GENETICS, 2026, 232(1), iyaf230
https://doi.org/10.1093/genetics/iyaf230
Advance Access Publication Date: 23 October 2025
Investigation

Impact of erroneous marker data on the accuracy of narrow-sense heritability

Christi Sagariya [1], Václav Bittner [1,2], Torsten Pook [3], Amit Roy [1], Milan Lstibůrek [1,*]

[1]Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha 6 165 00, Czech Republic
[2]Faculty of Science, Humanities and Education, Technical University of Liberec, Liberec 1 461 17, Czech Republic
[3]Animal Breeding and Genomics, Wageningen University & Research, P.O. Box 338, Wageningen, AH 6700, The Netherlands

*Corresponding author: Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha 6 165 00, Czech Republic.
Email: lstiburek@fld.czu.cz

Genomic relationship matrices computed from single nucleotide polymorphism (SNP) data are now widely used to estimate narrow-sense heritability (h^2), yet the impact of genotyping error on these estimates is not well understood. We used stochastic simulation and supporting algebra to examine this impact and its interplay with marker density. Starting from a diploid founder population with 300 additive quantitative trait loci, we simulated SNP panels with densities ranging from 6.25 to 50 SNPs per cM and traits with true h^2 of either 0.2 or 0.6. Genotypes were then altered at error rates $\varepsilon = 0 - 1$ under three error kernels. For each of 100 simulation replicates, we calculated the genomic relationship matrix using VanRaden's method and estimated h^2 with restricted maximum likelihood (REML). In the absence of error, low-density marker panels underestimated h^2. Sparse panels were also the most tolerant up to $\varepsilon \approx 0.1$ yet still underestimated h^2. Conversely, the densest panel recovered the true h^2 when $\varepsilon = 0$, but even a small error $\varepsilon > 0.01$ caused an upward bias. The analysis reveals that all distortions are attributable to: (i) a shift in the mean off-diagonal elements of the genomic relationship matrix with magnitude $(1 - \varepsilon)^2$ and (ii) a change in the ratio between the mean diagonal and mean off-diagonal elements of the genomic relationship matrix. When $\varepsilon \gtrsim 0.6$, every kernel pushed h^2 toward zero. Thus, even modest genotyping error can inflate or deflate additive genetic variance estimates. SNP panels therefore require rigorous laboratory quality control, error-aware imputation, and statistical models that account for genotype uncertainty when estimating h^2.

Keywords: genomic relationship matrix; genetic evaluation; allelic-based simulation; single nucleotide polymorphism; genotyping error

Introduction

The conventional numerator relationship matrix represents expected genetic relationships between individuals based on identity-by-descent probabilities, which are calculated recursively from a pedigree. In contrast, the genomic relationship matrix **G** captures the degree of genetic resemblance directly from DNA sequence data (VanRaden 2008). This approach offers multiple advantages, including accounting for historical relationships, separating nonadditive genetic variance components (Gamal El-Dien et al. 2016), and incorporating the Mendelian sampling term (Hill and Weir 2011). Consequently, it enhances the accuracy of breeding value estimation, genetic gain prediction, and increases selection efficiency for complex traits (Meuwissen et al. 2001).

With the growing availability of DNA marker data, **G** matrices are now widely used to estimate genetic parameters across various fields, including animal breeding (Legarra et al. 2009; Goddard et al. 2011), plant breeding (Bernardo and Yu 2007; Zapata-Valenzuela et al. 2013; Munoz et al. 2014), as well as conservation and evolutionary biology (Allendorf et al. 2010; Lee et al. 2010; Gienapp et al. 2017). Among these parameters, narrow-sense heritability h^2 is particularly significant and forms the focus of our investigation. Its estimation is typically performed by solving mixed-model equations (Henderson 1953), where **G** is incorporated into the random component. The **G** matrix is derived solely from genotypic data, integrating information from all marker loci. Consequently, the accuracy of **G** depends directly on the quality of marker genotypic data, which, in turn, affects the accuracy of h^2.

Several factors can compromise marker genotypic data quality, including issues related to DNA sequencing, sample quality, laboratory conditions, or human mistakes (Pompanon et al. 2005). Genotype mismatches may arise from mutations, null alleles, allelic dropouts, or false alleles (Wang J 2018). These errors impact estimating various parameters, including the genetic distance between markers in linkage mapping and recombination events (Goldstein et al. 1997), allele sharing in affected sib-pairs (Abecasis et al. 2001), and the impact of missing data on detecting linkage disequilibrium (LD) and haplotype generation (Hinrichs and Suarez 2005). Additionally, errors influence the identification of markers associated with low allele frequency (Hosking et al. 2004), the rate of nonrecombinant haplotypes inherited by offspring (Gordon et al. 1999), and inconsistencies with Mendelian

Received on 28 July 2025; accepted on 17 October 2025

inheritance, leading to false exclusion rates in family-based studies (Seaman and Holmans 2005).

Genotyping errors can arise at multiple sequencing and data processing stages. Wet lab issues such as DNA degradation, contamination, and polymerase chain reaction (PCR) amplification bias can lead to allelic dropout and unequal allele representation (Varma et al. 2007; Aird et al. 2011). During sequencing, platform-specific errors may introduce base substitutions, insertions and deletions, especially in homopolymer regions or context-sensitive motifs (Rang et al. 2018; Wenger et al. 2019). In subsequent steps, mapping bias and differences in variant caller sensitivity can distort genotype likelihoods, particularly for rare alleles and indels (Li 2014; Brandt et al. 2015). Postprocessing procedures such as genotype imputation and phasing, particularly when applied to low-coverage data, may introduce systematic miscoding even after stringent quality control (Browning and Browning 2016). Although standard quality control approaches, including Mendelian consistency checks and Hardy–Weinberg equilibrium (HWE) testing, are routinely implemented (Wiggans et al. 2009; Cheung et al. 2014), empirical studies still report locus-specific genotyping error rates ranging from less than one percent (Pook et al. 2019) to as high as fifteen percent (Bonin et al. 2004). These persistent errors pose a significant challenge for accurately estimating genetic parameters (Sobel et al. 2002; Forneris et al. 2015).

While the presence of genotyping errors is well-documented, their quantitative impact on genetic parameter estimation has been explored in a limited number of studies, primarily focusing on the reliability of genomic predictions. Akbarpour et al. (2021) simulated a genome resembling that of rainbow trout, introducing genotyping errors randomly without favoring any allele. They estimated genomic breeding values (GEBVs) by incorporating the genotype marker matrix into a mixed model to estimate single nucleotide polymorphism (SNP) marker effects. Since GEBVs are derived as the product of an individual's marker matrix and estimated marker effects, any genotyping error directly affects marker effect estimates, ultimately distorting genomic predictions. Their findings revealed that genotyping errors disrupted marker effects, causing bias in the regression coefficients between true and estimated GEBVs, regardless of trait genetic architecture. Even minimal errors significantly reduced genomic prediction accuracy. Furthermore, Wang X et al. (2019) focused on a common error in Genotyping-by-Sequencing (GBS) data where heterozygous genotypes are miscalled as homozygous, particularly at low sequencing depths. By simulating a livestock population, they demonstrated that statistically correcting these genotypes not only increased the accuracy of the genotype calls but also improved the reliability of subsequent genomic predictions. Similarly, Khalilisamani et al. (2021) simulated a genome similar to that of Black Tiger prawns, evaluating the effect of genotyping errors and family contribution types (equal vs. unequal) on genomic prediction accuracy. They applied a probability matrix to model genotypic errors and used the **G** matrix in mixed models to estimate breeding values. Their results revealed that genotyping errors had a greater impact on genomic prediction accuracy than family contribution types, regardless of marker subset size or h^2 value. Low-density SNP subsets were particularly vulnerable to genotyping errors, while high-density subsets were comparatively less affected.

Although the reliability of genomic evaluation has received growing attention (Goddard 2009; Wiggans et al. 2011; Misztal et al. 2020), the impact of genotyping errors on h^2 estimation remains unexplored. We use simulations to evaluate this impact, focusing on bias and precision. To investigate this, we simulate a population with varying marker densities and trait h^2, introduce different genotyping errors, and conduct genetic evaluations using **G** matrices. We assess how these errors affect the accuracy of h^2 estimates. We report that even minor genotyping errors can substantially bias h^2 estimates, especially at high marker densities. This bias increases nonlinearly with error magnitude. We provide mathematical reasoning to explain these patterns and discuss their implications for current practices in genomic evaluation, emphasizing the importance of accounting for genotyping errors.

Methods

The methodology of this study consists of three stages: population simulation, introduction of errors into the simulated marker dataset, and genetic evaluation. All stages of the simulation study were performed in the R system (R Core Team 2023). The entire process was repeated 100 times using independent stochastic simulations, and the means and standard deviations were calculated across these replicates.

Population simulation

The population simulation was conducted using the MoBPS software (Pook et al. 2020), starting with a founder population of 5,000 unrelated individuals. A genome was simulated with four linkage groups, each 100 cM in length. We considered 2,500, 5,000, and 20,000 SNP markers, corresponding to marker densities of 6.25, 12.5, and 50 markers per cM, respectively; these markers were evenly spaced across the genome. A predefined genetic architecture for a single quantitative trait included an additional 300 purely additive quantitative trait loci (QTLs). The initial allele frequencies for all markers and QTLs were randomly sampled from a Beta distribution with shape parameters $\alpha = 1$ and $\beta = 1$, resulting in a uniform distribution. Additive effects of QTLs were sampled from a normal distribution $N(0, \sigma_a^2)$, where σ_a^2 is the additive genetic variance. h^2 of the trait was set to 0.2 and 0.6.

To generate a LD structure, the founder population underwent 100 generations of random mating. From the final generation, a parental population of 100 individuals (denoted further as "parents") was randomly sampled. These individuals were then randomly mated to produce an offspring population of 1,000 individuals (denoted further as "offspring").

We emphasize that simulation was chosen deliberately over empirical data because it provides access to a known error-free reference, allowing us to isolate and quantify the impact of genotyping error. Unlike real datasets, where true relationships and error rates are unknown, simulations permit systematic evaluation across error kernels and magnitudes, with patterns further supported by mathematical reasoning.

Calculation of genetic relationships from marker data

Following VanRaden (2008), let $\mathbf{M} = (m_{ij} \in \mathbb{R}^{n \times m}) : m_{ij} \in \{-1; 0; 1\}$ be a genotype marker matrix, where n is the number of individuals (1,000 offspring with available phenotypic records assumed as a base for genetic evaluation in the current study) and m is the number of marker loci (described previously). The values −1 and 1 refer to homozygotes, while 0 corresponds to heterozygotes. Let $\mathbf{p}$ be a row vector whose ith element is defined as $2(p_i - 0.5)$, where p_i denotes the allele frequency of the second allele at locus i. The genomic relationship matrix $\mathbf{G} = (g_{ij} \in \mathbb{R}^{n \times n})$ can then be computed

(hereafter referred to as the first method) as

$$\mathbf{G}^{\text{method1}} = \frac{(\mathbf{M} - \mathbf{1}_n\mathbf{p})(\mathbf{M} - \mathbf{1}_n\mathbf{p})^\intercal}{2\sum p_i(1-p_i)} \quad (1)$$

Next, assume a diagonal matrix $\mathbf{D} = (d_{ii}) \in \mathbb{R}^{n\times n}$ such that

$$d_{i,i} = \begin{cases} 0 & \text{if } i \neq j \\ [m[2p_i(1-p_i)]]^{-1} & \text{if } i = j \end{cases} \quad (2)$$

Then, $\mathbf{G}$ can also be computed (referred to as the second method) as

$$\mathbf{G}^{\text{method2}} = (\mathbf{M} - \mathbf{1}_n\mathbf{p})\mathbf{D}(\mathbf{M} - \mathbf{1}_n\mathbf{p})^\intercal \quad (3)$$

The "miraculix" R package (Schlather 2020) was used to calculate $\mathbf{G}$ from $\mathbf{M}$.

Erroneous marker data

We introduced varying levels of error into the $\mathbf{M}$ matrix, where the proportion of altered elements, denoted as ε, ranged from $\varepsilon = 0$ (no error) to $\varepsilon = 1$ (all elements altered). The elements to be altered were selected randomly. Three scenarios for error introduction were considered, as summarized in Table 1. In Scenario 1, allelic substitution was based on the existing genotype in a respective element of $\mathbf{M}$. In Scenario 2, substitution probabilities were based on H-W genotypic frequencies calculated for each marker locus, which were fixed prior to introducing error. Finally, in Scenario 3, substitution probabilities were uniform across all elements of $\mathbf{M}$.

These scenarios can be interpreted as follows: In Scenario 1, substitutions deliberately promote alternative genotypes to replace existing ones. In Scenario 2, substitutions aim to preserve genotypes in line with the expected HWE frequencies at each locus. Meanwhile, Scenario 3 represents a neutral case where substitutions occur independently of the existing genotype or allelic frequency. After error introduction, we employed the ASRgenomics R package (Gezan et al. 2022) for quality filtering of $\mathbf{M}$ by applying a minor allele frequency threshold of 0.05. Under Scenario 2, where allele frequencies remain unchanged, approximately 5% of loci are expected to be removed, consistent with the initial uniform frequency distribution. In Scenarios 1 and 3, the introduced error tends to increase allele frequencies at rare loci (Table 1), reducing the number of loci falling below the threshold. Therefore, across all scenarios, the reduction of $\mathbf{M}$ due to filtering is at most 5%, and typically less in Scenarios 1 and 3 depending on the error rate ε.

Genetic evaluation

To estimate the h^2, we utilized the $\mathbf{G}$ matrix within a univariate mixed model implemented in the rrBLUP R package (Endelman 2011). The mixed model fitted by the restricted maximum likelihood (REML) is

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{a} + \mathbf{e} \quad (4)$$

where $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2\mathbf{G})$, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$, and σ_e^2 is the environmental variance. Specific $\mathbf{G}$ matrices were constructed to correspond to the respective errors introduced to the marker matrix $\mathbf{M}$. The estimate of h^2 was derived from the variance components as $\hat{\sigma}_a^2 / (\hat{\sigma}_a^2 + \hat{\sigma}_e^2)$. We then compared $\hat{h}^2$ with true (simulated) counterparts. To evaluate the effect of erroneous $\mathbf{M}$ on the ranking of individuals, we computed Spearman's rank correlation coefficient (ρ) between the true additive genetic values and their estimates.

Table 1. Incorporation of error into the matrix $\mathbf{M}$.

Error Scenario	Genotype	$\rightarrow A_1A_1$	$\rightarrow A_1A_2$	$\rightarrow A_2A_2$
1: Directional genotype perturbation	A_1A_1	0	2/3	1/3
	A_1A_2	1/3	1/3	1/3
	A_2A_2	1/3	2/3	0
2: HWE redraw	∀	$(1-p_i)^2$	$2p_i(1-p_i)$	p_i^2
3: Uniform redraw	∀	1/3	1/3	1/3

The table presents the probabilities of substituting specific elements of $\mathbf{M}$ under each of the three error scenarios (column 1). In Scenario 1, the probabilities depend on the existing genotypes, as listed in column 2. The error corresponds to redrawing the genotype uniformly at random from the full set $\{A_1A_1, A_1A_2, A_2A_2\}$. This definition allows the possibility that the resampled genotype coincides with the original one (e.g. $A_1A_2 \rightarrow A_1A_2$). In Scenario 2, substitution probabilities are based on the allelic frequency p_i in the ith locus. These probabilities are independent of the original genotype in $\mathbf{M}$. In Scenario 3, substitution probabilities are uniform across all elements of $\mathbf{M}$.

Results

Irrespective of the scenario, we observed a substantial bias introduced by the error in $\mathbf{M}$. As Fig. 1 illustrates, when no error is present ($\varepsilon = 0$ on the X-axes), there is a downward bias in h^2 estimates due to low SNP density, which is most pronounced at $h^2 = 0.6$. This bias diminishes with increasing marker density, becoming statistically insignificant ($\alpha = 0.05$) at a density of 50 SNPs per cM. Introducing the error ($\varepsilon > 0$ on the X-axes) results in a significant bias that follows a third-degree polynomial relationship with ε. Even at low error levels (2% – 5%), h^2 is consistently overestimated, particularly at high marker densities (≈ 50 markers/cM). The magnitude of this bias is strongly influenced by ε, SNP density, and h^2 levels.

In Scenario 1, $\hat{h}^2$ exhibits significant instability at a high marker density of 50 markers/cM, also following a cubic polynomial relationship with increasing ε. This instability contrasts with Scenarios 2 and 3 (Figs. 2 and 3), where bias remains minimal for $\varepsilon \leq 0.2$ and marker densities below 12.5 markers/cM. Scenario 1 demonstrates an upward bias that emerges more rapidly compared to Scenarios 2 and 3 and reaches the upper limit of h^2, followed by a sharp decline. Interestingly, Scenario 1 shows a subsequent upward trend in $\hat{h}^2$ values after they approach 0 at $\varepsilon = 0.8$, a pattern absent in Scenarios 2 and 3. The rank correlation coefficient ρ, reflecting the ranking of individuals based on additive genetic values, is also significantly impacted (Supplementary Figs. 1–3). In Scenario 1, ρ consistently declines and stabilizes at high marker densities. However, the decline in ρ is slower in Scenario 2 compared to Scenario 3, exhibiting distinct patterns of bias.

At the highest marker density under Scenario 1, the upward bias peaks at $\hat{h}^2 \approx 0.53$ for $\varepsilon = 0.4$, decreases to $\hat{h}^2 \approx 0.1$ at $\varepsilon = 0.8$, and then rises again to $\hat{h}^2 \approx 0.54$. A similar trend is observed at $h^2 = 0.6$. Notably, the differences between Scenario 2 and 3 are comparatively less pronounced, suggesting more similarity in the underlying factors driving their behavior (the error type). In contrast, Scenario 1 stands out with more pronounced biases and the rapid upward bias at high ε values. These findings highlight the substantial impact of even minor genotyping errors on the accuracy of $\hat{h}^2$, particularly under conditions characterized by high marker density.

To assess the robustness of our findings to founder population size, we performed an additional simulation under Scenario 2

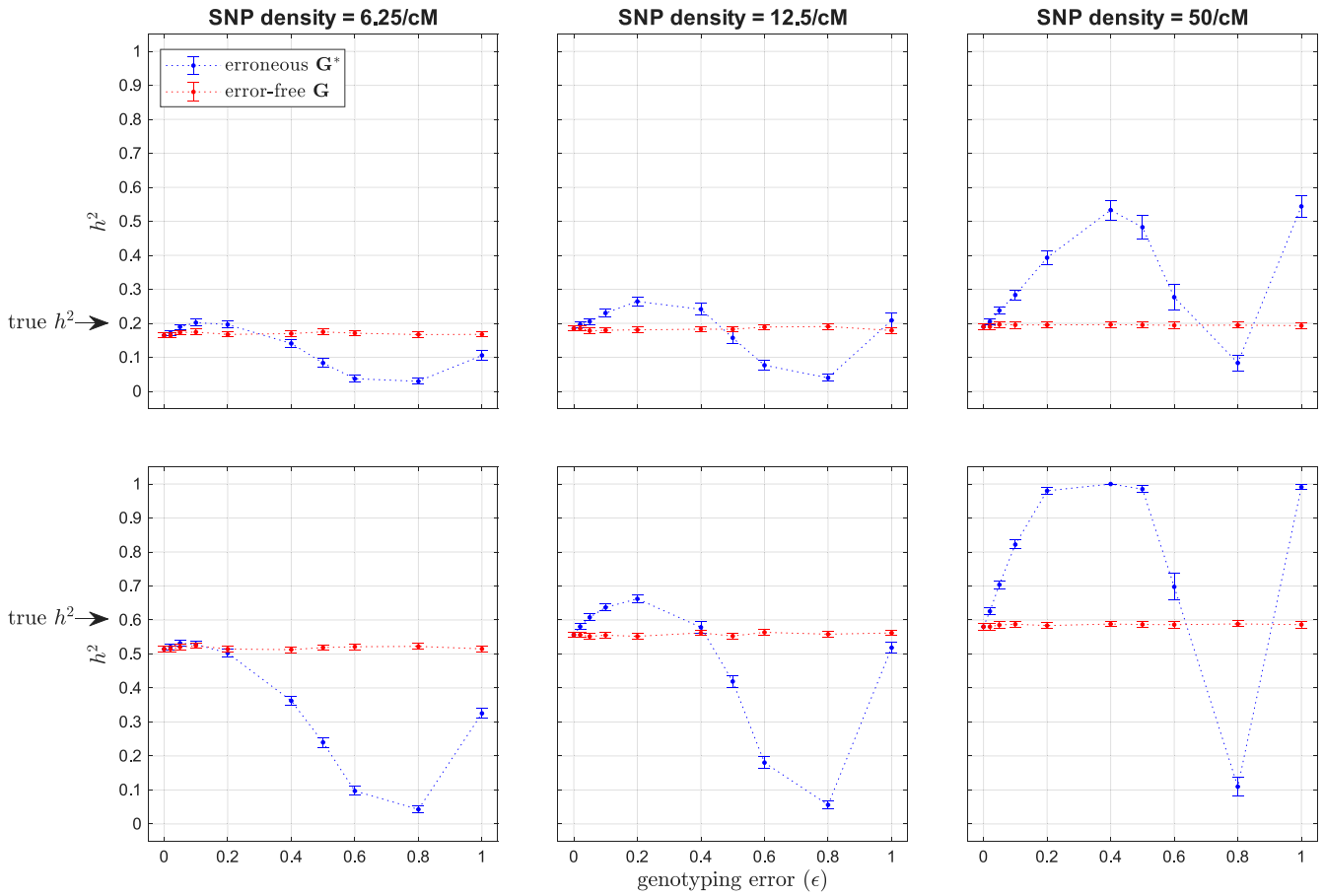

Fig. 1. Scenario 1 (substitution based on existing genotypes). The top row represents true $h^2 = 0.2$, and the bottom row true $h^2 = 0.6$. From left to right, graphs correspond to SNP densities of 6.25, 12.5, and 50 SNPs per cM. The X-axis indicates the percentage of modified genotypes (ε); the Y-axis shows h^2, with mean values across 100 replicates connected by dotted lines and 95% confidence intervals. Estimates assuming error-free **G** are represented by the lines parallel to the X-axis, while estimates based on erroneous **G*** are represented by the lines exhibiting a polynomial bias pattern.

with 1,000 founders. The bias pattern of h^2 as a function of error closely resembled that observed with 5,000 founders, consistent with mathematical expectations (Supplementary Fig. 4). To evaluate the role of genetic architecture, we repeated the analysis under Scenario 2 with 50 QTLs instead of 300. The resulting bias pattern was again consistent with the 300-QTL setting (Supplementary Fig. 5).

Regardless of the scenario, the $\hat{h}^2$ showed no statistically significant differences when **G** was calculated using method 2 (Equation 3).

Mathematical reasoning

Here, we attempt to interpret the bias pattern presented in Figs. 1–3. We begin by focusing on the results corresponding to the highest marker density tested, i.e. 50 markers per cM to isolate the explanation from the additional sources of bias associated with lower marker densities. We will address this additional bias in a later part of this section.

Recall that the elements of matrix **M**, prior to the introduction of error, represent the true genotypes for m biallelic loci across n individuals. Let $\mathbf{x} = (X_i \in -1, 0, 1)$ denote a general column vector of **M**, where $i = 1, \ldots, m$ refers to the marker loci.

Then $\mathbf{M}^* = (m^*_{ij})$ represents an erroneous alternative to **M**. Each element X^*_i is obtained by perturbing X_i. For each genotype entry, we introduce an error indicator $Z \sim \text{Bernoulli}(\varepsilon)$. If $Z = 0$, the entry is unchanged ($X^* = X$). If $Z = 1$, the entry is redrawn from the kernel in Table 1 (Scenario 1–3). Thus $X^* \in \{-1, 0, 1\}$ and $\varepsilon \in [0, 1]$ is the proportion of altered elements introduced earlier. Accordingly, using Equation (1), the matrix $\mathbf{G}^*$ is derived from $\mathbf{M}^*$.

Estimation of variance components using REML

The aim of the REML analysis is to optimize $\hat{\sigma}^2_a$ and $\hat{\sigma}^2_e$ to best match the model covariance matrix $\mathbf{V} = \sigma^2_a\mathbf{G} + \sigma^2_e\mathbf{I}$ with its empirical phenotypic covariance counterpart $\mathbf{S} = \mathbf{y}\mathbf{y}^\intercal$.

In REML analysis, two equations are solved: one for the total variance (mean diagonal element)

$$\frac{1}{n}\text{tr}(\mathbf{V}) = \frac{1}{n}\text{tr}(\mathbf{S}) \tag{5}$$

and one for the covariance (mean off-diagonal element):

$$\frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{V}_{ij} = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{S}_{ij} \tag{6}$$

(a derivation provided in Supplementary Material A).

Given the above structure of **V** and **S** and utilizing Equations (5) and (6), the following holds

$$\sigma^2_a\bar{d}^* + \sigma^2_e = C_1 \tag{7}$$

$$\sigma^2_a\bar{\delta}^*_{\text{off}} = C_2 \tag{8}$$

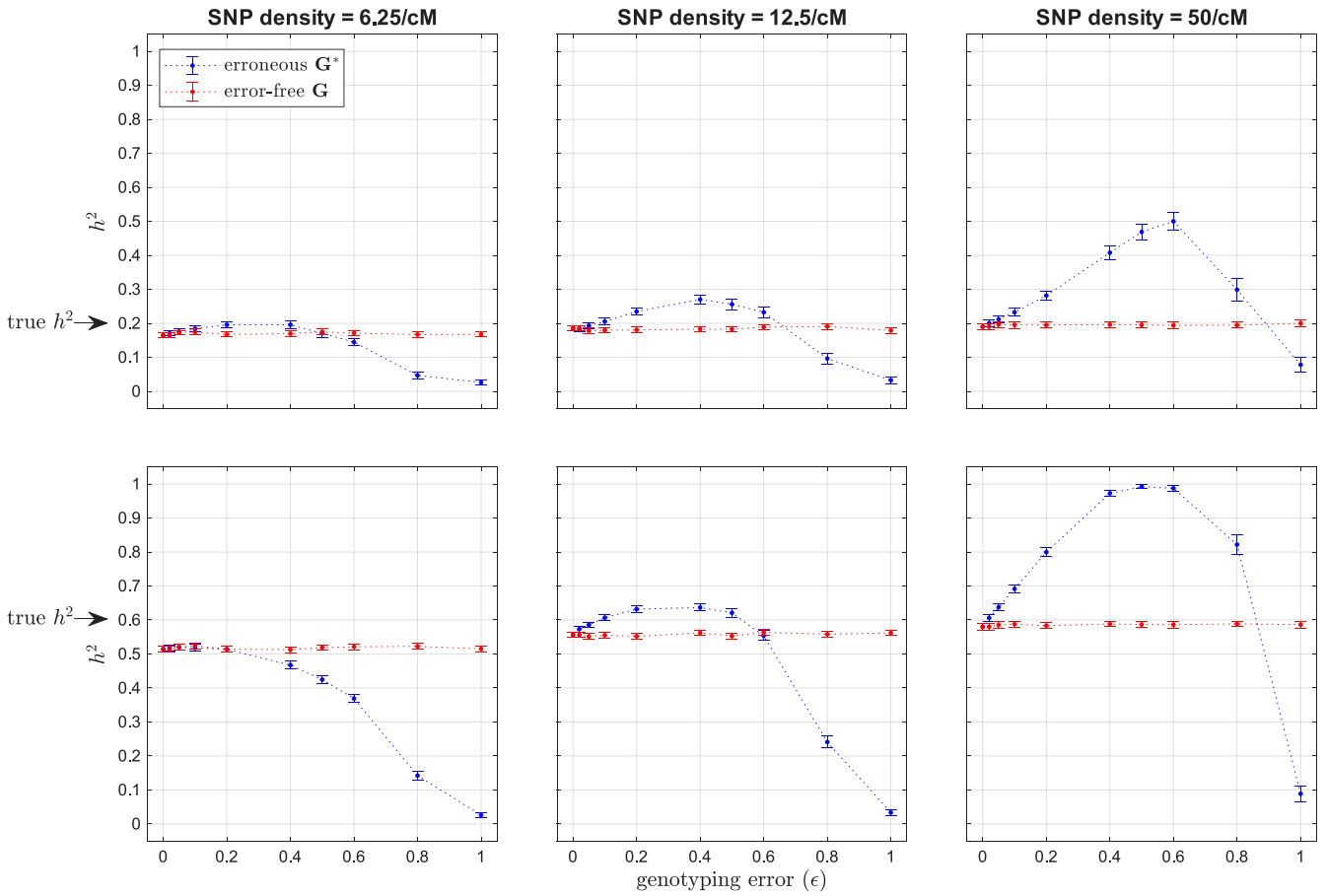

Fig. 2. Scenario 2 (substitution based on H-W frequencies). The top row represents true $h^2 = 0.2$, and the bottom row true $h^2 = 0.6$. From left to right, graphs correspond to SNP densities of 6.25, 12.5, and 50 SNPs per cM. The X-axis indicates the percentage of modified genotypes (ε); the Y-axis shows h^2, with mean values across 100 replicates connected by dotted lines and 95% confidence intervals. Estimates assuming error-free **G** are represented by the lines parallel to the X-axis, while estimates based on erroneous **G*** are represented by the lines exhibiting a polynomial bias pattern.

where $\bar{d}^* = \frac{1}{n}\mathrm{tr}(\mathbf{G}^*)$ (average diagonal element $\mathbf{G}^*$), $\bar{\delta}^*_{\mathrm{off}} = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{G}^*_{ij}$ (average off-diagonal element $\mathbf{G}^*$), and C_1, C_2 are constants derived from **S**. Errors in $\mathbf{G}^*$ influence $\bar{d}^*$ and $\bar{\delta}^*_{\mathrm{off}}$, which distort both $\hat{\sigma}^2_a$ and $\hat{\sigma}^2_e$ and consequently $\hat{h}^2$.

Impact of genotyping error

$\mathbf{G}^*$ is influenced by $\mathbf{M}^*$, following VanRaden's reasoning (Equation 1). Consequently, for all scenarios, three characteristics must be determined: expected value, variance, and covariance (Table 2). The expected value influences the centering of $\mathbf{G}^*$, variance impacts its diagonal elements, and covariance affects its off-diagonal elements.

Utilizing Equations (7) and (8), $\hat{h}^2$ can then be expressed as

$$\hat{h}^2(\varepsilon) = \frac{\hat{\sigma}^2_a(\varepsilon)}{\hat{\sigma}^2_a(\varepsilon) + \hat{\sigma}^2_e(\varepsilon)} = \frac{\frac{C_2}{\bar{\delta}^*_{\mathrm{off}}}}{\frac{C_2}{\bar{\delta}^*_{\mathrm{off}}} + \left(C_1 - \frac{C_2\bar{d}^*}{\bar{\delta}^*_{\mathrm{off}}}\right)} \tag{9}$$

Let us assume that the ratio of the average diagonal and off-diagonal elements is a function of the error (ε), such as

$$\kappa(\varepsilon) = \frac{\bar{d}^*}{\bar{\delta}^*_{\mathrm{off}}} \tag{10}$$

Next, we introduce $\gamma = \frac{C_1}{C_2}$, therefore

$$\hat{h}^2(\varepsilon) = \frac{\frac{1}{\bar{\delta}^*_{\mathrm{off}}}}{\frac{1}{\bar{\delta}^*_{\mathrm{off}}} + (\gamma - \kappa(\varepsilon))} = \frac{1}{\gamma\bar{\delta}^*_{\mathrm{off}} + (1 - \bar{d}^*)} \tag{11}$$

From the relationships above, it is clear that the effect of the error ε on $\hat{h}^2$ can be assessed by examining the changes in the mean values of the diagonal and off-diagonal elements of the matrix $\mathbf{G}^*(\varepsilon)$, particularly from their ratio. It follows from Equation (11) that for a sufficiently high marker density and small errors ($\varepsilon \ll 1$), where $\bar{d}^* \approx 1$ and $\gamma \approx$ const., the following holds

$$\hat{h}^2(\varepsilon) \approx \frac{1}{\gamma\bar{\delta}^*_{\mathrm{off}}} \tag{12}$$

From Table 2 and the derivation in Supplementary Material B, the off-diagonal elements of the matrix $\mathbf{G}^*(\varepsilon)$ can be approximated as

$$\bar{\delta}^*_{\mathrm{off}} \approx (1 - \varepsilon)^2\bar{\delta}_{\mathrm{off}} \tag{13}$$

It is clear that REML increasingly overestimates h^2 as the error ε grows, which is a key finding. Based on Table 2 and the derivation in Supplementary Material B, we conclude that when ε is not small

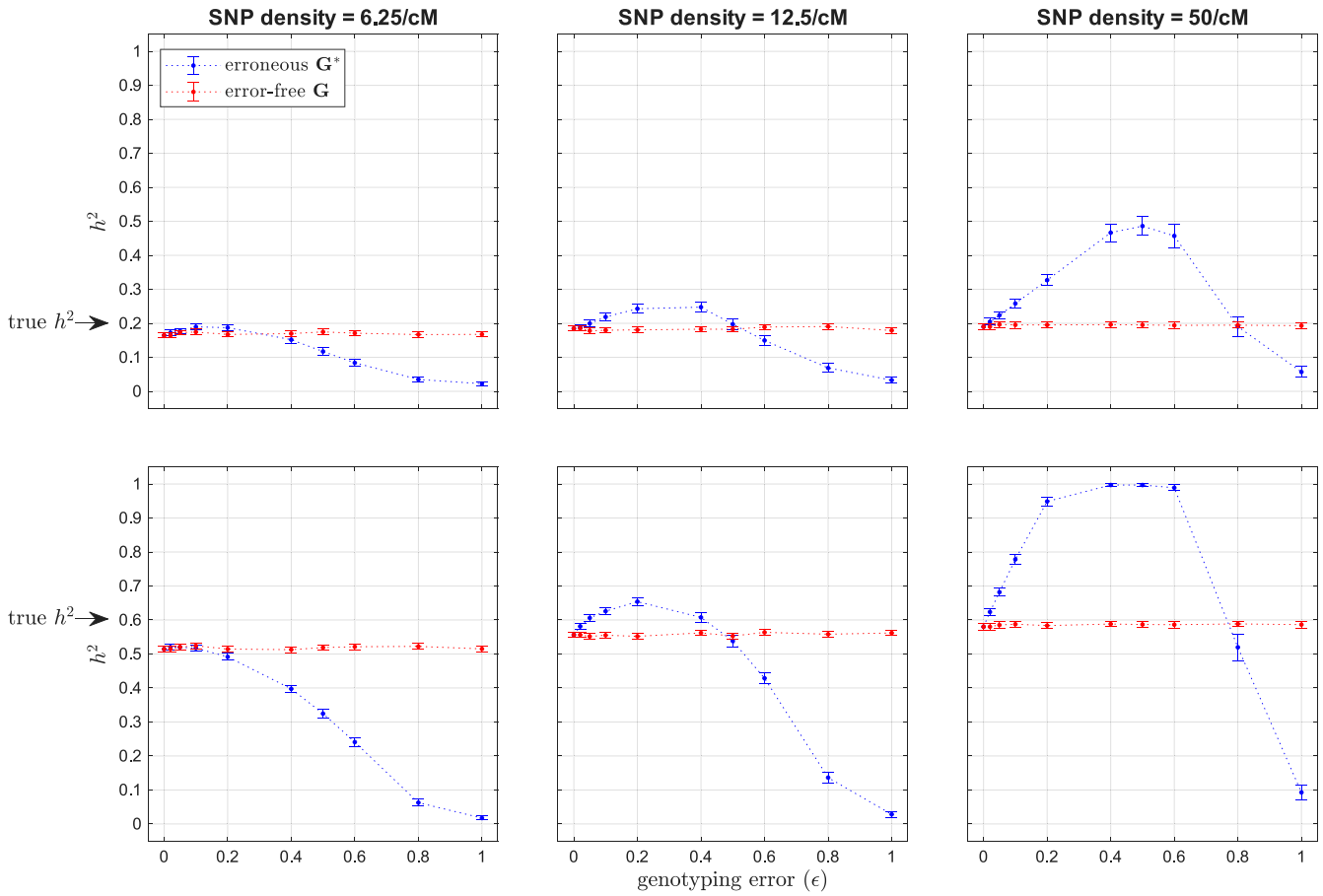

Fig. 3. Scenario 3 (uniform substitution probabilities). The top row represents true $h^2 = 0.2$, and the bottom row true $h^2 = 0.6$. From left to right, graphs correspond to SNP densities of 6.25, 12.5, and 50 SNPs per cM. The X-axis indicates the percentage of modified genotypes (ε); the Y-axis shows h^2, with mean values across 100 replicates connected by dotted lines and 95% confidence intervals. Estimates assuming error-free **G** are represented by the lines parallel to the X-axis, while estimates based on erroneous **G*** are represented by the lines exhibiting a polynomial bias pattern.

Table 2. Expected values, variances, and covariances of the three scenarios.

Scenario	$\mathbb{E}[X_i^*]$	$\mathbf{Var}[X_i^*]$	$\mathbf{Cov}[X_i^*, X_j^*]$
1	$2p_i - 1 + \frac{4\varepsilon}{3}(1 - 2p_i)$	$2p_i(1-p_i) + \varepsilon[\frac{2}{3} - 2p_i(1-p_i)] + \varepsilon(1-\varepsilon)(2p_i-1)^2$	$(1-\varepsilon)^2\mathrm{Cov}[X_i, X_j]$
2	$2p_i - 1$	$2p_i(1-p_i)$	$(1-\varepsilon)^2\mathrm{Cov}[X_i, X_j]$
3	$(2p_i - 1)(1-\varepsilon)$	$(1-\varepsilon)2p_i(1-p_i) + \frac{2}{3}\varepsilon + \varepsilon(1-\varepsilon)(2p_i-1)^2$	$(1-\varepsilon)^2\mathrm{Cov}[X_i, X_j]$

p_i is the allelic frequency, and ε is the proportion of altered elements in **M**.

the term $\kappa(\varepsilon)$ becomes relevant and further contributes to the inflation of h^2. We note that one of the conclusions in Supplementary Material B is the importance of considering marker density and incorporating it into the estimates, as shown in the approach described in the section Impact of marker density.

However, this increase cannot be unlimited. The reason lies in the solvability of the system of Equations (7) and (8). If σ_a^2 increases too much (due to a decrease in $\bar{\delta}^*_{\text{off}}$), then σ_e^2 would become negative, which is inadmissible. REML, as an optimization method, therefore constrains σ_a^2 to ensure that $\sigma_e^2 \geq 0$. This explains why h^2 estimates in all scenarios reach their maximum at $\varepsilon = 0.5$ at the latest.

With further increases in ε, a collapse of covariance occurs in all scenarios ($\bar{\delta}^*_{\text{off}} \to 0$). Information about the relatedness of individuals disappears, and $h^2 \to 0$. At $\varepsilon = 1$, a singular solution to the system of Equations (7) and (8) arises, where $\hat{h}^2 = 0$ or $\hat{h}^2 = 1$.

Impact of marker density

Denser marker coverage across all scenarios (Figs. 1–3) increases the sensitivity of $\hat{h}^2$ to ε. In particular, the initial rise of $\hat{h}^2(\varepsilon)$ is not as profound under low marker density. Thus, the density of marker coverage partly compensates for the distortion caused by introducing the error. This effect can be explained by the interaction between marker density (m) and genotyping errors (ε). Let us assume that

$$\hat{g}_{ij}(m) = g_{ij} \cdot \lambda(m) + \eta_{ij} \tag{14}$$

where

- g_{ij} is true genetic relationship between individuals i, j,
- $\hat{g}_{ij}(m)$ is the respective estimated genetic relationship utilizing m marker loci,

- $\lambda(m)$ captures marker efficiency ($0 < \lambda(m) < 1$ and $\lim_{m\to\infty}\lambda(m) = 1$), quantifying how well a given number of markers capture true genetic relationships. It ranges from 0 (no markers) to 1 (infinitely many markers) and reflects marker quantity rather than genotyping error,
- η_{ij} represents noise, $\eta_{ij} \sim \mathcal{N}(0, \sigma_\eta^2)$.

The expected mean value of the off-diagonal elements of matrix **G** can then be expressed as

$$\mathbb{E}[\delta_{\text{off}}(m)] = \lambda(m) \cdot \bar{\delta}_{\text{off}}(\infty) \tag{15}$$

Marker density (m) together with ε (using Equation 13) has the following influence on the ratio $\kappa(\varepsilon, m)$

$$\kappa(\varepsilon, m) = \frac{\bar{d}^*}{(1-\varepsilon)^2 \cdot \lambda(m) \cdot \bar{\delta}_{\text{off}}(\infty)} \tag{16}$$

Equations (15) and (16) show clearly that the smaller $\lambda(m)$, meaning the lower the efficiency of marker coverage, the more quickly the mean value of off-diagonal elements of $\mathbf{G}^*$ will decrease, while the ratio $\kappa(\varepsilon, m)$ will increase. The effect described above will arise even sooner because REML limits $\hat{\sigma}_a^2$ so that $\hat{\sigma}_e^2 \geq 0$. As a result, any increase in $\hat{h}^2$ is restricted or even eliminated by the rapid decline in covariance. Even when $\varepsilon = 0$ and the diagonal remains stable ($\bar{d} \approx 1$) we can write

$$\kappa(0, m) = \frac{\bar{d}}{\bar{\delta}_{\text{off}}(m)} = \frac{1}{\lambda(m) \cdot \bar{\delta}_{\text{off}}(\infty)} > \frac{1}{\bar{\delta}_{\text{off}}(\infty)} = \kappa(0, \infty) \tag{17}$$

REML then compensates for the higher κ by overestimating σ_e^2, which leads to the underestimation of h^2 corresponding to the **G** matrix without noise; see Figs. 1–3.

We can therefore conclude that marker panels with low density are more robust to errors ($\varepsilon < 0.1$), yet they systematically underestimate h^2 because $\lambda(m) < 1$. Marker panels with high-density deliver an accurate estimate of h^2 when $\varepsilon = 0$, but even small errors ($\varepsilon > 0.01$) lead to an overestimation of h^2.

Discussion

Consistent with earlier studies, our simulations show that SNP marker density strongly affects genomic prediction accuracy because dense panels tag causal variants more effectively and yield more reliable genomic relationship estimates (Lee et al. 2011; Wang J 2018; Dufflocq et al. 2019; Akbarpour et al. 2021; Khalilisamani et al. 2021). Our novel finding, focusing exclusively on h^2, is a pronounced nonlinear bias in h^2 when **G** is computed from erroneous SNP marker data; this distortion increases with marker density. Even low error rates can therefore skew h^2 estimates.

The three stylized error kernels adopted in our simulations were chosen to bracket the main classes of mistakes observed in real SNP datasets. (i) Directional genotype perturbation preferentially replaces the resident homozygote with the alternative allele. It therefore mirrors chemistry-specific base-calling errors and allele dropout that favor transitions over transversions, as well as cluster mis-centering on array platforms. (ii) HWE redraw preserves the locus-specific allele frequency while breaking genotype co-segregation, emulating imputation or phasing tools that redraw uncertain calls from population priors and quality control (QC) scripts that "heal" Mendelian inconsistencies by frequency-based relabeling. (iii) Uniform redraw assigns each genotype class with equal probability, representing nonsystematic mistakes such as barcode cross-talk, sample swaps, or low-level contamination that randomize genotype calls irrespective of allele frequency. Modeling these three kernels allows us to disentangle the effects of directional miscoding, frequency-preserving uncertainty, and purely stochastic noise, thereby providing lower- and upper-bound benchmarks for the distortions likely to arise in sequencing or array data.

In plain terms, the algebra shows that REML balances two summaries of the genomic relationship matrix **G**: the mean of its diagonal and off-diagonal elements. When a fraction ε of genotype calls is wrong, the diagonals are minimally affected, but every error diminishes the off-diagonals by roughly $(1-\varepsilon)^2$. REML "rescues" this lost covariance by inflating the additive variance estimate, and because the shrinkage is multiplicative, even a small error ($\varepsilon \approx 2-5\%$) can drive $\hat{h}^2$ noticeably upward, especially with dense SNP panels where the diagonals are already close to their maximum. Inflation peaks around $\varepsilon = 0.3$ to 0.5; beyond that point, the relatedness information collapses and REML must shrink the additive term, forcing $\hat{h}^2$ toward zero. This ratio effect explains the response curves in Figs. 1–3, the sharper bias under the directional genotype perturbation (where diagonals rise while off-diagonals fall), and the contrasting downward bias in sparse panels, where causal variants are under-tagged regardless of error. Low-density marker panels are robust to errors below 0.1 but systematically underestimate h^2, whereas high-density panels estimate h^2 accurately only when error is zero and begin to overestimate it at error levels as low as $\varepsilon \approx 0.01$.

Mitigating genotyping error in SNP data requires a truly end-to-end strategy that spans wet-lab practice, sequencing chemistry, and downstream bioinformatics. At the experimental stage, limiting freeze–thaw cycles, enforcing clean-bench workflows during DNA extraction, and quantifying nucleic acids with high-precision fluorometric assays (e.g. Qubit) minimize degradation and concentration bias. Library preparation artefacts can be reduced by employing high-fidelity polymerases and thermocycling protocols validated for even coverage (Aird et al. 2011; Laehnemann et al. 2016). Once sequence data are generated, pangenome or population-specific references have proven effective at rescuing reads that deviate markedly from a single linear reference, thereby improving variant recall and genotype accuracy (Clevenger et al. 2015). Routine quality control filters, including replicate concordance, scans for deviations from HWE, trio-based Mendelian checks, and manual inspection of genotype clusters, remain indispensable, especially for large cohorts (Bonin et al. 2004; Hosking et al. 2004; Guo et al. 2014).

Long-read platforms such as Oxford Nanopore and PacBio HiFi now approach Illumina-like single-base accuracy when coupled with short-read polishing, enabling orthogonal confirmation of complex variants (Wenger et al. 2019). In particular, recent advances in PacBio HiFi technology, which generates highly accurate (>99.9%) circular consensus sequencing (CCS) reads by repeatedly sequencing the same molecule, have reduced the need for short-read polishing altogether. HiFi reads combine the benefits of long reads, such as resolving structural variants, spanning repetitive regions, and enabling haplotype phasing, with the high accuracy previously restricted to short-read platforms, thereby supporting de novo assembly, variant calling, and epigenetic profiling without hybrid data. Machine-learning variant callers (DeepVariant, Clair, PEPPER–Margin–DeepVariant) further boost single nucleotide and indel precision in repetitive or GC-rich regions (Poplin et al. 2018; Shafin et al. 2021). Graph-based reference genomes

and unique molecular identifiers (UMIs) add complementary layers of error suppression by lowering mapping bias and permitting molecule-level consensus calling (Paten et al. 2017; Rakocevic et al. 2019). Finally, next-generation SNP arrays that incorporate rare alleles and multi-ethnic tag variants improve cross-population imputation accuracy, mitigating the minor allele biases (Bycroft et al. 2018). Taken together, these laboratory and computational advances, when combined with strong QC pipelines, can help lower the latent error rate. At the same time, it is important to note that not all variant classes benefit equally from these improvements. Low-pass sequencing SNPs and structural variants often have higher calling error, and our findings suggest that such errors could contribute to inflated h^2 estimates. This may explain observations where including structural variants, despite their lower calling accuracy, led to higher h^2 estimates in prediction models (Derks et al. 2025).

We modeled a single additive trait in an ideal, randomly mating population to gain clearer mathematical insight into how genotype miscoding alone distorts the **G** matrix and the h^2 estimates. Other factors worth exploring include nonrandom mating, nonadditive genetic effects, maternal effects, genotype-by-environment covariance and interaction, sexual dimorphism, and selection (Lynch and Walsh 1998). While our three kernels span directional, frequency-preserving, and purely random miscoding, they do not capture block-level or allele-specific dropout errors common in low-coverage sequencing and copy number variation (CNV)-rich regions; such errors could exacerbate the distortions we report and merit separate investigation. For the current dataset size, we employed the rrBLUP R package (Endelman 2011), which implements a standard REML-based variance component estimation approach (Kang et al. 2008). Comparable REML-based estimates would be expected from other widely used tools such as GCTA (Yang et al. 2011) or ASReml (Butler et al. 2023). Our simulations used 1,000 individuals and up to 20,000 SNPs, a scale chosen deliberately to balance clarity and computational feasibility. At 50 SNPs per cM with 1,000 offspring, the downward bias observed at lower marker densities disappears, yielding unbiased estimates in the absence of error. This reference point allowed us to isolate the effect of erroneous markers. Scaling up to larger populations or denser marker panels would not change these qualitative patterns, as demonstrated by the accompanying mathematical derivations.

In summary, our results show that even a residual two percent miscoding rate can significantly bias h^2; upcoming marker-based analyzes therefore need to move beyond the assumption of error-free genotypes. Analytical remedies, which merit further research, include (i) robust construction of the **G** matrix, for example by winsorizing extreme off-diagonal elements and scaling diagonals by their median; (ii) likelihood-based analyzes that weight loci by genotype certainty or embed a Bayesian prior for miscoding; (iii) approaches that bypass **G** altogether, such as LD adjusted REML or regressions with Huber or other M-estimators; and (iv) simulation-guided correction that maps bias as a function of marker density and error, coupled with explicit error imputation.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables. The simulation code is available at https://github.com/mlstiburek/genotyping_error_heritability.

Supplemental material available at GENETICS online.

Acknowledgment

We thank the editor and reviewers for their insightful comments, which substantially improved the manuscript.

Funding

This project was supported by the Faculty of Forestry and Wood Sciences at the Czech University of Life Sciences Prague, internal grant (IGA No 1312).

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Abecasis GR, Cherny SS, Cardon LR. 2001. The impact of genotyping error on family-based analysis of quantitative traits. Eur J Hum Genet. 9:130–134. https://doi.org/10.1038/sj.ejhg.5200594.

Aird D *et al.* 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12:1–14. https://doi.org/10.1186/gb-2011-12-2-r18.

Akbarpour T, Ghavi Hossein-Zadeh N, Shadparvar AA. 2021. Marker genotyping error effects on genomic predictions under different genetic architectures. Mol Genet Genomics. 296:79–89. https://doi.org/10.1007/s00438-020-01728-z.

Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of conservation genetics. Nat Rev Genet. 11:697–709. https://doi.org/10.1038/nrg2844.

Bernardo R, Yu J. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090. https://doi.org/10.2135/cropsci2006.11.0690.

Bonin A *et al.* 2004. How to track and assess genotyping errors in population genetics studies. Mol Ecol. 13:3261–3273. https://doi.org/10.1111/mec.2004.13.issue-11.

Brandt DY *et al.* 2015. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. G3–Genes Genom Genet. 5:931–941. https://doi.org/10.1534/g3.114.015784.

Browning BL, Browning SR. 2016. Genotype imputation with millions of reference samples. Am J Hum Genet. 98:116–126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Butler D, Cullis B, Gilmour A, Gogel B, Thompson R. 2023. ASReml-R Reference Manual Version 4.2. VSN International Ltd. Hemel Hempstead, HP2 4TP, UK.

Bycroft C *et al.* 2018. The UK Biobank resource with deep phenotyping and genomic data. Nature. 562:203–209. https://doi.org/10.1038/s41586-018-0579-z.

Cheung CY, Thompson EA, Wijsman EM. 2014. Detection of Mendelian consistent genotyping errors in pedigrees. Genet Epidemiol. 38:291–299. https://doi.org/10.1002/gepi.2014.38.issue-4.

Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. 2015. Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. Mol Plant. 8:831–846. https://doi.org/10.1016/j.molp.2015.02.002.

Derks M, Pook T, Chen J, Hawken R, Bouwman A. 2025. Utilizing low-pass sequence data to study the impact of structural variants on polygenic traits. Preprint. Research Square.

Dufflocq P, Pérez-Enciso M, Lhorente JP, Yáñez JM. 2019. Accuracy of genomic predictions using different imputation error rates in aquaculture breeding programs: a simulation study. Aquaculture. 503:225–230. https://doi.org/10.1016/j.aquaculture.2018.12.061.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome–US. 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024.

Forneris NS *et al.* 2015. Quality control of genotypes using heritability estimates of gene content at the marker. Genetics. 199:675–681. https://doi.org/10.1534/genetics.114.173559.

Gamal El-Dien O *et al.* 2016. Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. G3–Genes Genom Genet. 6:743–753. https://doi.org/10.1534/g3.115.025957.

Gezan SA, de Oliveira AA, Galli G, Murray D. 2022. User's manual for ASRgenomics v. 1.1. 0: an R package with complementary genomic functions. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Gienapp P *et al.* 2017. Genomic quantitative genetics to study evolution in the wild. Trends Ecol Evol. 32:897–908. https://doi.org/10.1016/j.tree.2017.09.004.

Goddard ME. 2009. View to the future: could genomic evaluation become the standard?. Interbull Bull. 39:83–88.

Goddard ME, Hayes BJ, Meuwissen TH. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 128:409–421. https://doi.org/10.1111/jbg.2011.128.issue-6.

Goldstein DR, Zhao H, Speed TP. 1997. The effects of genotyping errors and interference on estimation of genetic distance. Hum Hered. 47:86–100. https://doi.org/10.1159/000154396.

Gordon D, Heath S, Ott J. 1999. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Hum Hered. 49:65–70. https://doi.org/10.1159/000022846.

Guo Y *et al.* 2014. Illumina human exome genotyping array clustering and quality control. Nat Protoc. 9:2643–2662. https://doi.org/10.1038/nprot.2014.174.

Henderson CR. 1953. Estimation of variance and covariance components. Biometrics. 9:226–252. https://doi.org/10.2307/3001853.

Hill WG, Weir BS. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res. 93:47–64. https://doi.org/10.1017/S0016672310000480.

Hinrichs AL, Suarez BK. 2005. Genotyping errors, pedigree errors, and missing data. Genet Epidemiol. 29:S120–S124. https://doi.org/10.1002/(ISSN)1098-2272.

Hosking L *et al.* 2004. Detection of genotyping errors by Hardy–Weinberg equilibrium testing. Eur J Hum Genet. 12:395–399. https://doi.org/10.1038/sj.ejhg.5201164.

Kang HM *et al.* 2008. Efficient control of population structure in model organism association mapping. Genetics. 178:1709–1723. https://doi.org/10.1534/genetics.107.080101.

Khalilisamani N, Thomson P, Raadsma H, Khatkar M. 2021. Impact of genotypic errors with equal and unequal family contribution on accuracy of genomic prediction in aquaculture using simulation. Sci Rep–UK. 11:18318. https://doi.org/10.1038/s41598-021-97873-5.

Laehnemann D, Borkhardt A, McHardy AC. 2016. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform. 17:154–179. https://doi.org/10.1093/bib/bbv029.

Lee SH, Goddard ME, Visscher PM, Van Der Werf JH. 2010. Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genet Sel Evol. 42:1–14. https://doi.org/10.1186/1297-9686-42-22.

Lee SH, Wray NR, Goddard ME, Visscher PM. 2011. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 88:294–305. https://doi.org/10.1016/j.ajhg.2011.02.002.

Legarra A, Aguilar I, Misztal I. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 92:4656–4663. https://doi.org/10.3168/jds.2009-2061.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 30:2843–2851. https://doi.org/10.1093/bioinformatics/btu356.

Lynch M, Walsh B. 1998. Genetics and analysis of quantitative traits. Vol. 1. Sinauer Sunderland.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Misztal I, Lourenco D, Legarra A. 2020. Current status of genomic evaluation. J Anim Sci. 98:skaa101. https://doi.org/10.1093/jas/skaa101.

Munoz PR *et al.* 2014. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. Crop Sci. 54:1115–1123. https://doi.org/10.2135/cropsci2012.12.0673.

Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. Genome Res. 27:665–676. https://doi.org/10.1101/gr.214155.116.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nat Rev Genet. 6:847–859. https://doi.org/10.1038/nrg1707.

Pook T *et al.* 2019. Haploblocker: creation of subgroup-specific haplotype blocks and libraries. Genetics. 212:1045–1061. https://doi.org/10.1534/genetics.119.302283.

Pook T, Schlather M, Simianer H. 2020. MoBPS-modular breeding program simulator. G3–Genes Genom Genet. 10:1915–1918. https://doi.org/10.1534/g3.120.401193.

Poplin R *et al.* 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 36:983–987. https://doi.org/10.1038/nbt.4235.

Rakocevic G *et al.* 2019. Fast and accurate genomic analyses using genome graphs. Nat Genet. 51:354–362. https://doi.org/10.1038/s41588-018-0316-4.

Rang FJ, Kloosterman WP, de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19:90. https://doi.org/10.1186/s13059-018-1462-9.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Schlather M. 2020. Efficient calculation of the genomic relationship matrix. bioRxiv. pp. 1–7. https://doi.org/10.1101/2020.01.12.903146, preprint: not peer reviewed.

Seaman S, Holmans P. 2005. Effect of genotyping error on type-i error rate of affected sib pair studies with genotyped parents. Hum Hered. 59:157–164. https://doi.org/10.1159/000085939.

Shafin K *et al.* 2021. Haplotype-aware variant calling with PEPPER–Margin–DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 18:1322–1332. https://doi.org/10.1038/s41592-021-01299-w.

Sobel E, Papp JC, Lange K. 2002. Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet. 70:496–508. https://doi.org/10.1086/338920.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Varma A, Padh H, Shrivastava N. 2007. Plant genomic DNA isolation: an art or a science. Biotechnol J. 2:386–392. https://doi.org/10.1002/biot.v2:3.

Wang J. 2018. Estimating genotyping errors from genotype and reconstructed pedigree data. Methods Ecol Evol. 9:109–120. https://doi.org/10.1111/mee3.2018.9.issue-1.

Wang X *et al.* 2019. Improving genomic predictions by correction of genotypes from genotyping by sequencing in livestock populations. J Anim Sci Biotechnol. 10:8. https://doi.org/10.1186/s40104-019-0315-z.

Wenger AM *et al.* 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 37:1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

Wiggans G *et al.* 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J Dairy Sci. 92:3431–3436. https://doi.org/10.3168/jds.2008-1758.

Wiggans G, VanRaden P, Cooper T. 2011. The genomic evaluation system in the United States: Past, present, future. J Dairy Sci. 94:3202–3211. https://doi.org/10.3168/jds.2010-3866.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F. 2013. Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. G3–Genes Genom Genet. 3:909–916. https://doi.org/10.1534/g3.113.005975.

Editor: C. Yang