some AA or SF horses, with high percentages of AR ancestors, were closer to the AR cluster than AA or SF clusters.

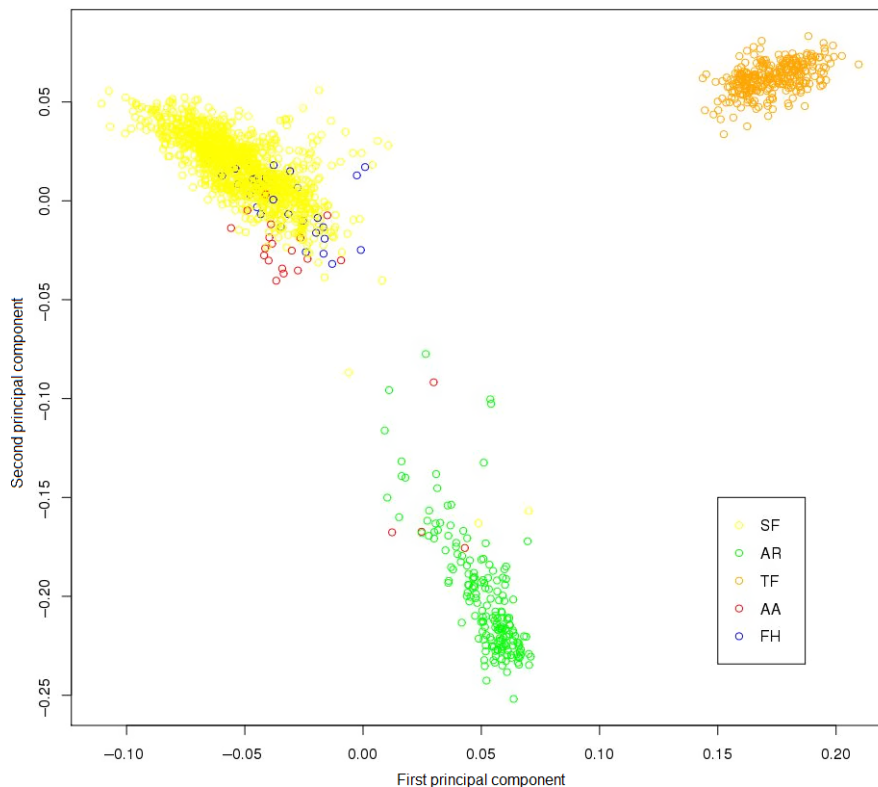


FIGURE 2 Plot of horses on the first two principal components for the 1,659 horses genotyped with the high-density chip. Yellow, green, orange, red and blue circles represent Selle Français, Arabs, Trotteur Français, Anglo-Arabs and foreign sport horses breed groups, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Linkage disequilibrium

Results for decay of linkage disequilibrium are shown in Figure 3. The linkage disequilibrium between the SNPs of the MD chip was higher than that between the SNPs of the HD chip. In both MD and HD chips, r_{LD}^2 strongly decreased within the first 5 kb in all breed groups. Thereafter it decreased at a lower rate, and differences between breed groups could be observed. Linkage disequilibrium levels were highest for AA, intermediate for FH, TF, and AR, and lowest for SF.

3.3 | Imputation accuracy at the breed and genome levels

The sizes of the reference and validation populations, CR and r^2 , per individual and per SNP, are provided for each strategy in Table 3. For the “within-breed” strategy, mean individual CR values ranged from 97.84% (AR) to 99.03% (TF). For the “multi-breed” strategy, mean individual CR values ranged from 97.37% (AR) to 98.89% (TF).

With both strategies, the mean individual CR was highest for TF, intermediate for SF and lowest for AR. With the “multi-breed” strategy, FH and AA had intermediate mean individual CR values.

The pattern of mean CR values per SNP for the different breed groups and both strategies was similar to that obtained per individual. Imputation accuracies were higher for TF than for the SF and AR breed groups.

For the “within-breed” strategy, mean r^2 values per SNP ranged from 0.80 (AR) to 0.89 (TF). For the “multi-breed” strategy, mean r^2 values ranged from 0.76 (AR) to 0.87 (TF). The pattern of mean r^2 values per individual was similar to that obtained per SNP for the different breed groups and both strategies.

Like CR, r^2 was lower for the “multi-breed” strategy compared with the “within-breed” strategy. The differences between the two strategies were larger for the AR breed group than for SF and TF breed groups.

3.4 | Effect of MAF on imputation accuracy

Figures 4 and 5 show the relationships between the mean CR and the mean r^2 per SNP for the 10 SNP bins based on the MAF, for both strategies in the AR, TF and SF breed groups. The patterns were clearly different for the two measures of accuracy: mean CR decreased with MAF (Figure 4), whereas mean r^2 increased (Figure 5). Both tended to plateau at a MAF of 0.50. The general trends of mean CR and mean r^2 against MAF were the same for all breed groups and strategies. In the TF and SF breed groups, mean CR tended to plateau at MAF > 0.10. For SNPs with low MAF (<0.10), the trend for mean CR was more variable.

3.5 | Imputation accuracy per chromosome

Figure 6 shows the mean r^2 values per chromosome and evidences lower mean r^2 values for chromosomes ECA1,

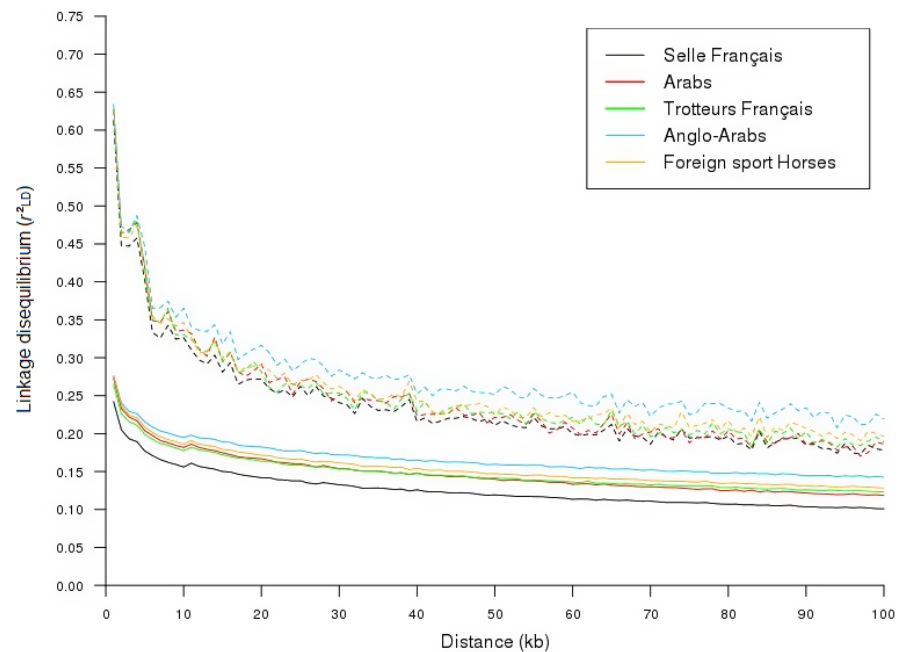


FIGURE 3 Linkage disequilibrium (r^2_{LD}) decay for the five breed groups, between the high-density genotypes (full lines) and the medium-density genotypes (dotted lines) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Mean, standard deviation (*SD*), minimum (*Min*) and maximum (*Max*) for the concordance rate (*CR*) and correlation between genotypes (r^2), per individual and per SNP, in AR (Arabs), TF (Trotteur Français) and SF (Selle Français) breed groups, for the “within-breed” and “multi-breed” strategies, in AA (Anglo-Arabs) and FH (foreign sport horses) breed groups, for the “multi-breed” strategy

	AR		TF		SF		AA	FH
	Within-breed	Multi-breed	Within-breed	Multi-breed	Within-breed	Multi-breed	Multi breed	Multi breed
Per individual								
Npop Ref ^a	119.3	1,599.3	194.6	1,561.6	756.0	1,281.0	1,637.0	1,627.0
Npop Val ^b	59.6	59.6	97.3	97.3	378.0	378.0	22.0	32.0
Mean CR	0.9784	0.9737	0.9903	0.9889	0.9871	0.9865	0.9813	0.9845
SD	0.0120	0.0125	0.0030	0.0032	0.0058	0.0051	0.0121	0.0073
Min	0.8997	0.9224	0.9687	0.9669	0.9169	0.9306	0.9307	0.9590
Max	0.9899	0.9899	0.9949	0.9941	0.9943	0.9942	0.9898	0.9918
Mean r^2	0.9417	0.9254	0.9905	0.9888	0.9870	0.9864	0.9818	0.9847
SD	0.0325	0.0352	0.0029	0.0032	0.0055	0.0049	0.0108	0.0069
Min	0.7188	0.7750	0.9885	0.9664	0.9235	0.9348	0.9381	0.9604
Max	0.9701	0.9707	0.9949	0.9943	0.9944	0.9942	0.9903	0.9918
Per SNP								
Mean CR	0.9769	0.9720	0.9896	0.9880	0.9866	0.9862	0.9797	0.9835
SD	0.0255	0.0321	0.0185	0.0198	0.0188	0.0191	0.0308	0.0253
Min	0.4811	0.1508	0.4685	0.4698	0.5024	0.4901	0.3500	0.3333
Max	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Mean r^2	0.8026	0.7561	0.8886	0.8687	0.8431	0.8322	0.8659	0.8792
SD	0.2009	0.2323	0.1897	0.2071	0.2075	0.2174	0.2020	0.1830
Min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Max	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Notes. ^aN pop Ref, Number of horses in the reference population. ^bN pop Val, Number of horses in the validation population.

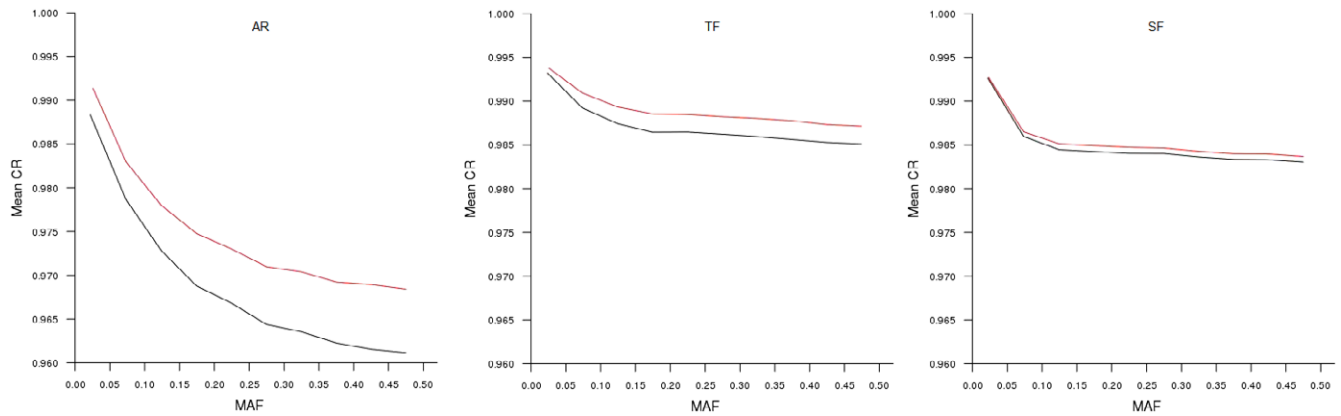


FIGURE 4 Relationship between the mean concordance rate and minor allele frequency in Arab, Trotteur Français and Selle Français breed groups for “within-breed” (red) and “multi-breed” (black) strategies [Colour figure can be viewed at wileyonlinelibrary.com]

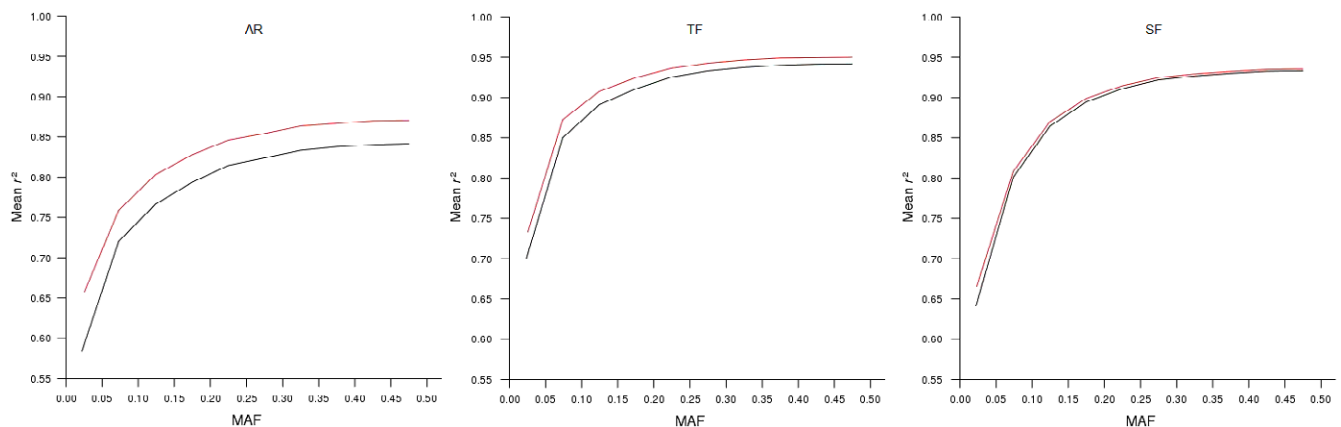


FIGURE 5 Relationship between the mean correlation (r^2) and minor allele frequency in Arab, Trotteur Français and Selle Français breed groups for “within-breed” (red) and “multi-breed” (black) strategies [Colour figure can be viewed at wileyonlinelibrary.com]

ECA2, ECA6, ECA12 and ECA20 for all breed groups and strategies.

Figure 7 summarizes the effects of MAF, linkage disequilibrium and SNP density on the mean imputation accuracy per chromosome measured by r^2 for “within-breed” imputation, for the three breed groups. Linkage disequilibrium (regression coefficients from 0.0065 to 0.0135 for 0.0100 of r^2_{LD}) had a positive effect, whereas SNP density (regression coefficients from -0.024 to -0.037 for 100 SNPs per Mb) had a negative effect in all breed groups. A positive effect of MAF was demonstrated for SF and TF (regression coefficients of 0.0076 to 0.0124 for 0.0100 point of MAF) but not for AR.

3.6 | Intra-chromosome imputation accuracy

Analyses of intra-chromosome imputation accuracy revealed for four chromosomes, ECA1, ECA2, ECA6 and ECA20, special patterns of regions with a very high SNP density at the ends of the chromosomes (after 145 Mb for ECA1, from 49 to 53 Mb, after 74 Mb for ECA2, after

32 Mb for ECA6 and after 28 Mb for ECA20). These regions contained on average 521 SNPs per Mb compared with 158 SNPs per Mb elsewhere. These HD regions showed low linkage disequilibrium with a mean r^2_{LD} of 0.089 compared with 0.105 elsewhere. However, the SNPs of these regions also had a low mean MAF (0.164 versus 0.211). For these chromosomes, the mean r^2 was lower than for other chromosomes in the SF and TF breed groups. At the intra-chromosomal level, r^2 was significantly lower in the regions of high SNP density (-0.62 (TF), -0.73 (SF) and -0.37 (AR): correlation between these two criteria measured by 1 Mb windows). The phenomenon is illustrated in Figure 8 for chromosome ECA20 in the SF breed group.

For the other chromosomes, we also observed a higher SNP density in regions of lower LD but with SNPs with high MAF in the three breed groups (correlation between MAF and SNP density higher than 0.40). The major factor explaining r^2 was MAF and then SNP density. In a multiple regression model in which r^2 is the independent variable and dependent MAF, SNP

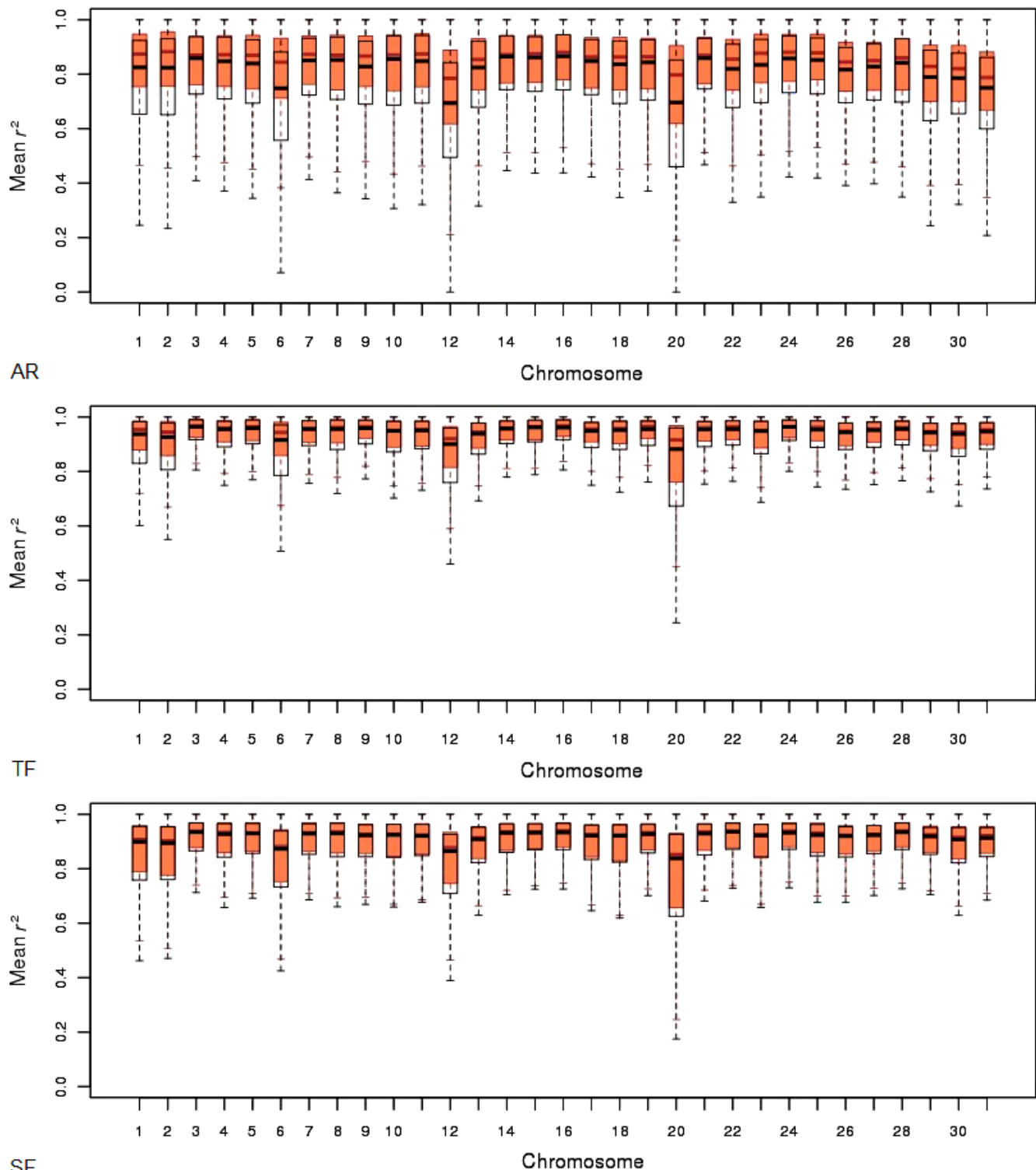


FIGURE 6 Mean correlation (r^2) per chromosome for “within-breed” (red) and “multi-breed” (black) strategies in Arabs, Trotteur Français and Selle Français breed groups [Colour figure can be viewed at wileyonlinelibrary.com]

density and r_{LD}^2 for each 1 Mb window for all these chromosomes, MAF was the first variable introduced for SF and TF with r -squared values of 5% and 6%, respectively. But the r -squared values (adding density and r_{LD}^2) remained low, 6% and 10%, respectively, so MAF,

linkage disequilibrium and SNP density did not explain much of the variation in SNP imputation accuracy. In the AR breed group, SNP density remained the major factor influencing r^2 . This is illustrated for chromosome ECA5 in Figure 9.

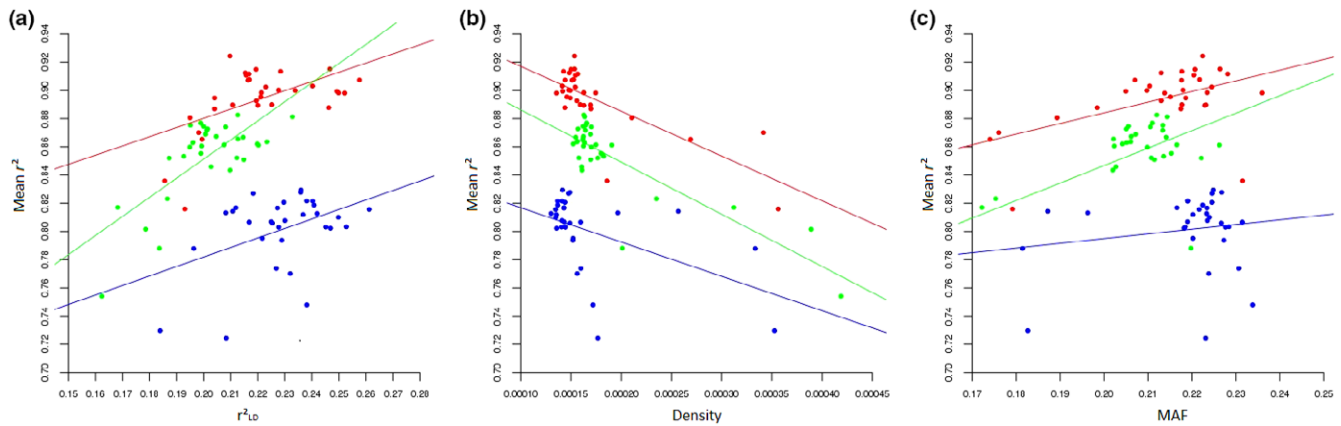


FIGURE 7 Plot of mean correlation (r^2) between imputed and true genotypes for the 31 chromosomes as function of mean linkage disequilibrium (r^2_{LD}) (a), SNP density (number of imputed SNPs/chromosome length) (b) and mean of minor allele frequency (c). Lines show the corresponding linear regression in Arab (blue), Trotteur Français (red) and Selle Français (green) breed groups for the “within-breed” strategy [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

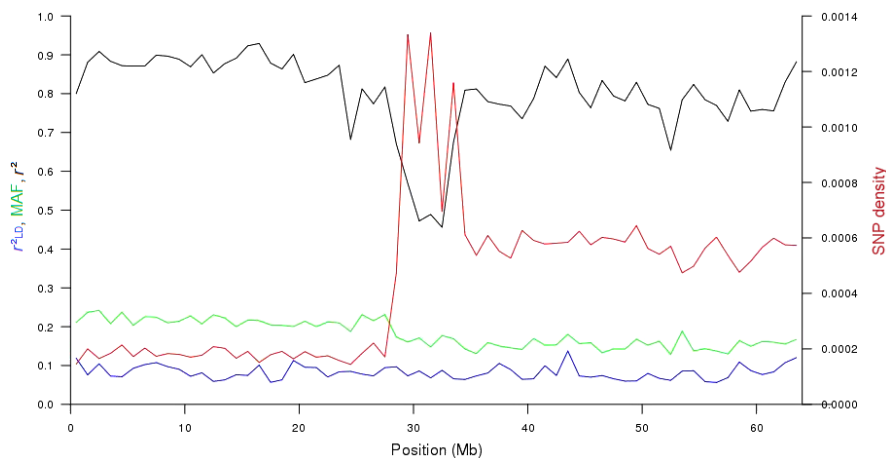


FIGURE 8 Correlation between true and imputed genotypes (r^2 , black), minor allele frequency (green), linkage disequilibrium (r^2_{LD} , blue) and SNP density (red) on ECA20 in the Selle Français breed group for the “within-breed” strategy. MAF, r^2_{LD} and r^2 were calculated in windows of 1 Mb along the chromosome [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

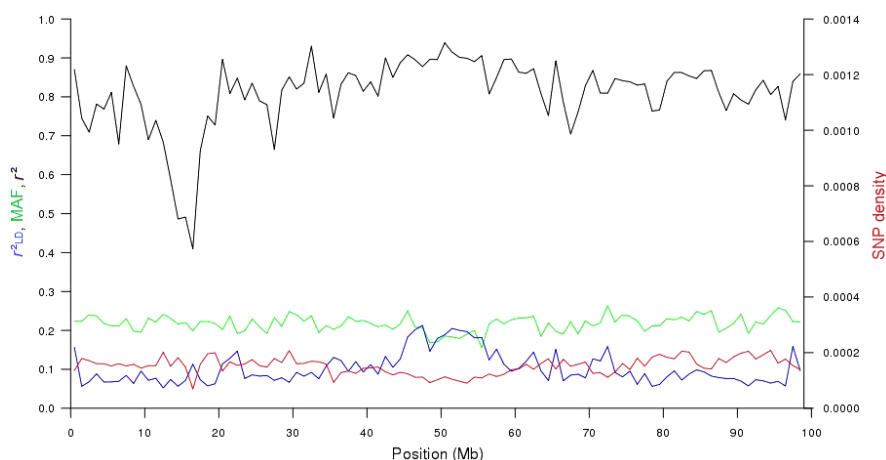


FIGURE 9 Correlation between true and imputed genotypes (r^2 , black), minor allele frequency (MAF, green), linkage disequilibrium (r^2_{LD} , blue) and SNP density (red) on ECA5 in the Arab breed group for the “within-breed” strategy. MAF, r^2_{LD} and r^2 were calculated in windows of 1 Mb along the chromosome [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

4 | DISCUSSION

The aim of this study was to assess the accuracy of imputation from MD to HD genotypes in five groups of horse breeds using two imputation strategies: “within-breed” and “multi-breed.”

Whatever the breed group and strategy used, the mean CR per breed group was greater than 97%. In cattle, an imputation accuracy >97% is interpreted as an accurate imputation (Hozé et al., 2013). The imputation accuracies we found are similar to those found in other studies on horses, with chips of varying density. McCoy and McCue

(2014) reported CR values per horse ranging from 82.2% to 100% for imputations from 54k and 65k chips to a MD (74k) chip for Quarter Horses, Standardbreds and Thoroughbreds. Another study found that imputation in Thoroughbreds was feasible: the proportion of genotypes correctly imputed per horse ranged from 79% to 98% from a very-low-density (1-6k) to a MD chip (65k) (Corbin et al., 2014). A study using whole genome sequencing (~13 million SNPs) from 44 Franches-Montagnes and Warmbloods imputed SNPs from MD (54k) genotypes to nearly 13 million SNPs with CR values per horse ranging from 85% to 99% (Frischknecht et al., 2014). A recent study designed two chips for genotype imputation: MNEc670k, consisting of ~670k SNPs, and MNEc2M, a next-generation HD SNP chip (~2 million SNPs). Genotype imputation accuracy from the MNEc670k SNP set to the MNEc2M SNP set ranged between 96.6% and 99.4% in the fifteen breeds tested (Schaefer et al., 2017).

In the breed groups considered in this study, the mean r^2 was investigated for two strategies. The imputation accuracies we reported, for “within-breed” and “multi-breed” strategies, are lower than those found by Pereira et al. (2017). They reported a mean r^2 per SNP of 0.98 for an imputation from a 54k chip to a 65k chip for racing Quarter Horses. However, McCoy & McCue (2014) reported an overall mean r^2 of 0.77 for imputations from a 54k chip to a 65k chip in three breeds: Quarter Horse, Standardbred and Thoroughbred.

We observed different imputation accuracies between the breed groups. For both strategies, the mean CR and the mean r^2 were lowest in AR, intermediate in SF and highest in TF breed groups. There are four possible reasons for these differences: the size of the reference population (Pausch et al., 2013), the breed homogeneity (Frischknecht et al., 2014), the importance of the relationship between validation and reference populations (Hayes et al., 2012) and the linkage disequilibrium (Corbin et al., 2014).

Previous studies have shown that imputation accuracy increases with larger reference population sizes (Hozé et al., 2013). In the present study, the numbers of horses with HD genotypes in the reference populations were 756 for SF, 195 for TF and 119 for AR breed groups, respectively. Consequently, mean CR was expected to be lower for the AR breed group, which has the smallest reference population. Yet, SF breed group, which had the largest reference population, did not show the highest mean CR. According to Hozé et al. (2013), the size of the reference population has a limited effect when the number of HD genotyped horses is greater than a minimum threshold which these authors estimated at 200–400 animals.

Arab and TF breed groups have closed studbooks, whereas the SF studbook is open to cross-breeding with FH and AA horses. Accordingly, the genetic diversity of

the SF breed group is higher than that of AR and TF breed groups, which could be seen from the cluster analysis. This may explain the lower imputation accuracy for SF compared to TF. Similarly, Frischknecht et al. (2014) showed that the variation in genotype imputation accuracies between horses was related to the level of admixture with introgressed Warmblood horses. They observed greater imputation accuracies in closed populations than in highly admixed populations.

Hayes et al. (2012) demonstrated that the accuracy of imputation can be increased if sires and other ancestors of the individuals to be imputed are included in the reference population. In our data, there was no sire–progeny relationship amongst the horses with HD genotypes. The closest relationship was half sibs. The sizes of the half-sib families differed between breed groups. The mean number of offspring per stallion was 2.31, 3.39 and 3.47 in the AR, SF and TF breed groups, respectively, which echoes the accuracy of imputation in each breed group.

In dairy cattle, Hozé et al. (2013) suggested that the level of linkage disequilibrium is not a major factor affecting imputation accuracy. Linkage disequilibrium levels were highest for AA, intermediate for FH, TF and AR and lowest for SF. This ranking is not in agreement with the levels of imputation accuracy between the five breed groups. According to these results, imputation accuracy was obviously affected by a combination of size of reference population, homogeneity of the breed, relationship between validation and reference populations and linkage disequilibrium.

The mean CR and the mean r^2 per horse were slightly lower with the “multi-breed” strategy than with “within-breed” strategy in all breed groups. The “multi-breed” strategy was designed to increase the size of the reference population, especially for the AR and TF breed groups. However, this increase did not compensate for the genetic distance between the breeds and consequently for the differences in allele frequencies which cause imputation errors. These results are in agreement with those of McCoy and McCue (2014), who demonstrated that a reference population that is breed-matched to the imputed population gives better imputation accuracies than a mixed reference population. This result favours the “within-breed” strategy.

Moreover, to check whether the validation populations selected to assess the accuracy of the imputation strategy were similar to the populations genotyped with the MD chip for which genotypes were to be imputed, the birth date and proportion of genotyped parents in the data sets were compared. These criteria were reviewed for each breed group, and we observed that the percentage of horses genotyped with the MD chip with parents genotyped with the MD chip was similar to the percentage of horses genotyped with the HD chip with parents genotyped with the

MD chip for SF, AR, AA and FH breed groups. Only, in the TF breed group, the percentage of horses genotyped with the MD chip with parents genotyped with the MD chip was higher than the percentage of horses genotyped with the HD chip with parents genotyped with the MD chip, because there were more stallions and mares in the MD data set. The validation populations used were therefore similar to the population genotyped with MD chip. Thus, the accuracy of the imputation of HD genotypes from MD genotypes could be assessed.

Variations in SNP imputation accuracy (CR and r^2) were observed. We investigated these variations by analysing the SNP map and SNP characteristics: MAF, linkage disequilibrium (r^2_{LD}) and SNP density.

r^2 is commonly used to estimate the imputation accuracy for alleles with low MAF to minimize the dependence on allele frequency (Sargolzaei et al., 2014). When MAF was low, r^2 was low; conversely, CR was high. The conflicting patterns are determined by the features of the two measures. CR includes good filling by chance, which is favourable for SNPs with low MAF, so CR tends to overestimate the imputation accuracy for low MAF SNPs. On the contrary, the correlation between the allele dosage (number of minor alleles) of the most likely imputed genotypes, and the allele dosage of the true genotypes (r^2) is greatly influenced by extreme values. A few imputation errors for a SNP with a low MAF can greatly reduce r^2 for this SNP (Ma, Brøndum, Zhang, Lund, & Su, 2013). The correlation r^2 is better for capturing the difference between imputation accuracy and correct filling by chance, which is important when estimating the imputation accuracy for markers with low MAF. Moreover, r^2 is directly linked to the efficiency of the regression used in GWAS and genomic selection. To summarize, r^2 was preferred to CR to investigate the causes of variation of the accuracy of imputation.

According to our analyses at the chromosome and intra-chromosomal levels, high imputation accuracy is usually related to low SNP density, high MAF and high linkage disequilibrium without necessarily clear biological explanation. The high SNP density regions on chromosomes ECA1, ECA2, ECA6 and ECA20 on the HD chip, with low linkage disequilibrium and rare alleles (lower MAF), seemed actually to cause the imputation difficulties experienced in our breed groups or at least did not improve imputation accuracy. Pereira et al. (2017) also observed lower imputation accuracy levels for chromosomes ECA6 and ECA12. Consistently with the findings of Corbin et al. (2014), we found that high linkage disequilibrium was generally linked to more accurate imputation, but was not the major factor involved for chromosomes other than ECA1, ECA2, ECA6 and ECA20 because MAF was of primary importance according to the r-squared values.

5 | CONCLUSION

In the present study, we showed that genotype imputation from MD to HD chips is feasible with a CR >97% and a correlation >0.93 in the five groups of horse breeds studied, and without loss in the quality of information. The investigation of the differences in imputation accuracies revealed complex interactions with the size of the reference population, the genetic diversity of the breed, the importance of the relationship between validation and reference populations, the SNP density of the chip and finally, to a lesser extent, the linkage disequilibrium and the MAF. Comparing the “within-breed” and the “multi-breed” strategies showed that increasing the size of the reference population by adding horses of different breeds, more or less unrelated, used in equine sports does not improve imputation accuracy. Hence, the “within-breed” strategy will be preferred for future imputations.

ACKNOWLEDGEMENT

This project was funded by the Institut Français du Cheval et de l'Équitation, the Institut National de la Recherche Agronomique (INRA) and the Fonds Eperon. We would like to thank the owners of genotyped horses and the Association Nationale du Selle Français, the Association Cheval Arabe, and LeTrot for supporting the project. We would also like to thank the GENTYANE platform at INRA Clermont-Ferrand for the quality of the genotyping service.

ORCID

Marjorie Chassier  <http://orcid.org/0000-0001-5386-1726>

REFERENCES

- Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: An R package for genome-wide association analysis. *Bioinformatics*, 23, 1294–1296. <https://doi.org/10.1093/bioinformatics/btm108>
- Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., ... Fritz, S. (2012). Genomic selection in French dairy cattle. *Animal Production Science*, 52, 115–120. <https://doi.org/10.1071/AN11119>
- Coleman, S., Zeng, Z., Wang, K., Luo, S., Khrebtukova, I., Mienaltowski, M. J., ... MacLeod, J. N. (2010). Structural annotation of equine protein-coding genes determined by mRNA sequencing. *Animal Genetics*, 41(Suppl. 2), 121–130. <https://doi.org/10.1111/j.1365-2052.2010.02118.x>
- Corbin, L. J., Kranis, A., Blott, S. C., Swinburne, J. E., Vaudin, M., Bishop, S. C., & Woolliams, J. A. (2014). The utility of low-density genotyping for imputation in the Thoroughbred horse. *Genetic Selection Evolution*, 46, 9. <https://doi.org/10.1186/1297-9686-46-9>
- Frischknecht, M., Neuditschko, M., Jagannathan, V., Drögemüller, C., Tetens, J., & Thaller, G. (2014). Imputation of sequence level

- genotypes in the Franches-Montagnes horse breed. *Genetic Selection Evolution*, 46, 63. <https://doi.org/10.1186/s12711-014-0063-7>
- Gualdrón Duarte, J. L., Bates, R. O., Ernst, C. W., Raney, N. E., Cantet, R. J. C., & Steibel, J. P. (2013). Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genetics*, 14, 38. <https://doi.org/10.1186/1471-2156-14-38>
- Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., & van der Werf, J. H. (2012). Accuracy of genotype imputation in sheep breeds. *Animal Genetics*, 43, 72–80. <https://doi.org/10.1111/j.1365-2052.2011.02208.x>
- Hozé, C., Fouilloux, M. N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., ... Croiseau, P. (2013). High-density marker imputation accuracy in sixteen French cattle breeds. *Genetic Selection Evolution*, 45, 33. <https://doi.org/10.1186/1297-9686-45-33>
- Leroy, G., Callède, L., Verrier, E., Mériaux, J. C., Ricard, A., Danchin-Burge, C., & Rognon, X. (2009). Genetic diversity of a large set of horse breeds raised in France assessed by microsatellite polymorphism. *Genetic Selection Evolution*, 41, 5. <https://doi.org/10.1186/1297-9686-41-5>
- Ma, P., Brøndum, R. F., Zhang, Q., Lund, M. S., & Su, G. (2013). Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *Journal of Dairy Science*, 96, 4666–4677. <https://doi.org/10.3168/jds.2012-6316>
- McCoy, A. M., & McCue, M. E. (2014). Validation of imputation between equine genotyping arrays. *Animal Genetics*, 45, 153. <https://doi.org/10.1111/age.12093>
- McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., ... Mickelson, J. R. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genetics*, 8, e1002451. <https://doi.org/10.1371/journal.pgen.1002451>
- Moghaddar, N., Gore, K. P., Daetwyler, H. D., Hayes, B. J., & van der Werf, J. H. (2015). Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. *Genetic Selection Evolution*, 47, 97. <https://doi.org/10.1186/s12711-015-0175-8>
- Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K. U., & Fries, R. (2013). Imputation of high-density genotypes in the Fleckvieh cattle production. *Genetic Selection Evolution*, 45, 3. <https://doi.org/10.1186/1297-9686-45-3>
- Pereira, G. L., Chud, T. C. S., Bernardes, P. A., Venturini, G. C., Chardulo, L. A. L., & Curi, R. A. (2017). Genotype imputation and accuracy evaluation in racing Quarter Horses genotyped using different commercial SNP panels. *Journal of Equine Veterinary Science*, 58, 89–96. <https://doi.org/10.1016/j.jevs.2017.07.012>
- Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, 15, 478. <https://doi.org/10.1186/1471-2164-15-478>
- Schaefer, R., Schubert, M., Bailey, E., Bannasch, D. L., Barrey, E., Bar-Gal, G. K., ... McCue, M. E. (2017). Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics*, 18, 656. <https://doi.org/10.1186/s12864-017-3943-8>
- Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., ... Lindblad-Toh, K. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326, 865–867. <https://doi.org/10.1126/science.1178158>

How to cite this article: Chassier M, Barrey E, Robert C, Duluard A, Danvy S, Ricard A. Genotype imputation accuracy in multiple equine breeds from medium- to high-density genotypes. *J Anim Breed Genet*. 2018;135:420–431. <https://doi.org/10.1111/jbg.12358>