Required marker properties for unbiased estimates of the genetic correlation between populations

Yvonne C.J. Wientjes*, Mario P.L. Calus*, Pascal Duenk*, Piter Bijma*

* Wageningen University & Research, Animal Breeding and Genomics, 6700 AH Wageningen, The Netherlands

Running title: Genetic correlation between populations

Key words: genetic correlation between populations, genomic relationships, marker-based relationships, multi-trait model

Author information:
Yvonne Wientjes
Wageningen University & Research
Animal Breeding and Genomics
P.O. box 338, 6700 AH Wageningen, the Netherlands
E-mail: yvonne.wientjes@wur.nl
Phone: +31 317 481 904

ABSTRACT

Populations generally differ in environmental and genetic factors, which can create differences in allele substitution effects between populations. Therefore, a single genotype may have different additive genetic values in different populations. The correlation between the two additive genetic values of a single genotype in both populations is known as the additive genetic correlation between populations and can differ from one. Our objective was to investigate whether differences in linkage disequilibrium (LD) and allele frequencies of markers and causal loci between populations affect bias of the estimated genetic correlation. We simulated two populations that were separated for 50 generations. Markers and causal loci were selected to either have similar or different allele frequencies in the two populations. Differences in consistency of LD between populations were obtained by using different marker density panels. Results showed that when the difference in allele frequencies of causal loci between populations was reflected by the markers, genetic correlations were only slightly underestimated using markers. This was even the case when LD patterns, measured by LD statistic *r*, were different between populations. When the difference in allele frequencies of causal loci between populations was not reflected by the markers, genetic correlations were severely underestimated. We conclude that for an unbiased estimate of the genetic correlation between populations, marker allele frequencies should reflect allele frequencies of causal loci so that marker-based relationships can accurately predict the relationships at causal loci, i.e. $E(\mathbf{G}_{\text{causal loci}}|\mathbf{G}_{\text{markers}}) \neq \mathbf{G}_{\text{markers}}$. Differences in LD between populations have little effect on the estimated genetic correlation.

INTRODUCTION

Alleles in different populations are often expressed in a different environment and a different genetic background. As a result of genotype by environment interaction and non-additive genetic effects, those differences result in different allele substitution effects between populations (Fisher 1918; Fisher 1930; Falconer 1952). In addition, the set of loci underlying a trait can differ between populations. Therefore, a single genotype may have different additive genetic values in different populations. For each population, the additive genetic value is the product of the genotype, measured as allele count at each locus, multiplied by the allele substitution effects for that population. The additive genetic correlation between two populations is the correlation between the two additive genetic values of a single genotype in both populations and may considerably differ from one.

Knowledge of the genetic correlation between populations helps to understand the differences and similarities between populations in genetic architecture of complex traits (De Candia *et al.* 2013; Brown *et al.* 2016). For both genomic prediction and genome-wide association studies, combining information from populations is an attractive approach to increase the prediction accuracy of estimated genetic values or the power to identify quantitative trait loci. This is especially the case when the number of individuals with genotypes and phenotypes in a population is limited. For both genomic prediction as well as genome-wide association studies, the genetic correlation between populations determines the added benefit of combining information from multiple populations (De Candia *et al.* 2013; Wientjes *et al.* 2015; Wientjes *et al.* 2016). Therefore, the genetic correlation between populations is an important parameter in human studies (e.g., De Candia *et al.* 2013; Yang *et al.* 2013), as well as in animal and plant breeding (e.g., Karoui *et al.* 2012; Lehermeier *et al.* 2015).

For estimating a genetic correlation between two populations, it is essential to know the relationships between individuals from the two populations. Traditionally, relationships between individuals are based on pedigree information, which is generally only available within population. The current availability of genome-wide marker panels has opened up new opportunities to estimate genetic correlations between populations of distantly related individuals, such as between breeds (e.g., Karoui *et al.* 2012; Carillier *et al.* 2014), lines (Huang *et al.* 2014), sub-populations (e.g., Lehermeier *et al.* 2015), or ethnicities (e.g., De Candia *et al.* 2013; Yang *et al.* 2013). Genetic correlations between populations can be estimated using methods based on genomic relationships (Karoui *et al.* 2012), random regression on genotypes (Sørensen *et al.* 2012; Krag *et al.* 2013), or summary statistics of genome-wide association studies (Bulik-Sullivan *et al.* 2015; Brown *et al.* 2016). Wientjes *et al.* (2017) showed that an unbiased estimate of the genetic correlation can be obtained from genomic relationships based on causal loci.

Because causal loci are generally unknown, genomic relationships have to be based on marker information. The strength and phase of linkage disequilibrium (LD) between causal loci and markers is different between populations in humans (Sawyer *et al.* 2005), livestock (e.g., Heifetz *et al.* 2005; Veroneze *et al.* 2013) and plants (Flint-Garcia *et al.* 2003; Lehermeier *et al.* 2014). Due to imperfect LD between causal loci and markers, not all genetic variance is explained by the markers which can distort the estimation of genetic correlations (Bulik-Sullivan *et al.* 2015; Gianola *et al.* 2015). However, in a simulation study where populations had different LD patterns, the genetic correlation between populations was accurately estimated based on marker information (Wientjes *et al.* 2015).

The objective of this study was to investigate whether differences in LD and allele frequencies of markers and causal loci between populations affect bias of the estimated genetic correlation. We simulated two populations that were separated for 50 generations

using scenarios differing in consistency of LD and in allele frequencies of markers and causal loci between the populations. We used different marker-based relationship matrices to estimate the genetic correlation.

MATERIALS AND METHODS

Population structure

Two populations were simulated using QMSim software (Sargolzaei and Schenkel 2009). The simulations were set-up to have the following two characteristics; 1) the two populations should have different LD patterns, as measured by the LD statistic *r*, and 2) a large number of loci should segregate in the last generation of which a part (>200 000) has similar allele frequencies in both populations and another part (>200 000) different allele frequencies in both populations. We simulated a historical population for 212 generations. The first generation (generation -211) contained 300 individuals. In the following 100 generations (generation -211 – -112), population size gradually decreased to 50 individuals to create LD. From generation -111 to generation -12, population size gradually increased to 300 individuals and was kept constant for the next 10 generations (generation -11 – -2). In the last generation of the historical population (generation -1), population size increased to 1800 individuals.

The last generation of the historical population was randomly divided into two equally sized populations (A and B) of 900 individuals. In the next generation, the size of both populations was increased to 1800 individuals and was kept constant for the following 40 generations (generation 1-40). Those reasonably large population sizes limited the drift of allele frequencies. Number of offspring was set to 10 and selection was at random, so the number of selected offspring per individual approximately followed a Poisson distribution, as assumed in the Wright-Fisher model of genetic drift. In the last 10 generations (generation 41-50), population size decreased to 120 individuals in each population to increase the extent of LD in each population, and the number of offspring was set to 20. In the entire simulation, the male to female ratio was 1:5, generations were not overlapping and mating was at random. All individuals from the last generation (2000) were used for the analyses.

122

123 **Genome size**

124 A genome of 10 chromosomes of one Morgan each was simulated. This genome size was a
125 balance between the computational effort of the analyses and the variation in relationships
126 between family members. By using fewer chromosomes, computational effort reduced, but
127 variation in relationships around their expectation based on the pedigree would have been
128 inflated (Hill 1993). Each chromosome contained 300 000 randomly spaced loci, with a
129 recurrent mutation rate of 0.00005 in the historical population. In the last generation of the
130 historical population, segregating loci were selected and mutation was stopped. The chosen
131 population size and mutation rate resulted in a U-shaped allele frequency distribution of loci
132 in the two populations, as commonly found in real populations.

133 In the last generation (generation 50), markers and 2000 causal loci were selected from all
134 segregating loci. Three marker panels were constructed: a High Density Panel (HDP) with
135 200 000 markers, a Low Density Panel (LDP) with 20 000 markers, and a Very Low Density
136 Panel (VLDP) with 2000 markers. Each of the smaller marker panels was a subset from the
137 larger marker panels. The different marker densities were used to represent differences in
138 consistency of LD between populations, since consistency in LD decreases when genomic
139 distance between markers and causal loci increases (De Roos et al. 2008).

140 Markers and causal loci were selected to either have similar or different allele frequencies
141 in population A and B. For both approaches, three selection criteria were used; namely (1) the
142 segregation in one or both populations, (2) the absolute difference in allele frequency between
143 population A (p_A) and population B (p_B), and (3) the difference in variance explained by a
144 locus between population A and B, when allele substitution effects would be the same in both
145 populations. The last criterion was mainly effective for loci with a low allele frequency, since

146 an apparently small difference in allele frequency can result in a relatively large difference in
147 variance explained for those loci.

148 For selecting markers with similar allele frequencies in the two populations, loci had to (1)
149 segregate in both populations, (2) $|p_A - p_B|$ should be less than 0.14, and (3)
150 $|2p_A(1-p_A) - 2p_B(1-p_B)|/[2\bar{p}_{AB}(1-\bar{p}_{AB})]$ should be less than 2, where $\bar{p}_{AB}$ was the average
151 of p_A and p_B. For selecting markers with different allele frequencies in the two populations,
152 (1) loci had to segregate in at least one population, (2) $|p_A - p_B|$ should be more than 0.14, and
153 (3) $|2p_A(1-p_A) - 2p_B(1-p_B)|/[2\bar{p}_{AB}(1-\bar{p}_{AB})]$ should be more than 1. The cut-off values
154 were chosen to either minimize or maximize the difference in allele frequencies between the
155 populations, while ensuring that enough loci in each replicate met the criteria. We aimed to
156 select marker panels with a uniform allele frequency distribution to reflect commercially
157 available marker chips (Matsuzaki *et al.* 2004; Matukumalli *et al.* 2009; Ramos *et al.* 2009;
158 Groenen *et al.* 2011). For this step, the loci that met the criteria were divided in 50 bins based
159 on average allele frequency over the two populations (i.e., allele frequencies of bin 1 ranged
160 from 0 – 0.02, of bin 2 from 0.02 – 0.04, etc.) and from each bin an equal number of loci was
161 randomly selected. When the number of loci was too small in the two extreme bins (0.00 –
162 0.02, and 0.98 – 1.00), the bins were combined with the neighboring bin.

163 For selecting causal loci, the same criteria and cut-off values were used as for markers,
164 with one exception. For the scenario where allele frequencies in the two populations were
165 similar, causal loci did not have to segregate in both populations, since some causal loci are
166 known to be at least partly population-specific (Kemper *et al.* 2015). As an additional
167 criterion, causal loci could not already be selected as marker. Causal loci were randomly
168 selected from all loci that met the criteria, and therefore their allele frequency pattern

169 followed an approximate U-shaped distribution as expected for causal loci (Yang *et al.* 2010;
170 Kemper and Goddard 2012).
171
172 **LD patterns and consistency of LD**
173 The LD pattern and consistency in LD between the populations was investigated. Within
174 each population and between all causal loci and markers less than 10 cM apart, the parameter
175 r was calculated (Hill and Robertson 1968):
176 $$r = \frac{\left(f_{11} f_{22} - f_{12} f_{21}\right)}{\sqrt{f_{1.} f_{2.} f_{.1} f_{.2}}},$$
177 where f_{11} is the haplotype frequency with allele 1 at the first locus and allele 1 at the second
178 locus, f_{22}, f_{12} and f_{21} are frequencies of the other possible haplotypes, $f_{1.}$ and $f_{2.}$ are the
179 frequencies of allele 1 and allele 2 at the first locus, and $f_{.1}$ and $f_{.2}$ are the frequencies of allele
180 1 and allele 2 at the second locus. The LD pattern within each population was represented by
181 the average r^2 for intervals of 0.1 cM distance between the markers. The consistency of LD
182 between the two populations was calculated as the correlation between r values of the two
183 populations for intervals of 0.1 cM, following De Roos *et al.* (2008).
184
185 **Phenotypes**
186 For each causal locus, allele substitution effects were sampled from a bivariate normal
187 distribution, with mean 0, standard deviation 1, and a correlation between the populations of
188 either 1, 0.8, 0.6, 0.4, 0.2 or 0. For each individual, its allele counts for the causal loci (coded
189 as 0, 1, and 2) were multiplied by the corresponding allele substitution effects and results
190 were summed over loci to calculate the additive genetic value (AGV) of the individual. The
191 AGV were scaled to a mean of 0 and variance of 1 across all individuals. Since allele
192 substitution effects were sampled independently from allele frequency, the correlation

193 between AGV of population 1 and 2 (i.e., genetic correlation) was similar to the correlation
194 between allele substitution effects (i.e., either 1, 0.8, 0.6, 0.4, 0.2 or 0). A normally-distributed
195 environmental effect was sampled for each individual to obtain a heritability of 0.3 in each
196 population. Phenotypes of all 2000 individuals in generation 50 were computed by summing
197 the AGV and the environmental effects.
198
199 **Estimating the genetic correlation**
200 The additive genetic correlation between populations was estimated using the following
201 bivariate model:

202 $$\begin{bmatrix}\mathbf{y}_A\\ \mathbf{y}_B\end{bmatrix}=\begin{bmatrix}\mathbf{x}_A & 0\\ 0 & \mathbf{x}_B\end{bmatrix}\begin{bmatrix}\mu_A\\ \mu_B\end{bmatrix}+\begin{bmatrix}\mathbf{Z}_A & 0\\ 0 & \mathbf{Z}_B\end{bmatrix}\begin{bmatrix}\mathbf{a}_A\\ \mathbf{a}_B\end{bmatrix}+\begin{bmatrix}\mathbf{e}_A\\ \mathbf{e}_B\end{bmatrix},$$

203 where $\mathbf{y}_k$ is a vector with phenotypes for population k (k= *A, B)*, $\mathbf{x}_k$ is an incidence vector
204 relating phenotypes to the mean in population k (μ_k), $\mathbf{Z}_k$ is an incidence matrix relating
205 phenotypes to estimated additive genetic values ($\mathbf{a}_k \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{G}_{AA} & \mathbf{G}_{AB}\\ \mathbf{G}_{BA} & \mathbf{G}_{BB}\end{bmatrix}\otimes\begin{bmatrix}\sigma_A^2 & \sigma_{AB}\\ \sigma_{AB} & \sigma_B^2\end{bmatrix}\right)$)
206 with $\otimes$ representing the Kronecker product function, and $\mathbf{e}_k$ are vectors with independent
207 residual effects. Genetic and residual variances were estimated using REML. The first
208 analyses were performed using ASReml software (Gilmour *et al.* 2015). For the scenarios
209 analyzed later, we switched to MTG2 (Lee and van der Werf 2016) to reduce computation
210 time. We verified that the estimated variance components were identical using both programs.
211 The genomic relationship matrix ($\mathbf{G}$) between all individuals was calculated as (Wientjes *et*
212 *al.* 2017):

213 $$\mathbf{G}=\begin{bmatrix}\mathbf{G}_{AA} & \mathbf{G}_{AB}\\ \mathbf{G}_{BA} & \mathbf{G}_{BB}\end{bmatrix}=\begin{bmatrix}\dfrac{\mathbf{W}_A\mathbf{W}_A'}{\sum 2p_{Ai}(1-p_{Ai})} & \dfrac{\mathbf{W}_A\mathbf{W}_B'}{\sqrt{\sum 2p_{Ai}(1-p_{Ai})}\sqrt{\sum 2p_{Bi}(1-p_{Bi})}}\\ \dfrac{\mathbf{W}_B\mathbf{W}_A'}{\sqrt{\sum 2p_{Ai}(1-p_{Ai})}\sqrt{\sum 2p_{Bi}(1-p_{Bi})}} & \dfrac{\mathbf{W}_B\mathbf{W}_B'}{\sum 2p_{Bi}(1-p_{Bi})}\end{bmatrix}$$

214

where $\mathbf{W}_k$ is a matrix with centered allele counts of all individuals from population *k*, and p_{ki} is the allele frequency for locus *i* in population *k*. Centered allele counts were calculated as $g_{ijk} - 2p_{ik}$, where g_{ijk} is the allele count of locus *i* for individual *j* from population *k*, coded as 0, 1 or 2. This **G** defines the relationships as standardized covariances between the genetic values of individuals (Wientjes *et al.* 2017). In all scenarios and in all 50 replicates, we calculated **G** using allele counts of 1) causal loci, 2) HDP markers, 3) LDP markers, or 4) VLDP markers.

The relationships at causal loci are the true relationships for that trait, that are approximated when using markers. Marker-based relationships are subject to sampling error, since markers are a subset of the genome. A way to account for this sampling error is by regressing **G** towards the pedigree relationship matrix (**A**) (Powell *et al.* 2010; Yang *et al.* 2010; Goddard *et al.* 2011), which is expected to reduce bias of estimated variance components (Yang *et al.* 2010). To investigate the effect of this regression, **G** matrices based on the three marker panels were regressed towards **A** and used for the scenarios with a correlation of 0.8 or 0.4.

Before regressing **G** towards **A**, the inbreeding level of each within-population block in **G** was rescaled to the inbreeding level in **A**, following (Powell *et al.* 2010):

$$\mathbf{G}^* = \left(1 - \overline{F_k}\right)\mathbf{G} + 2\overline{F_k}\,\mathbf{J},$$

where $\overline{F_k}$ is the average inbreeding coefficient of all individuals of population *k* based on the pedigree, and **J** is a matrix of ones. The rescaled $\mathbf{G}^*$ was regressed towards **A** following (Yang *et al.* 2010; Goddard *et al.* 2011):

$$\hat{\mathbf{G}} = \mathbf{A} + b\left(\mathbf{G}^* - \mathbf{A}\right),$$

with

$$b = \frac{Var\left(\mathbf{G}^* - \mathbf{A}\right)}{Var\left(\mathbf{G}^* - \mathbf{A}\right) + \frac{1}{n}},$$

where *n* is the number of markers. To set-up **A**, the pedigree of the last 10 generations was used, so that between-population **A** relationships were zero. The regression was done separately within each population per bin of pedigree relationships (<0.10, 0.10-0.25, 0.25-0.50, >0.5) and between populations, since regression coefficients are higher for higher pedigree relationships (Veerkamp *et al.* 2011; Wientjes *et al.* 2013). For the diagonal elements, only the inbreeding coefficients were regressed (Yang *et al.* 2010). Regression coefficients were all close to one for higher marker density panels (>0.99 for HDP and >0.97 for LDP). For VLDP markers, regression coefficients were lower; ~0.84 for between-population relationships, ~0.89, ~0.91, ~0.94 and ~0.96 for the four bins of within-population relationships, and ~0.93 for inbreeding coefficients.

Data availability

Supplemental Material, File S1, is available at FigShare. This file contains the input file used for QMSim, the Fortran-programs to select markers and causal loci for the different scenarios, the Fortran-program to simulate phenotypes and the seeds for the different programs in each of the replicates.

RESULTS

Characteristics of simulations

The criteria for selecting markers and causal loci resulted in clear differences between the scenarios with similar and different allele frequencies in the two populations (Figure 1). As intended, the allele frequency distribution was uniform for markers and U-shaped for causal loci (not shown). Therefore, the percentage of causal loci with a minor allele frequency below 0.05 was higher (on average 33% in each population) than the percentage of markers with a minor allele frequency below 0.05 (on average only 15% in each population). The decay of LD was similar in both populations (Figure 2), with a strong decay of LD at increasing distances between the loci at the 0 – 2 cM interval. The consistency of LD phase decreased rapidly at short distances (0 – 5 cM), and fluctuated around zero at distances larger than 5 cM.

Proportion of variance explained

The proportion of the phenotypic variance explained by the markers, known as the genomic heritability (De los Campos *et al.* 2015), was close to the simulated heritability for all scenarios (not shown). This implies that genetic variances were accurately estimated using all three marker panels.

Estimated genetic correlation

With relationships based on causal loci, all estimated genetic correlations were unbiased, irrespective of whether causal loci had similar or different allele frequencies in the two populations (Figure 3). This was also expected based on previous results (Wientjes *et al.* 2017).

With relationships based on markers, all estimated genetic correlations were biased. When marker-based relationships were not regressed towards the pedigree relationships, genetic

279 correlations were only slightly underestimated when the difference in allele frequencies of
280 causal loci between populations was reflected by the markers, i.e., when markers and causal
281 loci both had similar or different allele frequencies in the two populations (Figure 3A and 3C;
282 ~2.5% for HDP, ~3% for LDP, and ~11% for VLDP). The genetic correlation was much more
283 severely underestimated when the difference in allele frequencies of causal loci between
284 populations was not reflected by the markers (Figure 3B; ~28% for HDP, ~30% for LDP, and
285 ~41% for VLDP).

286 Across all scenarios, regressing **G** towards the pedigree relationship matrix only had a
287 small effect on the estimated genetic correlation (Figure 4). At a high marker density,
288 regressing **G** lowered the estimated genetic correlation. Therefore, the underestimation for
289 HDP and LDP markers increased from ~4% to ~9% when the difference in allele frequencies
290 of causal loci between populations was reflected by the markers, and from ~28% to ~32%
291 when the difference in allele frequencies of causal loci between populations was not reflected
292 by the markers. In contrast, regressing **G** resulted in higher estimated genetic correlations at
293 low marker density. For VLDP markers, the underestimation decreased from ~12% to ~8%
294 when the difference in allele frequencies of causal loci between populations was reflected by
295 the markers, and from ~41% to ~38% when the difference in allele frequencies of causal loci
296 between populations was not reflected by the markers. Thus, regressing **G** was only beneficial
297 for estimating the genetic correlation between populations when the marker density was low.

298 Standard errors across replicates for the estimated genetic correlation were generally small
299 for all scenarios (~0.02), and tended to be slightly larger for lower true genetic correlations.
300 Moreover, standard errors were slightly larger when the difference in allele frequencies of
301 causal loci between populations was not reflected by the markers (Figure 3B versus Figure 3A
302 and 3C). Regression of **G** towards the pedigree relationship matrix had no effect on the
303 standard error.

304

305 **Genomic relationships**

306 Genetic variance estimates are biased when the regression of true relationships on marker-
307 based relationships is not equal to one (Goddard *et al.* 2011). We investigated whether this
308 could explain the underestimation of the genetic correlation by considering the genomic
309 relationships at the causal loci as the true relationships for that trait. In Figure 5 and 6, we
310 plotted the relationships at the causal loci versus the unregressed relationships at the markers
311 for one of the replicates. The regression coefficients for within-population genomic
312 relationships were close to one, and were only slightly lower when causal loci had different
313 allele frequencies (Figure 6) compared to similar allele frequencies (Figure 5) in the two
314 populations. This means that the within-population relationships at the markers can quite
315 accurately predict the relationships at the causal loci.

316 Regression coefficients of between-population relationships deviated more from one,
317 especially at low marker density. When the difference in allele frequencies of causal loci
318 between populations was reflected by the markers, the regression coefficients were ~0.8 for
319 HDP and LDP, and 0.67 for VLDP (Figure 5). This means that the relationships at the
320 markers overpredict the relationships at the causal loci. When the difference in allele
321 frequencies of causal loci between populations was not reflected by the markers, regression
322 coefficients of between-population relationships were ~0.30 (Figure 6). Thus the
323 overprediction of between-population relationships using markers was much larger when the
324 difference in allele frequency of the causal loci between the populations was not reflected by
325 the markers.

326 The correlation between the relationships at the causal loci and at the markers, i.e., the
327 accuracy of the marker-based relationships, decreased when the density of the markers
328 decreased (Figure 5 and 6). When the difference in allele frequencies of causal loci between

329 populations was reflected by the markers, the correlation for within-population relationships
330 was ~0.91 for HDP and LDP, and ~0.88 for VLDP. The correlation for between-population
331 relationships was ~0.70 for HDP and LDP, and 0.60 for VLDP. The correlation between
332 relationships at causal loci and at markers was much lower when the difference in allele
333 frequencies of causal loci between populations was not reflected by the markers (within-
334 population relationships: ~0.66 for HDP and LDP, ~0.63 for VLDP; between-population
335 relationships: ~0.09 for HDP and LDP, ~0.08 for VLDP).

DISCUSSION

The objective of this study was to investigate whether differences in LD and allele frequencies of markers and causal loci between populations affect bias of the estimated genetic correlation between populations. Results showed that when a difference in allele frequencies of causal loci between populations was reflected by the markers, estimated genetic correlations were only slightly underestimated using markers. This was even the case when LD patterns, as measured by LD-statistic *r*, were different between populations. When the difference in allele frequencies of causal loci between populations was not reflected by the markers, genetic correlations were severely underestimated. Differences in LD and allele frequencies of causal loci between populations only had a very slight effect on the precision of the estimated genetic correlation.

Estimating the genetic correlation using marker-based relationships

Genetic variance and heritability estimates are known to be biased when the regression coefficient of the true relationships on the marker-based relationships is not equal to one, i.e., when $E(\mathbf{G}_{\text{causal loci}}|\mathbf{G}_{\text{markers}}) \neq \mathbf{G}_{\text{markers}}$ (Yang *et al.* 2010; Goddard *et al.* 2011; Yang *et al.* 2015). When this regression coefficient is below one, relationships at the markers show too much variation, resulting in an underestimation of the genetic variance. Yang *et al.* (2010) argued that a regression coefficient smaller than one can be a result of two effects; 1) sampling error on the relationships because the number of markers is finite, and 2) a difference in allele frequency distribution between causal loci and markers. In all our scenarios, the number of markers was finite and the allele frequency distribution was different for causal loci than for markers. However, within populations, the estimated genomic heritability (De los Campos *et al.* 2015) was close to the simulated trait heritability for all scenarios. This suggests that enough markers were used to constrain the sampling error on

361 within-population relationships to an acceptable level, and that our estimated genetic
362 variances were only slightly affected by the difference in allele frequency distribution
363 between causal loci and markers. Thus the underestimation of the genetic correlation between
364 populations is not a result of biased genetic variance estimates.

365 The relative sampling error as a result of using a finite number of markers was much larger
366 for between-population relationships than for within-population relationships, because more
367 markers are needed to accurately estimate the small between-population relationships
368 (Goddard *et al.* 2011). Moreover, the accuracy of predicting the between-population
369 relationships at the causal loci using markers was depending on the reflection of the
370 difference in allele frequency of causal loci between populations by the markers. Those two
371 effects can result in an underestimated genetic covariance between populations, which can
372 explain the slight underestimation of the genetic correlation in the scenarios where the
373 difference in allele frequencies of causal loci between the populations was reflected by the
374 markers, and the more severe underestimation in the scenarios where this was not the case.
375 The higher sampling error on between-population relationships can also explain the larger
376 underestimation of the genetic correlation for VLDP markers than for HDP and LDP markers.
377 Thus for estimating the genetic correlation between populations, it is important that the
378 difference in allele frequencies of causal loci between the populations is reflected by the
379 markers and that the number of markers is high.
380
381 **Regression of the maker-based relationships**
382 Regressing **G** towards the pedigree relationship matrix is a way to correct the marker-
383 based relationships for the sampling error as a result of using a finite number of markers
384 (Powell *et al.* 2010). The regression was strongest for VLDP markers, where it reduced the
385 underestimation of the genetic correlation. Those results agree with the findings that

386 regressing **G** is important when the number of markers is low (Yang *et al.* 2010) and supports
387 our statement that relationships at VLDP markers were affected by sampling error. However,
388 regressing **G** increased the underestimation of the genetic correlation with HDP and LDP
389 markers. The reason for this is not clear. It might be that the regression of **G** not only reduces
390 the sampling error, but also amplifies the effect of the difference in allele frequency
391 distribution of causal loci and markers.

392 In our study, regressing **G** towards **A** was detrimental for estimating the genetic correlation
393 when using HDP (200 000) or LDP (20 000) markers, where all regression coefficients were
394 close to one, and regressing was beneficial when using VLDP (2000) markers, where
395 regression coefficients were considerably below one. The simulated genome was about one
396 third of the genome of livestock species such as cattle and chicken (Ihara *et al.* 2004; Groenen
397 *et al.* 2009). This would indicate that regressing **G** is detrimental when using a genome-wide
398 total of 60 000 or more markers in livestock. Note that this number of markers will depend on
399 the consistency in LD between populations. Between-population relationships are all closer to
400 zero when consistency in LD between populations is lower (Goddard 2009). Those lower
401 relationships generally require more markers to reduce their relative sampling error to an
402 acceptable level (Yang *et al.* 2010). Hence, we think that the regression coefficients may be a
403 better indicator for deciding whether or not to regress **G**; when all regression coefficients are
404 close to one, e.g., above 0.95, it is probably better to not regress **G** towards **A** when estimating
405 the genetic correlation between populations.

406 The coefficients to regress **G** towards **A** were approximated using the number of markers
407 and the variation in $\mathbf{G}_{markers}$-**A**, assuming that the sampling error was only a result of using a
408 limited number of markers (Goddard *et al.* 2011). To investigate the impact of this
409 approximation and whether we could remove the observed underestimation of the genetic
410 correlation by rescaling $\mathbf{G}_{markers}$ such that $E(\mathbf{G}_{causal\ loci}|\mathbf{G}_{markers}) = \mathbf{G}_{markers}$, we repeated some

411 analysis using $b = \frac{Cov\left(\mathbf{G}_{\text{causal loci}} - \mathbf{A}, \mathbf{G}_{\text{markers}} - \mathbf{A}\right)}{Var\left(\mathbf{G}_{\text{markers}} - \mathbf{A}\right)}$ (Goddard *et al.* 2011) as regression

412 coefficient to regress **G** towards **A**. This regression requires the causal loci to be known,
413 which was the case in our simulations. We calculated *b* separately for within- and between-
414 population relationships, using 11 bins based on pedigree relationships within populations
415 (<0.05, 0.05-0.10, 0.10-0.15, 0.15-0.20, 0.20-0.25, 0.25-0.30, 0.30-0.35, 0.35-0.40, 0.40-0.50,
416 >0.50, self-relationships) and 3 bins based on genomic relationships between populations (<-
417 0.10, -0.10-0.10, >0.10), and used those *b*'s to rescale the relationships. As shown in Figure 7,
418 this rescaling almost completely removed the bias in genetic correlation estimates using HDP
419 and LDP markers. The genetic correlation was overestimated when using rescaled
420 relationships based on VLDP markers. This might be a result of the much larger sampling
421 error for VLDP markers compared to HDP and LDP markers, which could result in
422 underestimated *b* values. Thus, there appears to be a lower boundary for the number of
423 markers to calculate between-population genomic relationships that can be corrected using
424 regression. Altogether, those results confirm that for an unbiased estimate of the genetic
425 correlation between populations, the regression coefficient of true relationships on marker-
426 based relationships should be one.

427

428 **Consistency in LD**

429 We used different marker densities to represent differences in consistency in LD between
430 populations. We expected that a lower consistency in LD would reduce the estimated genetic
431 correlation between populations, because it reduces the correlation between (apparent) marker
432 effects. Surprisingly, our results showed that estimated genetic correlations were similar with
433 HDP and LDP markers, and only slightly lower with VLDP markers. This can be explained
434 by the potential of marker-based relationships to accurately predict the relationships at the

435 causal loci, which is essential to unbiasedly estimate the genetic (co)variances and the genetic
436 correlation between populations. A lower consistency in LD between populations results in a
437 lower variation in between-population relationships (Goddard 2009; Goddard *et al.* 2011).
438 Because a lower consistency in LD reduces the variation in between-population relationships
439 at both causal loci and markers, the regression coefficient of the relationships at the causal
440 loci on the relationships at the markers may not be affected much (Figure 5 and 6; HDP and
441 LDP markers). Therefore, the estimated genetic correlation between populations seems little
442 affected by the consistency in LD between the populations.

443 The consistency in LD between populations does affect the correlation between the
444 relationships at the causal loci and the marker-based relationships (Figure 5 and 6), i.e., the
445 accuracy of the marker-based relationships. For an unbiased estimate of the genetic
446 correlation between populations, the regression of true relationships on marker-relationships
447 should be one and marker-based relationships don't necessarily have to be accurate. This is in
448 contrast to estimating genetic values, as is done in genomic prediction, for which relationships
449 have to be accurate and have to show variation (Goddard *et al.* 2011). Thus, an unbiased
450 estimate of the genetic correlation between populations does not guarantee that accurate
451 genomic prediction across populations can be performed.

452

453 **LD structure**

454 The extent and consistency of LD in the simulated populations is comparable to the
455 patterns found in chicken and pig populations (Andreescu *et al.* 2007; Badke *et al.* 2012;
456 Veroneze *et al.* 2013; Veroneze *et al.* 2014). This simulated LD was much higher than
457 generally found in human populations (Pritchard and Przeworski 2001; Shifman *et al.* 2003).
458 Since marker density, and thereby the average LD between causal loci and nearest marker,

459 had no effect on the estimated genetic correlation, it is expected that the simulated LD pattern
460 did not affect the results.

461 We simulated causal loci randomly spread across the genome, which is not always the case
462 in real populations. When causal loci are enriched in regions with either high or low LD,
463 (co)variance estimates can be over- or underestimated (Speed *et al.* 2012; Yang *et al.* 2015).
464 However, we would expect a smaller impact of the heterogeneity of LD on the estimated
465 genetic correlation than on the heritability, since differences in LD across the genome affect
466 both the genetic variance and covariance estimates. This mechanism may also explain why
467 genetic correlation estimates between traits within a population are less affected by
468 incomplete LD between causal loci and markers than genetic variance estimates (Trzaskowski
469 *et al.* 2013).

470

471 **Genomic relationship matrix**

472 The current generation within each population was used as base population for our
473 genomic relationships, since we used current population-specific allele frequencies. This
474 means that between-population relationships are on average zero. When the consistency in LD
475 between the populations is not zero, due to the existence of a recent or distant common
476 ancestor, between-population relationships will show variation around zero (Goddard 2009).
477 That variation is essential in order to estimate the genetic correlation between populations,
478 and genetic correlation estimates are more precise when the variation in between-population
479 relationships is higher (Visscher *et al.* 2014).

480 Another commonly used multi-population **G** matrix is the matrix following Chen *et al.*
481 (2013). We repeated part of our analyses using that matrix, where the scaling factor of the
482 block between populations is $\sum 2\sqrt{p_{Ai}(1-p_{Ai})p_{Bi}(1-p_{Bi})}$ ($\mathbf{G}_{\text{Chen}}$) instead of
483 $\sqrt{\sum 2p_{Ai}(1-p_{Ai})}\sqrt{\sum 2p_{Bi}(1-p_{Bi})}$ ($\mathbf{G}_{\text{Wientjes}}$). In agreement with our previous study based

484 on causal loci (Wientjes *et al.* 2017), we found that genetic correlations were underestimated
485 using $\mathbf{G}_{\text{Chen}}$. This underestimation is mainly a result of effectively removing markers
486 segregating in only one population from the scaling factor of between-population
487 relationships. This underestimation increases when those markers were also removed from
488 within-population relationships, because it increased the bias in genetic variance estimates.
489 Moreover, $\mathbf{G}_{\text{Chen}}$ was more prone to singularities than $\mathbf{G}_{\text{Wientjes}}$. In $\mathbf{G}_{\text{Wientjes}}$, markers
490 segregating in only one population contributed to the scaling factor for between-population
491 relationships, which resulted in lower between-population relationships when the number of
492 markers segregating in only one population was higher. This resulted in a larger difference
493 between within- and between-population relationships in $\mathbf{G}_{\text{Wientjes}}$, which reduced the risk of
494 singularities.
495
496 **Implications**
497 Marker panels are generally composed to have intermediate allele frequencies across
498 multiple populations (Matsuzaki *et al.* 2004; Matukumalli *et al.* 2009; Groenen *et al.* 2011).
499 Therefore, markers tend to have a higher average minor allele frequency than causal loci
500 (Yang *et al.* 2010; Kemper and Goddard 2012). Moreover, the difference in allele frequencies
501 of causal loci between populations is probably not accurately represented by markers. Those
502 factors likely result in underestimated genetic correlations between populations using real
503 data, but the impact of each of the factors requires further research.
504
505 **Conclusion**
506 For an unbiased estimate of the genetic correlation between populations from marker
507 information, it is important that marker-based relationships accurately predict the
508 relationships at causal loci, i.e., $E(\mathbf{G}_{\text{causal loci}}|\mathbf{G}_{\text{markers}}) = \mathbf{G}_{\text{markers}}$. To achieve this, the difference

509 in allele frequencies of causal loci between the populations should be reflected by the
510 markers, and the number of markers should be sufficiently high to constrain the sampling
511 error on between-population relationships to an acceptable level. The consistency in LD
512 between populations has little effect on the bias of the estimated genetic correlation.
513

514 **ACKNOWLEDGMENTS**

515 This study was financially supported by NWO-TTW and the Breed4Food partners Cobb
516 Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster has been
517 made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR).

518

519 **LITERATURE CITED**

520

521 Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, *et al.*, 2007 Linkage
522 disequilibrium in related breeding lines of chickens. Genetics 177: 2161-2169.
523 Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab and J. P. Steibel, 2012 Estimation of
524 linkage disequilibrium in four US pig breeds. BMC Genom. 13: 1.
525 Brown, B. C., C. J. Ye, A. L. Price and N. Zaitlen, 2016 Transethnic genetic-correlation
526 estimates from summary statistics. Am. J. Hum. Genet. 99: 76-88.
527 Bulik-Sullivan, B., H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, *et al.*, 2015 An atlas of
528 genetic correlations across human diseases and traits. Nat. Genet. 47: 1236-1241.
529 Carillier, C., H. Larroque and C. Robert-Granié, 2014 Comparison of joint versus purebred
530 genomic evaluation in the French multi-breed dairy goat population. Genet. Sel. Evol.
531 46: 67.
532 Chen, L., F. Schenkel, M. Vinsky, D. Crews and C. Li, 2013 Accuracy of predicting genomic
533 breeding values for residual feed intake in Angus and Charolais beef cattle. J. Anim.
534 Sci. 91: 4669-4678.
535 De Candia, T. R., S. H. Lee, J. Yang, B. L. Browning, P. V. Gejman, *et al.*, 2013 Additive
536 genetic variation in schizophrenia risk is shared by populations of African and European
537 descent. Am. J. Hum. Genet. 93: 463-470.
538 De los Campos, G., D. Sorensen and D. Gianola, 2015 Genomic Heritability: What Is It?
539 PLoS Genet. 11: e1005048.
540 De Roos, A. P. W., B. J. Hayes, R. J. Spelman and M. E. Goddard, 2008 Linkage
541 disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle.
542 Genetics 179: 1503-1512.
543 Falconer, D. S., 1952 The problem of environment and selection. Amer. Nat. 86: 293-298.

544 Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian
545 inheritance. Trans. Roy. Soc. Edinburgh 52: 399-433.
546 Fisher, R. A., 1930 *The genetical theory of natural selection*. Oxford University Press,
547 Oxford, United Kingdom.
548 Flint-Garcia, S. A., J. M. Thornsberry and E. S. Buckler IV, 2003 Structure of linkage
549 disequilibrium in plants. Annu. Rev. Plant Biol. 54: 357-374.
550 Gianola, D., G. De los Campos, M. A. Toro, H. Naya, C.-C. Schön, *et al.*, 2015 Do molecular
551 markers inform about pleiotropy? Genetics 201: 23-29.
552 Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham and R. Thompson, 2015 *ASReml user*
553 *guide release 4.1*. VSN International Ltd, Hemel Hempstead.
554 Goddard, M. E., 2009 Genomic selection: Prediction of accuracy and maximisation of long
555 term response. Genetica 136: 245-257.
556 Goddard, M. E., B. J. Hayes and T. H. E. Meuwissen, 2011 Using the genomic relationship
557 matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128: 409-
558 421.
559 Groenen, M. A., P. Wahlberg, M. Foglio, H. H. Cheng, H.-J. Megens, *et al.*, 2009 A high-
560 density SNP-based linkage map of the chicken genome reveals sequence features
561 correlated with recombination rate. Genome Res. 19: 510-519.
562 Groenen, M. A., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, *et al.*, 2011 The
563 development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:
564 274.
565 Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, *et al.*, 2005 Extent and
566 consistency across generations of linkage disequilibrium in commercial layer chicken
567 breeding populations. Genetics 171: 1173-1181.

568 Hill, W. G. and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor.
569 Appl. Genet. 38: 226-231.
570 Hill, W. G., 1993 Variation in genetic identity within kinships. Heredity 71: 652-653.
571 Huang, H., J. J. Windig, A. Vereijken and M. P. Calus, 2014 Genomic prediction based on
572 data from three layer lines using non-linear regression models. Genet. Sel. Evol. 46: 75.
573 Ihara, N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, *et al.*, 2004 A comprehensive
574 genetic map of the cattle genome based on 3802 microsatellites. Genome Res. 14: 1987-
575 1998.
576 Karoui, S., M. Carabaño, C. Díaz and A. Legarra, 2012 Joint genomic evaluation of French
577 dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44: 39.
578 Kemper, K. E. and M. E. Goddard, 2012 Understanding and predicting complex traits:
579 Knowledge from cattle. Hum. Mol. Genet. 21: R45-R51.
580 Kemper, K. E., B. J. Hayes, H. D. Daetwyler and M. E. Goddard, 2015 How old are
581 quantitative trait loci and how widely do they segregate? J. Anim. Breed. Genet. 132:
582 121-134.
583 Krag, K., N. A. Poulsen, M. K. Larsen, L. B. Larsen, L. L. Janss, *et al.*, 2013 Genetic
584 parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a
585 Bayesian approach. BMC Genet. 14: 79.
586 Lee, S. H. and J. H. J. van der Werf, 2016 MTG2: an efficient algorithm for multivariate
587 linear mixed model analysis based on genomic information. Bioinformatics 32: 1420-
588 1422.
589 Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan, *et al.*, 2014 Usefulness of
590 multiparental populations of maize (*Zea mays* L.) for genome-based prediction.
591 Genetics 198: 3-16.

592 Lehermeier, C., C.-C. Schön and G. De los Campos, 2015 Assessment of genetic
593 heterogeneity in structured plant populations using multivariate whole-genome
594 regression models. Genetics 201: 323-337.
595 Matsuzaki, H., S. Dong, H. Loi, X. Di, G. Liu, *et al.*, 2004 Genotyping over 100,000 SNPs on
596 a pair of oligonucleotide arrays. Nat. Meth. 1: 109-111.
597 Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, *et al.*, 2009
598 Development and characterization of a high density SNP genotyping assay for cattle.
599 PLoS ONE 4: e5350.
600 Powell, J. E., P. M. Visscher and M. E. Goddard, 2010 Reconciling the analysis of IBD and
601 IBS in complex trait studies. Nat. Rev. Gen. 11: 800-805.
602 Pritchard, J. K. and M. Przeworski, 2001 Linkage disequilibrium in humans: Models and data.
603 Am. J. Hum. Genet. 69: 1-14.
604 Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, *et al.*,
605 2009 Design of a high density SNP genotyping assay in the pig using SNPs identified
606 and characterized by next generation sequencing technology. PLoS ONE 4: e6524.
607 Sargolzaei, M. and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for
608 livestock. Bioinformatics 25: 680-681.
609 Sawyer, S. L., N. Mukherjee, A. J. Pakstis, L. Feuk, J. R. Kidd, *et al.*, 2005 Linkage
610 disequilibrium patterns vary substantially among populations. Europ. J. Hum. Genet.
611 13: 677-686.
612 Shifman, S., J. Kuypers, M. Kokoris, B. Yakir and A. Darvasi, 2003 Linkage disequilibrium
613 patterns of the human genome across populations. Hum. Mol. Genet. 12: 771-776.
614 Sørensen, L. P., L. Janss, P. Madsen, T. Mark and M. S. Lund, 2012 Estimation of
615 (co)variances for genomic regions of flexible sizes: application to complex infectious
616 udder diseases in dairy cattle. Genet. Sel. Evol. 44: 18.

617 Speed, D., G. Hemani, Michael R. Johnson and David J. Balding, 2012 Improved heritability
618 estimation from genome-wide SNPs. Am. J. Hum. Genet. 91: 1011-1021.
619 Trzaskowski, M., O. S. P. Davis, J. C. DeFries, J. Yang, P. M. Visscher, *et al.*, 2013 DNA
620 evidence for strong genome-wide pleiotropy of cognitive and learning abilities. Behav.
621 Genet. 43: 267-273.
622 Veerkamp, R. F., H. A. Mulder, R. Thompson and M. P. L. Calus, 2011 Genomic and
623 pedigree-based genetic parameters for scarcely recorded traits when some animals are
624 genotyped. J. Dairy Sci. 94: 4189-4197.
625 Veroneze, R., P. S. Lopes, S. E. F. Guimarães, F. F. Silva, M. S. Lopes, *et al.*, 2013 Linkage
626 disequilibrium and haplotype block structure in six commercial pig lines. J. Anim. Sci.
627 91: 3493-3501.
628 Veroneze, R., J. W. Bastiaansen, E. F. Knol, S. E. Guimarães, F. F. Silva, *et al.*, 2014 Linkage
629 disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus
630 scrofa) populations. BMC Genet. 15: 126.
631 Visscher, P. M., G. Hemani, A. A. E. Vinkhuyzen, G.-B. Chen, S. H. Lee, *et al.*, 2014
632 Statistical power to detect genetic (co)variance of complex traits using SNP data in
633 unrelated samples. PLoS Genet 10: e1004269.
634 Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2013 The effect of linkage
635 disequilibrium and family relationships on the reliability of genomic prediction.
636 Genetics 193: 621-631.
637 Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten, *et al.*, 2015
638 Empirical and deterministic accuracies of across-population genomic prediction. Genet.
639 Sel. Evol. 47: 5.

Wientjes, Y. C. J., P. Bijma, R. F. Veerkamp and M. P. L. Calus, 2016 An equation to predict the accuracy of genomic values by combining data from multiple traits, breeds, lines, or environments. Genetics 202: 799-823.

Wientjes, Y. C. J., P. Bijma, J. Vandenplas and M. P. L. Calus, 2017 Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. Genetics 207: 503-515.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, *et al.*, 2015 Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47: 1114-1120.

Yang, L., B. M. Neale, L. Liu, S. H. Lee, N. R. Wray, *et al.*, 2013 Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: Genome-wide association study of both common and rare variants. Am. J. Med. Genet. 162: 419-430.

FIGURES

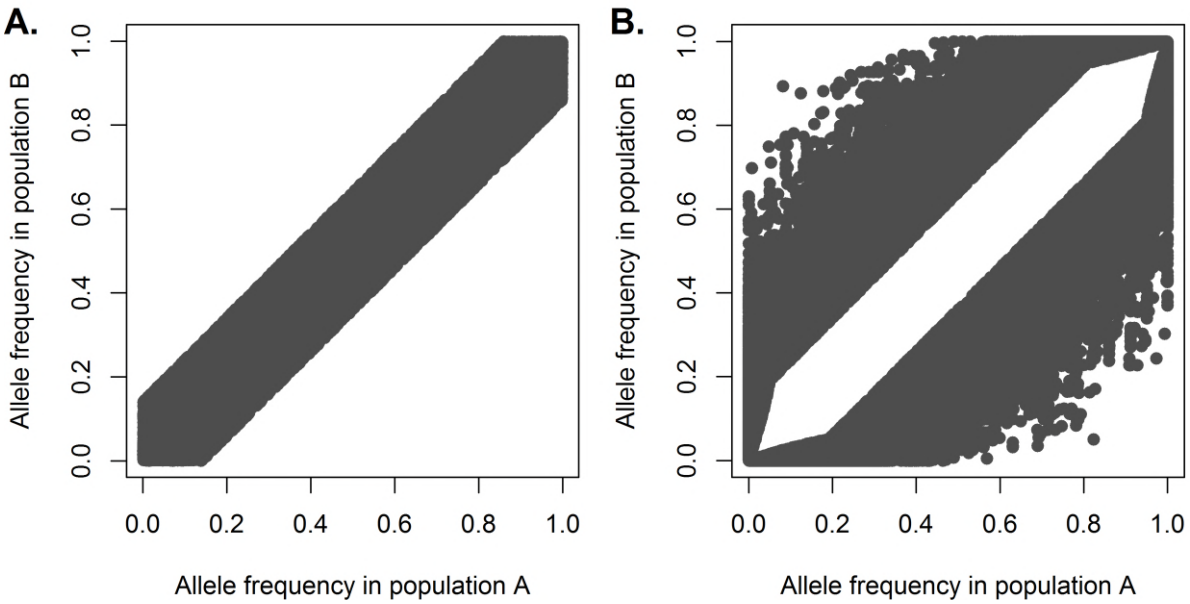
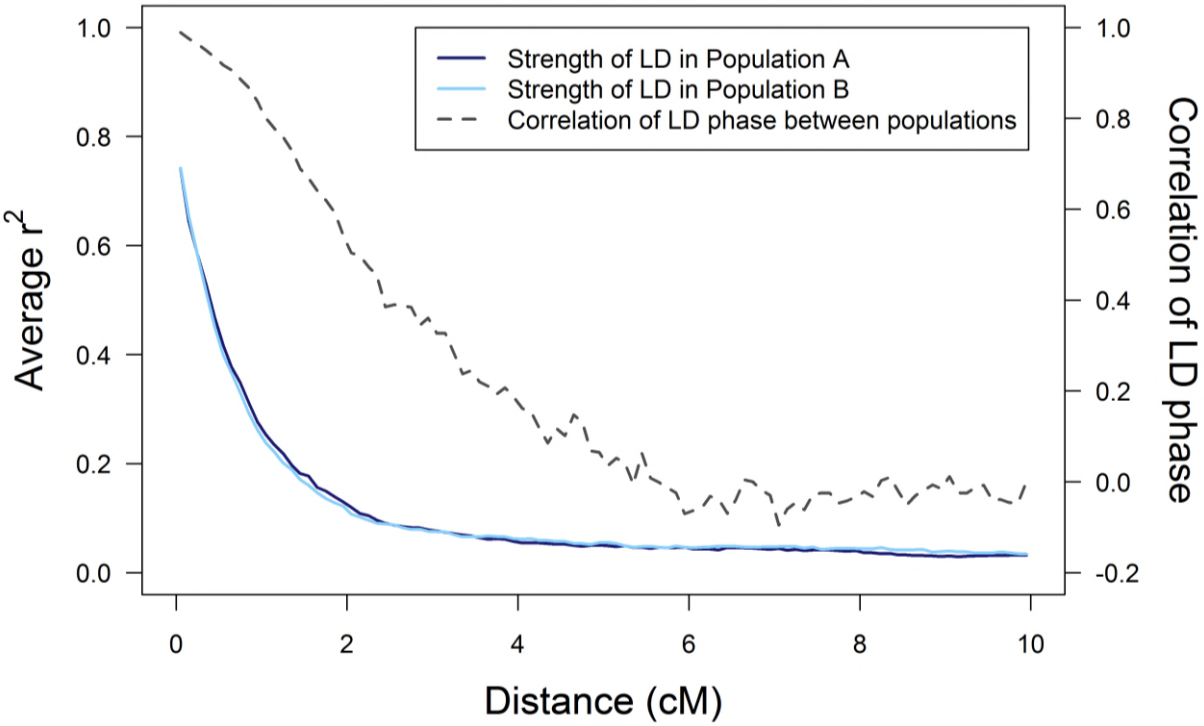


Figure 1 - Allele frequencies of markers for two populations using two selection approaches.

For one random replicate, allele frequencies of markers from both populations are plotted against each other when markers are selected to have (**A.**) similar allele frequencies in the two populations, or (**B.**) different allele frequencies in the two populations.



666

667 **Figure 2 - LD pattern in two populations and correlation of LD phase between the**

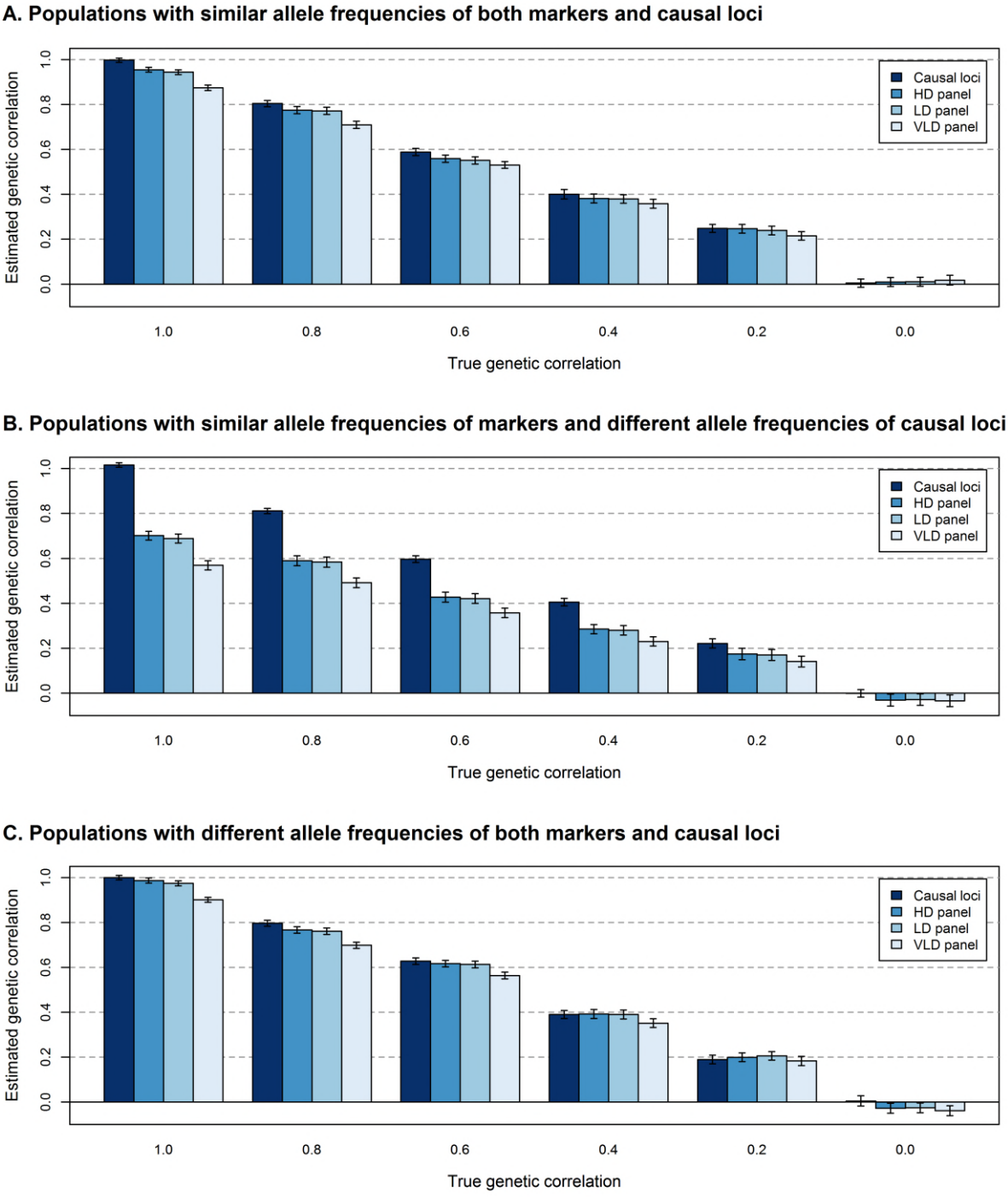
668 **populations.**

669 The average LD (r^2) between causal loci and markers for both populations, and the correlation

670 of LD-phase (correlation of r) between the populations, as a function of distance between

671 causal loci and markers for one random replicate.

672

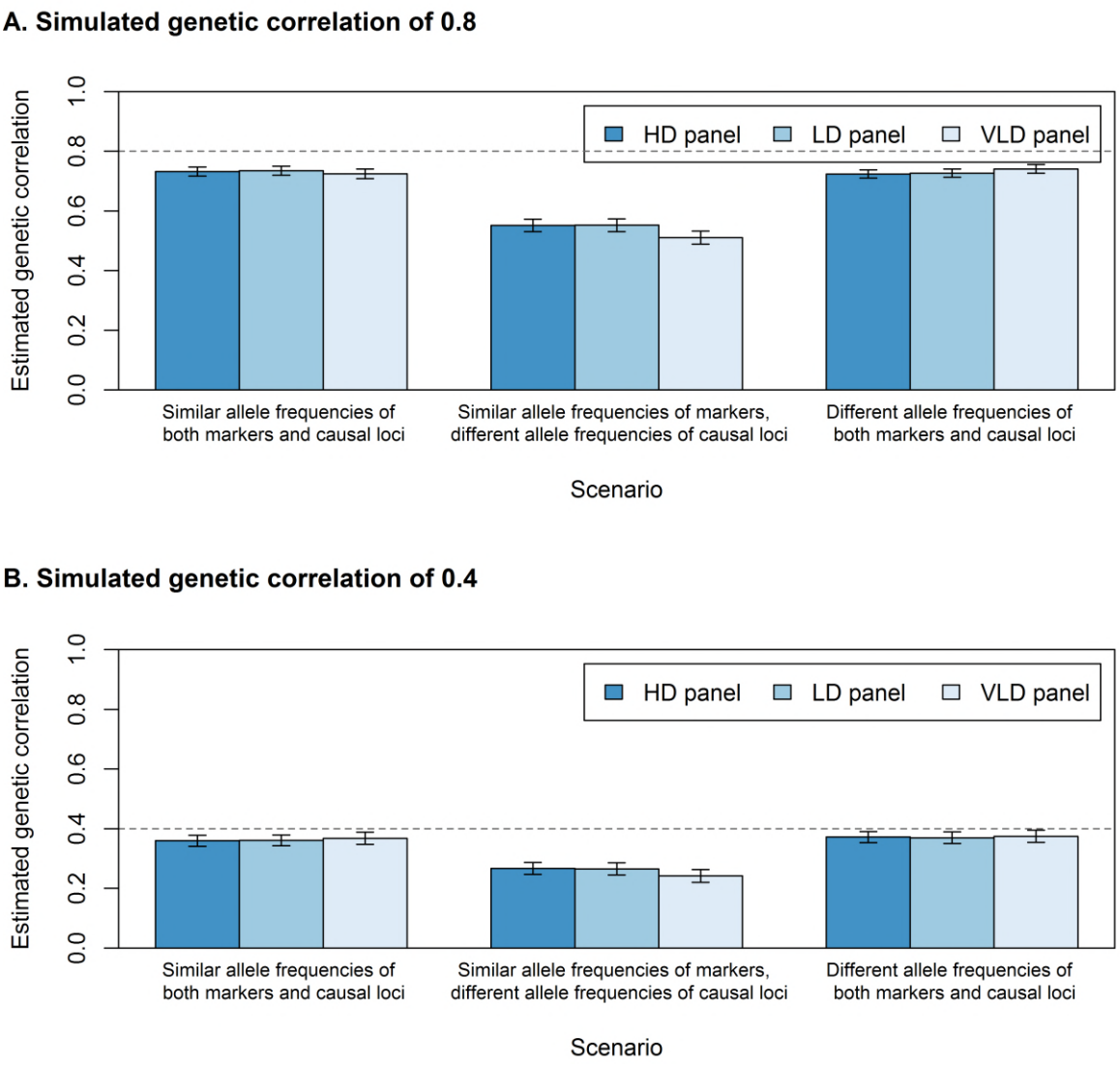


673

674 **Figure 3 - Estimated genetic correlations between populations without regressing the**
675 **genomic relationship matrix.**

676 The average estimated genetic correlation (± standard error) at different simulated genetic
677 correlations for the scenario where (**A.**) markers and causal loci have similar allele
678 frequencies in the two populations, (**B.**) markers have similar and causal loci different allele

679 frequencies in the two populations, or (**C.**) markers and causal loci have different allele
680 frequencies in the two populations, when the genomic relationship matrix is either based on
681 the genotypes of causal loci (2000), HDP (200 000), LDP (20 000), or VLDP (2000) markers
682 without regression towards the pedigree relationship matrix. Standard errors were calculated
683 as the standard deviation over replicates divided by the square root of the number of
684 replicates.
685



686

687 **Figure 4 - Estimated genetic correlations between populations with regression of the**
688 **genomic relationship matrix.**

689 The average estimated genetic correlation (± standard error) at a simulated genetic correlation
690 of (**A.**) 0.8 or (**B.**) 0.4 for the three scenarios with HDP (200 000), LDP (20 000), or VLDP
691 (2000) markers and regression of **G** towards the pedigree relationship matrix. Standard errors
692 were calculated as the standard deviation over replicates divided by the square root of the
693 number of replicates.

694

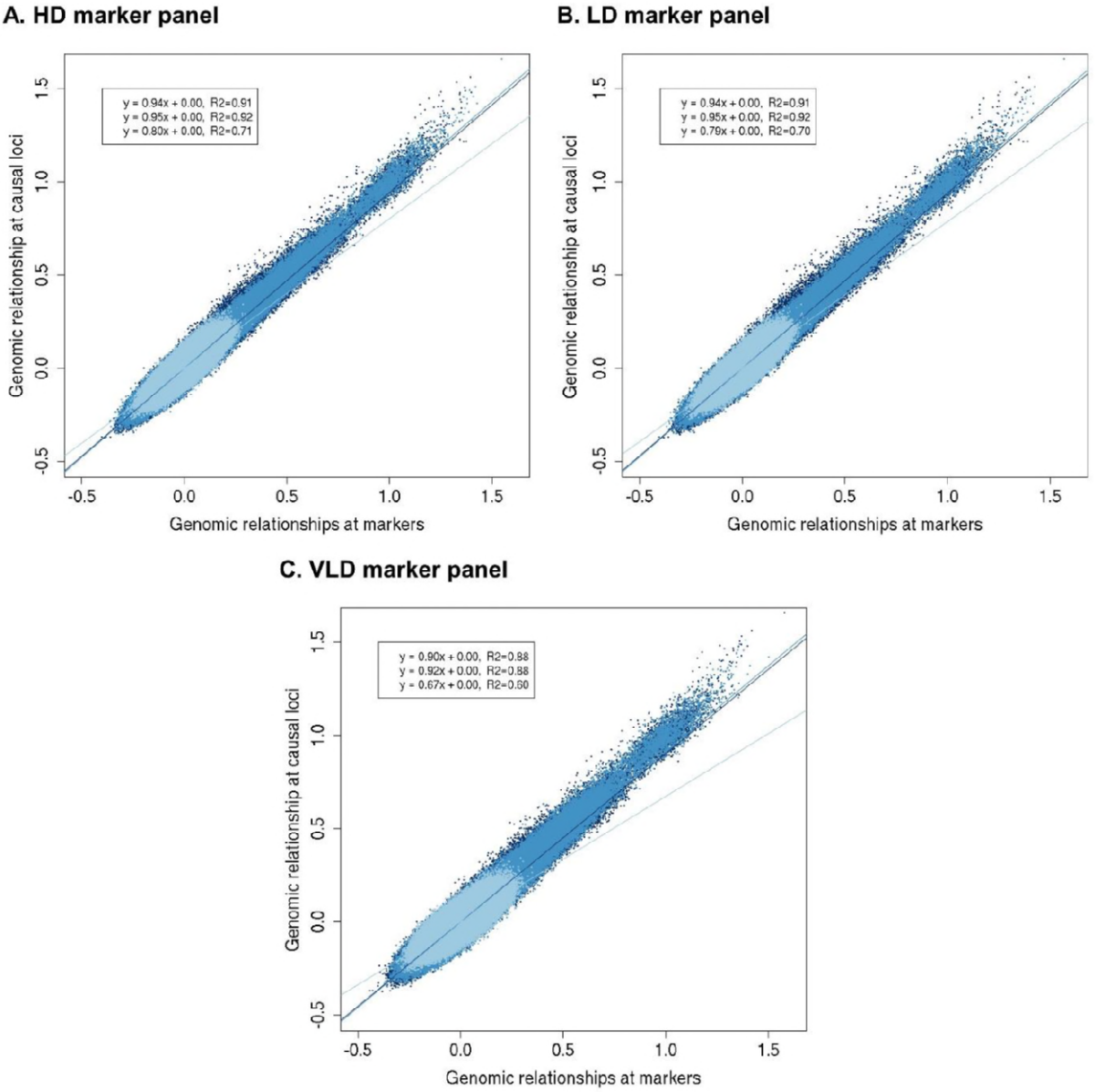
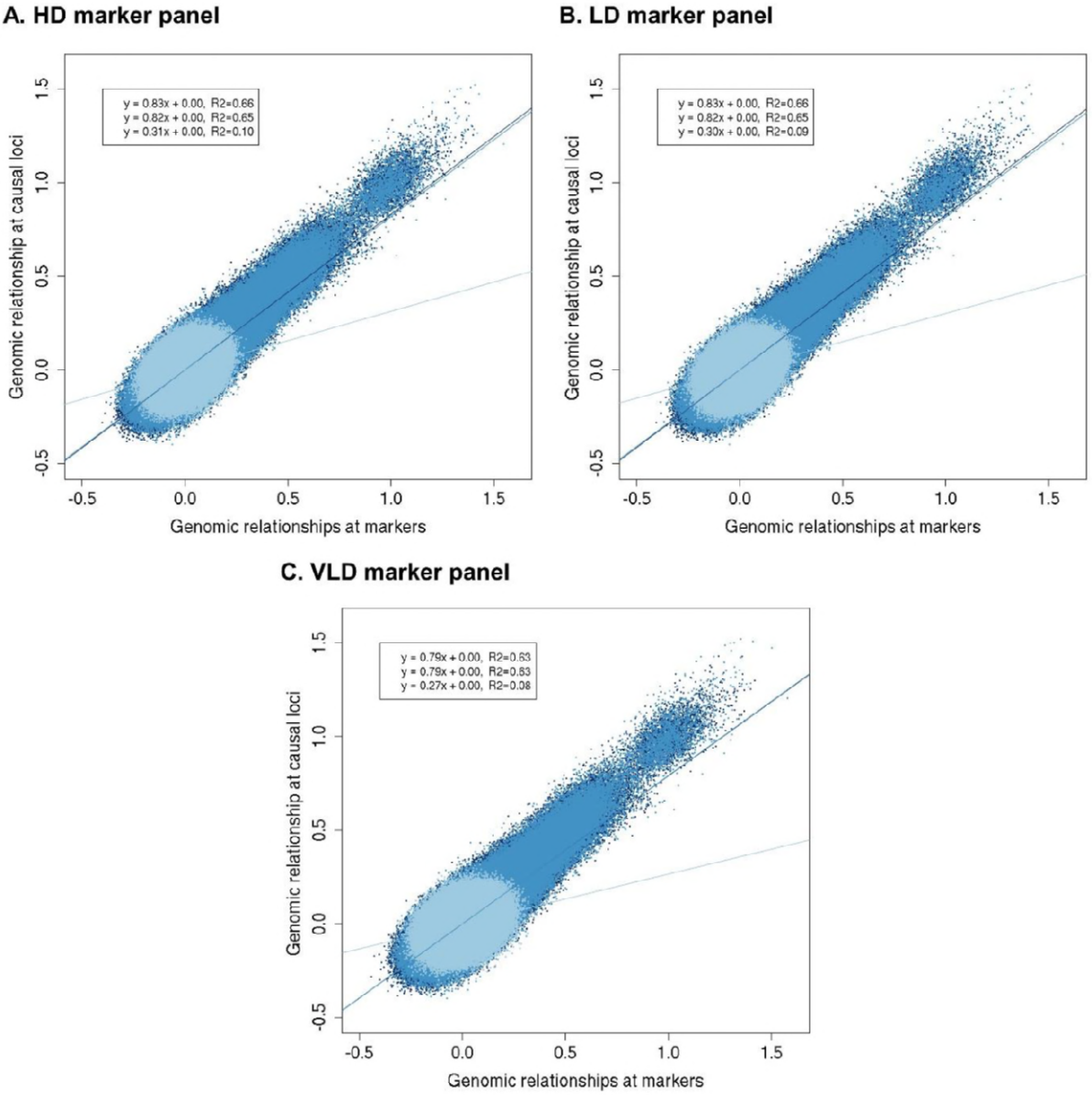


Figure 5 - Genomic relationships at causal loci versus markers when causal loci have similar allele frequencies in the two populations. The genomic relationships at the causal loci versus the genomic relationships based on (**A.**) HDP (200 000) markers, (**B.**) LDP (20 000) markers, or (**C.**) VLDP (2000) markers, when markers and causal loci have similar allele frequencies in the two populations for one replicate. Relationships in population A are represented in dark blue (equation 1 of regression line and correlation), relationships in population B are represented in medium blue (equation 2 of regression line and correlation), and relationships between population A and B are represented in light blue (equation 3 of regression line and correlation).



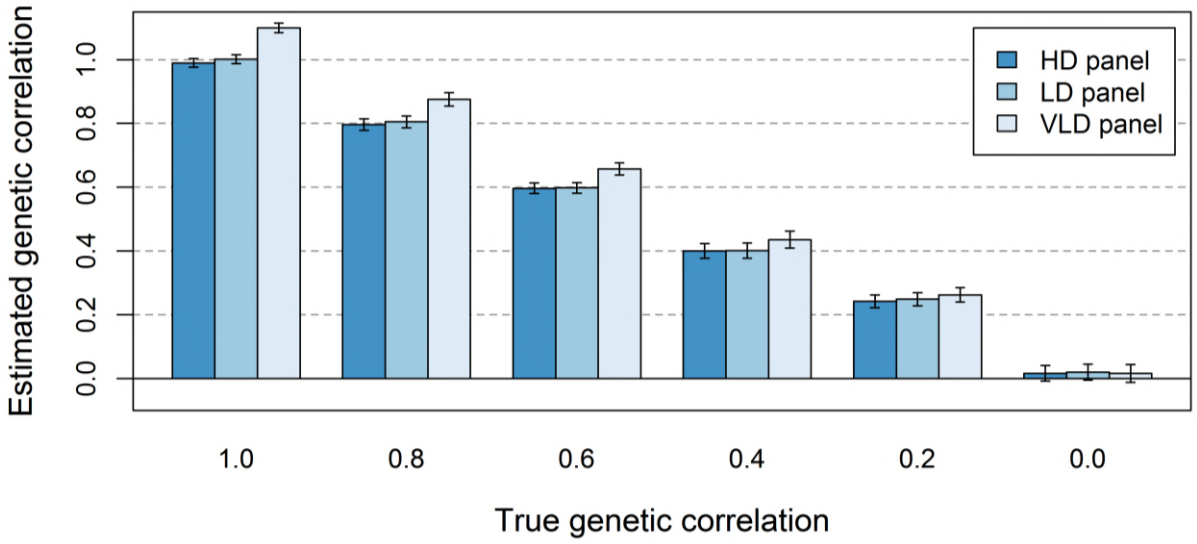
705

706 **Figure 6 - Genomic relationships at causal loci versus markers when causal loci have**
707 **different allele frequencies in the two populations.** The genomic relationships at the causal
708 loci versus the genomic relationships based on the (**A.**) HDP (200 000) markers, (**B.**) LDP (20
709 000) markers, or (**C.**) VLDP (2000) markers, when markers have similar and causal loci
710 different allele frequencies in the two populations for one replicate. Relationships in
711 population A are represented in dark blue (equation 1 of regression line and correlation),
712 relationships in population B are represented in medium blue (equation 2 of regression line

713 and correlation), and relationships between population A and B are represented in light blue

714 (equation 3 of regression line and correlation).

715



716

717 **Figure 7 - Estimated genetic correlations between populations after rescaling the**
718 **marker-based genomic relationship matrix.**

719 The average estimated genetic correlation (± standard error) at different simulated genetic
720 correlations for the scenario where markers and causal loci have similar allele frequencies in
721 the two populations when the genomic relationship matrix is either based on the genotypes of
722 HDP (200 000), LDP (20 000), or VLDP (2000) markers, after rescaling the marker-based
723 relationships using a regression coefficient based on the relationships at causal loci. Standard
724 errors were calculated as the standard deviation over replicates divided by the square root of
725 the number of replicates.

726

727

728

729