





Genetics and population analysis

Inferring the heritability of bacterial traits in the era of machine learning

T. Tien Mai ^{1,*}, John A. Lees ^{2,3}, Rebecca A. Gladstone ⁴ and Jukka Corander ^{4,5,6}

¹Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim 7034, Norway, ²European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton CB10 1SD, UK, ³Department of Infectious Disease Epidemiology, MRC Centre for Global Infectious Disease Analysis, Imperial College London, London W2 1PG, UK, ⁴Department of Biostatistics, University of Oslo, Oslo 0372, Norway, ⁵Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland and ⁶Pathogens and Microbes, Wellcome Sanger Institute, Hinxton CB10 1SD, UK

*To whom correspondence should be addressed.

Associate Editor: Nicola Mulder

Received on September 2, 2022; revised on January 18, 2023; editorial decision on February 24, 2023; accepted on March 3, 2023

Abstract

Quantification of heritability is a fundamental desideratum in genetics, which allows an assessment of the contribution of additive genetic variation to the variability of a trait of interest. The traditional computational approaches for assessing the heritability of a trait have been developed in the field of quantitative genetics. However, the rise of modern population genomics with large sample sizes has led to the development of several new machine learning-based approaches to inferring heritability. In this article, we systematically summarize recent advances in machine learning which can be used to infer heritability. We focus on an application of these methods to bacterial genomes, where heritability plays a key role in understanding phenotypes such as antibiotic resistance and virulence, which are particularly important due to the rising frequency of antimicrobial resistance. By designing a heritability model incorporating realistic patterns of genome-wide linkage disequilibrium for a frequently recombining bacterial pathogen, we test the performance of a wide spectrum of different inference methods, including also GCTA. In addition to the synthetic data benchmark, we present a comparison of the methods for antibiotic resistance traits for multiple bacterial pathogens. Insights from the benchmarking and real data analyses indicate a highly variable performance of the different methods and suggest that heritability inference would likely benefit from tailoring of the methods to the specific genetic architecture of the target organism.

Availability and implementation: The R codes and data used in the numerical experiments are available at: https://github.com/tienmt/her_MLs.

Contact: the.t.mai@ntnu.no

1 Introduction

Heritability is a fundamental quantity in genetic applications (Falconer, 1960; Lynch and Walsh, 1998) which specifies the contribution of additive genetic factors to the variation of a phenotype. In the narrow-sense, heritability is defined as the proportion of the variance of a phenotype explained by the additive genetic factors. Heritability can be used to compare the relative importance between genes and environment to the variability of traits, within and across populations. Together with GWAS (genome-wide association studies), the primary tool for discovering the genetic basis of a phenotype of interest, heritability has been playing as a more and more critical role in exploring the genetic architecture of complex traits.

Current investigations of heritability in the quantitative genetics literature have focused on using the linear mixed-effect model framework (Bonnet, 2016; Bulik-Sullivan *et al.*, 2015; Golan *et al.*, 2014; Speed *et al.*, 2012, 2017; Yang *et al.*, 2010; Zhou, 2017). In this framework, the effect sizes of genetic markers, usually SNPs (single nucleotide polymorphisms), are assumed to be independent and identically distributed random variables, and often the normal Gaussian distribution is used for computational reasons. The genomic restricted maximum likelihood (GREML) and method of moments are the most widely used methods for heritability inference in this model, and some corresponding popular software are GCTA (Yang *et al.*, 2011), LDSC (Bulik-Sullivan *et al.*, 2015) and LDAK (Speed *et al.*, 2012; Speed and Balding, 2019). Although the linear mixed-effect model provides various ways to interpret correlations among

covariates and traits, and is computationally tractable, it makes assumptions which do not necessarily accurately reflect the underlying genetics (Gorfine *et al.*, 2017; Holmes *et al.*, 2019; Janson *et al.*, 2017; Lee *et al.*, 2018; Li *et al.*, 2019; Speed *et al.*, 2017; Speed and Balding, 2019). Some comparisons of different methods in this direction for estimating heritability have been recently conducted, for example, in Zhou (2017), Evans *et al.* (2018), Weissbrod *et al.* (2018), Gorfine *et al.* (2017) and Holmes *et al.* (2019).

The study of heritability estimation in the statistical machine learning community has been started relatively recently and it has not yet received wider attention. Current machine learning approaches often focus on the high-dimensional linear regression model where some sparsity regularizations could be used on the number of the covariates. This is a natural model for GWAS in modeling the whole-genome level contributions of genetic variation (Falconer, 1960; Lynch and Walsh, 1998). The benefit of this model over the classical univariate approach in GWAS has been demonstrated for example in Wu *et al.* (2009) and Brzyski *et al.* (2017). Several machine learning methods for making heritability inference have been studied: a method of moments approach is proposed in Dicker (2014); a convex optimization strategy is investigated in Janson *et al.* (2017) through a singular value decomposition; maximum likelihood estimation is studied in Dicker and Erdogdu (2016); some adaptive procedures have also been theoretically studied in Verzelen and Gassiat (2018); two-step procedures based on high-dimensional regularized regression have been introduced in Gorfine *et al.* (2017) and Li *et al.* (2019); and, a strategy for aggregating heritabilities through multiple sample splitting is introduced in Mai *et al.* (2021). However, up to our knowledge, a systematic numerical comparison of these different methods for estimating heritability has not yet been conducted.

In this study, we provide a systematic summary of recent advances in machine learning methods for estimating heritability. More especially, we review and discuss the six above-mentioned different methods and compare them with GCTA method, a state-of-the-art method in quantitative genetics. The application is carried in a bacterial GWAS context for estimating the heritability of antibiotic resistant phenotypes. While estimating heritability in human GWAS has been studied in numerous works, the topic has not yet been considered widely in bacteria, for the only prominent examples see Lees *et al.* (2017, 2020) and Mallawaarachchi *et al.* (2022). This is partly because bacterial GWAS poses unique challenges compared to studies with human or eukaryotic DNA in general, originating from highly structured populations and more limited recombination that produce in considerable linkage disequilibrium across whole chromosomes.

Our article is structured as follows. In Section 2, the problem of heritability estimation is introduced and a systematic review of machine learning methods for estimating heritability is given. In Section 3, we briefly introduce the test datasets used in our evaluation. Results and discussion of different methods on test datasets are presented in Section 4.

2 Heritability inference using machine-learning methods

The following notations are used in the work. The $\ell_{q,0 < q < +\infty}$ norm of a vector $x \in \mathbb{R}^d$ is defined by $\|x\|_q = (\sum_{i=1}^d |x_i|^q)^{1/q}$. For a matrix $A \in \mathbb{R}^{n \times m}$, A_i denotes its i th row and A_j denotes its j th column. For any index set $S \subseteq \{1, \dots, d\}$, x_S denotes the subvector of x containing only the components indexed by S , and A_S denotes the submatrix of A forming by columns of A indexed by S .

2.1 The problem of heritability estimation

Let $y_i \in \mathbb{R}$ be the measured phenotype of subject i such that

$$y_i = X_i \beta + \varepsilon_i, i = 1, \dots, n \quad (1)$$

where $X_i = (X_{i1}, \dots, X_{ip}) \in \mathbb{R}^p$ is the vector of genotypes of subject i and p is the total number of variants; $\varepsilon_1, \dots, \varepsilon_n \in \mathbb{R}$ are unobserved

independent and identically distributed (iid) errors with $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2 > 0$; $\beta = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown p -dimensional parameter. We assume that $X_i, i = 1, \dots, n$ are iid random vectors and independent of ε_i with $\mathbb{E}(X_i) = 0$ and $p \times p$ positive definite covariance matrix $\text{cov}(X_i) = \Sigma$.

Under the model (1), for the i th observation it follows that $\text{Var}(y_i) = \text{Var}(X_i \beta) + \sigma_\varepsilon^2 = \beta^\top \Sigma \beta + \sigma_\varepsilon^2$. Our main focus is in estimating the narrow-sense heritability for the phenotype y defined as

$$h^2 = \frac{\beta^\top \Sigma \beta}{\beta^\top \Sigma \beta + \sigma_\varepsilon^2}. \quad (2)$$

In other words, this quantity computes the proportion of genetic differences present in the population variability of a trait. It can benefit modeling the underlying genetic architecture of a trait, because a heritability close to zero means that environmental factors cause most of the variability of the trait, while a heritability close to 1 indicates that the variability of the trait is nearly solely caused by the differences in genetic factors.

It is noted that as $\mathbb{E}[\|y\|_2^2/n] = \text{Var}(y)$, one can use $\|y\|_2^2/n$ as an unbiased estimator for the denominator of the heritability. Further, (2) can also be written as

$$h^2 = 1 - \frac{\sigma_\varepsilon^2}{\text{Var}(y)}. \quad (3)$$

And thus, an estimate of the noise-variance $\hat{\sigma}_\varepsilon^2$ [see e.g. Reid *et al.* (2016)] can be used to estimate h^2 rather than directly estimating the genetic variance $\beta^\top \Sigma \beta$.

Hereafter, we provide details for different methods for making heritability inference. We especially focus on methods that enable confidence intervals (CIs) to be computed.

2.2 Direct methods

We first recall some methods that can directly estimate heritability from the data without identifying the genetic basis of the phenotype.

2.2.1 Convex optimization approach

Using a singular value decomposition (s.v.d.) transformation, the work in Janson *et al.* (2017) proposes a method, called Eigenprism, to estimate the squared signal $\beta^\top \Sigma \beta$ and thus heritability by solving a convex optimization problem. They also prove the asymptotic normality of their estimator.

More specifically, let $X = UDV^\top$ be a singular value decomposition (s.v.d.) and put $z = U^\top y$. Let $\lambda_{i=1:n}$ denote the eigenvalues of XX^\top/p . The authors of Janson *et al.* (2017) consider the following convex optimization problem, denoted by P_1 ,

$$\begin{aligned} \arg \min_{w \in \mathbb{R}^n} \max & \left(\sum_{i=1}^n w_i^2, \sum_{i=1}^n w_i^2 \lambda_i^2 \right), \\ \text{such that} & \sum_{i=1}^n w_i = 0, \sum_{i=1}^n w_i \lambda_i = 1. \end{aligned}$$

Let w^* be the solution to the problem P_1 , then the heritability estimator is given by

$$\hat{h}_{\text{Eprism}}^2 = \frac{\sum_{i=1}^n w_i^* \lambda_i^2}{\|y\|_2^2/n}.$$

With P_1^* being the minimized objective function value, the $(1 - \alpha)$ -CI is given by $[\hat{h}_{\text{Eprism}}^2 \pm z_{1-\alpha/2} \sqrt{2P_1^*}]$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

The cost of this method stems mainly from the calculation of the s.v.d. of X . This will be expensive and slow for a genotype matrix with large dimensions. Moreover, we note that in practice the optimization in P_1 can fail sometimes and, in addition, this method only works with high-dimensional data where $p > n$.

2.2.2 Maximum likelihood estimation

Another direct method for estimating heritability is based on using the maximum likelihood method. In the paper (Dicker and Erdogdu, 2016), the authors derive consistency and asymptotic normality of the maximum likelihood estimation (MLE) under additional Gaussian assumptions. More specifically, the maximum likelihood problem is defined as

$$(\hat{\eta}, \hat{\sigma}_{MLE}^2) = \arg \max_{\eta, \sigma^2} \left\{ -\frac{\log(\sigma^2)}{2} - \frac{1}{2n} \log \det \left(\frac{\eta}{p} XX^\top + \mathbf{I} \right) - \frac{1}{2\sigma^2 n} y^\top \left(\frac{\eta}{p} XX^\top + \mathbf{I} \right)^{-1} y \right\}$$

and the heritability estimate is given by

$$\hat{h}_{MLE}^2 = 1 - \frac{\hat{\sigma}_{MLE}^2}{\text{Var}(y)}.$$

The authors also studied the consistency and asymptotic normality of this MLE estimator. As a result, the $(1 - \alpha)$ -CI is given by $[\hat{h}_{MLE}^2 \pm z_{1-\alpha/2}/\sqrt{2n}]$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

This method appears to be quite efficient computationally as it only requires handling a matrix inversion of dimension $n \times n$.

2.2.3 Moments method

Heritability estimation based on method-of-moment has been proposed and studied in Dicker (2014) and Verzelen and Gassiat (2018). Several estimators have been proposed in these works. However, when the covariance matrix Σ is non-estimable or expensive to estimate (as often is the case in practice), the reference (Dicker, 2014) proposed an estimator as follows, with $S = X^\top X/n$, $\hat{m}_1 = \text{trace}(S)/p$, $\hat{m}_2 = \frac{1}{p} \text{trace}(S^2) - \frac{p}{n} \hat{m}_1^2$, and put

$$\hat{\sigma}^2 = \left(1 + \frac{p\hat{m}_1^2}{(n+1)\hat{m}_2} \right) \frac{\|y\|^2}{n} - \frac{\hat{m}_1}{n(n+1)\hat{m}_2} \|X^\top y\|^2, \\ \hat{\tau}^2 = -\frac{p\hat{m}_1^2}{(n+1)\hat{m}_2} \frac{\|y\|^2}{n} + \frac{\hat{m}_1}{n(n+1)\hat{m}_2} \|X^\top y\|^2,$$

and the heritability estimate is then

$$\hat{h}_{Moment}^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}.$$

The asymptotic properties of the moment estimator were proven in Dicker (2014) and some non-asymptotic results were derived in Verzelen and Gassiat (2018). These results allow to obtain the $(1 - \alpha)$ -CI approximately as $[\hat{h}_{Moment}^2 \pm \log(1/2)\sqrt{\hat{p}/n}]$.

It is noted that this method requires to compute a $p \times p$ matrix S which is very costly for data with large dimensions.

2.3 Plug-in Lasso type approaches

We now discuss some naive plug-in methods that are based on sparsity penalized regression methods. These methods typically assume that there is a small subset of biomarkers (in the genotype matrix) that will be important to the phenotype and influence its variability.

2.3.1 Scaled Lasso

The paper (Verzelen and Gassiat, 2018) studied the problem of heritability estimation through using a variance estimation, see formula (3), in high-dimensional sparse regression from the scaled Lasso method (Sun and Zhang, 2012). The scaled Lasso (also known as square-root Lasso) is defined as

$$(\hat{\beta}_{SL}, \hat{\sigma}_{SL}) = \arg \min_{\beta, \sigma} \frac{1}{2n\sigma} \|y - X\beta\|_2^2 + \frac{n\sigma}{2} + \lambda \|\beta\|_1,$$

where $\lambda > 0$ is the tuning parameter.

The heritability estimate is

$$\hat{h}_{Slasso}^2 = 1 - \frac{\hat{\sigma}_{SL}^2}{\text{Var}(y)},$$

and its honest $(1 - \alpha)$ -CI, given in Verzelen and Gassiat (2018), is given by $[\hat{h}_{Slasso}^2 \pm \log(1/2)(k\sqrt{\hat{p}}/n + 1/\sqrt{n})]$, where $k := \|\hat{\beta}_{SL}\|_0$ the number of non-zero components in $\hat{\beta}_{SL}$. It is noted that this CI is rather honest in the sense that its width tends to be quite large. A sharp confidence interval for this estimator has not yet been constructed.

We note that the scaled Lasso method shares the same spirit as Lasso which returns a very sparse model. Therefore, heritability estimation for a phenotype with a polygenic basis by this method tends to lead to underestimation.

2.3.2 Elastic Net

From (2), a direct heritability estimate can be obtained by using a Lasso type method. More precisely, let $S = \{j : \hat{\beta}_j \neq 0\}$ where $\hat{\beta}$ is an estimate from a Lasso-type method, we can calculate the heritability as in equation (2) with $\hat{\Sigma}_S = X_S X_S^\top / (n - 1)$,

$$\hat{h}^2 = \frac{\hat{\beta}_S^\top \hat{\Sigma}_S \hat{\beta}_S}{\text{Var}(y)}.$$

Here, we focus on using the Elastic Net estimate for the regression parameters: $\hat{\beta} = \hat{\beta}_{Enet}$. The Elastic Net has been shown to be especially useful when the variables are dependent (Zou and Hastie, 2005) (LD structure), which is particularly relevant in bacterial genome data (Earle et al., 2016). The corresponding estimator is defined as

$$\hat{\beta}_{Enet} := \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \beta^\top X_i) + \lambda \left[\frac{(1 - \alpha)}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right],$$

where $\ell(\cdot)$ is the negative log-likelihood for an observation. Elastic Net is tuned by $\alpha \in [0, 1]$, that bridges the gap between Lasso ($\alpha = 1$) and ridge regression ($\alpha = 0$). The tuning parameter $\lambda > 0$ controls the overall strength of the penalty and can be chosen by using cross-validation. For example, 10-fold cross-validation is often used in practice.

The paper (Lees et al., 2020) showed that Elastic Net is a promising method for bacterial GWAS data where the authors suggest using a small value of α , e.g. $\alpha = 0.01$. However, we would like to note that the CI for heritability estimated by this plug-in method is not available yet. It is worth noting that a GWAS analysis using high-dimensional sparse regression, such as the Elastic net discussed above, would already provide the estimated effect sizes corresponding to the selected covariates. Therefore, estimating the heritability by using these effect sizes would bring insight on understanding both the genetic architecture as well as the genetic contribution to a trait.

2.4 Boosting heritability method

We now present some advanced approaches in machine learning for making inference about heritability. The core idea of these types of methods is to combine a covariate selection step with an estimation step, known as the selective inference approach. The selection step aims at either reducing the dimension of the problem or removing irrelevant biomarkers, after which a heritability estimation step is applied. These approaches have been shown not only to improve the computational aspects of the estimation procedure, but to also yield more accurate results.

In this approach, the original data (y, X) is randomly divided into two disjoint datasets $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$ with equal sample sizes.

The HERRA method introduced in Gorfine et al. (2017) is first based on a screening method [e.g. as in Fan and Lv (2008)] to reduce the number of covariates below the sample size. With the remaining covariates, the data are randomly split into two equally sized subsets. Then, a Lasso-type estimator is employed on $(y^{(1)}, X^{(1)})$ to select a small number of important variables. After that, the authors use the least squares estimator on $(y^{(2)}, X^{(2)})$ with only the selected covariates (from the Lasso-type estimator) to obtain an estimate of the noise-variance. The procedure is repeated where the role of $(y^{(1)}, X^{(1)})$ and

$(y^{(2)}, X^{(2)})$ is switched to obtain another estimate of the noise-variance. Finally, heritability is calculated as in the formula (3) where the noise-variance is the mean of the two estimated noise-variances.

The work in Li et al. (2019) has also proposed a ‘two-stage’ approach with sample splitting. More particularly, the data is also randomly split into two disjoint datasets with equal sizes. On the first dataset $(y^{(1)}, X^{(1)})$, they use a sparse regularization method based on Elastic net to first reduce the model by selecting the relevant variables. Then, on the second dataset $(y^{(2)}, X^{(2)})$, only the selected variables are used to estimate the heritability through a method of moments approach (Dicker, 2014) or GCTA method.

These post-selection approaches guarantee that sparse regularization and variance estimation are carried out on independent datasets and thus the heritability will not be overestimated. However, both methods in Gorfine et al. (2017) and Li et al. (2019) heavily depend on the way data is split. One can avoid this dependence by performing the sample splitting and inference procedure many times (e.g. 100 times) and aggregating the corresponding results. This is to make sure that the different latent structures possibly residing in the sample are properly taken into account in both the selection and estimation steps. This is the core idea of the ‘Boosting heritability’ strategy in Mai et al. (2021).

The work in Mai et al. (2021) recently introduces a generic strategy for heritability inference, termed as ‘Boosting Heritability’, which generalized the ideas from the post-selection approaches by Gorfine et al. (2017) and Li et al. (2019). Boosting heritability, detailed in Algorithm 1, uses in particular a multiple sample splitting strategy which is shown to lead in general to a stable and more reliable heritability estimate. More importantly, this procedure also provides an informative interval of estimated heritabilities that shows the range of the target heritability would belong to. We call this interval a ‘reliable’ interval.

Algorithm 1 Boosting heritability (Mai et al., 2021)

Step 0: Using a screening method, such as the sample correlation (Fan and Lv, 2008), to remove 25% least associated covariates.

Repeat B times from step 1 to step 4,

Step 1: With the remaining covariates, divide the sample uniformly at random into two equal parts.

Step 2: On the first part of the data, use Elastic net to select the important covariates.

Step 3: On the second subset with only selected covariates from Step 2, estimate the heritability by using (3) where the noise-variance is estimated using the least square method.

Step 4: Repeat Step 2 and Step 3 by changing the role of the first and second subset.

Final → The final heritability estimate is the mean of the estimated heritabilities at each repeat.

3 Test datasets

We investigate performance of the different methods using four public bacterial datasets suitable for GWAS and simulated phenotypes based on the genetic architecture present in the data. We focus on estimating heritability of the antibiotic resistance phenotypes. Table 1 summarizes our test datasets. The heritability of the antibiotic resistance phenotype is expected to be high, meaning that the variability stems primarily from the observed genetic differences among these bacteria.

3.1 *Streptococcus pneumoniae*: MA data

Streptococcus pneumoniae (the pneumococcus) is a common nasopharyngeal commensal that can cause invasive pneumococcal disease. Here, we consider two datasets for this bacterium, abbreviated as the MA and Maela data.

The MA dataset consists of 616 *S.pneumoniae* genomes from isolates collected from healthy children in an asymptomatic nasopharyngeal colonization survey in Massachusetts between 2001 and 2007. The genomic data and phenotypes are publicly available through the publication (Croucher et al., 2015). After initial data filtering using a minor allele frequency (5%) and removing missing data greater than 10%, we obtain a genotype matrix of 603 samples with 89 703 SNPs (Chewapreecha et al., 2014). We consider resistance to penicillin antibiotics as the phenotype (Croucher et al., 2013). The genome-wide association studies of this phenotype were conducted in Chewapreecha et al. (2014). The results are given in Figure 3. It is noted that the main mechanism of resistance to penicillin can be explained by the causal SNPs in the penicillin binding proteins *pbp2x*, *pbp2b* and *pbp1a*, see Chewapreecha et al. (2014).

3.2 *Streptococcus pneumoniae*: Maela data

The Maela dataset is a large *S.pneumoniae* dataset which consists of 3069 whole genomes produced from randomly selected isolates from a longitudinal nasopharyngeal colonization study of infants and a subset of their mothers, performed between 2007 and 2010 in a rural refugee camp on the Thailand-Myanmar border (Chewapreecha et al., 2014; Lees et al., 2016). The genomic data and penicillin MICs are publicly available from Chewapreecha et al. (2014). Using a minor allele frequency threshold (5%) and removing missing data greater than 10%, we obtain a genotype matrix with 121 014 SNPs.

We use a continuous phenotype corresponding to the inhibition zone diameters measured in the lab. These inhibition zone diameters are in practice used to define whether a sample is ‘Sensitive’ or ‘Resistant’ to an antibiotic, for some antibiotics, an ‘Intermediate’ designation is also given which we treat as resistant. We consider resistances to three different antibiotics as the phenotypes: tetracycline, penicillin and co-trimoxazole. The results are given in Figure 4. The genetic loci associated with these antibiotic resistances have been examined in genome-wide association studies in Lees et al. (2016). The tetracycline resistance is conferred by the *tetM* gene and

Table 1. Summary of test datasets

Dataset name	Bacteria	Antibiotic resistant phenotype(s) to	No. of samples	No. of genetic features	Reference
MA	<i>Streptococcus pneumoniae</i>	Penicillin	603	89 703	Croucher et al. (2015)
Maela	<i>Streptococcus pneumoniae</i>	Tetracycline, Co-trimoxazole, Penicillin	3069	121 014	Chewapreecha et al. (2014) and Lees et al. (2016)
<i>E.coli</i>	<i>Escherichia coli</i>	Amoxicillin, Cefotaxime, Cefazidime, Cefuroxime, Ciprofloxacin, Gentamicin	1509	121 779	Kallonen et al. (2017)
NG	<i>Neisseria gonorrhoeae</i>	Azithromycin, Cefixime, Ciprofloxacin, Penicillin, Tetracycline	1595	20 486	Schubert et al. (2019) and Unemo et al. (2016)

the co-trimoxazole resistance is conferred by the SNPs in the *dyr* gene (Maskell *et al.*, 2001).

3.3 *Escherichia coli*: *E.coli* data

Escherichia coli is a common colonizer of the human gut but is also a leading cause of blood stream infections, in which antibiotic resistance is increasing. The *E.coli* data from Kallonen *et al.* (2017) consists of 1509 isolates from a systematic survey of blood stream infections conducted in England between 2001 and 2012 with an alignment of 121 779 SNPs (after initial data filtering with a minor allele frequency threshold 5% and removing missing data greater than 10%).

We consider resistances to six different antibiotics as the phenotypes: amoxicillin, cefotaxime, ceftazidime, cefuroxime, ciprofloxacin, gentamicin reported as categorical phenotypes ‘resistant’, ‘intermediate’, ‘sensitive’ as in Kallonen *et al.* (2017). The results are given in Table 2.

3.4 *Neisseria gonorrhoeae*: NG data

Neisseria gonorrhoeae is a sexually transmitted pathogen in which antibiotic resistance is rapidly evolving, leading to multidrug resistance (MDR) and some extremely drug resistant (XDR) strains. The NG data has been analyzed in these studies (Grad *et al.*, 2016; Schubert *et al.*, 2019; Unemo *et al.*, 2016). These 1595 clinical samples were from surveillance in the USA (2000–2013), Canada (1989–2003, selected for decreased susceptibility to cephalosporin) and the UK (2004–2013). We obtain a genotype matrix with 20 486 SNPs (after initial data filtering using a minor allele frequency threshold 5% and removing missing data greater than 10%). We consider resistances to five different antibiotics as the phenotypes: azithromycin, cefixime, ciprofloxacin, penicillin, tetracycline. The results are given in Table 3.

4 Results and discussion

4.1 Simulations

4.1.1 Simulation settings

As the basis for systematically evaluating the performance of the different methods, we use a subset of the Maela dataset (see Section 3) to create a semi-synthetic dataset that incorporates levels of population structure and LD closely reflecting those present in natural

populations (see Fig. 1). This subset corresponds to a genotype matrix of 3051 samples and 5000 SNPs. Using this real genotype matrix, we simulate the responses/phenotypes through the linear model defined in (1). For choosing the causal SNPs (non-zero effect sizes), we follow the penicillin resistance-like setting (Dewé *et al.*, 2019; Lees *et al.*, 2016): to select all SNPs from three genes (*pbpX*, *pbp1A*, *penA*) as causal.

Given the chosen SNPs, regression coefficients β^0 are either drawn from the normal distribution $\mathcal{N}(0, 1)$ or Student t_3 distribution. Because the true covariance of the genotype matrix is not given, we need to re-normalize the coefficient β^0 as $\beta = \beta^0 \sqrt{\sigma_e^2 h^2 / (\beta^{0T} \Sigma \beta^0 (1 - h^2))}$ to assure that the true corresponding heritability is approximating our target. Here, $h^2 = 0.8$ is our target heritability and Σ is the sample covariance matrix of the genotype matrix and the noise variance is fixed as $\sigma_e^2 = 1$.

In simulations, the true covariance matrix of the genotype matrix is unknown, so phenotypes are approximated based on a given heritability using model (1). To establish a benchmark for comparison, the noise variance is set to $\sigma_e^2 = 1$, and an estimator is calculated using formula (3). This estimator, denoted by ‘oracle’, is based on the true simulated values and cannot be used with real data. The methods are tested using the following settings: ‘wholegenes’ (whole genome analysis), ‘subsample1500’ (1500 randomly selected samples), ‘subsample500’ (500 randomly selected samples), ‘causalgenes’ (only true causal genes, Eprism does not work in this setting) and ‘t-effect’ (effect sizes simulated from Student t_3 distribution). Additionally, the GCTA (mixed-effect) model is used to simulate phenotypes and is denoted as ‘GCTA.model’ (oracle does not work in this setting). Results from 50 replications are shown in Figure 2.

4.1.2 Simulation results

As seen in the simulation results presented in Figure 2, the ‘oracle’ estimator consistently demonstrates the highest level of accuracy and therefore serves as an effective benchmark for comparison. It is also evident that both the Elastic net and Lasso methods tend to underestimate the target heritability. However, the Elastic net method does provide a more reliable lower bound for the underlying heritability. The moment method, on the other hand, is found to be particularly unreliable in this context, failing to produce acceptable results across all settings. The Eprism approach, which is based on

Table 2. Heritability estimation of antibiotic resistances in *E.coli* data (CIs are in parentheses)

	Amoxicillin	Cefotaxime	Ceftazidime	Cefuroxime	Ciprofloxacin	Gentamicin
Enet	0.56	0.44	0.26	0.23	0.75	0.31
Eprism	0.99 (0.76–1.00)	Failed	Failed	0.55 (0.34–0.77)	Failed	Failed
MLE	0.00 (0.00–0.04)	0.44 (0.40–0.47)	0.34 (0.30–0.37)	0.32 (0.28–0.35)	0.90 (0.86–0.94)	0.34 (0.31–0.38)
Moment	0.77 (0.55–0.99)	0.03 (0.00–0.20)	0.03 (0.00–0.20)	0.12 (0.00–0.28)	0.10 (0.00–0.26)	0.05 (0.00–0.21)
S_Lasso	0.42 (0.15–0.70)	0.17 (0.00–0.36)	0.14 (0.02–0.26)	0.11 (0.03–0.19)	0.80 (0.49–1.00)	0.13 (0.03–0.22)
GCTA	0.82 (0.77–0.86)	0.72 (0.65–0.79)	0.69 (0.62–0.76)	0.31 (0.22–0.41)	Failed	0.41 (0.31–0.51)
BoostHER	0.67 (0.64–0.70)	0.53 (0.47–0.58)	0.42 (0.33–0.51)	0.40 (0.34–0.45)	0.89 (0.87–0.92)	0.45 (0.39–0.50)

Table 3. Heritability estimation of antibiotic resistances in NG data (CIs are in parentheses)

	Azithromycin	Cefixime	Ciprofloxacin	Penicillin	Tetracycline
Enet	0.69	0.78	0.91	0.68	0.73
Eprism	0.99 (0.82–0.99)	0.74 (0.56–0.92)	0.28 (0.10–0.45)	0.99 (0.82–0.99)	0.99 (0.82–0.99)
MLE	0.80 (0.76–0.83)	0.87 (0.83–0.91)	0.98 (0.94–0.99)	0.80 (0.76–0.83)	0.85 (0.81–0.89)
Moment	0.08 (0.01–0.14)	0.99 (0.93–0.99)	0.19 (0.12–0.26)	0.06 (0.00–0.12)	0.58 (0.52–0.65)
S_Lasso	0.33 (0.25–0.40)	0.69 (0.51–0.86)	0.95 (0.72–0.99)	0.44 (0.33–0.55)	0.37 (0.29–0.45)
GCTA	0.81 (0.77–0.85)	0.77 (0.73–0.82)	0.85 (0.82–0.88)	0.73 (0.68–0.78)	0.81 (0.77–0.84)
BoostHER	0.70 (0.66–0.74)	0.84 (0.81–0.85)	0.96 (0.95–0.97)	0.77 (0.74–0.79)	0.81 (0.79–0.83)

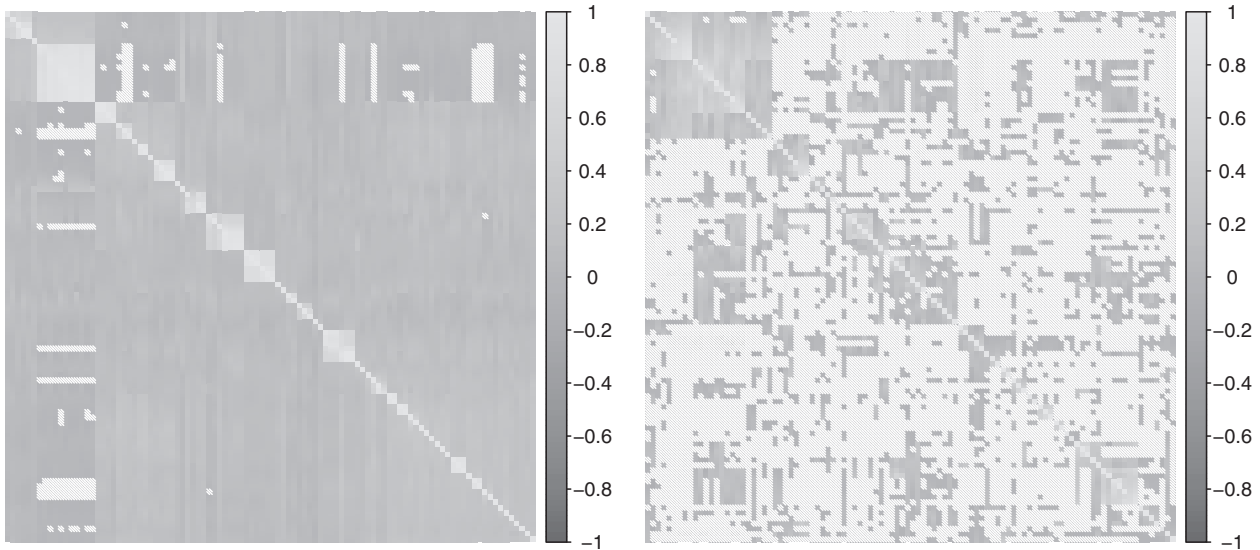


Fig. 1. Sample covariance matrices of the 100 random SNPs (right) and 100 samples (left) in the genotype matrix shows the complex dependence structure present in the *S.pneumoniae* Maela data

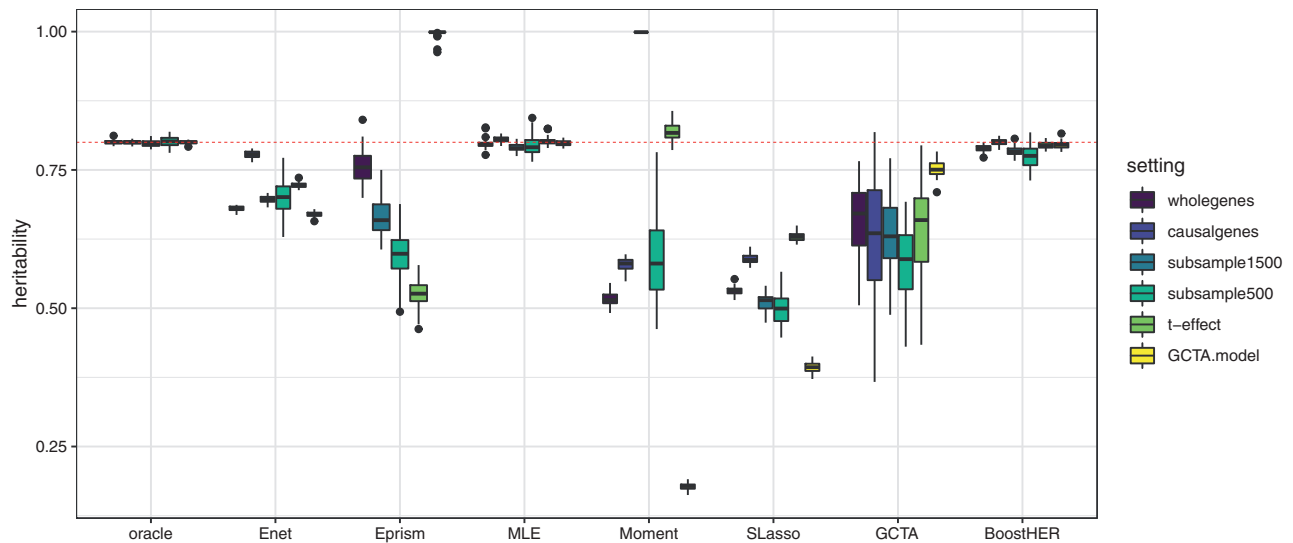


Fig. 2. Simulation results over 50 data replicates with Penicillin-like setting in Maela data, the true heritability is 0.8 (red-dashed line). Settings: 'wholegenomes': run on whole genomes, 'causalgenes': run only with true causal genes, 'subsample1500': run with 1500 randomly selected samples, 'subsample500': run with 500 randomly selected samples, 't-effect': the effect sizes are simulated from Student t_3 distribution, 'GCTA.model': the phenotypes are simulated from the GCTA model

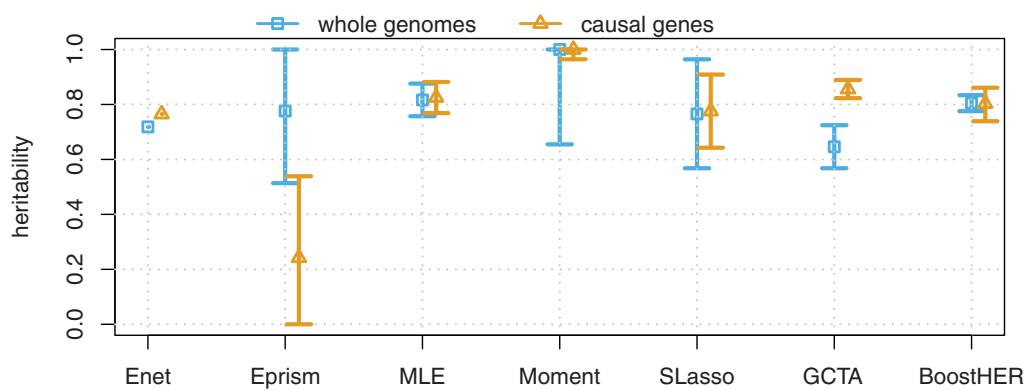


Fig. 3. Heritability estimation of antibiotic resistance in MA data

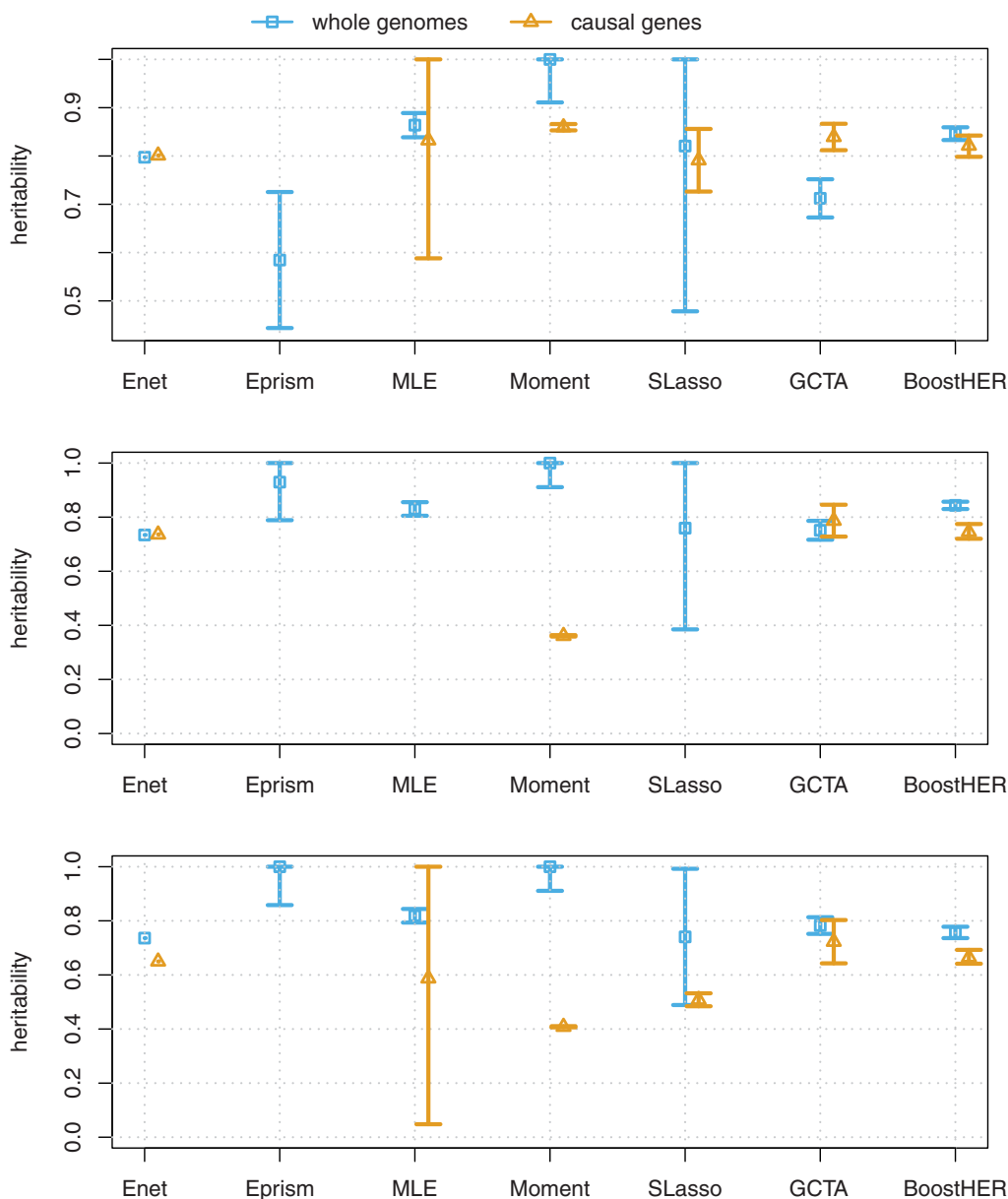


Fig. 4. Heritability estimations of antibiotic resistances in Maela data. The top plot is for Penicillin, the middle is for Tetracycline and the bottom plot is for Co-trimoxazole

convex optimization, is also observed to be quite unstable in its performance.

On the other hand, both the maximum likelihood estimation (MLE) and Boosting methods consistently provide accurate approximations of the target heritability for bacterial genome data across all the simulation settings that were considered. The GCTA method, however, is observed to consistently underestimate the target heritability, with a high degree of variability. This may be due to the fact that the GCTA method was specifically designed for mixed-effect models with different underlying assumptions about the effect sizes.

It has been observed that using the Elastic net (Enet) method to estimate heritability provides a reliable lower bound for the heritability and serves as a benchmark for comparison. However, studies such as Qian *et al.* (2020) and Mai *et al.* (2021) have shown that Enet tends to underestimate the true heritability due to a bias known to affect Lasso-type approaches. This is because coefficients associated with weak effects are shifted toward zero, even though these weak effects may still play a significant role in the overall genetic

variation of a trait. On the other hand, the scaled Lasso (SLasso) approach often results in even lower heritability estimates than Enet. This is because it only selects a small number of covariates, resulting in a model that can only explain a limited amount of variation in the phenotype.

4.2 Results with real data

The results on real datasets are given in Figures 3, 4 and Tables 2, 3.

Overall, we see that direct approaches like convex optimization (Eprism) and moment method (Moment) are not able to deal with these data and return unstable results quite often. On the other hand, the maximum likelihood method (MLE) yields consistent results in line with the outcome from Elastic Net (Enet) and GCTA methods.

The approach by combining selection and estimation steps together with multiple sample splitting as done in the Boosting heritability (BoostHer) method always returns reliable results. More

Table 4. Running times of different methods on four real datasets in seconds

Data	Enet	Eprism	MLE	Moment	SLasso	GCTA	BoostHER
MA	26.1	38.2	0.8	162.7	10	0.3	82.0
Maela	171.6	288.4	16.8	323.8	86.5	17.6	410.7
<i>E.coli</i>	82.8	68.8	2.1	477.3	33.8	0.8	119.9
NG	18.1	33.4	2.6	8.7	5.6	3.4	64.4

specifically, its results are always at a higher value than the lower bounds given by the Enet method, and it can still work well when either MLE or GCTA method fails, as seen from Table 2.

We also consider the estimation with respect to the causal genes for MA and Maela data. We observe that for MA data the estimation of heritability with only causal genes is slightly improved, see Figure 3. While considering causal genes for Maela data does not gain an improvement, and in the case of Co-trimoxazole, it even lower downs the estimation, see Figure 4. Note that this is desirable, since in practice we may not know all the causal genes, and thus we would like to obtain good results from using all predictors.

Regarding the uncertainty quantification, we can see that the CIs of MLE and the ‘reliable’ intervals of BoostHer are stable. More particularly, their widths are similar to those from the GCTA method. In contrast, other methods come with wider intervals and thus they can be harder to interpret. As an example, the CIs for penicillin resistance heritability in Maela data (Fig. 4) are as follows: the width for CI of GCTA is 7.91%; of MLE is 5.02%; and of BoostHer is 2.64%; while the width of CI of Eprism is 28.16% and of SLasso is 52.16%. Since heritability is between 0 and 1, the latter two intervals are of limited value for interpretation.

4.2.1 Running time

The indicative running times of all considered methods on four tested datasets are given in Table 4. The codes were executed on a Linux Redhat 64-bit operating system using a CPU with Intel-E7-4850v3 processor and 3TB of RAM, with the splitting step utilizing 10 CPU cores for parallelization. Overall, the maximum likelihood (MLE) method is the fastest method and also returns trustworthy results. The moment method seems to be computationally expensive, while its results are highly unstable and unreliable.

5 Conclusions

In this study, we have conducted a thorough examination of multiple techniques for estimating heritability in bacteria, focusing on their precision and calibration of uncertainty. We have compared a diverse set of methods, including both traditional and newer techniques. Our findings revealed that, as anticipated, the maximum likelihood method is the fastest of all the methods we evaluated, while the method of moments is generally the slowest. However, it is important to note that none of the methods we tested had running times that would be considered prohibitive for practical use.

In our simulations, the maximum likelihood method demonstrated consistently strong performance. However, when applied to real data cases, its behavior was more mixed. This is likely caused by sensitivity to model assumptions, which tend to lead to a lack of robustness of MLE in general when the data deviate from these assumptions.

Our analysis also revealed that certain methods, such as Eprism, method of moments and SLasso, consistently displayed poor performance and are not recommended for estimating heritability in bacteria. Conversely, the multiple sample splitting technique used in the BoostHer method emerged as the most reliable and accurate approach in our experiments. Overall, our findings provide useful insights for researchers and practitioners looking to determine heritability in bacteria.

It is important to note that our findings have implications for the estimation of heritability in other organisms as well. The patterns of linkage disequilibrium (LD), population structure and existing studies of heritability are all quite different in humans and other organisms compared to bacteria. Therefore, it is crucial to consider the unique characteristics of the organism and the data when choosing a method for heritability inference. Furthermore, it is worth noting that the results of our computational experiments may not generalize to all possible scenarios and further research is needed to fully understand the applicability of these methods in different contexts.

Another investigation of heritability estimation for bacteria recently appeared, where linear mixed models were compared with the Elastic net and LD-score regression (Mallawaarachchi et al., 2022). The study found that linear mixed models showed poor correlation with the ground truth and typically overestimated heritability to a large degree, while Elastic net and LD-score regression methods were found to perform well. This observation is consistent with our findings, where we found that multiple sample splitting as in BoostHer appears overall as the most reliable and accurate approach in our experiments. It is worth noting that this recent study highlights the need for further research in the field of heritability estimation for bacteria, as it is clear that there is a lack of consensus on the best approach to use. The combined results of our study and the aforementioned study call for more research in the field of heritability estimation in bacteria to facilitate future studies of genetic architectures in bacteria.

Acknowledgements

The authors thank the associate editor and three anonymous referees who kindly reviewed the earlier version of this manuscript and provided valuable suggestions and enlightening comments. J.A.L. acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union.

Author contributions

The Tien Mai (Conceptualization [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Resources [equal], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), John A. Lees (Data curation [equal], Investigation [supporting], Resources [equal], Writing—original draft [equal], Writing—review & editing [equal]), Rebecca A. Gladstone (Data curation [Lead], Resources [equal]), and Jukka Corander (Conceptualization [supporting], Funding acquisition [lead], Project administration [lead], Supervision [lead], Validation [equal], Writing—original draft [supporting], Writing—review & editing [equal]). All the authors read and approved the final manuscript.

Funding

This work was supported by the European Research Council [SCARABEE, 742158]; the Norwegian Research Council [309960] through the Centre for Geophysical Forecasting at NTNU to T.T.M.

Conflict of Interest: The authors declare no conflict of interest.

Data availability

The data underlying this article are available in the article.

References

- Bonnet,A. (2016) Heritability estimation in high-dimensional mixed models: theory and applications. PhD Thesis, Universite Paris, Saclay.
- Brzyski,D. *et al.* (2017) Controlling the rate of GWAS false discoveries. *Genetics*, **205**, 61–75.
- Bulik-Sullivan,B.K. *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.
- Chewapreecha,C. *et al.* (2014) Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.*, **10**, e1004547.
- Croucher,N.J. *et al.* (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.*, **45**, 656–663.
- Croucher,N.J. *et al.* (2015) Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Sci. Data*, **2**, 150058.
- Dewé,T.C. *et al.* (2019) Genomic epidemiology of penicillin-non-susceptible *Streptococcus pneumoniae*. *Microbial Genomics*, **5**, e000305.
- Dicker,L.H. (2014) Variance estimation in high-dimensional linear models. *Biometrika*, **101**, 269–284.
- Dicker,L.H. and Erdogdu,M.A. (2016) Maximum likelihood for variance estimation in high-dimensional linear models. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, PMLR, Cadiz, Spain, Vol. 51, pp. 159–167.
- Earle,S.G. *et al.* (2016) Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.*, **1**, 1–8.
- Evans,L.M. *et al.*; Haplotype Reference Consortium. (2018) Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.*, **50**, 737–745.
- Falconer,D.S. (1960) *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh, London.
- Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **70**, 849–911.
- Golan,D. *et al.* (2014) Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA*, **111**, E5272–E5281.
- Gorfine,M. *et al.* (2017) Heritability estimation using a regularized regression approach (HERRA): applicable to continuous, dichotomous or age-at-onset outcome. *PLoS One*, **12**, e0181269.
- Grad,Y.H. *et al.* (2016) Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000–2013. *J. Infect. Dis.*, **214**, 1579–1587.
- Holmes,J.B. *et al.* (2019) Summary statistic analyses can mistake confounding bias for heritability. *Genet. Epidemiol.*, **43**, 930–940.
- Janson,L. *et al.* (2017) Eigenprism: inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **79**, 1037–1065.
- Kallonen,T. *et al.* (2017) Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of st131. *Genome Res.*, **27**, 1437–1449.
- Lee,J.J. *et al.* (2018) The accuracy of LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genet. Epidemiol.*, **42**, 783–795.
- Lees,J.A. *et al.* (2016) Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.*, **7**, 12797.
- Lees,J.A. *et al.* (2017) Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife*, **6**, e26255.
- Lees,J.A. *et al.* (2020) Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio*, **11**, e01344–20.
- Li,X. *et al.* (2019) Reliable heritability estimation using sparse regularization in ultrahigh dimensional genome-wide association studies. *BMC Bioinformatics*, **20**, 219.
- Lynch,M. and Walsh,B. (1998) *Genetics and Analysis of Quantitative Traits*, Vol. 1. Sinauer, Sunderland, MA.
- Mai,T.T. (2021) Boosting heritability: estimating the genetic component of phenotypic variation with multiple sample splitting. *BMC Bioinformatics*, **22**, 1–16.
- Mallawaarachchi,S. *et al.* (2022) Genome-wide association, prediction and heritability in bacteria with application to *Streptococcus pneumoniae*. *NAR Genomics Bioinformatics*, **4**, lqac011.
- Maskell,J.P. *et al.* (2001) Multiple mutations modulate the function of dihydrofolate reductase in trimethoprim-resistant *Streptococcus pneumoniae*. *Antimicrob. Agents Chemother.*, **45**, 1104–1108.
- Qian,J. *et al.* (2020) A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.*, **16**, e1009141.
- Reid,S. *et al.* (2016) A study of error variance estimation in lasso regression. *Stat. Sin.*, **26**, 35–67.
- Schubert,B. *et al.* (2019) Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nat. Microbiol.*, **4**, 328–338.
- Speed,D. and Balding,D.J. (2019) SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.*, **51**, 277–284.
- Speed,D. *et al.* (2012) Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.*, **91**, 1011–1021.
- Speed,D. *et al.*; UCLEB Consortium. (2017) Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*, **49**, 986–992.
- Sun,T. and Zhang,C.H. (2012) Scaled sparse linear regression. *Biometrika*, **99**, 879–898.
- Unemo,M. *et al.* (2016) The novel 2016 who *Neisseria gonorrhoeae* reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. *J. Antimicrob. Chemother.*, **71**, 3096–3108.
- Verzelen,N. and Gassiat,E. (2018) Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, **24**, 3683–3710.
- Weissbrod,O. *et al.* (2018) Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet.*, **103**, 89–99.
- Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.
- Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.
- Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
- Zhou,X. (2017) A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann. Appl. Stat.*, **11**, 2027–2051.
- Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.