Supplementary Information

Contents

A Simulation Studies
A.1 Breeding Program Simulation using the WHEAT data ........................................... 2
A.2 Cross-Population Simulation using the HUMAN data ........................................... 4

B Real-World Data Analyses
B.1 WHEAT Data ................................. 5
B.2 MICE Data ................................... 6
B.3 Cross-Validation and Decay Curve in the WHEAT and MICE data ......................... 8

C Kinship and $F_{ST}$ .......................... 9

D Relationship between Squared Predictive Correlation and $F_{ST}^2$ ......................... 12
A Simulation Studies

A.1 Breeding Program Simulation using the WHEAT data

<table>
<thead>
<tr>
<th>Causal Variants</th>
<th>Generation</th>
<th>$F_{ST}$</th>
<th>$\bar{\rho}$</th>
<th>$\hat{\rho}_D$</th>
<th>$\hat{\rho}_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>0.003</td>
<td>0.54</td>
<td>0.61</td>
<td>0.63</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.027</td>
<td>0.50</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.055</td>
<td>0.31</td>
<td>0.49</td>
<td>0.48</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.001</td>
<td>0.50</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.026</td>
<td>0.34</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.052</td>
<td>0.24</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>0.001</td>
<td>0.46</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>200</td>
<td>2</td>
<td>0.027</td>
<td>0.26</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>200</td>
<td>3</td>
<td>0.053</td>
<td>0.19</td>
<td>0.23</td>
<td>0.18</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.001</td>
<td>0.44</td>
<td>0.35</td>
<td>0.36</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>0.027</td>
<td>0.25</td>
<td>0.29</td>
<td>0.28</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td>0.055</td>
<td>0.20</td>
<td>0.18</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table A.1. Predictive correlations for the simulations shown in Figure 1 in the paper; the training population for the genomic prediction model is composed by 200 varieties from 2002–2007 WHEAT data. $\bar{\rho}$ is the average predictive correlation for a given generation, training population size and number of causal variants; and $F_{ST}$ is the corresponding average $F_{ST}$. $\hat{\rho}_D$ is the decay curve estimate of $\bar{\rho}$, and is only available if the generation average falls within the span of the decay curve. $\hat{\rho}_L$ is the corresponding estimate from the linear extrapolation.
<table>
<thead>
<tr>
<th>Causal Variants</th>
<th>Generation</th>
<th>$\hat{F}_{ST}$</th>
<th>$\hat{\rho}$</th>
<th>$\hat{\rho}_D$</th>
<th>$\hat{\rho}_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>1</td>
<td>0.018</td>
<td>0.58</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>200</td>
<td>2</td>
<td>0.041</td>
<td>0.47</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>200</td>
<td>3</td>
<td>0.066</td>
<td>0.40</td>
<td>–</td>
<td>0.46</td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>0.088</td>
<td>0.36</td>
<td>–</td>
<td>0.42</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>0.111</td>
<td>0.30</td>
<td>–</td>
<td>0.38</td>
</tr>
<tr>
<td>200</td>
<td>6</td>
<td>0.127</td>
<td>0.27</td>
<td>–</td>
<td>0.35</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>0.141</td>
<td>0.25</td>
<td>–</td>
<td>0.33</td>
</tr>
<tr>
<td>200</td>
<td>8</td>
<td>0.151</td>
<td>0.20</td>
<td>–</td>
<td>0.31</td>
</tr>
<tr>
<td>200</td>
<td>9</td>
<td>0.158</td>
<td>0.19</td>
<td>–</td>
<td>0.30</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0.165</td>
<td>0.15</td>
<td>–</td>
<td>0.28</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.019</td>
<td>0.62</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>0.047</td>
<td>0.50</td>
<td>0.48</td>
<td>0.47</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td>0.077</td>
<td>0.46</td>
<td>–</td>
<td>0.41</td>
</tr>
<tr>
<td>1000</td>
<td>4</td>
<td>0.106</td>
<td>0.40</td>
<td>–</td>
<td>0.35</td>
</tr>
<tr>
<td>1000</td>
<td>5</td>
<td>0.126</td>
<td>0.33</td>
<td>–</td>
<td>0.31</td>
</tr>
<tr>
<td>1000</td>
<td>6</td>
<td>0.139</td>
<td>0.30</td>
<td>–</td>
<td>0.28</td>
</tr>
<tr>
<td>1000</td>
<td>7</td>
<td>0.150</td>
<td>0.25</td>
<td>–</td>
<td>0.26</td>
</tr>
<tr>
<td>1000</td>
<td>8</td>
<td>0.157</td>
<td>0.20</td>
<td>–</td>
<td>0.24</td>
</tr>
<tr>
<td>1000</td>
<td>9</td>
<td>0.164</td>
<td>0.19</td>
<td>–</td>
<td>0.23</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>0.168</td>
<td>0.15</td>
<td>–</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table A.2. Predictive correlations for the simulations shown in Figure 2 in the paper; the training population for the genomic prediction model is composed by the 800 varieties available after the second round of selection in the simulation. The notation is the same as in Table A.1.
## A.2 Cross-Population Simulation using the HUMAN data

<table>
<thead>
<tr>
<th>Training Population</th>
<th>Target Population</th>
<th>Causal Variants</th>
<th>( F_{ST} )</th>
<th>( \hat{\rho}_P )</th>
<th>( \hat{\rho}_D )</th>
<th>( \hat{\rho}_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>Europe</td>
<td>5</td>
<td>0.068</td>
<td>0.68</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Middle east</td>
<td>5</td>
<td>0.076</td>
<td>0.67</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>America</td>
<td>5</td>
<td>0.154</td>
<td>0.69</td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Africa</td>
<td>5</td>
<td>0.156</td>
<td>0.64</td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>5</td>
<td>0.174</td>
<td>0.78</td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td>Asia</td>
<td>Europe</td>
<td>20</td>
<td>0.068</td>
<td>0.49</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Middle east</td>
<td>20</td>
<td>0.076</td>
<td>0.32</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>America</td>
<td>20</td>
<td>0.154</td>
<td>0.48</td>
<td></td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Africa</td>
<td>20</td>
<td>0.156</td>
<td>0.59</td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>20</td>
<td>0.174</td>
<td>0.43</td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>Asia</td>
<td>Europe</td>
<td>100</td>
<td>0.068</td>
<td>0.09</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Middle east</td>
<td>100</td>
<td>0.076</td>
<td>0.12</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>America</td>
<td>100</td>
<td>0.154</td>
<td>0.02</td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Africa</td>
<td>100</td>
<td>0.156</td>
<td>0.15</td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>100</td>
<td>0.174</td>
<td>0.03</td>
<td></td>
<td>−0.05</td>
</tr>
<tr>
<td>Asia</td>
<td>Europe</td>
<td>2000</td>
<td>0.068</td>
<td>0.13</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Middle east</td>
<td>2000</td>
<td>0.076</td>
<td>0.14</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>America</td>
<td>2000</td>
<td>0.154</td>
<td>0.24</td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Africa</td>
<td>2000</td>
<td>0.156</td>
<td>0.03</td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>2000</td>
<td>0.174</td>
<td>0.03</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Asia</td>
<td>Europe</td>
<td>10000</td>
<td>0.068</td>
<td>0.15</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Middle east</td>
<td>10000</td>
<td>0.076</td>
<td>0.21</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>America</td>
<td>10000</td>
<td>0.154</td>
<td>0.02</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Africa</td>
<td>10000</td>
<td>0.156</td>
<td>0.22</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>10000</td>
<td>0.174</td>
<td>−0.18</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>Asia</td>
<td>Europe</td>
<td>50000</td>
<td>0.068</td>
<td>0.28</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Middle east</td>
<td>50000</td>
<td>0.076</td>
<td>0.11</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>America</td>
<td>50000</td>
<td>0.154</td>
<td>0.00</td>
<td></td>
<td>−0.07</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>50000</td>
<td>0.174</td>
<td>−0.10</td>
<td></td>
<td>−0.07</td>
</tr>
</tbody>
</table>

Table A.3. Predictive correlations for the simulations shown in Figure 3 in the paper. \( \hat{\rho}_P \) is the predictive correlation for the target population from the full training population. \( \hat{\rho}_D \) is the decay curve estimate of \( \hat{\rho}_P \), and is only available if the target population falls within the span of the decay curve. \( \hat{\rho}_L \) is the corresponding estimate from the linear extrapolation.
B  Real-World Data Analyses

B.1  WHEAT Data

Figure B.1. Decay curves for grain yield, height, flowering time and grain protein content estimated from the French wheat varieties in the WHEAT data. The blue circles are the \( \hat{\rho}_D^{(m)} \) used to build the curve, and the red point is \( \hat{\rho}_{D}^{(0)} \). The blue line is the mean decay trend, with a shaded 95% confidence interval, and the dashed blue line is the linear interpolation provided by the \( \hat{\rho}_L \). Gray squares are the \( \hat{\rho}_{CV} \) computed using hold-out cross-validation. The red squares labelled GBR and DEU correspond to the \( \hat{\rho}_P \) for the British and German varieties, and the red brackets are the respective 95% confidence intervals.
B.2 MICE Data

Figure B.2. Decay curves for weight estimated from the 4 largest families in the MICE data, labelled F005, F008, F010 and F016. The red squares in each panel correspond to the predictive correlations for the populations not used for estimating the decay curve; the red brackets are 95% confidence intervals. Formatting is the same as in Figure B.1.
Figure B.3. Decay curves for growth rate estimated from the 4 largest families in the MICE data, labelled F005, F008, F010 and F016. The red squares in each panel correspond to the predictive correlations for the populations not used for estimating the decay curve; the red brackets are 95% confidence intervals. Formatting is the same as in Figure B.1.
### B.3 Cross-Validation and Decay Curve in the WHEAT and MICE data

<table>
<thead>
<tr>
<th>Trait</th>
<th>Training Population</th>
<th>( \hat{F}_{ST} )</th>
<th>( \hat{\rho}_{CV} )</th>
<th>( \hat{\rho}_{D} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEAT, Yield</td>
<td>France</td>
<td>0.006</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>WHEAT, Height</td>
<td>France</td>
<td>0.006</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td>WHEAT, Flowering time</td>
<td>France</td>
<td>0.006</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>WHEAT, Grain protein content</td>
<td>France</td>
<td>0.006</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>MICE, Weight</td>
<td>F005</td>
<td>0.001</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>F008</td>
<td>0.001</td>
<td>0.56</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>F010</td>
<td>0.001</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>F016</td>
<td>0.001</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>MICE, Growth rate</td>
<td>F005</td>
<td>0.001</td>
<td>0.27</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>F008</td>
<td>0.001</td>
<td>0.34</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>F010</td>
<td>0.001</td>
<td>0.40</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>F016</td>
<td>0.001</td>
<td>0.22</td>
<td>0.23</td>
</tr>
</tbody>
</table>

**Table B.4.** Predictive correlations from the decay curves and from cross-validation for the analyses shown in Figures B.1, B.2 and B.3. \( \hat{F}_{ST} \) and \( \hat{\rho}_{CV} \) are the mean genetic distance and mean predictive correlation from the 40 runs of hold-out cross-validation; \( \hat{\rho}_{D} \) is the predictive correlation estimated by the decay curve at genetic distance \( \hat{F}_{ST} \).
C Kinship and $F_{ST}$

<table>
<thead>
<tr>
<th>Data</th>
<th>Subset</th>
<th>ms</th>
<th>COR($\hat{F}_{ST}^{(m)}$, $\bar{k}^{(m)}$)</th>
<th>$\log_{10}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEAT</td>
<td>France</td>
<td>401</td>
<td>−0.9894</td>
<td>−672.10</td>
</tr>
<tr>
<td>MICE</td>
<td>F005</td>
<td>601</td>
<td>−0.9982</td>
<td>−1467.58</td>
</tr>
<tr>
<td>MICE</td>
<td>F008</td>
<td>601</td>
<td>−0.9982</td>
<td>−1467.58</td>
</tr>
<tr>
<td>MICE</td>
<td>F010</td>
<td>601</td>
<td>−0.9906</td>
<td>−1038.57</td>
</tr>
<tr>
<td>MICE</td>
<td>F016</td>
<td>601</td>
<td>−0.9948</td>
<td>−1192.05</td>
</tr>
<tr>
<td>HUMAN</td>
<td>Asia</td>
<td>601</td>
<td>−0.9998</td>
<td>−2038.97</td>
</tr>
</tbody>
</table>

Table C.5. Correlation between $\hat{F}_{ST}^{(m)}$ and $\bar{k}^{(m)}$ in the data sets and training populations used in the paper. The p-values are computed using the exact t-test for the correlation coefficient [2] and adjusted for multiplicity via FDR [1].

Figure C.4. ($F_{ST}^{(m)}$, $\bar{k}^{(m)}$) pairs generated from the French wheat varieties in the WHEAT data.
Figure C.5. $(F_{ST}^{(m)}, \bar{\tilde{k}}^{(m)})$ pairs generated from the 4 largest families in the MICE data, labelled F005, F008, F010 and F016.
Figure C.6. \((F_{ST}^{(m)}, \bar{k}^{(m)})\) pairs generated from the Asian individuals in the HUMAN data.
D Relationship between Squared Predictive Correlation and $F^2_{ST}$

[4] used a simulated dairy cattle population, created simulating both phenotypes and genotypes, suggested that squared predictive correlation has a stronger linear relationship with squared mean kinship than predictive correlation does with mean kinship. Predictive correlation was computed using GBLUP as a genomic prediction model.

In the context of this paper, this is equivalent to testing whether the $\left(\hat{\rho}^{(m)}_D\right)^2$ have a stronger linear relationship with the $\left(\hat{F}^{(m)}_{ST}\right)^2$ than the $\hat{\rho}^{(m)}_D$ do with the $\hat{F}^{(m)}_{ST}$; we have shown that $F^{(m)}_{ST}$ and $\bar{k}^{(m)}$ are almost perfectly linearly correlated so they can be used interchangeably for this purpose. We regress the $\hat{\rho}^{(m)}_D$ on the $\hat{F}^{(m)}_{ST}$ and measure the $R^2$ coefficient of the resulting linear model, denoted as $R^2_{\text{LINEAR}}$. Similarly, we regress the $\left(\hat{\rho}^{(m)}_D\right)^2$ on the $\left(\hat{F}^{(m)}_{ST}\right)^2$ and measure $R^2_{\text{QUADRATIC}}$. Both are reported in Tables D.6 and D.7 for all the analyses with real and simulated phenotypes.

To test whether there is a significant difference between $R^2_{\text{LINEAR}}$ and $R^2_{\text{QUADRATIC}}$ we perform a permutation two-sample $t$-test as described in [3], using 10000 permutations. The resulting p-value is 0.784, hence we conclude that the difference between the relationship we consider in this paper and that suggested in [4] is not significant.
<table>
<thead>
<tr>
<th>Data</th>
<th>Trait</th>
<th>Training Population</th>
<th>$R^2_{\text{LINEAR}}$</th>
<th>$R^2_{\text{QUADRATIC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEAT</td>
<td>Yield</td>
<td>France</td>
<td>0.575</td>
<td>0.634</td>
</tr>
<tr>
<td></td>
<td>Height</td>
<td>France</td>
<td>0.371</td>
<td>0.424</td>
</tr>
<tr>
<td></td>
<td>Flowering Time</td>
<td>France</td>
<td>0.412</td>
<td>0.410</td>
</tr>
<tr>
<td></td>
<td>Grain protein content</td>
<td>France</td>
<td>0.681</td>
<td>0.681</td>
</tr>
<tr>
<td>MICE</td>
<td>Weight</td>
<td>F005</td>
<td>0.056</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F008</td>
<td>0.246</td>
<td>0.236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F010</td>
<td>0.537</td>
<td>0.463</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F016</td>
<td>0.311</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>F005</td>
<td>0.446</td>
<td>0.437</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F008</td>
<td>0.426</td>
<td>0.404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F010</td>
<td>0.013</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F016</td>
<td>0.384</td>
<td>0.372</td>
</tr>
</tbody>
</table>

Table D.6. $R^2_{\text{LINEAR}}$ and $R^2_{\text{QUADRATIC}}$ for the data analyses on real phenotypes.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Sample Size</th>
<th>Causal Variants</th>
<th>$R^2_{\text{LINEAR}}$</th>
<th>$R^2_{\text{QUADRATIC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic selection</td>
<td>200</td>
<td>10</td>
<td>0.387</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>50</td>
<td>0.307</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>200</td>
<td>0.122</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1000</td>
<td>0.263</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>800</td>
<td>0.284</td>
<td>0.293</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1000</td>
<td>0.351</td>
<td>0.352</td>
</tr>
<tr>
<td>Cross-population</td>
<td>435</td>
<td>5</td>
<td>0.123</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>435</td>
<td>20</td>
<td>0.175</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>435</td>
<td>100</td>
<td>0.565</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>435</td>
<td>2000</td>
<td>0.131</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>435</td>
<td>10000</td>
<td>0.023</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>435</td>
<td>50000</td>
<td>0.256</td>
<td>0.118</td>
</tr>
</tbody>
</table>

Table D.7. $R^2_{\text{LINEAR}}$ and $R^2_{\text{QUADRATIC}}$ for the data used in the simulation studies.
References


