mathematics

MDPI

Article

Reliable Genetic Correlation Estimation via Multiple Sample Splitting and Smoothing

The Tien Mai

Department of Mathematical Sciences, Norwegian University of Science and Technology,
7034 Trondheim, Norway; the.t.mai@ntnu.no

Abstract: In this paper, we aim to investigate the problem of estimating the genetic correlation between two traits. Instead of making assumptions about the distribution of effect sizes of the genetic factors, we propose the use of a high-dimensional linear model to relate a trait to genetic factors. To estimate the genetic correlation, we develop a generic strategy that combines the use of sparse penalization methods and multiple sample splitting approaches. The final estimate is determined by taking the median of the calculations, resulting in a smoothed and reliable estimate. Through simulations, we demonstrate that our proposed approach is reliable and accurate in comparison to naive plug-in methods. To further illustrate the advantages of our method, we apply it to a real-world example of a bacterial GWAS dataset, specifically to estimate the genetic correlation between antibiotic resistant traits in *Streptococus pneumoniae*. This application not only validates the effectiveness of our method but also highlights its potential in real-world applications.

Keywords: antimicrobial resistance; genetic correlation; sparse regression; multiple sample splitting

MSC: 62J07; 62H20; 62F35

check for updates

Citation: Mai, T.T. Reliable Genetic Correlation Estimation via Multiple Sample Splitting and Smoothing. *Mathematics* **2023**, *11*, 2163. https://doi.org/10.3390/math11092163

Academic Editors: Jordi Martorell-Marugán and Pedro Carmona-Sáez

Received: 20 March 2023
Revised: 23 April 2023
Accepted: 3 May 2023
Published: 4 May 2023

1. Introduction

Genome-wide association studies (GWAS) have been instrumental in demonstrating that various complex traits may be influenced by common genetic variants, as exemplified in studies such as [1–3]. Recently, there has been a growing interest in understanding the relationship of common genetic variants across pairs of traits, as seen in, for example, [4–8]. This understanding can have a wide range of benefits, such as in epidemiological and etiological studies [9,10], as well as in genetic risk prediction [11,12]. By quantifying the genetic variants across pairs of traits, it can provide useful insights in understanding the underlying genetic mechanisms of the complex traits and can be applied in various fields of genetics.

In the field of genetics, the concept of heritability of a trait has been widely studied and is considered a key quantitative measurement [13,14]. Building upon the heritability notation, the concept of genetic correlation between two phenotypes has been defined in order to capture the correlation between causal loci [4,6,15]. Essentially, this metric is used to measure the extent to which genetic variants across the genome contribute to the correlation between two phenotypes. Understanding the genetic correlation between different traits is informative as it can provide insights into the underlying polygenic genetic architecture of complex traits. It can also be used to identify genetic factors that contribute to the correlation between two traits and to help in understanding the underlying genetic mechanisms. Additionally, genetic correlation can also be used to suggest a role for pleiotropic genetic effects [16], which is when a gene has multiple effects on different traits. Furthermore, the estimation of genetic correlation is an important step in many genetic studies, such as genome-wide association studies (GWAS), as it can help to identify genetic loci that are associated with multiple traits, which can be useful in the study of complex diseases. Moreover, genetic correlation can also be used to identify genetic loci that are associated with the same trait but in different populations. This can provide insights into the underlying genetic architecture of the trait and can help to identify genetic loci that are associated with the trait in different populations.

Traditional approaches to estimating genetic correlation are based on exploring a linear mixed-effect model framework, in which the effects of genetic variants are assumed to be random (usually assumed to be normally distributed with 0-mean). This is opposite to the fixed effect assumption embedded in the framework of quantitative genetics theory [13,17–21]. Two popular approaches to making inferences about genetic correlation in a linear mixed model are the following: maximum likelihood estimation [22–24], moment method [7,25] and linkage disequilibrium score regression [4,26]. A comprehensive comparison and discussion of these approaches in the context of the random-effect model can be found in a very recent review by [15,27].

In this paper, we focused on utilizing a high-dimensional linear regression model to understand the relationship between phenotype and genotype in genetic studies, particularly in the context of genome-wide association studies (GWAS). One of the key advantages of this approach is that it does not rely on any assumptions about the distribution of effect sizes. This is a particularly natural model for investigating the entire genome in GWAS, and has been shown to be more efficient than traditional univariate methods in GWAS. The use of this model has been previously demonstrated in studies, such as [13,17,28–30], as a means to improve the understanding of the genetic basis of complex traits. Furthermore, the use of a high-dimensional linear regression model allows for the analysis of a large number of genetic variations at once, which can help to identify the most important genetic factors influencing the trait of interest. In this way, we aim to make a contribution to the field by highlighting the benefits of using this model over the traditional univariate approach in GWAS.

Built upon recent advances in machine learning approaches for statistical genetics, we propose an aggregation approach based on multiple sample splittings [31–33]. The main ingredient in our approach is the selective inference framework [34–38]. More specifically, our approach is a two-stage strategy, which involves, first, splitting samples into two parts and performing variable selection, via a sparse regularization (such as Lasso), and, then, applying partial regression in the second part to provide valid inferences under the selected model. This is to ensure that the different latent structures possibly residing in the sample are properly taken into account in both the selection and estimation steps. We obtain a numerically stable smoothed estimate by taking the median of the estimates over multiple random splits.

The idea of "splitting and smoothing" different estimates to yield an estimate with improved statistical properties is the central feature of the generic boosting approach widely used in machine learning, such as AdaBoost [39]. It is noted that the estimation based on a single split is highly unstable and, thus, difficulties arise in separating true signals from noises. This has been observed when using a single tree in the bagging algorithm [40]. In order to reduce this variability, we propose a multi-sample splitting scheme, in which the data is randomly split multiple times and the estimation procedure repeated accordingly. The final estimation is obtained via taking the median of the resulting estimates to obtain the smoothed estimate. The multiple sample splitting approach has previously been proposed in the statistics community, such as in [36,41], and was successfully used in GWAS in [42–44].

In order to evaluate the effectiveness of our proposed methods, numerical simulations were conducted. These simulations allowed us to assess the performance of our approach under various conditions and provided valuable insights into its capabilities. Additionally, to further illustrate the utility of our framework, we applied our procedure to a specific case study involving bacterial GWAS. Specifically, we used our methods to estimate the genetic correlations of antibiotic resistant phenotypes in bacteria. This application is particularly noteworthy as, while there has been a considerable amount of research on estimating genetic correlation in human GWAS, the topic has received relatively little attention in the context of bacteria. The results of our case study not only demonstrate the applicability of our methods to this understudied area but also provide valuable insights into the genetic factors underlying antibiotic resistance in bacteria. Overall, the numerical simulations and case study serve to both validate our methods and showcase their potential in real-world applications.

The structure of the paper is as follows. In Section 2, we formally introduce our model and provide a clear and precise definition for the concept of genetic correlation, which is the focus of our study. This section serves as a foundation for the rest of the paper and sets the stage for the subsequent sections. In Section 3, we delve deeper into the proposed method and provide a detailed explanation of the steps involved in the analysis. This section is of particular importance as it presents the core of our work and the key contributions of the paper. The methods are explained in a clear and easy-to-follow manner and are supplemented with relevant mathematical derivations, where appropriate. To demonstrate the effectiveness and practical utility of our method, we conduct numerical studies with simulations in Section 4 and a real data application in Section 5. These sections provide a thorough evaluation of our method and serves to validate its performance under various scenarios. Finally, in Section 6 we provide a discussion of the results, highlighting the key findings of our study and their implications. We also provide a conclusion that summarizes the main contributions of our paper and highlights the potential future directions of research in this area.

2. Model

We study the problem of genetic correlation estimation based on individual-level GWAS data. More specifically, we observe two traits y and z of n samples and a genotype matrix X of size $n \times p$ (often SNPs). Each trait is modeled as a linear combination of p genetic variants $X_{.j}$ and an error term (environmental and unmeasured genetic effects)

$$\begin{aligned} y &= X\beta_y + \varepsilon, \\ z &= X\beta_z + \gamma, \end{aligned} \tag{1}$$

where β_y, β_z are vectors of SNP effect sizes of length p. We assume that $X_{i.}$ are i.i.d random vector with 0-mean and covariance matrix Σ and that the random noises ε_i, γ_i are independent of X with 0-mean and the variances $\sigma_\varepsilon^2, \sigma_\gamma^2$.

In this study, a key departure from the typical approach in the field is the lack of assumptions regarding the distribution of effects β_y and β_z. Most research in this area, particularly within the framework of linear mixed models, relies on such assumptions to inform the analysis. However, by forgoing these assumptions, our work allows for a more flexible and inclusive approach to understanding the underlying relationships. This unique perspective may provide new insights and better understanding of the data. Additionally, this method can be useful in the case of non-normally distributed data and can be less restrictive. This can be especially beneficial in cases where the distribution of the effects may not be known or may be difficult to accurately estimate. Overall, this approach provides a fresh perspective on the data, and can yield new insights into the underlying relationships being studied.

2.1. *Genetic Correlation*

From the problem formulation (1), the covariance between the phenotypes y and z can be explained as a summation of the genetic covariance and environmental covariance as follows:

$$\begin{aligned}\mathrm{Cov}(y,z) &= \mathbb{E}(yz) - \mathbb{E}(y)\mathbb{E}(z) \\ &= \mathbb{E}\langle X\beta_y + \varepsilon, X\beta_z + \gamma\rangle \\ &= \beta_z^\top \mathbb{E}(X^\top X)\beta_y + \mathbb{E}(\varepsilon\gamma) \\ &= \beta_z^\top \Sigma\beta_y + \mathrm{Cov}(\varepsilon,\gamma).\end{aligned}$$

As a consequence, the *genetic correlation* between two traits y and z is defined by

$$\mathrm{g.Cor}(y,z) = \beta_z^\top \Sigma\beta_y. \tag{2}$$

This can then be further normalized to obtain the genetic correlation scaling between -1 and 1; where -1 indicates perfect negative correlation and 1 shows perfect positive correlation. More precisely, we define

$$\mathrm{sg.Cor}(y,z) = \frac{\beta_z^\top \Sigma\beta_y}{\sqrt{\beta_z^\top \Sigma\beta_z}\sqrt{\beta_y^\top \Sigma\beta_y}}, \tag{3}$$

which provides a standardized genetic correlation scale between -1 and 1 and, thus, can be used to compare across pairs of traits and facilitates comparisons across genomic regions.

2.2. Plug-In Lasso Type Estimation

We can obtain naive estimates of genetic correlation (2) and its normalization (3) by using some estimates $\hat{\beta}_z, \hat{\beta}_y$ of the effect sizes and $\hat{\Sigma}$ of the covariance matrix. More specifically, by using the Lasso method, one can obtain the non-zero estimated effect sizes of the selected covariates, and one can also use these covariates to obtain a sample covariance matrix.

More precisely, let $S = \{j : \hat{\beta}_{y_j} \neq 0\} \cup \left\{j : \hat{\beta}_{z_j} \neq 0\right\}$ where $\hat{\beta}_z, \hat{\beta}_y$ are estimates from the Lasso method,

$$\begin{aligned}\hat{\beta}_y &:= \hat{\beta}_{y\mathrm{Lasso}} = \arg\min_{\beta_y} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \beta_y^\top x_i) + \lambda_y||\beta_y||_1, \\ \hat{\beta}_z &:= \hat{\beta}_{z\mathrm{Lasso}} = \arg\min_{\beta_z} \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, \beta_z^\top x_i) + \lambda_z||\beta_z||_1.\end{aligned} \tag{4}$$

Here, $\ell(a,b)$ is the negative log-likelihood for an observation; for example, for the linear Gaussian case it is $\frac{1}{2}(a-b)^2$, and for logistic regression it is $-a \cdot b + \log(1+e^b)$. The positive tuning parameters λ_y, λ_z control the overall strength of the penalties, and we used 10-fold cross-validation to choose a suitable value for each of them. The Lasso was implemented in the `R` package `glmnet` [45].

Now, we can calculate the quantity of interest by plugging in these Lasso estimates, with $\hat{\Sigma}_S = X_S X_S^\top / n$,

$$\widehat{\mathrm{g.Cor}}(y,z) = \hat{\beta}_{zS}^{\top} \hat{\Sigma}_S \hat{\beta}_{yS}.$$

Other quantities can also be estimated directly via plugging in Lasso outputs.

However, as shown in simulations in Section 4, these naive plug-in approaches create a huge bias. Moreover, uncertainty quantification for this kind of approach is difficult to obtain and, actually, is not known up to the present time. To overcome these problems, we, next, present a reliable approach via multiple sample splitting and aggregation which provides an accurate estimate, together with some reliable uncertainty quantification.

3. Method

In this section, we introduce our generic strategy to estimate genetic correlation, which consists of a multi-sample splitting strategy and a sparse regularization step, followed by an estimation step.

3.1. Estimating via Multiple Sample Splitting and Aggregation

First, without loss of generality, let us assume that the sample size is even for simplicity. The original dataset (y, z, X) is randomly divided into two disjoint datasets $\mathcal{D}_1 = (y^{(1)}, z^{(1)}, X^{(1)})$ and $\mathcal{D}_2 = (y^{(2)}, z^{(2)}, X^{(2)})$ with equal sample sizes. Sample splitting is a useful approach that can help to eliminate overfitting when variable selection and subsequent estimation are performed on the same dataset [36,43,46].

First, we apply a variable selection step on data $\mathcal{D}_1$, as in (4), where we propose using Lasso as a default alternative, to select the most relevant covariates (to reduce dimension of the model). Other variable selection methods could be used, such as, for example, the sure independence screening (SIS) procedure [47]. Denote $S_y \subset \{1, \ldots, p\}$ and $S_z \subset \{1, \ldots, p\}$ the subset of important predictors obtained, respectively, for the traits y and z, where $|S_y| < n/2$ and $|S_z| < n/2$.

Then, using data $\mathcal{D}_2$, we fit linear regression on $(y^{(2)}, X^{(2)}_{S_y})$ and $(z^{(2)}, X^{(2)}_{S_z})$ to obtain unbiased estimates of the regression coefficients β_y and β_z. The estimation of genetic correlation between two traits is calculated, as in (2), where the sample covariance matrix is used. Thus, it is guaranteed that variable selection and estimation are performed on independent samples.

Note that, in order to avoid crucial dependence on the particular sample splitting employed, we propose performing the sample splitting and inference procedure many times (e.g., 100 times) and aggregating the corresponding results via taking the median. Our proposed strategy is summarized in Algorithm 1. Another important point in our proposed method, which is different to that of other works, is the aggregating of the final result via the median, inspired from reference [48]. This is further confirmed in our simulation studies in Section 4.

Algorithm 1 B.CORE Algorithm

```
Repeat $B$ times from step 1 to step 4,
    Step 1: Divide the sample uniformly at random into two equal parts.
    Step 2: In the first part of the data, use a high-dimensional variable selection method,
such as lasso, to select the important covariates.
    Step 3: Then, on the second subset with only selected covariates from Step 2, apply the
ordinary least-squares method to obtain unbiased estimates of the regression coefficients
and the genetic correlation.
    Step 4: Repeat Step 2 and Step 3 by changing the role of the first and second subsets.
Output The output is the median of all 2B estimated quantities.
```

In addition, switching the roles of the data subsets in **Step 4** helps in obtaining another estimation, which, thus, leads to a more stable estimation. It is noted that the main cost for our Algorithm 1 is in fitting a penalized regression (**Step 2**) for variable selection in the setting where $p \gg n$. It is, however, important to note that there have been recent advancements in methods for quickly analyzing large GWAS data using penalized regression, as in, for example, [49]. Additionally, the process can easily be made more efficient by implementing multiple repetitions in parallel.

Furthermore, as suggested by the work in [19], the ultra-high dimensionality could also be reduced to a relatively large scale before applying Algorithm 1 by using a screening method [47]. A marginal-type screening technique, such as SIS, could be adopted, as long as the number of SNPs involved is not large, e.g., 10,000. For higher numbers of SNPs, a joint-type screener, such as the ITRRS, should be used, so that the LD between the SNPs is considered and the truly associated SNPs better selected.

Theoretical foundations for using the splitting and smoothing/aggregation method to estimate regression coefficients have been established in previous research for both linear models (as outlined in [32]) and generalized linear models (as in [33]).

3.2. A 95% Reliable Interval

It is important to note that by repeating sample splitting, we are able to obtain a range of estimates for the quantity of interest. This is beneficial because it allows us to construct a meaningful interval of the estimated values. More specifically, by using the outputs from Algorithm 1, which is $2B$ estimated quantities of interest, we can construct a 95% reliable interval by taking the interval from the 2.5 to 97.5% quantiles of the outputs. This is an important feature of our proposed method, as it allows us to not only estimate the genetic correlation, but also to quantify the uncertainty of the estimate. It is well known that interval estimates are more informative than point estimates, as they provide a range of plausible values for the true parameter.

The simulation results presented in Section 4 provide further evidence of the accuracy of our proposed method. These simulations show that the 95% reliable interval is highly accurate, in the sense that it always contains the true value of interest. This is a strong indication that our proposed method is able to provide accurate and reliable estimates of genetic correlation.

Moreover, by providing interval estimates, our method allows for a more comprehensive understanding of the underlying relationships, as it allows for a more thorough assessment of the uncertainty of the estimates. This is particularly important when dealing with complex traits, where the underlying distribution of effects may not be easily identifiable or may not conform to a known distribution. This feature of our method provides a more robust and inclusive approach to understanding.

3.3. Extension with Traits on Different Samples

We now describe a model in the general setting where the traits are not observed on the same set of samples. More specifically, two traits y and z are collected from two different sample sets of n_1 and n_2 sizes. Each is modeled as a linear combination of p genetic variants $X_{.j}$ and there is an error term (environmental and unmeasured genetic effects) as follows:

$$y_{n_1 \times 1} = X_{n_1 \times p} \beta_{y,p \times 1} + \varepsilon_{n_1 \times 1},$$
$$z_{n_2 \times 1} = X_{n_2 \times p} \beta_{z,p \times 1} + \gamma_{n_2 \times 1}.$$

As in [8], the Algorithm 1 can be used to estimate the genetic correlation between these two traits. More specifically, multiple sample splitting is conducted on each sample set to estimate the effect size. The sample covariance matrix is, then, estimated by merging these sample sets.

4. Numerical Studies

4.1. Simulation Studies

Experimental Designs

In this work, we used a real data set of 3069 *Streptococus pneumoniae* genomes, collected from an infant cohort study conducted in a refugee camp on the Thailand–Myanmar border [50,51]. This dataset provided us with an unique opportunity to study the genetic correlation of antibiotic resistant traits in bacteria and to create semi-synthetic datasets incorporating the levels of population structure and linkage disequilibrium present in natural populations. As seen in Figure 1, we used a fully-observed genotype matrix of 3051 samples and 5000 SNPs to conduct our analysis, which allowed us to gain a better understanding of the genetic factors associated with antibiotic resistance in *Streptococcus pneumoniae*.

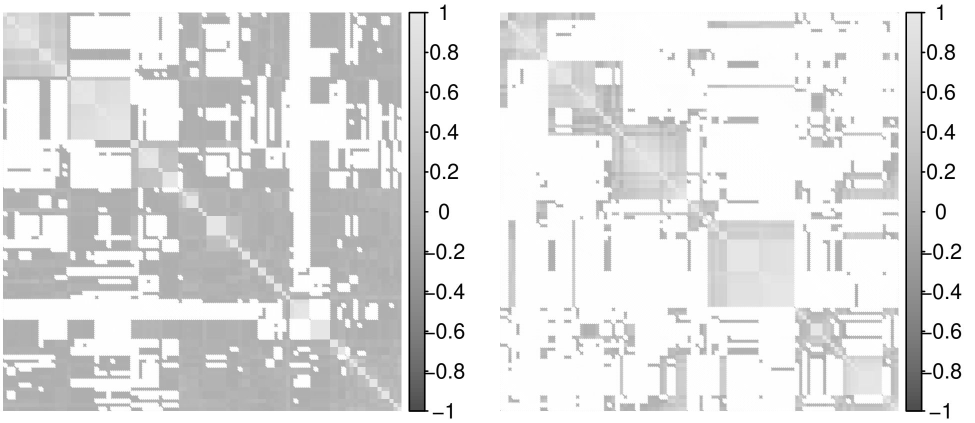

Figure 1. Sample correlation matrices of the 100 random SNPs (**right**) and 100 samples (**left**) in the genotype matrix displays the complex dependence structure presented in the real Maela data.

Given the genotypes, we considered the following settings for choosing the causal SNPs (non-zero effect sizes) under varying genetic architectures, and the phenotypes were simulated as in model (1):

Setting I: the genetic basis overlap, i.e. β_z and β_y, had 50% non-zero components in common. Then, we had the following scenario:

(a). y and z are polygenic and β_z and β_y have 1000 non-zero components.
(b). y and z are sparse and β_z and β_y have 50 non-zero components.
(c). y is polygenic and β_z has 1000 non-zero components, while z is sparse and β_y has 50 non-zero components.

Setting II: genetic basis non-overlap, i.e. β_z and β_y have no common non-zero components. Then, we have the following scenario:

(a). y and z are polygenic and β_z and β_y have 1000 non-zero components.
(b). y and z are sparse and β_z and β_y have 50 non-zero components.
(c). y is polygenic and β_z has 1000 non-zero components, while z is sparse and β_y has 50 non-zero components.

Given the above setting in choosing causal SNPs, the non-zero coefficients of β_z and β_y were drawn from the normal distribution $\mathcal{N}(0,1)$. Noise followed the normal distribution $\mathcal{N}(0,1)$. We further varied the size of the samples by uniformly subsampling the samples between 1000, 2000 or full samples. For each setup, we generated 50 simulation runs and reported the mean and the standard deviation of the absolute errors for each method across the simulation runs.

We performed simulations to compare the accuracy of our proposed method to other estimators of genetic correlation under different genetic architectures. We compared Lasso, B.CORE (mean) and B.CORE (median). Repeated sample splittings were performed $B = 50$ times. The Lasso was used with 10-fold cross validation to choose the tuning parameter λ.

4.2. Simulation Results

As demonstrated in the simulation results, presented in Figures 2–5, our proposed method of aggregating the outputs from multiple sample splitting using the median was more accurate than when using the mean. This finding was expected, as our target was a correlation quantity [48] and, also, for the real dataset, the samples would no longer be independent. Additionally, we also found that the median-aggregation led to a much smaller standard deviation, compared to the mean. This is an important aspect, as a smaller standard deviation indicates that the results are more consistent and less variable, making the estimates more reliable.

This conclusion was also supported by the fact that, when we observe the distribution of the estimates obtained from both the median and mean aggregations, the median aggregation yielded more precise and concentrated estimates around the true value, while the mean aggregation yielded estimates that were more dispersed. Moreover, we were not assuming any specific underlying distribution for the data and, hence, the median was a more robust statistic. It was less sensitive to outliers and skewness in the data, which can be present in real datasets. In addition, it is worth noting that our proposed method allows for a more flexible and inclusive approach to understanding the underlying relationships. This is particularly important when dealing with complex traits, where the underlying distribution of effects may not be easily identifiable or may not conform to a known distribution.

In comparison with the plug-in Lasso method, we found that our approach quite often yielded better results. This was more often the case when the same size was small. Most importantly, our approach returned interval estimations which facilitated, for example, confident intervals. It is challenging to obtain such results with the Lasso method. The standard deviation for Lasso and the mean-aggregation seemed comparable.

In order to determine the optimal number of sample splittings, B, to use in practice, we conducted a series of simulations where we varied B across different simulation settings. The results, presented in Figure 6, show that the number of sample splittings can greatly affect the estimation results in various settings. As seen from the figure, the accuracy of the estimates improved as the number of sample splittings increased, but it reached a plateau after a certain point. This was because, as the number of sample splittings increased, the estimation became more precise and the variability of the estimates decreased. However, at some point, additional sample splittings did not provide much improvement.

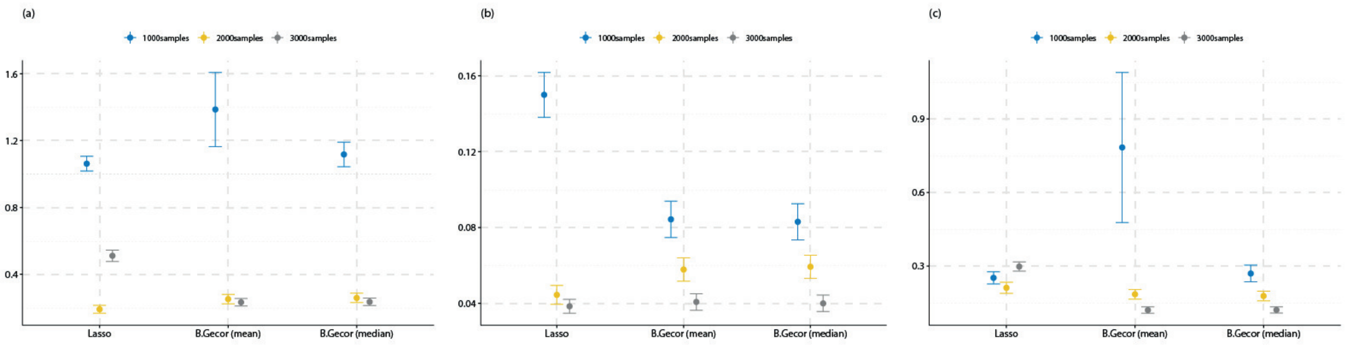

Figure 2. Simulation results on the absolute errors in estimating genetic correlation for Setting I with overlapping genetic basis between the traits. (Subfigures (**a**–**c**) corresponding to the scenario in the setting.)

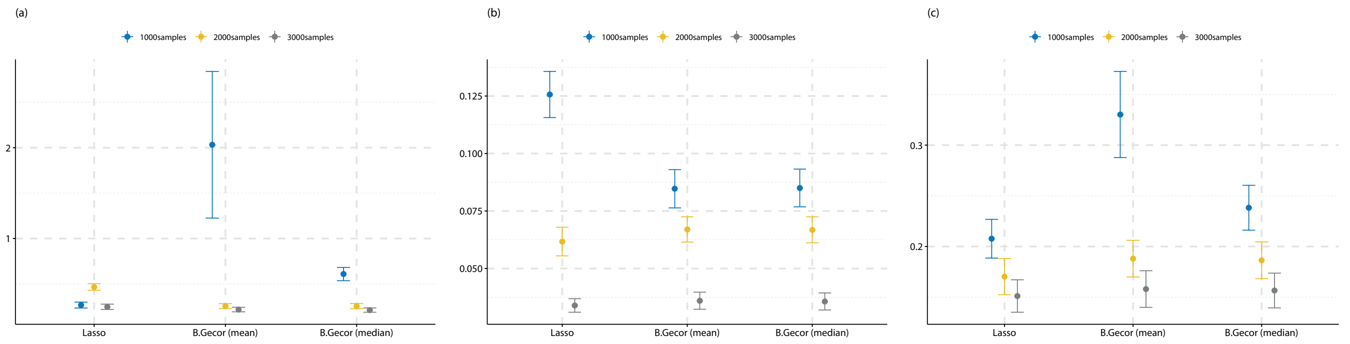

Figure 3. Simulation results on the absolute errors in estimating genetic correlation for Setting II with non-overlapping genetic basis between the traits. (Subfigures (**a**–**c**) corresponding to the scenario in the setting.)

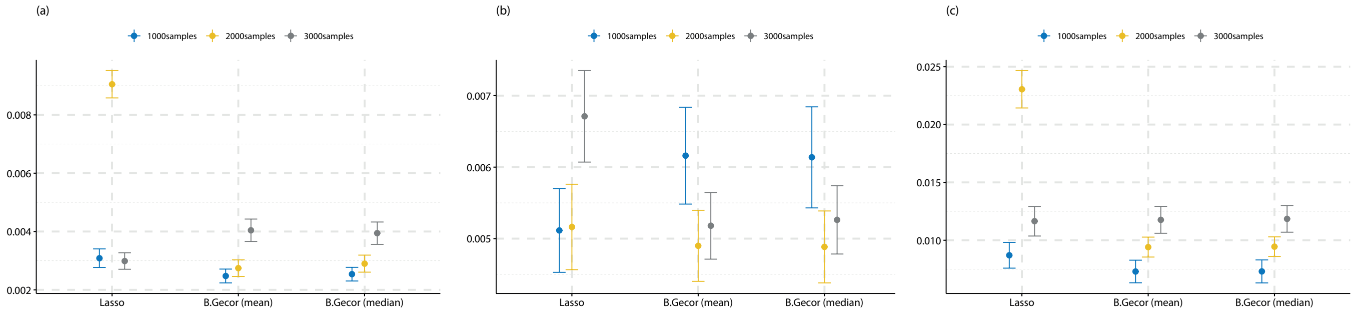

Figure 4. Simulation results on the absolute errors in estimating the standardized genetic correlation for Setting I with overlapping genetic basis between the traits. (Subfigures (**a**–**c**) corresponding to the scenario in the setting.)

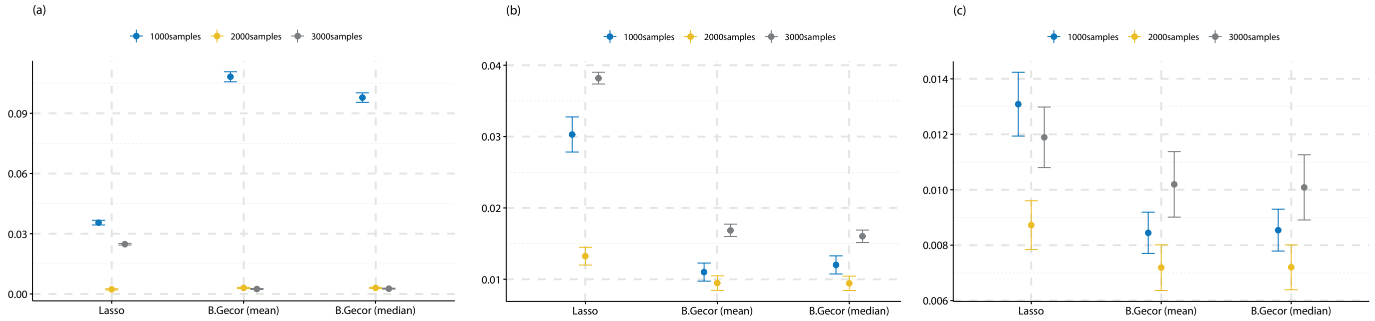

Figure 5. Simulation results on the absolute errors in estimating the standardized genetic correlation for Setting II with non-overlapping genetic basis between the traits. (Subfigures (**a**–**c**) corresponding to the scenario in the setting.)

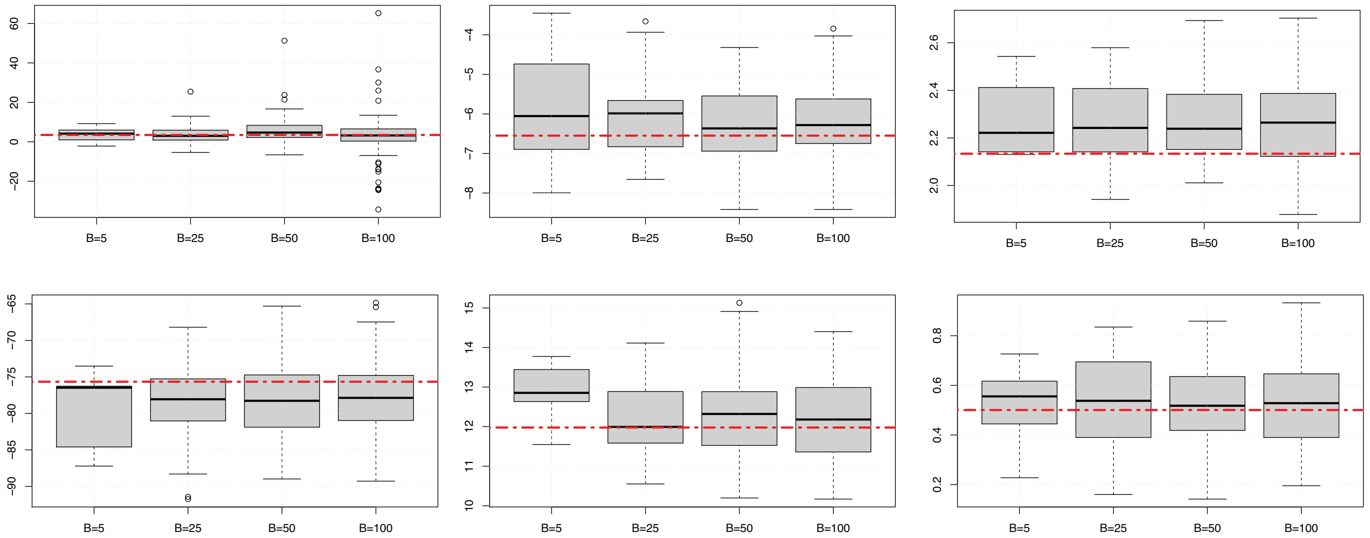

Figure 6. Simulation results when varying the number of sample splittings B (5, 25, 50, 100) in different settings (red line is the true value). **Top row** plots are for Setting I; **bottom row** plots are for Setting II.

Based on the results of our simulations, a value of $B = 50$ is recommended for practical use. This value strikes a balance between computation time and accuracy. However, when computational resources permit, a larger number of B can achieve even more stable results. For example, a larger number of B leads to more precise estimates, which can be useful in situations where the sample size is small or the correlation is weak. Additionally, a larger number of B is useful when there is a high degree of uncertainty in the data, as it allows for a more robust estimation of the genetic correlation.

In addition to evaluating the effectiveness of our proposed method, we also conducted experiments to investigate the effect of different sample splitting sizes. In particular, we varied the proportion of samples allocated to each split (α) and evaluated the performance of our method under different conditions. The results, presented in Figure 7, show that the minimum mean absolute error was achieved when $\alpha = 0.5$, suggesting that equal-size splitting is a rational choice in practice. This outcome can be explained by the fact that when the sample size was not large enough, the estimation of genetic correlation was sensitive to the size of sample splitting, and equal-size splitting was a practical choice. Equal-size splitting ensured that each split had sufficient sample size and, thus, the estimation was more stable and accurate. Furthermore, equal-size splitting also ensured that the sample splitting was random and unbiased, which was crucial for the validity of our inferences. Additionally, it is worth noting that equal-size splitting is also computationally efficient. As the sample size increases, the computation time also increases, and equal-size splitting can balance the computation time. This is important when dealing with large datasets when computational resources are limited.

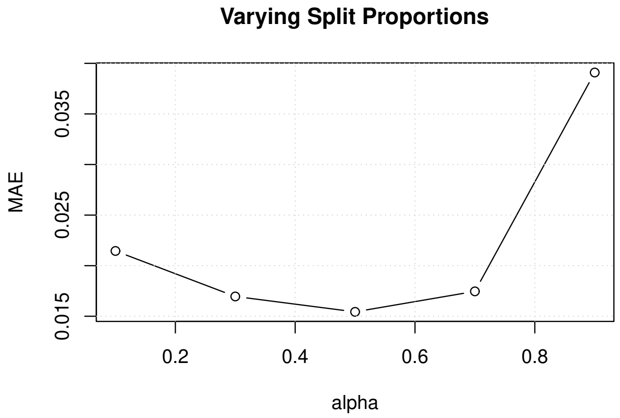

Figure 7. Average mean absolute error (MAE) of the median predictor at split proportion alpha (α) from 0.1 to 0.9.

5. Real Data Application

In this study, we aimed to examine the genetic correlations between resistance to different antibiotics in *Streptococcus pneumoniae*. To accomplish this, we used the Maela data set, which is a collection of 3069 Streptococcus pneumoniae genomes from an infant cohort study conducted in a refugee camp on the Thailand–Myanmar border. This dataset has been previously used in genome-wide association studies to identify genetic loci associated with antibiotic resistance [50,51]. We began by applying standard population genomic procedures to the data, such as using a minor allele frequency threshold and removing missing data, to obtain a genotype matrix with 121014 SNPs. We, then, considered resistances to five different antibiotics (Chloramphenicol, Erythromycin, Tetracycline, Penicillin and Co-trimoxazole) as the phenotypes.

The results of our analysis are presented in Figure 8. Our findings indicated that there were small to moderate genetic correlations between Chloramphenicol resistance and each of the other antibiotics considered. However, we observed higher genetic correlations between Erythromycin resistance, Tetracycline resistance and Co-trimoxazole resistance. Notably, the most significant genetic correlation was found to be between Penicillin resistance and Co-trimoxazole resistance, with a correlation coefficient of 0.45 (95% CI: 0.40–0.49).

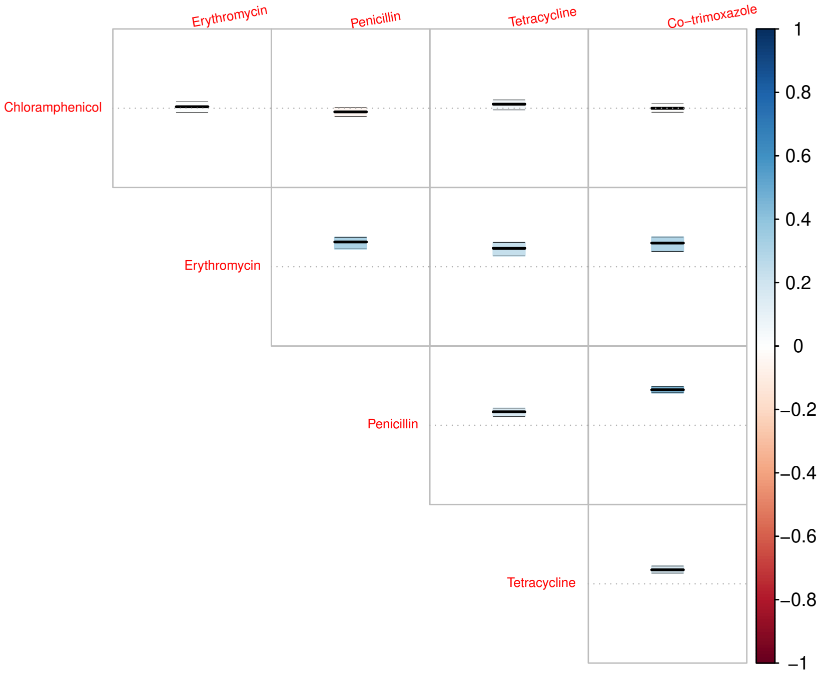

Figure 8. Estimation results of the normalized genetic correlation between five different antibiotic resistances in Maela real dataset.

Our results provide valuable insights into the genetic factors underlying antibiotic resistance in Streptococcus pneumoniae. They also demonstrate the potential of our proposed approach in identifying genetic correlations between different traits in real-world datasets. Furthermore, these results can be used to understand the underlying genetic mechanisms of antibiotic resistance in Streptococcus pneumoniae and could be useful in the development of new treatments and strategies to combat antibiotic resistance.

6. Conclusions

Estimating genetic correlation is a challenging statistical problem that is gaining increasing attention from researchers in various fields. It is a vital concept in the field of genetics that could allow us to understand the relationship between different traits and identify genetic loci that are associated with multiple traits. It is a powerful metric that provides valuable insights into the underlying genetic architecture of complex traits and has been widely studied under various models to better understand the correlation between causal loci.

In this work, we present a novel computational strategy for estimating genetic correlation efficiently and accurately. Our approach does not rely on any assumptions about the distribution of effect sizes, allowing for a more flexible and inclusive approach in understanding the underlying relationships. The proposed approach is built upon recent advances in machine learning approaches for statistical genetics, utilizing a selective inference framework and multiple sample splittings to provide valid inferences under the selected model. Furthermore, we demonstrated the efficacy of our method through numerical simulations and an application to a real dataset, where we estimated the genetic correlation between antimicrobial resistant traits in Streptococcus pneumoniae.

Overall, our proposed method is a valuable tool for researchers studying genetic correlation and can be applied to a wide range of complex traits and populations. It is an important step forward in understanding the underlying genetic mechanisms of complex traits and can provide valuable insights into the genetic basis of different diseases.

Funding: This research was funded by the Norwegian Research Council grant number 309960 through the Centre for Geophysical Forecasting at NTNU.

Data Availability Statement: The R codes and data used in the numerical experiments are available at: https://github.com/tienmt/bcore (accessed on 2 May 2023).

Acknowledgments: The author is grateful to four anonymous reviewers for their expert analysis and helpful suggestions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Giambartolomei, C.; Vukcevic, D.; Schadt, E.E.; Franke, L.; Hingorani, A.D.; Wallace, C.; Plagnol, V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **2014**, *10*, e1004383. [CrossRef] [PubMed]
2. Pickrell, J.K.; Berisa, T.; Liu, J.Z.; Ségurel, L.; Tung, J.Y.; Hinds, D.A. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **2016**, *48*, 709. [CrossRef]
3. Mancuso, N.; Shi, H.; Goddard, P.; Kichaev, G.; Gusev, A.; Pasaniuc, B. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **2017**, *100*, 473–487. [CrossRef] [PubMed]
4. Bulik-Sullivan, B.; Finucane, H.K.; Anttila, V.; Gusev, A.; Day, F.R.; Loh, P.R.; Duncan, L.; Perry, J.R.; Patterson, N.; Robinson, E.B.; et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **2015**, *47*, 1236. [CrossRef] [PubMed]
5. Furlotte, N.A.; Eskin, E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics* **2015**, *200*, 59–68. [CrossRef]
6. Shi, H.; Mancuso, N.; Spendlove, S.; Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **2017**, *101*, 737–751. [CrossRef]
7. Lu, Q.; Li, B.; Ou, D.; Erlendsdottir, M.; Powles, R.L.; Jiang, T.; Hu, Y.; Chang, D.; Jin, C.; Dai, W.; et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.* **2017**, *101*, 939–964. [CrossRef]
8. Guo, Z.; Wang, W.; Cai, T.T.; Li, H. Optimal estimation of genetic relatedness in high-dimensional linear models. *J. Am. Stat. Assoc.* **2019**, *114*, 358–369. [CrossRef]
9. Davey Smith, G.; Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **2003**, *32*, 1–22. [CrossRef]
10. Davey Smith, G.; Hemani, G. Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **2014**, *23*, R89–R98. [CrossRef]
11. Purcell, S.M.; Wray, N.R.; Stone, J.L.; Visscher, P.M.; O'donovan, M.C.; Sullivan, P.F.; Sklar, P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **2009**, *460*, 748–752. [PubMed]
12. Maier, R.; Moser, G.; Chen, G.B.; Ripke, S.; Absher, D.; Agartz, I.; Akil, H.; Amin, F.; Andreassen, O.A.; Anjorin, A.; et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* **2015**, *96*, 283–294. [CrossRef] [PubMed]
13. Lynch, M.; Walsh, B. *Genetics and Analysis of Quantitative Traits*; Sinauer: Sunderland, MA, USA, 1998; Volume 1.
14. Bürger, R. *The Mathematical Theory of Selection, Recombination, and Mutation*; John Wiley & Sons: Hoboken, NJ, USA, 2000.
15. Van Rheenen, W.; Peyrot, W.J.; Schork, A.J.; Lee, S.H.; Wray, N.R. Genetic correlations of polygenic disease traits: From theory to practice. *Nat. Rev. Genet.* **2019**, *20*, 567–581. [CrossRef]
16. Solovieff, N.; Cotsapas, C.; Lee, P.H.; Purcell, S.M.; Smoller, J.W. Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **2013**, *14*, 483–495. [CrossRef]
17. Falconer, D.S. *Introduction to Quantitative Genetics*; Oliver and Boyd: Edinburgh, UK; London, UK, 1960.
18. Lee, J.J.; McGue, M.; Iacono, W.G.; Chow, C.C. The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genet. Epidemiol.* **2018**, *42*, 783–795. [CrossRef]
19. Gorfine, M.; Berndt, S.I.; Chang-Claude, J.; Hoffmeister, M.; Le Marchand, L.; Potter, J.; Slattery, M.L.; Keret, N.; Peters, U.; Hsu, L. Heritability Estimation using a Regularized Regression Approach (HERRA): Applicable to continuous, dichotomous or age-at-onset outcome. *PLoS ONE* **2017**, *12*, e0181269. [CrossRef]
20. Janson, L.; Barber, R.F.; Candes, E. EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2017**, *79*, 1037–1065. [CrossRef]
21. Golan, D.; Rosset, S. Mixed models for case-control genome-wide association studies: Major challenges and partial solutions. In *Handbook of Statistical Methods for Case-Control Studies*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 495–514.
22. Loh, P.R.; Bhatia, G.; Gusev, A.; Finucane, H.K.; Bulik-Sullivan, B.K.; Pollack, S.J.; de Candia, T.R.; Lee, S.H.; Wray, N.R.; Kendler, K.S.; et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **2015**, *47*, 1385. [CrossRef]
23. Lee, S.H.; Yang, J.; Goddard, M.E.; Visscher, P.M.; Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **2012**, *28*, 2540–2542. [CrossRef] [PubMed]
24. Lee, S.H.; Ripke, S.; Neale, B.M.; Faraone, S.V.; Purcell, S.M.; Perlis, R.H.; Mowry, B.J.; Thapar, A.; Goddard, M.E.; Witte, J.S.; et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **2013**, *45*, 984.

25. Golan, D.; Lander, E.S.; Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E5272–E5281. [CrossRef]
26. Speed, D.; Balding, D.J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **2019**, *51*, 277. [CrossRef] [PubMed]
27. Zhang, Y.; Cheng, Y.; Jiang, W.; Ye, Y.; Lu, Q.; Zhao, H. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. *Briefings Bioinform.* **2021**, *22*, bbaa442. [CrossRef] [PubMed]
28. Wu, T.T.; Chen, Y.F.; Hastie, T.; Sobel, E.; Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **2009**, *25*, 714–721. [CrossRef]
29. Brzyski, D.; Peterson, C.B.; Sobczyk, P.; Candès, E.J.; Bogdan, M.; Sabatti, C. Controlling the rate of GWAS false discoveries. *Genetics* **2017**, *205*, 61–75. [CrossRef] [PubMed]
30. Lees, J.A.; Mai, T.T.; Galardini, M.; Wheeler, N.E.; Horsfield, S.T.; Parkhill, J.; Corander, J. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *MBio* **2020**, *11*, e01344-20. [CrossRef]
31. Dai, C.; Lin, B.; Xing, X.; Liu, J.S. False discovery rate control via data splitting. *J. Am. Stat. Assoc.* **2022**, 1–38. [CrossRef]
32. Fei, Z.; Zhu, J.; Banerjee, M.; Li, Y. Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach. *Biometrics* **2019**, *75*, 551–561. [CrossRef]
33. Fei, Z.; Li, Y. Estimation and Inference for High Dimensional Generalized Linear Models: A Splitting and Smoothing Approach. *J. Mach. Learn. Res.* **2021**, *22*, 2681–2712.
34. Tian, X. Prediction error after model search. *Ann. Stat.* **2020**, *48*, 763–784. [CrossRef]
35. Tian, X.; Taylor, J. Selective inference with a randomized response. *Ann. Stat.* **2018**, *46*, 679–710. [CrossRef]
36. Fan, J.; Guo, S.; Hao, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2012**, *74*, 37–65. [CrossRef] [PubMed]
37. Lee, J.D.; Sun, D.L.; Sun, Y.; Taylor, J.E. Exact post-selection inference, with application to the lasso. *Ann. Stat.* **2016**, *44*, 907–927. [CrossRef]
38. Tibshirani, R.J.; Taylor, J.; Lockhart, R.; Tibshirani, R. Exact post-selection inference for sequential regression procedures. *J. Am. Stat. Assoc.* **2016**, *111*, 600–620. [CrossRef]
39. Freund, Y.; Schapire, R.E. Experiments with a New Boosting Algorithm. In Proceedings of the 13th International Conference on Machine Learning, 1996, ICML'96, Bari, Italy, 3–6 July 1996; pp. 148–156.
40. Bühlmann, P.; Yu, B. Analyzing bagging. *Ann. Stat.* **2002**, *30*, 927–961. [CrossRef]
41. Meinshausen, N.; Meier, L.; Bühlmann, P. P-values for high-dimensional regression. *J. Am. Stat. Assoc.* **2009**, *104*, 1671–1681. [CrossRef]
42. Renaux, C.; Buzdugan, L.; Kalisch, M.; Bühlmann, P. Hierarchical inference for genome-wide association studies: A view on methodology with software. *Comput. Stat.* **2020**, *35*, 1–40. [CrossRef]
43. Buzdugan, L.; Kalisch, M.; Navarro, A.; Schunk, D.; Fehr, E.; Bühlmann, P. Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics* **2016**, *32*, 1990–2000. [CrossRef]
44. Mai, T.T.; Turner, P.; Corander, J. Boosting heritability: Estimating the genetic component of phenotypic variation with multiple sample splitting. *BMC Bioinform.* **2021**, *22*, 1–16. [CrossRef]
45. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
46. Li, X.; Wu, D.; Cui, Y.; Liu, B.; Walter, H.; Schumann, G.; Li, C.; Jiang, T. Reliable heritability estimation using sparse regularization in ultrahigh dimensional genome-wide association studies. *BMC Bioinform.* **2019**, *20*, 219. [CrossRef] [PubMed]
47. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 849–911. [CrossRef]
48. Lugosi, G.; Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **2019**, *19*, 1145–1190. [CrossRef]
49. Qian, J.; Tanigawa, Y.; Du, W.; Aguirre, M.; Chang, C.; Tibshirani, R.; Rivas, M.A.; Hastie, T. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* **2020**, *16*, e1009141. [CrossRef]
50. Chewapreecha, C.; Marttinen, P.; Croucher, N.J.; Salter, S.J.; Harris, S.R.; Mather, A.E.; Hanage, W.P.; Goldblatt, D.; Nosten, F.H.; Turner, C.; et al. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet.* **2014**, *10*, e1004547. [CrossRef] [PubMed]
51. Lees, J.A.; Vehkala, M.; Välimäki, N.; Harris, S.R.; Chewapreecha, C.; Croucher, N.J.; Marttinen, P.; Davies, M.R.; Steer, A.C.; Tong, S.Y.; et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **2016**, *7*, 12797. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.