Fine-mapping from summary data with the "Sum of Single Effects" model

Yuxin Zou[1], Peter Carbonetto[2,3], Gao Wang[4*], Matthew Stephens[1,2*]

1 Department of Statistics, University of Chicago, Chicago, IL, USA
2 Department of Human Genetics, University of Chicago, Chicago, IL, USA
3 Research Computing Center, University of Chicago, Chicago, IL, USA
4 Department of Neurology and the Gertrude. H. Sergievsky Center, Columbia University, New York, NY, USA

* wang.gao@columbia.edu
* mstephens@uchicago.edu

Abstract

In recent work, Wang *et al* introduced the "Sum of Single Effects" (SuSiE) model, and showed that it provides a simple and efficient approach to fine-mapping genetic variants from individual-level data. Here we present new methods for fitting the SuSiE model to summary data, for example to single-SNP z-scores from an association study and linkage disequilibrium (LD) values estimated from a suitable reference panel. To achieve this we introduce a simple strategy that could be used to extend *any* individual-level data method to deal with summary data. In essence, this strategy replaces the usual regression likelihood with an analogous likelihood based on summary data, exploiting the close connection between the two. Our strategy also has the benefit of dealing automatically with non-invertible LD matrices, which arise frequently in fine-mapping applications, and can complicate inference. We highlight other common practical issues in fine-mapping with summary data, including problems caused by inconsistencies between the z-scores and LD estimates, and we develop diagnostics to identify these inconsistencies. We also present a new refinement procedure that improves model fits in some data sets, and hence improves overall reliability of the SuSiE fine-mapping results. Simulation studies show that SuSiE applied to summary data is competitive, in both speed and accuracy, with the best available fine-mapping methods for summary data.

Introduction

Fine-mapping is the process of narrowing down genetic association signals to a small number of potential causal variants [1–4], and it plays an important part in the effort to understand the genetic causes of diseases [5, 6]. However, fine-mapping is a difficult problem due to the strong and complex correlation patterns ("linkage disequilibrium", or LD) that exist among nearby genetic variants. Many different methods and algorithms have been developed to tackle the fine-mapping problem [2, 7–19]. In recent work, Wang *et al* [17] introduced a new approach to fine-mapping, *SuSiE* (short for "SUm of SIngle Effects"), which has several advantages over existing approaches: it is more computationally scalable; and it provides a new, simple way to calculate "credible sets" of putative causal variants [2, 20]. However, the algorithms in [17] also have an important limitation: they require individual-level genotype and phenotype data. In

contrast, many other fine-mapping methods require only access to summary data, such as z-scores from single-SNP association analyses and an estimate of LD patterns from a suitable reference panel [7, 8, 11–13, 15, 16]. This is useful because individual-level data are often difficult to obtain, both for practical reasons, such as the need to obtain many data sets collected by many different researchers, and for reasons to do with consent and privacy. On the other hand, summary data are much easier to obtain, and many publications share such summary data [21].

In this paper, we introduce new variants of *SuSiE* for performing fine-mapping from summary data. First, we describe a variant, *SuSiE-suff*, that works with a particular type of summary data—*sufficient statistics* (explained below)—and yields exactly the same results as applying *SuSiE* to individual-level data. The limitation of *SuSiE-suff* is that the sufficient statistics are often not fully available. Therefore, we introduce another variant, *SuSiE-RSS*, which does not exactly reproduce the results of applying *SuSiE* to individual-level data, but has the advantage that it works with types of summary data that are more commonly available, namely z-scores from single-SNP association tests and LD estimates from suitable reference panels. These summary statistics are frequently made available when a genome-wide association study (GWAS) is published, so *SuSiE-RSS* greatly expands the potential for the *SuSiE* fine-mapping approach to be applied to existing genetic association studies.

A more general contribution of our work is to highlight and clarify the close connection between individual-level data methods and summary data methods. This close connection arises because, as we show here, the models usually used for summary data yield a likelihood, which we call the *RSS-Z* likelihood (14), that is very similar in form to the likelihood from an individual-level regression model. Here we exploit this connection to develop a single algorithm that can fit the *SuSiE* model to both individual-level data and summary data. However, this connection between the individual-level and summary data likelihoods is quite general, and not specific to the *SuSiE* model. Our work therefore provides a template that could be used to extend *any* individual-data method to work with summary data. Our approach also automatically deals with non-invertible LD matrices, which arise frequently in fine-mapping, and we argue, through both theory and example, that it provides a simpler and better solution to this issue than some existing approaches.

In addition to extending *SuSiE* to work with summary data, we introduce several other methodological innovations. Some of these innovations are not specific to *SuSiE* and could be used to adapt other statistical methods for individual-level data to deal with summary data. We describe methods for identifying "allele flips"—alleles that are (erroneously) encoded differently in the study and reference data—and other inconsistencies in the summary data. We illustrate through an example how a single allele flip can lead to inaccurate fine-mapping results, emphasizing the importance of careful quality control when performing fine-mapping using summary data. We also introduce a new refinement procedure for *SuSiE* that sometimes improves the estimates from the original fitting procedure.

Methods

We begin with some background and notation. Let $\boldsymbol{y} \in \mathbb{R}^N$ denote the phenotypes of N individuals in a genetic association study, and let $\boldsymbol{X} \in \mathbb{R}^{N \times J}$ denote their corresponding genotypes at J genetic variants (SNPs). To simplify the presentation, we assume the $\boldsymbol{y}$ are quantitative and approximately normally distributed, and that both $\boldsymbol{y}$ and the columns of $\boldsymbol{X}$ are centered to have mean zero, which avoids the need for an intercept term in (1) below [22]. We elaborate on treatment of binary and case-control phenotypes in the Discussion below.

Fine-mapping is usually performed by fitting the multiple linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{b} = (b_1, \ldots, b_J)^\intercal$ is a vector of multiple regression coefficients, $\boldsymbol{e}$ is an N-vector of error terms distributed as $\boldsymbol{e} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$, with (typically unknown) residual variance $\sigma^2 > 0$, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and $\mathcal{N}_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the r-variate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

In this multiple regression framework, the question of which SNPs are affecting $\boldsymbol{y}$ becomes a problem of "variable selection"; that is, identifying which elements of $\boldsymbol{b}$ are not zero. While many methods exist for variable selection in multiple regression, fine-mapping has some special features—typically very high correlations among some columns of $\boldsymbol{X}$ and very sparse $\boldsymbol{b}$—that make Bayesian methods with sparse priors a preferred approach (e.g., [7–9]). These methods specify a sparse prior for $\boldsymbol{b}$, and perform inference by approximating the posterior distribution $p(\boldsymbol{b} \mid \boldsymbol{X}, \boldsymbol{y})$. In particular, the evidence for SNP j having a non-zero effect is often summarized by the "posterior inclusion probability" (PIP),

$$\mathrm{PIP}_j := \Pr(b_j \neq 0 \mid \boldsymbol{X}, \boldsymbol{y}). \tag{2}$$

The Sum of Single Effects (*SuSiE*) model

The key idea behind *SuSiE* [17] is to write $\boldsymbol{b}$ as a sum,

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l, \tag{3}$$

in which each vector $\boldsymbol{b}_l = (b_{l1}, \ldots, b_{lJ})^\intercal$ is a "single effect" vector; that is, a vector with exactly one non-zero element. The representation (3) allows that $\boldsymbol{b}$ has at most L non-zero elements, where L is a user-specified upper bound on the number of effects. (For example if single-effect vectors $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ have non-zero element at the same SNP j, $\boldsymbol{b}$ will have fewer than L non-zeros.)

The special case $L = 1$ corresponds to the assumption that a region has exactly one causal SNP (which we use as shorthand for a SNP with a non-zero effect); in [17] this special case is called the "single effect regression" (SER) model. The SER is particularly convenient because posterior computations are analytically tractible [9]; consequently, despite its clear limitations, the SER has been widely used [2, 23–25].

For $L > 1$, Wang *et al* [17] introduced a simple model fitting algorithm, which they called Iterative Bayesian Stepwise Selection (IBSS). In brief, IBSS iterates through the single-effect vectors $l = 1, \ldots, L$, at each iteration fitting $\boldsymbol{b}_l$ while keeping the other single-effect vectors fixed. By construction, each step thus involves fitting an SER, which, as noted above, is straightforward. Wang *et al* [17] show that IBSS can be understood as computing an approximate posterior distribution $p(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2)$, and that the algorithm iteratively optimizes an objective function known as the "evidence lower bound" (ELBO).

Summary data for fine-mapping

Motivated by the difficulties in accessing the individual-level data $\boldsymbol{X}, \boldsymbol{y}$ from most studies, researchers have developed fine-mapping approaches that work with more available "summary data". Here we develop methods that use various combinations of the following summary data:

(i) Vectors $\hat{\boldsymbol{b}} = (\hat{b}_1, \ldots, \hat{b}_J)^\intercal$ and $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_J)^\intercal$ containing estimates of marginal association for each SNP j, and corresponding standard errors, from a simple linear regression:

$$\hat{b}_j := \boldsymbol{x}_j^\intercal \boldsymbol{y} / (\boldsymbol{x}_j^\intercal \boldsymbol{x}_j), \tag{4}$$

$$\hat{s}_j := [(\boldsymbol{y} - \boldsymbol{x}_j \hat{b}_j)^\intercal (\boldsymbol{y} - \boldsymbol{x}_j \hat{b}_j) / (N \boldsymbol{x}_j^\intercal \boldsymbol{x}_j)]^{1/2}. \tag{5}$$

An alternative to $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}$ is the vector $\hat{\boldsymbol{z}} = (\hat{z}_1, \ldots, \hat{z}_J)^\intercal$ of z-scores:

$$\hat{z}_j := \hat{b}_j / \hat{s}_j. \tag{6}$$

Many studies provide $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{s}}$ (see [21] for examples), and many more provide the z-scores, or data that can be used to compute the z-scores (e.g., $\hat{z}_j$, can be recovered from the p-value and the sign of $\hat{b}_j$ [26]).

(ii) An estimate, $\hat{\boldsymbol{R}}$, of the in-sample LD matrix, $\boldsymbol{R}$. To be precise, $\boldsymbol{R}$ is the $J \times J$ SNP-by-SNP sample correlation matrix,

$$\boldsymbol{R} := \boldsymbol{D}_X^{-1/2} \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{D}_X^{-1/2} \tag{7}$$

where $\boldsymbol{D}_X := \mathrm{diag}(\boldsymbol{X}^\intercal \boldsymbol{X})$ is a diagonal matrix that ensures the diagonal entries of $\boldsymbol{R}$ are all 1. Usually the estimate $\hat{\boldsymbol{R}}$ is taken to be an "out-of-sample" LD matrix—that is, the sample correlation matrix of the same J SNPs in a suitable reference panel, chosen to be genetically similar to the study population, possibly with additional shrinkage or banding steps to improve accuracy [14].

(iii) Optionally, the sample size N and the sample variance of $\boldsymbol{y}$. (Since $\boldsymbol{y}$ is centered, the sample variance of $\boldsymbol{y}$ is simply $\boldsymbol{y}^\intercal \boldsymbol{y}/N$). Knowing these quantities is obviously equivalent to knowing $\boldsymbol{y}^\intercal \boldsymbol{y}$ and N, so for brevity we will use the latter. These quantities are not required, but they can be helpful as we will see later.

SuSiE with summary data

We now describe methods to fit the *SuSiE* model (1–3) using summary data in various forms.

First, we consider a special type of summary data, called "sufficient statistics," which contain exactly the same information as the individual-level data $\boldsymbol{X}, \boldsymbol{y}$. For example, it turns out that the summary data $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, \boldsymbol{R}, \boldsymbol{y}^\intercal \boldsymbol{y}, N$ are sufficient statistics (and crucially, $\boldsymbol{R}$ must be the in-sample LD matrix, and not an estimate of it). For these sufficient statistics, we describe an algorithm, *SuSiE-suff*, that *exactly* recapitulates the results that would be obtained by running *SuSiE* on the original data $\boldsymbol{X}, \boldsymbol{y}$.

Second, we consider the case where we have access to summary data—for example, $\hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}$—but not sufficient statistics. In this case, it is not possible to exactly recapitulate the results that would be obtained from the individual-level data $\boldsymbol{X}, \boldsymbol{y}$, but approximations are possible. Building on *SuSiE-suff*, we develop a second method, *SuSiE-RSS*, for summary data that are not sufficient statistics.

SuSiE with sufficient statistics: *SuSiE-suff*

The IBSS algorithm of [17] fits the *SuSiE* model to individual-level data $\boldsymbol{X}, \boldsymbol{y}$. The data enter the *SuSiE* model only through the likelihood, which from (1) is

$$\ell(\boldsymbol{b}, \sigma^2; \boldsymbol{X}, \boldsymbol{y}) = (2\pi\sigma^2)^{-N/2} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{b}^\intercal \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{b}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{b}) \right\}. \tag{8}$$

This likelihood clearly depends on the data only through $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, N$, and so these are sufficient statistics. Indeed, careful inspection of the IBSS algorithm in [17] confirms that it depends on the data only through these sufficient statistics. Thus, by rearranging the computations we can obtain a variant of IBSS that fits the *SuSiE* model using these sufficient statistics; see Algorithm 1 in Detailed Methods. We call this method *SuSiE-suff*.

It is easy to show that the sufficient statistics $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, N$ can be computed from the more familiar quantities $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, \boldsymbol{R}, \boldsymbol{y}^\intercal\boldsymbol{y}, N$; see Lemma 1 in Detailed Methods. Thus these more familiar quantities are also sufficient, and can be used in *SuSiE-suff*.

When *SuSiE-suff* is applied to sufficient statistics it produces exactly the same result as the original IBSS algorithm applied to $\boldsymbol{X}, \boldsymbol{y}$. However, the computational complexity of the two algorithms is different: whereas the original algorithm requires $O(NJ)$ operations per iteration, *SuSiE-suff* requires $O(J^2)$ operations per iteration. (The number of iterations should be the same.) When $N \gg J$, which is often the case in fine-mapping applications, *SuSiE-suff* will usually be faster. However, computing the $J \times J$ matrix $\boldsymbol{R}$ (or $\boldsymbol{X}^\intercal\boldsymbol{X}$) — which is required for *SuSiE-suff* but not for the original algorithm — is expensive, requiring $O(NJ^2)$ operations. (We note that convenient and efficient software exists to do this computation, including PLINK [27] and LDstore [28].) In practice then choosing between *SuSiE* and *SuSiE-suff* may depend on one's preferred workflow—whether or not one prefers to precompute $\boldsymbol{R}$.

Note that if σ^2 is fixed then the likelihood (8), as a function of $\boldsymbol{b}$, depends on the data $\boldsymbol{X}, \boldsymbol{y}$ only through $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}$. So if σ^2 is fixed then *SuSiE-suff* requires only $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}$ to produce the same results as *SuSiE*.

In what follows it is helpful to make the dependence of the likelihood on the sufficient statistics explicit. We define

$$\ell_{\text{suff}}(\boldsymbol{b}, \sigma^2; \boldsymbol{S}_{xx}, \boldsymbol{s}_{xy}, s_{yy}, N) := (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}(s_{yy} - 2\boldsymbol{b}^\intercal\boldsymbol{s}_{xy} + \boldsymbol{b}^\intercal\boldsymbol{S}_{xx}\boldsymbol{b})\right\}. \quad (9)$$

so that the two likelihoods are the same so long as the sufficient statistics $\boldsymbol{S}_{xx}, \boldsymbol{s}_{xy}, s_{yy}$ are obtained correctly from $\boldsymbol{X}, \boldsymbol{y}$:

$$\ell_{\text{suff}}(\boldsymbol{b}, \sigma^2; \boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, N) = \ell(\boldsymbol{b}, \sigma^2; \boldsymbol{X}, \boldsymbol{y}).$$

SuSiE with non-sufficient summary statistics: *SuSiE-RSS*

We now describe an extension of *SuSiE* for non-sufficient summary data, specifically for $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{R}}$, which are the single-SNP z-scores (6) and an estimated LD matrix. Our strategy for doing this is quite general, and could be used to adapt *any* multiple regression method for individual-level data $\boldsymbol{X}, \boldsymbol{y}$ to work for $\hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}$. The approach is also easily adapted to work with $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{R}}$ instead; see Detailed Methods.

Following several existing fine-mapping methods (e.g., [7, 8, 12, 16]), we use the following model:

$$\hat{\boldsymbol{z}} \mid \boldsymbol{z}, \hat{\boldsymbol{R}} \sim \mathcal{N}_J(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}}), \quad (10)$$

where $\boldsymbol{z} = (z_1, \ldots, z_J)^\intercal$ is the unobserved vector of scaled effects, also called the noncentrality parameters (NCPs),

$$z_j := \frac{b_j\sqrt{\boldsymbol{x}_j^\intercal\boldsymbol{x}_j}}{\sigma}. \quad (11)$$

We refer to (10) as the "*RSS-Z* model"; the RSS stands for "regression with summary statistics," as in [29], and the Z indicates that this is a model for z-scores.

The *RSS-Z* model can be derived from the multiple regression model (1) with some additional assumptions [7, 29]. In brief, the assumptions are: (i) the correlation of the

response, $\boldsymbol{y}$, with any single variant $\boldsymbol{x}_j$ is small; and (ii) $\hat{\boldsymbol{R}}$ is a good approximation to $\boldsymbol{R}$. It is also important that the same samples are used to compute z_j for each SNP j (using genotype imputation if necessary). See [29] for more details and discussion.

If the matrix $\hat{\boldsymbol{R}}$ is invertible, the *RSS-Z* model (10) has a probability density,

$$p(\hat{\boldsymbol{z}} \mid \boldsymbol{z}, \hat{\boldsymbol{R}}) = |2\pi\hat{\boldsymbol{R}}|^{-1/2} \exp\{-\tfrac{1}{2}(\hat{\boldsymbol{z}} - \hat{\boldsymbol{R}}\boldsymbol{z})^\intercal \hat{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{z}} - \hat{\boldsymbol{R}}\boldsymbol{z})\}. \tag{12}$$

The complication is that $\hat{\boldsymbol{R}}$ is frequently not invertible in fine-mapping studies. For example, if $\hat{\boldsymbol{R}}$ is the sample correlation matrix from a reference panel, then $\hat{\boldsymbol{R}}$ will be non-invertible if the number of individuals in the panel is less than J, or if any two SNPs are in complete LD in the panel. In such cases, the *RSS-Z* model does not have a density (with respect to the Lebesgue measure). Methods using (10) have therefore required workarounds to deal with this issue. One workaround is to modify $\hat{\boldsymbol{R}}$ to be invertible by adding a small, positive constant to the diagonal [7]. In another approach, the data are transformed into a lower-dimensional space [30, 31]. This is equivalent to replacing $\hat{\boldsymbol{R}}^{-1}$ in (12) with the pseudoinverse, or Moore-Penrose inverse, of $\hat{\boldsymbol{R}}$ (Detailed Methods).

Here we take a simpler alternative approach, which, as far as we know has not been previously used in this setting. Our approach is based on the fact that, when (12) is considered as a function of $\boldsymbol{z}$, the inverse of $\hat{\boldsymbol{R}}$ affects only the constant of proportionality. That is,

$$p(\hat{\boldsymbol{z}} \mid \boldsymbol{z}, \hat{\boldsymbol{R}}) = \text{constant} \times \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}\boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}) \tag{13}$$

where the constant does not depend on $\boldsymbol{z}$. Thus, inference under the *RSS-Z* model is equivalent to using the following likelihood:

$$\ell_{\textit{RSS-Z}}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) := \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}\boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}). \tag{14}$$

Since (14) is well-defined even for non-invertible $\hat{\boldsymbol{R}}$, we can use it as a likelihood for $\boldsymbol{z}$ even when $\hat{\boldsymbol{R}}$ is non-invertible. Mathematically, this can be justified by considering a sequence of invertible matrices $\hat{\boldsymbol{R}}_\lambda$ such that $\lim_{\lambda\to 0} \hat{\boldsymbol{R}}_\lambda = \hat{\boldsymbol{R}}$, and noting that $\lim_{\lambda\to 0} \ell_{\textit{RSS-Z}}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) \to \ell_{\textit{RSS-Z}}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}})$; see also Proposition 2 in Detailed Methods. (This justification requires $\hat{\boldsymbol{R}}$ be positive semi-definite, which is true for any empirical LD matrix computed from a panel, and is desirable in any case because otherwise the likelihood (14) can be unbounded.) This approach is different from approaches that modify $\hat{\boldsymbol{R}}$ or $\hat{\boldsymbol{R}}^{-1}$ (Detailed Methods), and has some advantages: it is simpler; it is computationally more attractive because it does not involve an expensive inversion or factorization of a (possibly very large) $J \times J$ matrix; and it preserves the property that results under the SER model do not depend on LD (see Results, and Remark 5 in Detailed Methods). Also note that this approach can be combined with modifications to $\hat{\boldsymbol{R}}$, such as the LD regularization approaches discussed below.

The final step of our approach is to connect the summary data likelihood $\ell_{\textit{RSS-Z}}$ (14) with the full data likelihood ℓ_{suff} (9). Indeed, $\ell_{\textit{RSS-Z}}$ is a special case of ℓ_{suff}: making the substitutions $\boldsymbol{S}_{xx} \leftarrow \hat{\boldsymbol{R}}$, $\boldsymbol{s}_{xy} \leftarrow \hat{\boldsymbol{z}}$, $s_{yy} \leftarrow 1$, $N \leftarrow 1$, $\sigma^2 \leftarrow 1$ and $\boldsymbol{b} \leftarrow \boldsymbol{z}$ in (9) gives

$$\ell_{\textit{RSS-Z}}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) \propto \ell_{\text{suff}}(\boldsymbol{z}, 1; \hat{\boldsymbol{R}}, \hat{\boldsymbol{z}}, 1, 1). \tag{15}$$

(We set $\boldsymbol{y}^\intercal\boldsymbol{y} = N = 1$ for concreteness, but these choice are arbitrary because (15) holds for any settings of $\boldsymbol{y}^\intercal\boldsymbol{y}, N$.) This equivalence of likelihoods implies that *any algorithm that uses sufficient statistics to fit a Bayesian multiple regression model with some prior ϕ for the regression cofficients $\boldsymbol{b}$, can be directly applied to fit the RSS-Z model with the same prior ϕ used for $\boldsymbol{z}$.* Applying this logic to *SuSiE*, we can fit the *RSS-Z* model with

the *SuSiE* prior on $\boldsymbol{z}$ by simply applying the *SuSiE-suff* algorithm with the substitutions used in (15). We call this approach *SuSiE-RSS*.

We emphasize that although *SuSiE-RSS* uses the same algorithm as *SuSiE-suff*, it uses different inputs. Thus, *SuSiE-RSS* will not give the same result as applying *SuSiE-suff* to the sufficient statistics or applying *SuSiE* to the individual data. However, our approach makes explicit the connections between inference from individual data and summary data, and allows a single algorithm, *SuSiE-suff*, to be used to fit both types of data.

New refinement procedure for more accurate CSs

As noted in [17], the IBSS algorithm can sometimes converge to a poor solution (local optimum of the ELBO). Although this is rare, it can produce misleading results when it does occur; in particular it can produce false positive CSs (*i.e.*, CSs in which all the SNPs have true effects that are zero). To address this issue, we developed a simple refinement procedure for escaping local optima. The procedure is heuristic, and not guaranteed to eliminate all convergence issues, but in practice it often helps in those rare cases where the original IBSS has problems. The refinement procedure applies equally to both individual-level data and summary data.

In brief, the refinement procedure involves two steps: first, fit a *SuSiE* model by running the IBSS algorithm to convergence; second, for each CS identified from the fitted *SuSiE* model, refine the fit by rerunning IBSS after first removing all SNPs in the CS (which forces the algorithm to seek alternative explanations for observed associations), then refine the fit one more time, again using IBSS, but this time with all SNPs. If these refinement steps improve the objective function, the new solution is accepted; otherwise, the original solution is kept. This process is repeated until the refinement steps no longer make any improvements to the objective. By construction, this refinement procedure always produces a solution whose objective is at least as good as the original IBSS solution. For more details, see Detailed Methods.

Because the refinement procedure reruns IBSS for each CS discovered in the initial round of model fitting, the computation increases with the number of CSs found. In data sets with many CSs the refinement procedure may be quite time-consuming.

Other improvements to fine-mapping with summary data

Here we introduce additional methods to improve accuracy of fine-mapping with summary data. These methods can be applied to other fine-mapping methods as well as *SuSiE*.

Regularization to improve consistency of the estimated LD matrix

Accurate fine-mapping requires $\hat{\boldsymbol{R}}$ to be an accurate estimate of $\boldsymbol{R}$. When $\hat{\boldsymbol{R}}$ is computed from a reference panel, the reference panel should not be too small [28], and should be of similar ancestry to the study sample. Even when a suitable panel is used, there will inevitably be differences between $\hat{\boldsymbol{R}}$ and $\boldsymbol{R}$. A common way to improve estimation of covariance matrices is to use regularization [32], replacing $\hat{\boldsymbol{R}}$ with $\hat{\boldsymbol{R}}_\lambda$,

$$\hat{\boldsymbol{R}}_\lambda := (1-\lambda)\hat{\boldsymbol{R}}_0 + \lambda \boldsymbol{I}, \tag{16}$$

where $\hat{\boldsymbol{R}}_0$ is the sample correlation matrix computed from the reference panel, and $\lambda \in [0,1]$ controls the amount of regularization. This strategy has previously been used in fine-mapping from summary data (e.g., [8, 33, 34]), but in previous work λ was

usually fixed at some arbitrarily small value, or chosen using cross validation. Here, we estimate λ by maximizing the likelihood under the null ($\boldsymbol{z} = \mathbf{0}$),

$$\hat{\lambda} := \underset{\lambda \in [0,1]}{\operatorname{argmax}} \mathcal{N}(\hat{\boldsymbol{z}}; \mathbf{0}, (1-\lambda)\hat{\boldsymbol{R}}_0 + \lambda \boldsymbol{I}). \tag{17}$$

The estimated $\hat{\lambda}$ reflects the consistency between the observed z-scores, $\hat{\boldsymbol{z}}$, and the LD matrix $\hat{\boldsymbol{R}}_0$; if the two are consistent with one another $\hat{\lambda}$ will be close to zero.

Detecting and removing large inconsistencies in summary data

Regularizing $\hat{\boldsymbol{R}}$ can help address subtle inconsistencies between $\hat{\boldsymbol{R}}$ and $\boldsymbol{R}$. However, regularization cannot adequately deal with large inconsistencies in the summary data, which, in our experience, occur often. One common source of such inconsistencies is an "allele flip," in which the alleles of a SNP are encoded one way in the study sample (used to compute $\hat{\boldsymbol{z}}$), and in a different way in the reference panel (used to compute $\hat{\boldsymbol{R}}$). Large inconsistencies can also arise from using z-scores that were obtained using different samples at different SNPs (which should be avoided by performing genotype imputation [29]). Anecdotally we have found large inconsistencies like these often lead *SuSiE* to converge very slowly, as well as producing misleading results. We have therefore developed diagnostics to help users detect such anomalous data.

Under the *RSS-Z* model (10), the conditional distribution of $\hat{z}_j$ given the other z-scores is

$$\hat{z}_j \mid \hat{\boldsymbol{R}}, \boldsymbol{z}, \hat{\boldsymbol{z}}_{-j} \sim \mathcal{N}((z_j - \boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j})/\Omega_{jj}, 1/\Omega_{jj}), \tag{18}$$

where $\boldsymbol{\Omega} := \hat{\boldsymbol{R}}^{-1}$, $\hat{\boldsymbol{z}}_{-j}$ denotes the vector of observed z-scores excluding $\hat{z}_j$, and $\boldsymbol{\Omega}_{j,-j}$ denotes the jth row of $\boldsymbol{\Omega}$ excluding Ω_{jj}. This conditional distribution depends on the unknown z_j. However, provided that the effect of SNP j is small (*i.e.*, $z_j \approx 0$), or that SNP j is in strong LD with other SNPs, which implies $1/\Omega_{jj} \approx 0$, we can approximate this by

$$\hat{z}_j \mid \hat{\boldsymbol{R}}, \hat{\boldsymbol{z}}_{-j} \sim \mathcal{N}(-\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}/\Omega_{jj}, 1/\Omega_{jj}). \tag{19}$$

An initial quality control check can be performed by plotting the observed $\hat{z}_j$ against its conditional expectation in (19), with large deviations potentially indicating anomalous z-scores. Since computing these conditional expectations involves the inverse of $\hat{\boldsymbol{R}}$, this matrix must be invertible. When $\hat{\boldsymbol{R}}$ is not invertible, we replace $\hat{\boldsymbol{R}}$ with the regularized (and invertible) matrix $\hat{\boldsymbol{R}}_\lambda$ following the steps described above. The distribution (19) has previously been used in z-score imputation [35] and GWAS quality control [36].

A more quantitative measure of the discordance of $\hat{z}_j$ with its expectation under the model can be obtained by computing standardized differences between the observed and expected values,

$$t_j := \sqrt{\Omega_{jj}}(\hat{z}_j + \boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}/\Omega_{jj}). \tag{20}$$

SNPs j with largest (in magnitude) t_j are most likely to violate the *RSS-Z* model assumptions, and are therefore the top candidates for follow-up care. When any such candidates are detected, the user should check the data pre-processing steps and fix any errors that cause inconsistencies in summary data. If there is no way to fix the errors, removing the anomalous SNPs is a possible workaround. Sometimes removing a single SNP is enough to resolve the discrepancies—for example, a single allele flip can result in inconsistent z-scores among many other SNPs in LD with the allele-flip SNP. We have also developed a likelihood-ratio statistic based on (19) specifically for identifying allele flips; see Detailed Methods. After one or more SNPs are removed, one should consider re-running these diagnostics on the filtered summary data to search for additional inconsistencies that may have been missed in the first round.

Computing these diagnostics requires inverting or factorizing a $J \times J$ matrix, and may therefore involve large computational expense—potentially a greater expense than the fine-mapping itself—when J, the number of SNPs, is large.

Results

Fine-mapping with inconsistent summary data and a non-invertible LD matrix: an illustration

A technical issue that arises when developing fine-mapping methods for summary data (z-scores $\hat{\boldsymbol{z}}$ and LD matrix $\hat{\boldsymbol{R}}$) is that the LD matrix is often not invertible. Several approaches to dealing with this have been suggested, including modifying the LD matrix to be invertible, transforming the data into a lower-dimensional space, or replacing the inverse with what is known as the "pseudoinverse." In *SuSiE-RSS* we take a simpler approach: we use the *RSS-Z* likelihood (14) even when $\hat{\boldsymbol{R}}$ is not invertible. We summarize the theoretical relationships between these methods in Proposition 2 (Detailed Methods). Here illustrate the practical advantage of our approach in a toy example.

Consider a very simple situation with two SNPs, in strong LD, with observed z-scores $\hat{\boldsymbol{z}} = (6, 7)$. Both SNPs are significant, but SNP 2 is more significant. Under the assumption that exactly one of these SNPs has an effect—which, as a reminder, allows for exact posterior computatons under the SER model—SNP 2 is the better candidate, and should have a higher PIP. Further, the PIPs should be unaffected by LD between the SNPs (see Remark 5 in Detailed Methods). However, the transformation and pseudoinverse approaches, which are used by msCAVIAR [37] and in previous fine-mapping analyses [30, 31], do not guarantee that either of these properties are satisfied. For example, suppose the two SNPs are in complete LD in the reference panel, so $\hat{\boldsymbol{R}}$ is a 2×2 (non-invertible) matrix with all entries equal to 1. Note that $\hat{\boldsymbol{R}}$ is inconsistent with the observed $\hat{\boldsymbol{z}}$ because complete LD betweeen SNPs implies their z-scores should be identical. (This could happen if the LD in the reference panel used to compute $\hat{\boldsymbol{R}}$ is slightly different from the LD in the association study.) The tranformation approach effectively adjusts the observed data $\hat{\boldsymbol{z}}$ to be consistent with the LD matrix before drawing inferences; here it would adjust $\hat{\boldsymbol{z}}$ to $\hat{\boldsymbol{z}} = (6.5, 6.5)$, removing the observed difference between the SNPs and forcing them to be equally significant, which seems undesirable. The pseudoinverse approach turns out to be equivalent to the tranformation approach (see Detailed methods), and so behaves the same way. In contrast, our approach avoids this behaviour and correctly maintains the second SNP as the better candidate; applying *SuSiE-RSS* to this toy example yields PIPs 0.001 and 0.998, and a single CS containing the second SNP only. To reproduce this result, see the examples accompanying the `susie_rss` function in the `susieR` R package.

Effect of allele flips on accuracy of fine-mapping: an illustration

When fine-mapping is performed using $\hat{\boldsymbol{z}}$ from a study sample and LD matrix $\hat{\boldsymbol{R}}$ from a (different) reference sample, it is crucial that the same alleles encodings are used in each sample. In our experience, "allele flips," in which different allele encoding are used in the two samples, are a common source of problems. Here we use a simple simulation to illustrate this issue, and the steps we have implemented to help identify and correct the problem.

We simulated a fine-mapping data set with 1,002 SNPs, in which one out of the 1,002 SNPs was causal, and we deliberately used different allele encodings in the study sample and reference panel for one of the non-causal SNP; see Detailed Methods for

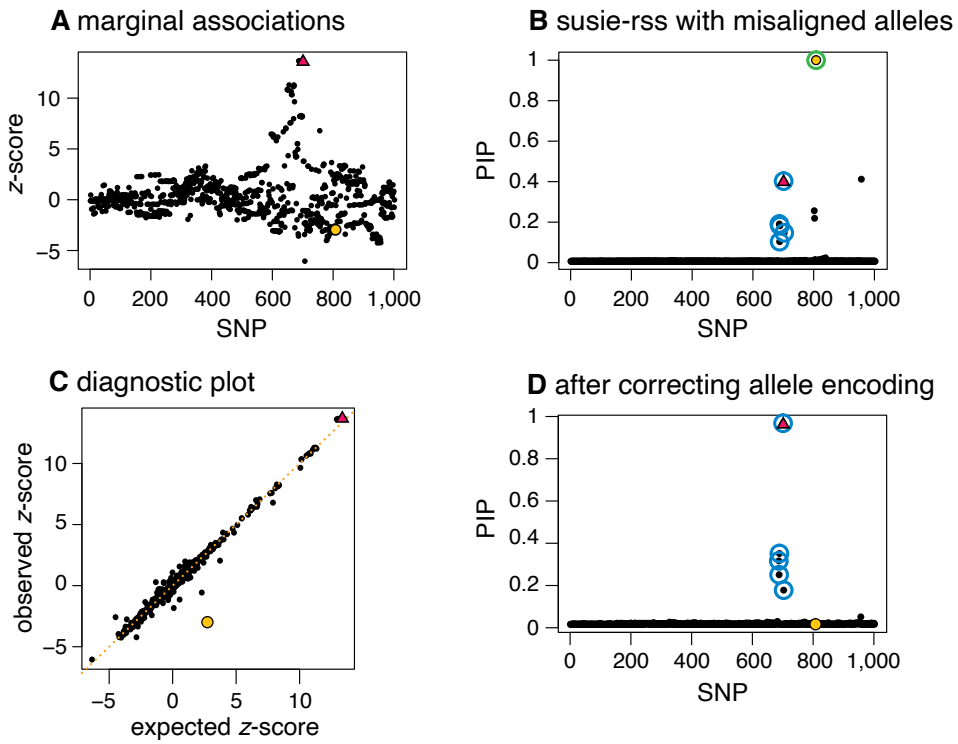

Figure 1. Example illustrating importance of identifying and correcting allele flips in fine-mapping. In this simulated example, one SNP (red triangle) affects the phenotype, and one SNP (yellow circle) has a different allele encoding in the study sample (*i.e.*, the data used to compute the z-scores) and the reference panel (*i.e.*, the data used to compute the LD matrix). Panel A shows the z-scores for all 1,002 SNPs. Panel B summarizes the results of running *SuSiE-RSS* on the summary data: *SuSiE-RSS* identifies a true positive CS (blue) containing the true causal SNP; and a false positive CS (green) that incorrectly contains the mismatched SNP. The mismatched SNP is also incorrectly estimated to have an effect on the phenotype with very high probability (PIP = 1.00). The diagnostic plot (Panel C) compares the observed z-scores against the expected z-scores under the *RSS-Z* model. In this plot, the mismatched SNP (yellow circle) shows the largest difference between observed and expected z-scores, and therefore appears furthest away from the diagonal. After fixing the allele encoding and recomputing the summary data, *SuSiE-RSS* identifies a single true positive CS containing the true-causal SNP (red triangle), and the formerly mismatched SNP is (correctly) not included in a CS (Panel D). This example is implemented as a vignette in the `susieR` package.

more details on this simulation. The causal SNP is among the SNPs with the highest z-scores (the red triangle in Fig. 1, Panel A), and *SuSiE-RSS* correctly includes this causal SNP in a CS (Panel B). However, *SuSiE-RSS* also wrongly includes the allele-flip SNP in a second CS (Panel B). This happens because the LD between the allele-flip SNP and other SNPs is incorrectly estimated. Figure 1, Panel C shows a diagnostic plot comparing each z-score against its expected value under the *RSS-Z* model. The allele-flip SNP stands out as a likely outlier (yellow circle), and our likelihood ratio statistic identifies this SNP as a likely allele flip: $\mathrm{LR} = 8.2 \times 10^3$ for the allele-flip SNP, whereas all of the likelihood ratios are less than 1 among the remaining 262 SNPs with z-scores greater than 2 in magnitude. After correcting the allele encoding to be the same in both study and reference samples, *SuSiE-RSS* infers a single CS containing the causal SNP, and the allele-flip SNP is no longer included in a CS; see Fig. 1, Panel D.

Note that these diagnostic tools are not sufficiently accurate to automatically identify and correct allele flips or other errors. Nonetheless they can be helpful to assist users with identifying errors in computational pipelines.

Simulations using UK Biobank genotypes

To systematically compare our new methods with existing methods for fine-mapping, we simulated fine-mapping data sets using the UK Biobank imputed genotypes [38]. The UK Biobank imputed genotypes are well suited to illustrate fine-mapping with summary data due to the large sample size, and the high density of available genetic variants after imputation. We randomly selected 200 regions on autosomal chromosomes for fine-mapping, such that each region contained roughly 1,000 SNPs (average size: 390 kb). Due to the high density of SNPs, these data sets often contain very strong correlations among SNPs; on average, a data set contained 30 SNPs with correlation exceeding 0.9 with at least one other SNP, and 14 SNPs with correlations exceeding 0.99 with at least one other SNP.

For each of the 200 regions, we simulated a quantitative trait under the multiple regression model (1) with $\boldsymbol{X}$ comprising genotypes of 50,000 randomly selected UK Biobank samples, and with 1, 2 or 3 causal variants explaining a total of 0.5% of variation in the trait. In total, we simulated $200 \times 3 = 600$ data sets. We computed summary data from these real genotypes and the synthetic phenotypes. To compare how choice of LD matrix affects fine-mapping, we used three different LD matrices: the in-sample LD matrix computed from the 50,000 individuals ($\boldsymbol{R}$), and two out-of-sample LD matrices computed from randomly-sampled reference panels of 500 or 1,000 individuals, denoted $\hat{\boldsymbol{R}}_{500}$ and $\hat{\boldsymbol{R}}_{1000}$, respectively. The reference panels were sampled to have no overlap with the $n = 50,000$ study sample, but were sampled from the same population, and so this simulation mimics a situation where the reference sample is well matched to the study sample.

Refining *SuSiE* model fits improves fine-mapping performance

We begin by demonstrating the benefits of our new refinement procedure for improving *SuSiE* model fits. Figure 2 shows an example drawn from our simulations where the regular IBSS algorithm converges to a poor solution and our refinement procedure improves the solution. The example has two causal SNPs in moderate LD with one another, which have opposite effects that partially cancel out each others' marginal associations (Panel A). This example is challenging because the SNP with the strongest marginal association (SMA) is not in high LD with either causal SNP (it is in moderate LD with Causal SNP 1, and low LD with Causal SNP 2). Although [17] showed that the IBSS algorithm can sometimes deal well with such situations, that does not happen in this case: the IBSS algorithm yields three CSs, two of which are false positives that

do not contain a causal SNP (Panel B). Applying our refinement procedure solves the problem; it yields a solution with higher objective function (ELBO), and with two CSs, each containing one of the causal SNPs (Panel C).

Although this sort of problem was not common in our simulations, it occurred often enough that the refinement procedure yielded a noticeable improvement in performance across many simulations (Fig. 2, Panel D). In our remaining experiments we therefore used the refinement procedure for fitting *SuSiE* models.

Impact of LD accuracy on fine-mapping

We used simulations to compare *SuSiE-RSS* with both *SuSiE-suff* (which produces the same results as running *SuSiE* on individual-level data) and several other fine-mapping methods for summary data: FINEMAP [12], DAP-G [14,16] and CAVIAR [7]. All these methods are based on the same *RSS-Z* model, and differ in the priors used, and in the approach taken to compute posterior quantities. For these simulations, *SuSiE-RSS*, FINEMAP and DAP-G are all very fast, usually taking no more than a few seconds per data set (see Table 1 in Detailed Methods); by contrast, CAVIAR is much slower because it exhaustively evaluates all causal SNP configurations. (Other Bayesian fine-mapping methods for summary data include PAINTOR [8], JAM [15] and CAVIARBF [11]. FINEMAP has been shown to be faster and at least as accurate as PAINTOR and CAVIARBF [12]. JAM is comparable in accuracy to FINEMAP [15] and is most beneficial when jointly fine-mapping multiple genomic regions, which we did not consider here.)

We compared methods based on both their posterior inclusion probabilities (PIPs) [39] and credible sets (CSs) [2,17]. These quantities have different advantages. PIPs have the advantage that they are returned by most methods, and can be used to assess familiar quantities such as power and false discovery rates. CSs have the advantage that, when the data support multiple causal signals, this is explicitly reflected in the number of CSs reported. Uncertainty in which SNP are causal is reflected in the size of the CSs.

First, we assessed the performance of summary-data methods using the in-sample LD matrix. Using the in-sample LD matrix, *SuSiE-suff* can provide the same results as *SuSiE* on the individual-level data, so we use those results as a benchmark. The results show that *SuSiE-suff*, *SuSiE-RSS*, FINEMAP and DAP-G have remarkably similar performance, as measured by both PIPs (Fig. 3) and CSs (the "in-sample LD" columns in Fig. 5). The main difference between the methods is that DAP-G produced some "high confidence" (high PIP) false positives, which hinders its ability to produce very low FDR values. Further, all four methods produced CSs whose coverage was close to the target level of 95% (Panel A in Fig. 5). Both *SuSiE-RSS* and FINEMAP require the user to specify an upper bound on the number of causal SNPs L; setting this upper bound to the true number of causal SNPs ("$L = \text{true}$" in the figures) did not improve their performance, demonstrating that, with in-sample LD matrix, these methods are robust to overstating this bound. CAVIAR performed notably less well than the other methods for the PIP computations. (The CSs computed by CAVIAR are defined differently from CSs computed by other methods, so we excluded CAVIAR from the CS comparisons.)

Next, we compared the summary data methods using different out-of-sample LD matrices, again using *SuSiE-suff* with in-sample LD as a benchmark. For every method, we computed out-of-sample LD matrices using two different panel sizes ($n = 500, 1000$) and three different values for the regularization parameter, λ ($\lambda = 0$, or no regularization; $\lambda = 0.001$; and λ estimated as described in Methods). As might be expected, the performance of *SuSiE-RSS*, FINEMAP and DAP-G all degraded with out-of-sample LD compared with in-sample LD; see Figures 4 and 5. Notably, the CSs

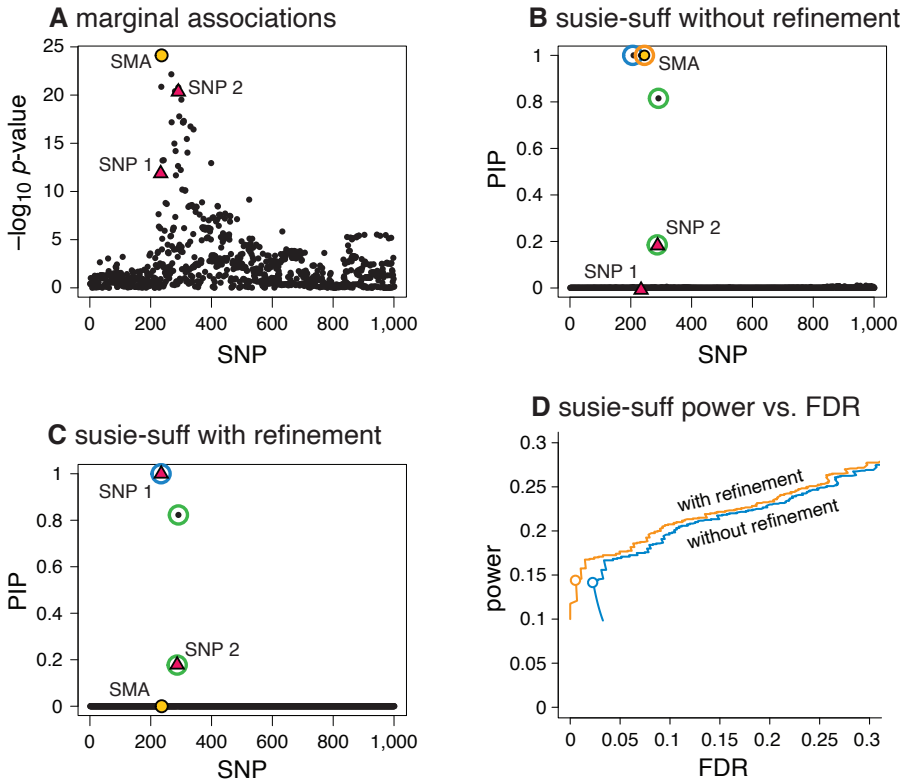

Figure 2. Refining *SuSiE* model fits improves fine-mapping accuracy. Panels A, B and C show a single example, drawn from our simulations, that illustrates how refining a *SuSiE-suff* model fit improves fine-mapping accuracy. In this example, there are 1,001 candidate SNPs, and two SNPs (the red triangles labeled "SNP 1" and "SNP 2") explain variation in the simulated phenotype. In this example, the strongest marginal association (yellow circle labeled "SMA") is not one of the two causal SNPs. Without refinement, the IBSS-suff algorithm (with default settings) returns a *SuSiE-suff* fit identifying three 95% CSs (blue, green and orange circles); two of the CSs (blue, orange) are false positives containing no true effect SNP, one of these CSs contains the SMA (orange), and no CS includes SNP 1. After running the refinement procedure, the fit is much improved, as measured by the "evidence lower bound" (ELBO); it increases the ELBO by 19.06 (−70837.09 vs. −70818.03). The new *SuSiE-suff* fit identifies two 95% CSs (blue, green), each containing a true causal SNP (Panel C), and neither contains the SMA. Panel D summarizes the improvement in fine-mapping across all simulations; it shows power and false discovery rate (FDR) for *SuSiE-suff* with and without using the refinement procedure as the PIP threshold for reporting causal SNPs is varied from 0 to 1. These quantities are calculated as FDR $:= \frac{\mathrm{FP}}{\mathrm{TP+FP}}$ and power $:= \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, where FP, TP, FN, TN denote, respectively, the number of false positives, true positives, false negatives and true negatives. (This plot is the same as a precision-recall curve after flipping the x-axis because precision $= \frac{\mathrm{TP}}{\mathrm{TP+FP}} = 1 - \mathrm{FDR}$ and recall = power.) Open circles are drawn at a PIP threshold of 0.95.

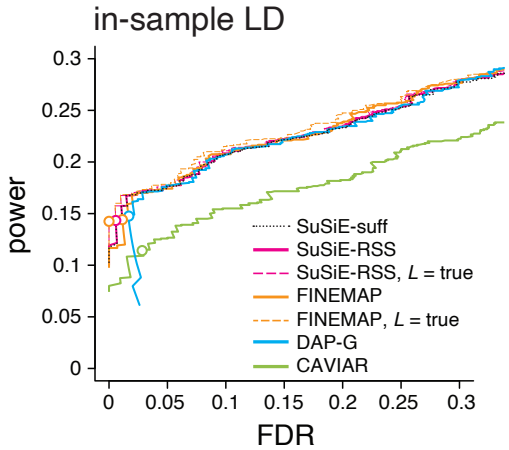

Figure 3. Discovery of causal SNPs using posterior inclusion probabilities—in-sample LD. Each curve shows power vs. FDR in identifying causal SNPs when the method (*SuSiE-RSS*, FINEMAP, DAP-G or CAVIAR) is provided with the in-sample LD matrix. FDR and power are calculated from all 600 simulations as the PIP threshold is varied from 0 to 1. Open circles are drawn at a PIP threshold of 0.95. Two variants of FINEMAP and *SuSiE-RSS* are also compared: when L, the maximum number of estimated causal SNPs, is set to the true value; and when L is larger than the true number. Power and FDR are virtually identical for *SuSiE-suff* and *SuSiE-RSS* so these two curves almost completely overlap in the plot.

no longer met the 95% target coverage (Panel A in Fig. 5). In all cases, performance was notably worse with the smaller reference panel, emphasising the importance of using sufficiently large reference panels [28]. Regarding regularization, *SuSiE-RSS* and DAP-G performed similarly at all levels of regularization, and so do not appear to require regularization; in contrast, FINEMAP required regularization, with an estimated regularization parameter λ, to compete with *SuSiE-RSS* and DAP-G. Since estimating the regularization parameter is somewhat computationally burdensome, *SuSiE-RSS* and DAP-G have an advantage here. All three methods benefited more from large panel size than from regularization, again emphasising the importance of sufficiently large reference panels. Interestingly, CAVIAR's performance was relatively insensitive to choice of LD matrix; however the other methods clearly outperformed CAVIAR with the larger ($n = 1,000$) reference panel.

The fine-mapping results with out-of-sample LD matrix also expose another interesting result: if FINEMAP and *SuSiE-RSS* are provided with the true number of causal SNPs ($L =$ true), their results improve (Figure 4, Panels C vs. D and Panels E vs. F). This is particularly noticeable for the small reference panel (light green lines). We interpret this result as indicating a tendency of these methods to react to misspecification of the LD matrix by sometimes including additional (false positive) signals. Specifying the true L reduces their tendency to do this because it limits the number of signals that can be included. This suggests that restricting the number of causal SNPs, L, may make fine-mapping results more robust to misspecification of the LD matrix (even for methods that are robust to overstatement of L when the LD matrix is accurate). Alternatively, priors or penalties that favor smaller L may also help. Indeed, when none of the methods are provided with information about the true number of causal SNPs, DAP-G slightly outperforms FINEMAP and *SuSiE-RSS*, possibly reflecting a tendency for DAP-G to favour models with smaller number of SNPs (either due to the differences in prior or differences in approximate posterior inference). Further

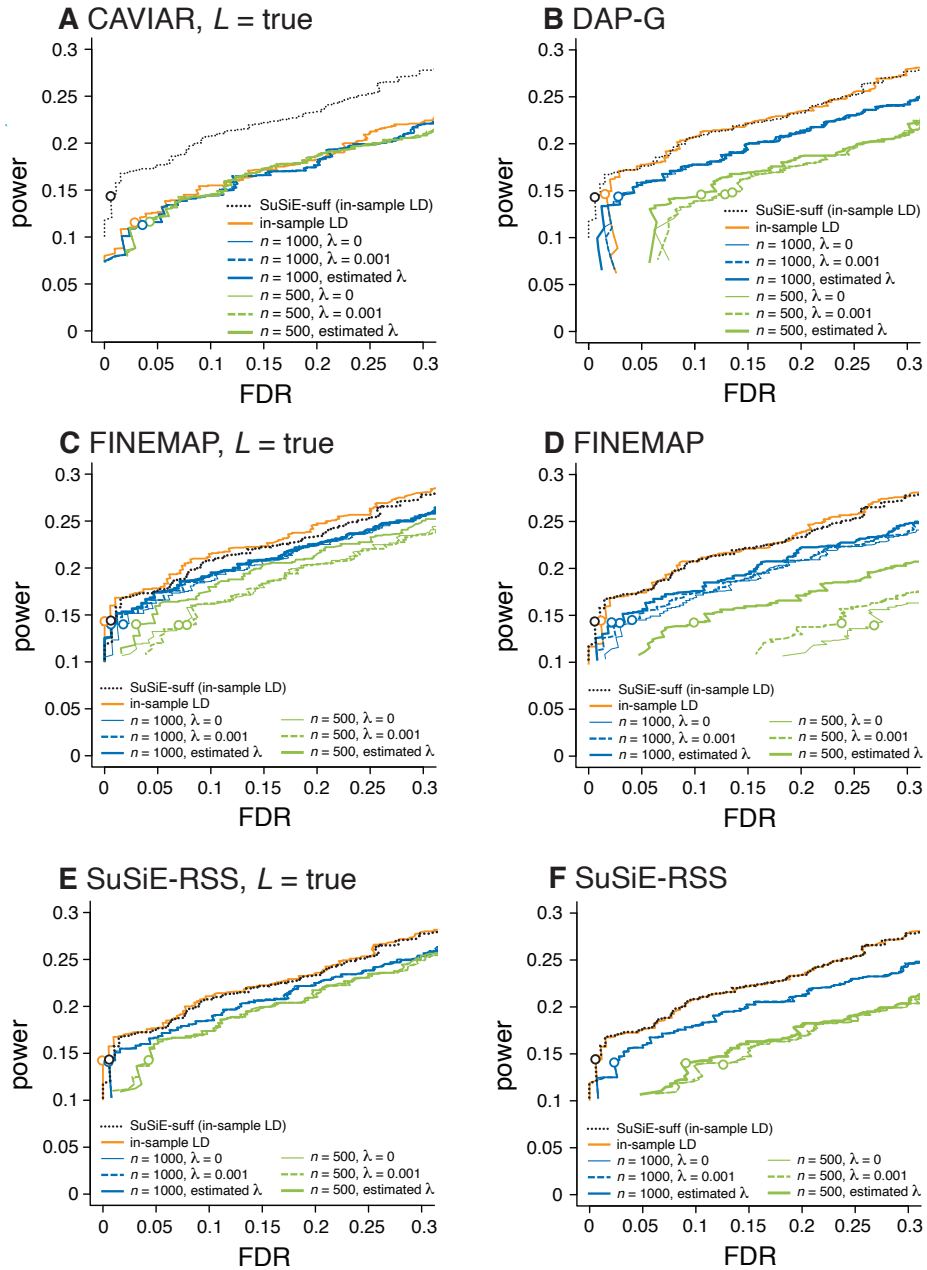

Figure 4. Discovery of causal SNPs using posterior inclusion probabilities—out-of-sample LD. Plots compare power vs FDR for different methods, across all 600 simulations, as the PIP threshold varies from 0 to 1. Open circles indicate results at PIP threshold of 0.95. Each plot compares performance of one method (CAVIAR, DAP-G, FINEMAP or *SuSiE-RSS*) when provided with different LD estimates: in-sample, $\hat{\boldsymbol{R}} = \boldsymbol{R}$; and out-of-sample LD from a reference panel with either 1,000 samples, $\hat{\boldsymbol{R}} = \hat{\boldsymbol{R}}_{1000}$ or 500 samples $\hat{\boldsymbol{R}} = \hat{\boldsymbol{R}}_{500}$. For out-of-sample LD, different levels of the regularization paremeter λ are also compared: $\lambda = 0$, $\lambda = 0.001$, and estimated λ (see Methods). Panels C–F show results for two variants of FINEMAP and *SuSiE-RSS*: in Panels C and E, the maximum number of causal SNPs, L, is set to the true value ("L = true"); in Panels D and F, L is fixed larger than the true value ($L = 4$ for FINEMAP; $L = 10$ for *SuSiE-RSS*). The power vs. FDR curve for *SuSiE-suff* with in-sample LD is shown in every panel to provide a baseline for comparison (dotted black line). Note that some power vs. FDR curves overlap almost completely (e.g., results for *SuSiE-RSS* with different LD regularization levels λ) and so not all are visible.

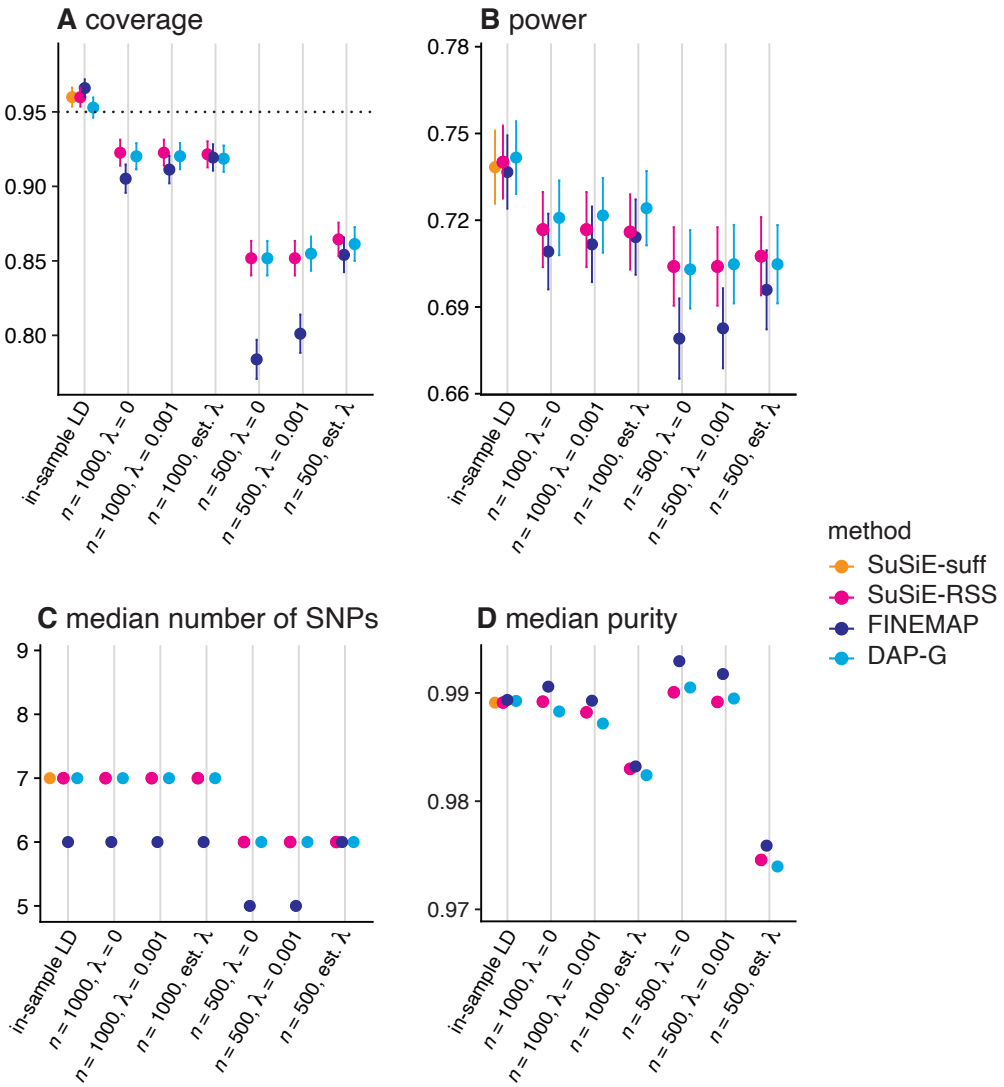

Figure 5. Assessment of 95% credible sets from *SuSiE-suff*, *SuSiE-RSS*, FINEMAP and DAP-G with different LD estimates. We evalulated CSs using: (A) coverage, the proportion of CSs that contain a true causal SNP; (B) power, the proportion of true causal SNPs included in a CS; (C) median number of SNPs in each CS; and (D) median purity, where "purity" is defined as the smallest absolute correlation among all pairs of SNPs within a CS. Following [17], we discarded any CSs with purity less than 0.5. These statistics are taken as the mean (A, B) or median (C, D) over all simulations; error bars in A and B show 2 times the standard error. The target coverage of 95% is shown as a dotted horizontal line in Panel A.

study of this issue may lead to methods that are more robust to misspecified LD.

Discussion

We have presented extensions of the *SuSiE* fine-mapping method to acccommodate summary data, with a focus on marginal z-scores and an out-of-sample LD matrix computed from a reference panel. Our approach provides a general template for how to extend any full-data regression method to analyse summary data: develop a full-data algorithm that works with sufficient statistics, then apply this algorithm directly to summary data. Although it is simple, as far as we are aware this template is novel, and it avoids the need for any special treatment of non-invertible LD matrices.

In simulations, we found that our new method, *SuSiE-RSS*, is competitive—in both accuracy and computational cost—with the best existing methods for fine-mapping from summary data, DAP-G and FINEMAP. Whatever method is used, our results underscore the importance of accurately computing the out-of-sample LD matrix from an appropriate and large reference panel (see also [28]). Indeed, for the best performing methods, performance depended more on choice of LD matrix than on choice of method. We also emphasize the importance of computing z-scores at different SNPs from the exact same samples, using genotype imputation if necessary [40]. It is also important to ensure that alleles are consistently encoded in study and reference samples.

Although our derivations and simulations focused on z-scores computed from quantitative traits with a simple linear regression, in practice it is common to apply summary data fine-mapping methods to z-scores computed in other ways, e.g., using logistic regression on a binary or case-control trait, or using linear mixed models to deal with population stratification and relatedness. The multivariate normal assumption on z-scores, which underlies all the methods considered here, should also apply to these settings, although as far as we are aware theoretical derivation of the precise form (10) is lacking in these settings (although see [41]). Since the model (10) is already only an approximation, one might expect that the additional effect of such issues might be small, particularly compared with the effect of allele flips or small reference panels. Nonetheless, since our simulations show that model misspecification can hurt performance of existing methods, further research to improve robustness of fine-mapping methods to model misspecification would be welcome.

Detailed methods

SuSiE-suff model and algorithm

SuSiE-suff fits the *SuSiE* model using a modified IBSS algorithm [17], in which the computations are rearranged so that they only require sufficient statistics. Here we describe these computations in detail.

The single effect regression (SER) model with summary statistics

The single effect regression (SER) model is defined in [17] (see also [9]) as a multiple regression model in which exactly one variable has a non-zero effect on the outcome. It is a special case of the *SuSiE* model, when $L = 1$. Posterior computations with the SER model form the basis for the *SuSiE* model fitting algorithm, IBSS, and hence form the basis for the *SuSiE*-suff model fitting algorithm. Here we show how posterior quantities under the SER model are computed using summary statistics.

Formally, the SER model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{21}$$

$$\boldsymbol{e} \sim \mathcal{N}_N(0, \sigma^2 \boldsymbol{I}_N) \tag{22}$$

$$\boldsymbol{b} = \boldsymbol{\gamma} b \tag{23}$$

$$\boldsymbol{\gamma} \sim \text{Multinomial}(1, \boldsymbol{\pi}) \tag{24}$$

$$b \sim \mathcal{N}(0, \sigma_0^2). \tag{25}$$

Here, $\boldsymbol{y} = (y_1, \ldots, y_N)^\intercal \in \mathbb{R}^N$ denotes the phenotypes of N individuals, $\boldsymbol{X} \in \mathbb{R}^{N \times J}$ denotes their corresponding genotypes at J genetic variants (SNPs), $\boldsymbol{b} = (b_1, \ldots, b_J)^\intercal$ denotes a vector of regression coefficients, $\boldsymbol{e}$ is an N-vector of error terms, $\sigma^2 > 0$ is the residual variance parameter, and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. To simplify the presentation, we assume $\boldsymbol{y}$ and the columns of $\boldsymbol{X}$ are centered to have mean zero, which avoids the need for an intercept term [22]. $\mathcal{N}(\mu, \sigma^2)$ denotes the univariate normal distribution with mean μ and variance σ^2, $\mathcal{N}_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the r-variate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, and Multinomial$(n, \boldsymbol{p})$ denotes the multinomial distribution with n trials and category probabilities $\boldsymbol{p} = (p_1, \ldots, p_J)$. Thus, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J) \in \{0, 1\}^J$ is a binary vector of length J in which exactly one element is 1 and the rest are 0, and so $\boldsymbol{b}$ is a vector with exactly one non-zero element (except for the special case when $b = 0$). The scalar b represents the value of the one non-zero element in $\boldsymbol{b}$ (the "single effect"). The prior inclusion probabilities, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$, which we assume are fixed and known, give the prior probability that each genetic variant has the non-zero effect. The prior variance of the single effect, σ_0^2, and the residual variance, σ^2, are hyperparameters that can be pre-specified or, more commonly, estimated.

Given settings of the hyperparameters σ^2, σ_0^2, the posterior distribution of $\boldsymbol{b}$ under the SER model is worked out in [17]. We summarize it here, with a focus on computations with the available summary statistics $\boldsymbol{X}^\intercal\boldsymbol{X}$ and $\boldsymbol{X}^\intercal\boldsymbol{y}$.

Proposition 1. *Consider the SER model (21–25) with known σ_0^2 and σ^2. The posterior distribution of $\boldsymbol{b} = \boldsymbol{\gamma} b$ can be expressed in terms of univariate least-squares estimates of b_j, $\hat{b}_j := \boldsymbol{x}_j^\intercal \boldsymbol{y} / \boldsymbol{x}_j^\intercal \boldsymbol{x}_j$, and their variances, $s_j^2 := \sigma^2 / \boldsymbol{x}_j^\intercal \boldsymbol{x}_j$. Specifically, the posterior distribution of $\boldsymbol{\gamma}$ is*

$$\boldsymbol{\gamma} \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2, \sigma_0^2 \sim \text{Multinomial}(1, \boldsymbol{\alpha}) \tag{26}$$

and the posterior distribution of b given $\boldsymbol{\gamma}$ is

$$b \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2, \sigma_0^2, \gamma_j = 1 \sim \mathcal{N}(\mu_{1j}, \sigma_{1j}^2), \tag{27}$$

where we define

$$\sigma_{1j}^2 := \frac{1}{1/\sigma_0^2 + 1/s_j^2} \tag{28}$$

$$\mu_{1j} := \sigma_{1j}^2 \hat{b}_j / s_j^2 \tag{29}$$

$$\alpha_j := \frac{\pi_j \text{BF}_j}{\sum_{j'=1}^J \pi_{j'} \text{BF}_{j'}} \tag{30}$$

$$\begin{aligned} \text{BF}_j &:= \text{BF}(\boldsymbol{x}_j, \boldsymbol{y}; \sigma^2, \sigma_0^2) \\ &:= \frac{p(\boldsymbol{y} \mid \boldsymbol{x}_j, \sigma^2, \sigma_0^2)}{p(\boldsymbol{y} \mid \boldsymbol{x}_j; \sigma^2, b = 0)} \\ &= \sqrt{\frac{s_j^2}{\sigma_0^2 + s_j^2}} \times \exp\left(\frac{\hat{b}_j^2}{2 s_j^2} \times \frac{\sigma_0^2}{\sigma_0^2 + s_j^2}\right). \end{aligned} \tag{31}$$

Note that the α_j's (30) *are the PIPs* (2) *under the SER model;* $\text{PIP}_j = \alpha_j$, $j = 1, \ldots, J$.

Proposition 1 shows that the posterior for $\boldsymbol{b}$ under the *SER* model can be computed from the sufficient statistics because the sufficient statistics only enter into the posterior expressions via the least-squares estimates $\hat{b}_j$ and variances s_j^2; in particular, $\hat{b}_j$ and s_j^2 only need $\boldsymbol{x}_j^\intercal \boldsymbol{x}_j$, the jth diagonal entry of the matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$, and $\boldsymbol{x}_j^\intercal \boldsymbol{y}$, the jth entry of the vector $\boldsymbol{X}^\intercal \boldsymbol{y}$. We define SER-suff as the function that returns the posterior distribution of $\boldsymbol{b}$ under the SER model given the sufficient statistics:

$$\text{SER-suff}(\boldsymbol{X}^\intercal \boldsymbol{X}, \boldsymbol{X}^\intercal \boldsymbol{y}; \sigma^2, \sigma_0^2) := (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2), \tag{32}$$

where $\boldsymbol{\mu}_1 := (\mu_{11}, \ldots, \mu_{1J})^\intercal$, $\boldsymbol{\sigma}_1^2 := (\sigma_{11}^2, \ldots, \sigma_{1J}^2)$ and $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_J)$.

Remark 1. *Although we write the posterior* SER-suff *as a function of* $\boldsymbol{X}^\intercal \boldsymbol{X}$*, it actually only depends on the diagonal elements of this matrix.*

Likewise, the likelihood for σ_0^2 and σ^2 under the *SER* model can be computed using only the sufficient statistics since it can be expressed as a weighted sum of the BFs:

$$\begin{aligned}
\ell_{SER}(\sigma^2, \sigma_0^2; \boldsymbol{X}, \boldsymbol{y}) &:= p(\boldsymbol{y} \mid \boldsymbol{X}, \sigma^2, \sigma_0^2) \\
&= p(\boldsymbol{y} \mid \boldsymbol{X}, \sigma^2, b = 0) \textstyle\sum_{j=1}^J \pi_j \text{BF}_j \\
&= \mathcal{N}_N(\boldsymbol{y}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_N) \textstyle\sum_{j=1}^J \pi_j \text{BF}_j \\
&:= \ell_{\text{SER-suff}}(\sigma^2, \sigma_0^2; \boldsymbol{X}^\intercal \boldsymbol{X}, \boldsymbol{X}^\intercal \boldsymbol{y}).
\end{aligned} \tag{33}$$

Following [17], we compute the maximum-likelihood estimate of σ_0^2 by maximizing this likelihood via numerical optimization.

Data preprocessing

As stated above, $\boldsymbol{y}$ and the columns of $\boldsymbol{X}$ are assumed to be centered, and accordingly the sufficient statistics should be computed using a centered $\boldsymbol{X}, \boldsymbol{y}$. If the sufficient statistics have been computed from an $\boldsymbol{X}, \boldsymbol{y}$ that have not been centered, the sufficient statistics can be modified after the fact so that they correspond to a centered $\boldsymbol{X}$ and $\boldsymbol{y}$. Denoting the unmodified data as $\boldsymbol{X}, \boldsymbol{y}$ and the centered data as $\tilde{\boldsymbol{X}} := \boldsymbol{X} - \boldsymbol{1}_N \bar{\boldsymbol{x}}^\intercal, \tilde{\boldsymbol{y}} := \boldsymbol{y} - \bar{y}\boldsymbol{1}_N$, where $\bar{\boldsymbol{x}} = \boldsymbol{X}^\intercal \boldsymbol{1}_N / N$ is the vector of column means, $\bar{y} := \sum_{i=1}^N y_i$, and $\boldsymbol{1}_N$ is a column vector of ones of length N, the centering calculations for the sufficient statistics are

$$\begin{aligned}
\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}} &= \boldsymbol{X}^\intercal \boldsymbol{X} - N \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^\intercal \\
\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{y}} &= \boldsymbol{X}^\intercal \boldsymbol{y} - N \bar{y} \bar{\boldsymbol{x}} \\
\tilde{\boldsymbol{y}}^\intercal \tilde{\boldsymbol{y}} &= \boldsymbol{y}^\intercal \boldsymbol{y} - N \bar{y}^2.
\end{aligned}$$

Similarly, the sufficient statistics may also be modified *post hoc* to be as if they were computed using a column-standardized $\boldsymbol{X}$; that is, an $\boldsymbol{X}$ in which each column has unit variance. Denoting the standardized (and centered) matrix as $\hat{\boldsymbol{X}}$, and assuming $\boldsymbol{X}$ is centered, the calculations are

$$\begin{aligned}
\hat{\boldsymbol{X}}^\intercal \hat{\boldsymbol{X}} &= \boldsymbol{D} \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{D} \\
\hat{\boldsymbol{X}}^\intercal \boldsymbol{y} &= \boldsymbol{D} \boldsymbol{X}^\intercal \boldsymbol{y}.
\end{aligned}$$

where $\boldsymbol{D}$ denotes the $J \times J$ diagonal matrix with diagonal entries $d_{jj} = \sqrt{(N-1)/(\boldsymbol{x}_j^\intercal \boldsymbol{x}_j)}$. In these formulae we have assumed that unbiased sample variances ($N - 1$ denominator) are used to standardize columns of $\boldsymbol{X}$.

Algorithm 1 Iterative Bayesian stepwise selection using sufficient statistics (IBSS-suff)

Require: Sufficient statistics $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}$ and, optionally, $\boldsymbol{y}^\intercal\boldsymbol{y}$, N (for estimating σ^2).
Require: Number of effects, L; initial estimates of hyperparameters $\sigma^2, \boldsymbol{\sigma}_0^2$.
Require: Initial estimates of the posterior mean single effects, $\bar{\boldsymbol{b}}_l$, for $l = 1, \ldots, L$.

repeat
  $\bar{\boldsymbol{\rho}} \leftarrow \boldsymbol{X}^\intercal\boldsymbol{y} - \boldsymbol{X}^\intercal\boldsymbol{X}\sum_{l=1}^L \bar{\boldsymbol{b}}_l$ ▷ Compute expected residuals.
  for l in $1, \ldots, L$ **do**
    $\bar{\boldsymbol{\rho}}_l \leftarrow \bar{\boldsymbol{\rho}} + \boldsymbol{X}^\intercal\boldsymbol{X}\bar{\boldsymbol{b}}_l$ ▷ Disregard lth single effect in residuals.
    $\sigma_{0l}^2 \leftarrow \operatorname{argmax}_{\sigma_0^2} \ell_{\text{SER-suff}}(\sigma^2, \sigma_0^2; \boldsymbol{X}^\intercal\boldsymbol{X}, \bar{\boldsymbol{\rho}})$ ▷ Update σ_{0l}^2 (optional).
    $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}^2) \leftarrow \text{SER-suff}(\boldsymbol{X}^\intercal\boldsymbol{X}, \bar{\boldsymbol{\rho}}; \sigma^2, \sigma_{0l}^2)$ ▷ Fit SER to residuals.
    $\bar{\boldsymbol{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_{1l}$ ▷ "$\circ$" denotes element-wise multiplication.
    $\bar{\boldsymbol{b}}_l^2 \leftarrow \boldsymbol{\alpha}_l \circ (\boldsymbol{\mu}_{1l} \circ \boldsymbol{\mu}_{1l} + \boldsymbol{\sigma}_{1l}^2)$ ▷ Compute posterior second moments.
    $\bar{\boldsymbol{\rho}} \leftarrow \bar{\boldsymbol{\rho}}_l - \boldsymbol{X}^\intercal\boldsymbol{X}\bar{\boldsymbol{b}}_l$ ▷ Update expected residuals.
  end for
  $\sigma^2 \leftarrow \frac{1}{N}\text{ERSS}_{\text{suff}}(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, \bar{\boldsymbol{b}}_1, \ldots, \bar{\boldsymbol{b}}_l, \bar{\boldsymbol{b}}_1^2, \ldots, \bar{\boldsymbol{b}}_L^2)$ ▷ Optional.
until convergence criterion satisfied
return $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}^2, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}^2$.

Fitting the *SuSiE* model with sufficient statistics: IBSS-suff

Using the results regarding the SER model with summary statistics, it is straightforward to modify the original IBSS algorithm [17] to work with sufficient statistics; see Algorithm 1. Additional notation used in Algorithm 1 includes: $\bar{\boldsymbol{b}}_l$, the expected value of $\boldsymbol{b}_l$ with respect to the approximate posterior distribution, $q(\boldsymbol{b})$; and $\bar{\boldsymbol{b}}_l^2 = (\bar{b_{l1}^2}, \ldots, \bar{b_{lJ}^2})^\intercal$, the vector of posterior second moments $\bar{b_{lj}^2} := \mathbb{E}_q[b_{lj}^2]$. A key change in implementation is that the original IBSS algorithm keeps track of the posterior mean residuals $\bar{\boldsymbol{r}} := \mathbb{E}_q[\boldsymbol{X}^\intercal\boldsymbol{y} - \boldsymbol{b}] = \boldsymbol{X}^\intercal\boldsymbol{y} - \sum_{l=1}^L \bar{\boldsymbol{b}}_l$ whereas IBSS-suff updates $\bar{\boldsymbol{\rho}} := \boldsymbol{X}^\intercal\bar{\boldsymbol{r}}$. See [17] for development and justification for the IBSS algorithm, and details of implementation, including preprocessing steps, and calculation of the credible sets.

The only missing piece to the IBSS-suff algorithm is the expression for the expected residual sum of squares (ERSS) under the variational approximation to the posterior, $q(\boldsymbol{b})$, which is needed to estimate σ^2. Again, the expression can be written in terms of the sufficient statistics:

$$\begin{aligned}
&\text{ERSS}(\boldsymbol{X}, \boldsymbol{y}, \bar{\boldsymbol{b}}_1, \ldots, \bar{\boldsymbol{b}}_L, \bar{\boldsymbol{b}}_1^2, \ldots, \bar{\boldsymbol{b}}_L^2]) \\
&\quad := \mathbb{E}_q[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2] \\
&\quad = \|\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{b}}\|^2 - \sum_{l=1}^L \bar{\boldsymbol{b}}_l\boldsymbol{X}^\intercal\boldsymbol{X}\bar{\boldsymbol{b}}_l + \sum_{l=1}^L\sum_{j=1}^J (\boldsymbol{x}_j^\intercal\boldsymbol{x}_j)\bar{b_{lj}^2} \\
&\quad = \boldsymbol{y}^\intercal\boldsymbol{y} - 2\bar{\boldsymbol{b}}^\intercal\boldsymbol{X}^\intercal\boldsymbol{y} + \bar{\boldsymbol{b}}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\bar{\boldsymbol{b}} - \sum_{l=1}^L \bar{\boldsymbol{b}}_l^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\bar{\boldsymbol{b}}_l + \sum_{l=1}^L\sum_{j=1}^J (\boldsymbol{x}_j^\intercal\boldsymbol{x}_j)\bar{b_{lj}^2} \\
&\quad := \text{ERSS}_{\text{suff}}(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, \bar{\boldsymbol{b}}_1, \ldots, \bar{\boldsymbol{b}}_L, \bar{\boldsymbol{b}}_1^2, \ldots, \bar{\boldsymbol{b}}_L^2).
\end{aligned} \tag{34}$$

The second ERSS definition makes explicit that it is a function of the sufficient statistics.

Computing the sufficient statistics

Lemma 1 (Computing sufficient statistics)**.** *The sufficient statistics* $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}$ *can be computed from the the summary data* $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, \boldsymbol{R}, \boldsymbol{y}^\intercal\boldsymbol{y}, N$ *from the following formulae:*

$$\hat{\sigma}_j^2 = \frac{\boldsymbol{y}^\intercal\boldsymbol{y}}{\hat{b}_j^2/\hat{s}_j^2 + N - 2} \tag{35}$$

$$\boldsymbol{x}_j^\intercal\boldsymbol{y} = \hat{\sigma}_j^2\hat{b}_j/\hat{s}_j^2 \tag{36}$$

$$\boldsymbol{x}_j^\intercal\boldsymbol{x}_j = \hat{\sigma}_j^2/\hat{s}_j^2 \tag{37}$$

$$\boldsymbol{D}_X = \mathrm{diag}(\boldsymbol{x}_1^\intercal\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J^\intercal\boldsymbol{x}_J) \tag{38}$$

$$\boldsymbol{X}^\intercal\boldsymbol{X} = \boldsymbol{D}_X^{1/2}\boldsymbol{R}\boldsymbol{D}_X^{1/2}. \tag{39}$$

Proof. The correlation coefficient r_j^2 for a simple linear regression of $\boldsymbol{y}$ on SNP j is:

$$r_j^2 := \frac{\hat{b}_j^2/\hat{s}_j^2}{\hat{b}_j^2/\hat{s}_j^2 + N - 2}. \tag{40}$$

The estimated residual variance from the simple linear regression model is therefore

$$\begin{aligned}\hat{\sigma}_j^2 &= \frac{\boldsymbol{y}^\intercal\boldsymbol{y}(1 - r_j^2)}{N-2} \\ &= \frac{\boldsymbol{y}^\intercal\boldsymbol{y}}{\hat{b}_j^2/\hat{s}_j^2 + N - 2}.\end{aligned} \tag{41}$$

The remaining expressions follow straightforwardly from standard statistical formulas for the sample correlation and point estimation in linear regression. □

Remark 2. *The formulae (35–39) are ordered in such a way that they suggest a step-by-step procedure for reconstructing the sufficient statistics* $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}$ *from the summary data* $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, \boldsymbol{R}, \boldsymbol{y}^\intercal\boldsymbol{y}, N$.

The *RSS-Z* model

As described in the main text, *SuSiE-RSS* simply applies *SuSiE-suff* with $\boldsymbol{X}^\intercal\boldsymbol{X} = \hat{\boldsymbol{R}}, \boldsymbol{X}^\intercal\boldsymbol{y} = \hat{\boldsymbol{z}}$ and $\sigma^2 = 1$. (We also arbitrarily set $\boldsymbol{y}^\intercal\boldsymbol{y} = N = 1$, but these values do not affect the statistical inferences.) This approach is therefore fitting the following model:

$$\hat{\boldsymbol{z}} \sim \mathcal{N}_J(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}}) \tag{42}$$

$$\boldsymbol{z} = \sum_{l=1}^{L} \boldsymbol{z}_l \tag{43}$$

$$\boldsymbol{z}_l = \boldsymbol{\gamma}_l z_l \tag{44}$$

$$\boldsymbol{\gamma}_l \sim \mathrm{Multinomial}(1, \boldsymbol{\pi}) \tag{45}$$

$$z_l \sim \mathcal{N}(0, \omega_l^2). \tag{46}$$

Here, L is a user-specified upper bound on the number of non-zero effect variables, and ω_l^2 is the prior variance for the lth effect.

This model assumes that the non-centrality parameters (NCPs) are exchangeable. This is a common assumption (e.g., [42, 43]), and implies that effect sizes for SNPs with low minor allele frequencies (MAFs) tend to be larger than those with large MAFs. The equivalent prior on the effects, assuming $\sigma^2 = 1$, would be $b_l \sim N(0, \omega_l^2/\boldsymbol{x}_j^\intercal\boldsymbol{x}_j)$ so that SNPs with lower MAFs are more likely to have larger effects *a priori*.

Extension to $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{s}}$

We have focussed on the *RSS-Z* model for z-scores because z-scores are more commonly available, and because most existing methods for summary data use this model. However, similar ideas motivate a more general *RSS* model for $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}$ [29],

$$\hat{\boldsymbol{b}} \mid \hat{\boldsymbol{s}} \sim \mathcal{N}(\hat{\boldsymbol{S}}\hat{\boldsymbol{R}}\hat{\boldsymbol{S}}^{-1}\boldsymbol{b}, \hat{\boldsymbol{S}}\hat{\boldsymbol{R}}\hat{\boldsymbol{S}}), \tag{47}$$

where $\hat{\boldsymbol{S}} := \mathrm{diag}(\hat{\boldsymbol{s}})$ denotes the diagonal matrix with diagonal entries $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_J)$. This model has likelihood

$$\ell_{\mathrm{RSS}}(\boldsymbol{b}; \hat{\boldsymbol{b}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}) := \exp(-\tfrac{1}{2}\boldsymbol{b}^\intercal \hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{R}}\hat{\boldsymbol{S}}^{-1}\boldsymbol{b} + \boldsymbol{b}^\intercal \hat{\boldsymbol{S}}^{-2}\hat{\boldsymbol{b}}), \tag{48}$$

This is the same as $\ell_{\mathrm{suff}}(\boldsymbol{b}, \sigma^2; \boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, N)$ (up to a constant of proportionality) after making substitutions $\boldsymbol{X}^\intercal\boldsymbol{X} = \hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{R}}\hat{\boldsymbol{S}}^{-1}$, $\boldsymbol{X}^\intercal\boldsymbol{y} = \hat{\boldsymbol{S}}^{-2}\hat{\boldsymbol{b}}$ and $\sigma^2 = 1$. Thus, we can fit this model by applying *SuSiE-suff* with these settings. The difference between this model and the *RSS-Z* model is that the *RSS* model assumes that the effects b_j are exchangeable, whereas the *RSS-Z* model assumes the NCPs z_j are exchangeable. Note that if the genotypes are standardized to have unit variance before computing $\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}$ then the *RSS* approach will be equivalent to *RSS-Z*, but will produce effect estimates on the phenotype scale rather than on the NCP scale.

Dealing with non-invertible LD matrix

If $\hat{\boldsymbol{R}}$ is not invertible, then model (10) does not have a density, which complicates defining the likelihood. Here we draw connections between different methods for dealing with this issue. We also show that the alternative approach that uses the *RSS-Z* likelihood (14) without modifying $\hat{\boldsymbol{R}}$ has some benefits; in particular, the *RSS-Z* likelihood produces the same result as making small adjustments to the LD matrix to make it invertible (Proposition 2).

We assume that although $\hat{\boldsymbol{R}}$ may not be invertible it is nonetheless a valid covariance matrix. That is, it is symmetric and positive semidefinite (PSD), which means that all its eigenvalues are non-negative. This is guaranteed so long as $\hat{\boldsymbol{R}}$ is a sample correlation matrix. However, it may be violated if $\hat{\boldsymbol{R}}$ is obtained by modifying a sample correlation matrix, for example by setting small correlations to zero. Any $J \times J$ symmetric PSD matrix $\hat{\boldsymbol{R}}$ with rank $r \leq J$ has an eigenvalue decomposition of the form

$$\hat{\boldsymbol{R}} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\intercal, \tag{49}$$

where $\boldsymbol{\Lambda}$ is an $r \times r$ diagonal matrix with the r positive eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_r > 0$ along its diagonal, and $\boldsymbol{Q}$ is a $J \times r$ matrix whose columns are the r eigenvectors corresponding to the r non-zero eigenvalues, and $\boldsymbol{Q}^\intercal\boldsymbol{Q} = \boldsymbol{I}_r$.

One can modify $\hat{\boldsymbol{R}}$ to make it invertible by simply adding a small diagonal element. Indeed, for any $\lambda \in (0, 1)$, the matrix

$$\hat{\boldsymbol{R}}_\lambda := (1 - \lambda)\hat{\boldsymbol{R}} + \lambda\boldsymbol{I} \tag{50}$$

will be invertible.

When $\hat{\boldsymbol{R}}$ is not invertible, the distribution (10) becomes degenerate, which means that some values of $\hat{\boldsymbol{z}}$ become impossible. In particular, with probability one, $\hat{\boldsymbol{z}} \in \mathrm{range}(\boldsymbol{Q})$; that is, $\hat{\boldsymbol{z}} = \boldsymbol{Q}\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha}$.

Definition 1 (Consistency of $\hat{\boldsymbol{z}}$ with $\hat{\boldsymbol{R}}$)**.** *We say that $\hat{\boldsymbol{z}}$ is* consistent *with $\hat{\boldsymbol{R}}$ if $\hat{\boldsymbol{z}} \in \mathrm{range}(\boldsymbol{Q})$. Otherwise, if $\hat{\boldsymbol{z}} \notin \mathrm{range}(\boldsymbol{Q})$ we say $\hat{\boldsymbol{z}}$ is* inconsistent *with $\hat{\boldsymbol{R}}$.*

Note that, if $\hat{\boldsymbol{R}}$ is invertible, then $\text{range}(\boldsymbol{Q}) = \mathbb{R}^J$, and so $\hat{\boldsymbol{z}}$ will always be consistent with $\hat{\boldsymbol{R}}$. Further, if $\hat{\boldsymbol{z}}$ was generated from the model (10), then it will be consistent with $\hat{\boldsymbol{R}}$ (with probability 1). However, in practice the model (10) is only an approximation, and so in practice $\hat{\boldsymbol{z}}$ may be inconsistent with $\hat{\boldsymbol{R}}$. Even when $\hat{\boldsymbol{R}} = \boldsymbol{R}$, $\hat{\boldsymbol{z}}$ could be inconsistent with $\hat{\boldsymbol{R}}$.

Now we consider four approaches to dealing with a non-invertible $\hat{\boldsymbol{R}}$.

Approach 1a: Substitute the non-invertible $\hat{\boldsymbol{R}}$ with the invertible matrix $\hat{\boldsymbol{R}}_\lambda$, for some small λ, in (10). This is the approach used in [7, 8]. The *RSS-Z* model becomes $\hat{\boldsymbol{z}} \sim \mathcal{N}_J(\hat{\boldsymbol{R}}_\lambda \boldsymbol{z}, \hat{\boldsymbol{R}}_\lambda)$, which has a density because $\hat{\boldsymbol{R}}_\lambda$ is invertible, yielding likelihood

$$\ell_{1a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) := \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}_\lambda \boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}). \tag{51}$$

Approach 1b: Use the *RSS-Z* likelihood (14) even though $\hat{\boldsymbol{R}}$ is non-invertible, so the likelihood is

$$\ell_{1b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) := \ell_{\textit{RSS-Z}}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) = \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}). \tag{52}$$

Note that the likelihood exists, and is easily computed, even when $\hat{\boldsymbol{R}}$ is not invertible. FINEMAP [12] effectively uses this approach, but does not allow configurations $\gamma \subseteq \{1, \ldots, J\}$ where $\hat{\boldsymbol{R}}_\gamma$ is not invertible.

Approach 2a: Set the covariance in the *RSS-Z* model (10) to $\hat{\boldsymbol{R}}_\lambda$ for some small λ, so that $\hat{\boldsymbol{z}} \sim N(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}}_\lambda)$. Note that this approach differs from 1a because it uses $\hat{\boldsymbol{R}}$ instead of $\hat{\boldsymbol{R}}_\lambda$ for the mean. This yields the following likelihood:

$$\ell_{2a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) := \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \hat{\boldsymbol{R}}_\lambda^{-1} \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \hat{\boldsymbol{R}}_\lambda^{-1} \hat{\boldsymbol{z}}). \tag{53}$$

Approach 2b: Project $\hat{\boldsymbol{z}}$ into a lower-dimensional subspace, $\tilde{\boldsymbol{z}} := \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}}$, which ensures that $\tilde{\boldsymbol{z}} \in \mathbb{R}^r$ has a probability density, $\tilde{\boldsymbol{z}} \sim \mathcal{N}_r(\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\intercal \boldsymbol{z}, \boldsymbol{I}_r)$. Thus we have

$$\begin{aligned} p(\tilde{\boldsymbol{z}} \mid \boldsymbol{z}, \hat{\boldsymbol{R}}) &\propto \exp\{-\tfrac{1}{2}(\tilde{\boldsymbol{z}} - \boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\intercal \boldsymbol{z})^\intercal (\tilde{\boldsymbol{z}} - \boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\intercal \boldsymbol{z})\} \\ &\propto \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\tilde{\boldsymbol{z}}) \\ &= \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{Q}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}}), \end{aligned} \tag{54}$$

and therefore the likelihood is

$$\ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) = \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{Q}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}}). \tag{55}$$

The same likelihood is obtained by replacing $\hat{\boldsymbol{R}}^{-1}$ in (12) with the Moore-Penrose inverse of $\hat{\boldsymbol{R}}$, which is $\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^\intercal$. This is the approach used by msCAVIAR [37].

We summarize the connections between these four approaches in the following proposition.

Proposition 2. *(a) As $\lambda \to 0$, Approach 1a becomes equivalent to Approach 1b, and Approach 2a becomes equivalent to Approach 2b; that is,*

$$\lim_{\lambda \to 0} \ell_{1a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) = \ell_{1b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}); \tag{56}$$

$$\lim_{\lambda \to 0} \ell_{2a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) = \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}). \tag{57}$$

(b) Approaches 1b and 2b are equivalent—i.e., $\ell_{1b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) = \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}})$—*if and only if $\hat{\boldsymbol{z}}$ is consistent with $\hat{\boldsymbol{R}}$.*

(c) *If $\hat{\boldsymbol{z}}$ is inconsistent with $\hat{\boldsymbol{R}}$, Approach 2b behaves discontinuously,*

$$\lim_{\lambda \to 0} \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) \neq \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}), \tag{58}$$

but at the limit it is equivalent to Approach 1b,

$$\lim_{\lambda \to 0} \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) = \ell_{1b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}). \tag{59}$$

Proof. (a) As $\lambda \to 0$, $\hat{\boldsymbol{R}}_\lambda \to \hat{\boldsymbol{R}}$, so (56) is trivially satisfied. To prove (57), we define $\boldsymbol{B} := \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}$ with pseudoinverse $\boldsymbol{B}^\dagger = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\intercal$. With these definitions, we can write $\ell_{2a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda)$ as

$$\begin{aligned} \ell_{2a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) = \exp(&-\tfrac{1}{2}\boldsymbol{z}^\intercal \boldsymbol{B}\boldsymbol{B}^\intercal((1-\lambda)\boldsymbol{B}\boldsymbol{B}^\intercal + \lambda \boldsymbol{I})^{-1}\boldsymbol{B}\boldsymbol{B}^\intercal \boldsymbol{z} \\ &+ \boldsymbol{z}^\intercal \boldsymbol{B}\boldsymbol{B}^\intercal((1-\lambda)\boldsymbol{B}\boldsymbol{B}^\intercal + \lambda \boldsymbol{I})^{-1}\hat{\boldsymbol{z}}). \end{aligned} \tag{60}$$

In the limit as $\lambda \to 0$, $\boldsymbol{B}^\intercal((1-\lambda)\boldsymbol{B}\boldsymbol{B}^\intercal + \lambda \boldsymbol{I})^{-1} \to \boldsymbol{B}^\dagger$ (Theorem 3.4 in [44]). Therefore,

$$\begin{aligned} \lim_{\lambda \to 0} \ell_{2a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) &= \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \boldsymbol{B}\boldsymbol{B}^\dagger \boldsymbol{B}\boldsymbol{B}^\intercal \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{B}\boldsymbol{B}^\dagger \hat{\boldsymbol{z}}) \\ &= \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}\boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{Q}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}}) \\ &= \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}). \end{aligned} \tag{61}$$

(b) $\ell_{1b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) = \ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}})$ for all $\boldsymbol{z}$ if and only if $\boldsymbol{Q}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}$. Since $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{Q}^\intercal$ is an orthogonal projector onto range($\boldsymbol{Q}$), $\hat{\boldsymbol{z}} \in \text{range}(Q) \Leftrightarrow \boldsymbol{Q}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}$ (see [45], Chapter 6).

(c) First we prove (59). Since $\hat{\boldsymbol{R}}_\lambda$ is full rank, the $J \times J$ matrix of eigenvectors $\boldsymbol{Q}_\lambda$ satisfies $\boldsymbol{Q}_\lambda \boldsymbol{Q}_\lambda^\intercal = \boldsymbol{I}_J$. Therefore,

$$\ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) = \exp(-\tfrac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}_\lambda \boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}) = \ell_{1a}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda), \tag{62}$$

so the result follows from (56). The result (58) then follows from part (b).

□

Remark 3. *Proposition 2 (a) and (b) together imply that if $\hat{\boldsymbol{z}}$ is consistent with $\hat{\boldsymbol{R}}$ then, for sufficiently small λ, all four approaches should give the same results (ignoring numerical errors that occur in floating-point computations). However, because the RSS-Z model is only an approximation, $\hat{\boldsymbol{z}}$ will often be inconsistent with $\hat{\boldsymbol{R}}$. It is this fact that causes the methods to give different results.*

Irrelevance of null SNPs

Given that the different approaches to dealing with non-invertible $\hat{\boldsymbol{R}}$ may give different results, it is natural to ask which approach is preferable. Here we argue that Approaches 1a and 1b are preferable to Approaches 2a and 2b because Approaches 1a and 1b always satisfy a simple property that we call "irrelevance of null SNPs." (Approaches 2a and 2b may sometimes satisfy this property, but they are not guaranteed to do so.) This property is also satisifed by the full data likelihood and implies, among other things, that inference under the SER model is independent of the LD matrix.

We motivate this property by observing that the individual-data multiple regression likelihood (1) has the following simple property: if a SNP j has no effect (a "null SNP"), then its genotypes $\boldsymbol{x}_j$ do not appear in the likelihood. As a result, genotypes at

null SNPs do not impact inference for other SNPs. We call this property "irrelevance of null SNPs."

In the summary-data setting, we do not directly observe the genotypes, so we need to formulate an analogous property. Therefore, to translate these ideas to the summary-data likelihoods with $\boldsymbol{z}, \hat{\boldsymbol{R}}$, we ask instead whether the likelihood has the following property: if $z_j = 0$, then all $\hat{R}_{jj'}$, for $j' = 1, \ldots, J$, do not appear in the likelihood. If, for example, $\hat{R}_{jj'}$ is the correlation between SNPs j and j' obtained from a suitable reference panel, then this property implies that the genotypes for SNP j have no impact on the summary-data likelihood when $z_j = 0$. (The same can be said for regularized LD matrices of the form (16).)

To formalize these ideas, we introduce some notation. We use γ to denote a subset of SNPs, $\gamma \subseteq \{1, \ldots, J\}$, and we let $\boldsymbol{z}_\gamma$ denote the elements of the vector $\boldsymbol{z}$ corresponding to the SNPs in γ. The remaining elements are denoted by $\boldsymbol{z}_{-\gamma}$. Similarly, we use $\hat{\boldsymbol{R}}_\gamma$ to denote the matrix containing only the rows and columns of $\hat{\boldsymbol{R}}$ in γ, and $\hat{\boldsymbol{R}}_{-\gamma}$ denotes the matrix containing only the rows and columns of $\hat{\boldsymbol{R}}$ that are not in γ. We then define irrelevance of null SNPs as follows.

Definition 2 (Irrelevance of null SNPs)**.** *Let $\ell(\boldsymbol{z})$ be any likelihood for $\boldsymbol{z}$ (which implicitly depends on $\hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}$, and the model parameters). For any subset $\gamma \subseteq \{1, \ldots, J\}$, let $\ell^\gamma(\boldsymbol{z}_\gamma)$ denote the likelihood for $\boldsymbol{z}_\gamma$ when the remaining elements $\boldsymbol{z}_{-\gamma}$ are set to zero; that is,*

$$\ell^\gamma(\boldsymbol{z}_\gamma) := \ell(\boldsymbol{z}_\gamma, \boldsymbol{z}_{-\gamma} = \boldsymbol{0}). \tag{63}$$

We say the likelihood $\ell(\boldsymbol{z})$ satisfies the irrelevance of null SNPs property if, for all γ, $\ell^\gamma(\boldsymbol{z}_\gamma)$ depends on $\hat{\boldsymbol{R}}$ only through $\hat{\boldsymbol{R}}_\gamma$.

Remark 4. *We have framed the definition in terms of likelihoods for the non-centrality parameters $\boldsymbol{z}$, but a similar definition could be obtained for the effects $\boldsymbol{b}$ by replacing $\boldsymbol{z}$ with $\boldsymbol{b}$.*

The multiple regression likelihood based on individual-level data (8) satisfies the irrelevance of null SNPs property. Likelihoods ℓ_{1a} (51) and ℓ_{1b} (52), which are both based on $\ell_{\text{RSS-}Z}$, but with different choices of $\hat{\boldsymbol{R}}$, also satisfy this property, as summarized by the following proposition.

Proposition 3. *$\ell_{\text{RSS-}Z}$ satisifies irrelevance of null SNPs (Definition 2).*

Proof. Setting $\boldsymbol{z}_{-\gamma} = \boldsymbol{0}$ in (14) yields

$$\ell_{\text{RSS-}Z}(\boldsymbol{z}_\gamma, \boldsymbol{z}_{-\gamma} = \boldsymbol{0}) = \exp(-\tfrac{1}{2}\boldsymbol{z}_\gamma^\intercal \hat{\boldsymbol{R}}_\gamma \boldsymbol{z}_\gamma + \boldsymbol{z}_\gamma^\intercal \hat{\boldsymbol{z}}_\gamma). \tag{64}$$

□

The irrelevance of null SNPs has the following simple implication: to assess support for the hypothesis $H_\gamma : \boldsymbol{z}_\gamma = 0$, one only needs the genotypes corresponding to the non-null SNPs. Indeed, if $p_\gamma(\boldsymbol{z}_\gamma)$ denotes any prior on $\boldsymbol{z}_\gamma$ under H_γ then, in the absence of nuisance parameters, the Bayes Factor for H_γ vs. H_0 is

$$\text{BF}_\gamma := \frac{\int \ell^\gamma(\boldsymbol{z}_\gamma)\, p_\gamma(\boldsymbol{z}_\gamma)\, d\boldsymbol{z}_\gamma}{\ell^\gamma(\boldsymbol{z}_\gamma = \boldsymbol{0})}. \tag{65}$$

This result implies that BF_γ depends only on $\hat{\boldsymbol{z}}_\gamma, \hat{\boldsymbol{R}}_\gamma$, whatever priors are used for each γ (assuming that the prior p_γ does not depend on the null genotypes). This result is easily extended to integrate out additional nuisance parameters (e.g., σ in the multiple regression model) in both the numerator and denominator.

A similar result is shown for specific priors in [12], and is exploited in FINEMAP. Our analysis here emphasizes that this is, fundamentally, due to properties of the likelihood, and is not confined to specific priors.

Remark 5. *Applying this result to the special case that γ contains a single SNP j,* i.e., $\gamma = \{j\}$, BF_γ *depends on the genotypes only through SNP j; in particular, it does not depend on the LD between SNPs. Thus, irrelevance of null SNPs implies that fitting a SER does not depend on LD. Since $\ell_{\text{RSS-}Z}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}})$ (14) satisfies irrelevance of null SNPs, inference under the SER model with the RSS-Z likelihood has the desirable property that it does not depend on LD.*

Optimization of λ in regularized LD matrix

To solve (17), we used the Brent-Dekker algorithm [46], which is implemented in R by the `optimize` function. This algorithm performs a 1-d search over $\lambda \in [0, 1]$. The main computational expense of this optimization step is the eigenvalue decomposition of $\hat{\boldsymbol{R}}_0$. Computing the eigenvalue decomposition has computational complexity $O(J^3)$, and therefore can impose a substantial computational burden on the overall fine-mapping analysis when J is large. In practice, we found that the regularization typically provided only a small improvement to the *SuSiE* fine-mapping results, so in the software we set $\lambda = 0$ by default to avoid this potentially large computational expense.

Likelihood ratio test for detecting allele flips

Based on the conditional distribution (19) and the standardized differences (20), we developed a likelihood ratio test to detect allele flips. When $\hat{\boldsymbol{R}} = \boldsymbol{R}$, t_j should be approximately standard normal; $t_j \sim \mathcal{N}(0, 1)$. However, when $\hat{\boldsymbol{R}}$ is estimated from a reference panel, even without errors such as allele flips, the empirical distribution of standardized differences will be longer tailed than the standard normal. This suggests that a more flexible distribution should be used to model the standardized differences. We used a mixture of normals to model the empirical conditional distribution,

$$\hat{z}_j \mid \hat{\boldsymbol{z}}_{-j}, \hat{\boldsymbol{R}} \sim \sum_{k=1}^{K} w_k \mathcal{N}(-\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}/\Omega_{jj}, \sigma_k^2/\Omega_{jj}), \tag{66}$$

where $\sigma_1, \ldots, \sigma_K$ are prespecified standard deviations such that $\sigma_1 < \cdots < \sigma_K$, and $\boldsymbol{w} = (w_1, \ldots, w_K)$ are mixture proportions (that is, they are all non-negative and sum to 1). In our analyses, we chose the σ_k's such that $\sigma_1 = 0.8$, $\sigma_K = 2 \times \max\{|t_1|, \ldots, |t_J|\}$ and $\sigma_{k+1} = 1.05 \times \sigma_k$. We estimated $\boldsymbol{w}$ by maximum likelihood, using summary data for all SNPs. Computing the maximum-likelihood estimate of $\boldsymbol{w}$ is a convex optimization problem and can be solved efficiently using `mixsqp` [47]. We then used the maximum-likelihood estimates of the mixture weights, $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_K)$, to compute a likelihood ratio for each SNP j,

$$\mathrm{LR}_j := \frac{\sum_{k=1}^{K} \hat{\omega}_k \mathcal{N}(\hat{z}_j; \boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}/\Omega_{jj}, \sigma_k^2/\Omega_{jj})}{\sum_{k=1}^{K} \hat{\omega}_k \mathcal{N}(\hat{z}_j; -\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}/\Omega_{jj}, \sigma_k^2/\Omega_{jj})}. \tag{67}$$

This test can only identify errors in SNPs j with z-scores that are large in magnitude. Therefore, after estimating the mixture weights, we focus on SNPs j with $|\hat{z}_j| > 2$; SNPs j with $|\hat{z}_j| > 2$ and the largest likelihood ratios LR_j are the top candidates for being allele flips.

SuSiE refinement procedure

The refinement procedure is outlined in Algorithm 2. This procedure will work with any individual-level data or summary data accepted by *SuSiE*. In the algorithm, when we say the *SuSiE* model fit s is initialized to fit t, specifically we mean that the posterior

means $\bar{\boldsymbol{b}}_l$ for each single effect $l = 1, \ldots, L$ are initialized from t. (Refer to Algorithm 1 in [17], and Algorithm 1 in this paper.) The default initialization is $\bar{\boldsymbol{b}}_l = 0$. In all cases the default initialization is used for σ^2 and $\boldsymbol{\sigma_0^2} = (\sigma_{01}^2, \ldots, \sigma_{0L}^2)$.

Algorithm 2 *SuSiE* refinement procedure

Require: A *SuSiE* model fit, s, with $K \geq 1$ credible sets, $CS_1, \ldots, CS_K$.

```
Compute ELBO at s, F ← ELBO(s)
repeat
    for k = 1 to K do
        π̃ ← π
        Set prior weights to 0 for all SNPs in CS_k; π̃_j ← 0 for all j ∈ CS_k
        Fit SuSiE model, t_k, using prior weights π̃ and default initialization
        Fit SuSiE model, s_k, using prior weights π, initialized at t_k
        Compute ELBO at s_k, F_k ← ELBO(s_k)
    end for
    k* ← argmax_k ELBO(s_k)
    if F_{k*} > F then
        s ← s_{k*}
    end if
until F_{k*} ≤ F
```

Details of calculations for toy example

In the toy example (see "Fine-mapping with inconsistent summary data and a non-invertible LD matrix: an illustration" in the Results), we assumed $\hat{\boldsymbol{R}}$ is the 2×2 rank-1 matrix of all ones. Then the eigenvalue decomposition of $\hat{\boldsymbol{R}}$ is $\hat{\boldsymbol{R}} = \boldsymbol{Q\Lambda Q}^\intercal$ with $\boldsymbol{\Lambda} = 1$, $\boldsymbol{Q} = (\sqrt{1/2}, \sqrt{1/2})^\intercal$, and $\boldsymbol{QQ}^T$ is the 2×2 matrix with all entries set to $1/2$. In the likelihood $\ell_{2b}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}})$ (see eq. 55), this has the effect that $\boldsymbol{QQ}^\intercal \hat{\boldsymbol{z}}$ is the average of the observed z-scores; that is, $\boldsymbol{QQ}^\intercal \hat{\boldsymbol{z}} = ((\hat{z}_1 + \hat{z}_2)/2, (\hat{z}_1 + \hat{z}_2)/2)^\intercal$.

The *SuSiE-RSS* results for this toy example with $\hat{\boldsymbol{z}} = (6, 7)$ were generated by running `susie_rss` with the default settings (`susieR` version 0.11.48).

Details of simulations

We evaluated the fine-mapping methods on summary data sets generated using real genotypes and simulated phenotypes. For genotype data, we used version 3 of the imputed genotypes from the UK Biobank resource [48]. These data are well suited for fine-mapping because of their large sample size (approximately 500,000 genotyped individuals) and the high density of available genetic variants after genotype imputation [38].

Following [49], we took steps to filter out genotype samples, resulting in a candidate set of 274,549 samples. In detail, we considered only genotype samples marked as "White British" (to limit confounding due to population structure). We then removed any samples from the White British subset that met one or more of the following criteria: mismatch between self-reported and genetic sex; outlier based on heterozygosity and/or rate of missing genotypes; has at least one close relative in the same data set (based on the UK Biobank's kinship and "relatedness" calculations); or does not have a measurement of standing height.

Using this collection of 50,000 genotype samples, we generated $3 \times 200 = 600$ data sets for fine-mapping: 200 non-overlapping regions on autosomal chromosomes, each large enough to contain roughly 1,000 genotyped SNPs, and 3 data sets for each region. A SNP was included in a region if it satisfied all of the following criteria: SNP with at

most two alleles; minor allele frequency of 1% or greater; and information score, which quantifies imputation quality, of 0.9 or greater. 998 SNPs were included in a region on average. The smallest region contained 998 SNPs, and the largest contained 1,001 SNPs. The average size of a region in base pairs was 390 kb.

For each of the 200 randomly chosen regions, we generated three data sets by following a procedure similar to [17]. Our procedure is briefly described here. We simulated phenotypes $\boldsymbol{y}$ under the multiple regression model (1) in which $\mathbf{X}$ was the centered and standardized matrix of 50,000 genotypes. We simulated three sets of phenotype data from the same $\mathbf{X}$ by setting the number of causal SNPs to be 1, 2 or 3. The causal SNPs were chosen uniformly at random among the available SNPs in the region. The causal (non-zero) SNP effects b_j were drawn randomly from the standard normal, then the residual variance σ^2 was adjusted so that the genotypes at all SNPs in the region explained 0.5% of the variance in $\boldsymbol{y}$. The outcomes $\boldsymbol{y}$ were then simulated as $y_i = x_{i1}b_1 + \cdots + x_{iJ}b_J + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$. We then calculated z-scores, $\boldsymbol{z}$, and the in-sample LD matrix $\boldsymbol{R}$ from $\boldsymbol{X}$ and the simulated $\boldsymbol{y}$.

To investigate the impact of misspecification of the LD matrix, we randomly sampled 500 and 1,000 individuals (not overlapping with the 50,000 samples used to compute the z-scores), and computed two "out-of-sample" LD matrices, denoted by $\hat{\boldsymbol{R}}_{500}$ and $\hat{\boldsymbol{R}}_{1000}$, respectively. Because the sample sizes were not large (at most 50,000), it was feasible to compute all in-sample and out-of-sample LD matrices using the function `cor` in `R`.

Fine-mapping methods

Together with *SuSiE-suff* and *SuSiE-RSS* (`susieR` version 0.11.48), we also assessed performance of FINEMAP [12] (version 1.4), CAVIAR [7] (version 2.2) and DAP-G [14, 16] (git commit id 875ba40). All these methods accept summary data of the same form as *SuSiE-RSS*; that is, their inputs are the z-scores $\boldsymbol{z}$ and LD matrix $\hat{\boldsymbol{R}}$. These methods are based on the same multiple linear regression model as *SuSiE-RSS*, differing in the choice of priors, the approach used to compute posterior probabilities, and definition of a credible set.

For simulations using in-sample LD matrices only, we called function `susie_suff_stats` from the `susieR` package with with `L = 10, max_iter = 1000, estimate_residual_variance = FALSE`. We fit *SuSiE* models both with and without the iterative refinement procedure; that is, with `refine = TRUE` and `refine = FALSE`. In "Assessment of fine-mapping methods and LD regularization approaches" the *SuSiE-suff* results are from model fitting with refinement.

Likewise, we called `susie_rss` with both `refine = TRUE` and `refine = FALSE`. All *SuSiE-RSS* results presented in "Assessment of fine-mapping methods and LD regularization approaches" are from calling `susie_rss` with `refine = TRUE`. We set the input argument `L` (the maximum number of non-zero effects in the regression model) to either 10 or, for the "*SuSiE-RSS*, $L =$ true" results, we set `L` to value used to simulate the phenotypes ($L = 1$, 2, or 3). We also set `max_iter = 1000`. The remaining optional arguments in `susie_suff_stat` and `susie_rss` were kept at their default settings.

We ran the FINEMAP program with flags `--sss --n-causal-snps 4`; with these options, FINEMAP used shotgun stochastic search to explore causal configurations, restricting to configurations with at most 4 causal SNPs. For the "FINEMAP, $L =$ true" results, we instead called FINEMAP with `--sss --n-causal-snps L`, where `L` was the number of causal SNPs used in the simulation (1, 2 or 3). Credible sets were introduced in newer versions of FINEMAP. In FINEMAP version 1.4, a credible set was defined conditioned on the number of causal SNPs, k: "For a specific k, FINEMAP takes the k-SNP causal configuration with highest posterior probability and then asks, for the lth SNP in that set, which are the other candidates that could possibly replace that SNP in this causal configuration. The lth credible set shows the best candidate SNPs and their

posterior probability of being in a k-SNP causal configuration that additionally contains $k-1$ SNPs. Note that the $k-1$ SNPs are chosen to have highest posterior probability in their credible set." FINEMAP outputted a set of results for each $k = 1, \ldots, L$. We kept the credible sets from the k with the highest posterior probability.

We ran DAP-G program using the default settings. Note that the default maximum number of causal SNPs in DAP-G (the "maximum model size") is J, the total number of SNPs. The DAP-G software outputs "signal clusters" [16], not credible sets. However, we were able to compute credible sets from the DAP-G output following this description from [16]: "For a signal whose local fdr $\leq t$, it is straightforward to construct a $(1-t)\%$ Bayesian credible set by selecting a minimum subset of SNPs such that their cumulative SNP-level PIPs reach $1-t$." We implemented this calculation in R to generate the DAP-G credible sets.

We ran CAVIAR with flags `-g 0.001 -c L` so that all SNPs had a prior inclusion probability of 1/1000, and the maximum number of causal SNPs was L, where L was the value used to simulate the phenotypes (1, 2 or 3). The remaining CAVIAR parameters were kept at their default settings.

Computing environment

All simulations were run on Linux machines (Scientific Linux 7.4) with Intel Xeon E5-2680v4 ("Broadwell") processors. *SuSiE-suff* and *SuSiE-RSS* were run in R 3.6.1 [50] linked to the OpenBLAS 0.2.19 optimized numerical libraries. At most 2 GB of memory was needed to run *SuSiE-suff* and *SuSiE-RSS* on the simulated data sets, and at most 10 GB was needed to run DAP-G, FINEMAP and CAVIAR. All methods and other computations were run without mutlithreading (one CPU). Runtime statistics for running the methods on summary data with in-sample LD matrices are given in Table 1.

Table 1. Runtimes on simulated data sets with in-sample LD matrix. Average runtimes are taken over all 600 simulations. All runtimes are in seconds. All runtimes include time taken to read the data and write the results to files.

method	min.(s)	average(s)	max.(s)
SuSiE-suff, no refinement	0.40	1.40	18.61
SuSiE-suff, with refinement	1.44	4.81	62.34
SuSiE-RSS, no refinement	0.39	1.31	20.42
SuSiE-RSS, with refinement	1.43	4.64	74.15
SuSiE-RSS, with refinement, $L =$ true	2.00	4.71	10.23
DAP-G	0.66	5.70	371.76
FINEMAP	1.67	16.11	39.27
FINEMAP, $L =$ true	1.00	12.92	42.93
CAVIAR, $L =$ true	3.54	1,516.91	4,831.95

We used the Dynamic Statistical Comparisons system (`https://github.com/stephenslab/dsc`) to perform the simulations. All code implementing the simulations, and the raw and compiled results generated from our simulations, are available at `https://github.com/stephenslab/dsc_susierss`, and were deposited on Zenodo [51].

Software availability

The *SuSiE*, *SuSiE-suff* and *SuSiE-RSS* methods are implemented in the R package `susieR`. It is available for download at `https://github.com/stephenslab/susieR`, and on CRAN at `https://cran.r-project.org/package=susieR`.

Acknowledgments

This work was supported in part by NIH National Human Genome Research Institute grant R01HG002585 and by a grant from the Gordon and Betty Moore Foundation to M. Stephens. G. Wang was supported by NIH National Institute on Aging grant U01AG072572 and funding from the Thompson Family Foundation (TAME-AD). We thank Kaiqian Zhang for her contributions to the development and testing of the `susieR` package. We thank the University of Chicago Research Computing Center and the Center for Research Informatics for providing high-performance computing resources used to run the numerical experiments. This research has been conducted using the UK Biobank Resource under Application Number 27386.

References

1. Hutchinson A, Asimit J, Wallace C. Fine-mapping genetic associations. Human Molecular Genetics. 2020;29(R1):R81–R88.
2. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics. 2012;44(12):1294–1301.
3. Kote-Jarai Z, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, Dadaev T, Jugurnauth-Little S, et al. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression. Human Molecular Genetics. 2013;22(12):2520–2528.
4. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 2018;19(8):491–504.
5. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human Molecular Genetics. 2015;24(R1):R111–R119.
6. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. American Journal of Human Genetics. 2017;101(1):5–22.
7. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014;198(2):497–508.
8. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genetics. 2014;10(10):e1004722.
9. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics. 2007;3(7):e114.
10. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics. 2012;44(4):369–375.
11. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics. 2015;200(3):719–736.

12. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32(10):1493–1501.

13. Li Y, Kellis M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Research. 2016;44(18):e144.

14. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. American Journal of Human Genetics. 2016;98(6):1114–1129.

15. Newcombe PJ, Conti DV, Richardson S. JAM: a scalable Bayesian framework for joint analysis of marginal SNP effects. Genetic Epidemiology. 2016;40(3):188–201.

16. Lee Y, Luca F, Pique-Regi R, Wen X. Bayesian multi-SNP genetic association analysis: control of FDR and use of summary statistics. bioRxiv. 2018;10.1101/316471.

17. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society, Series B. 2020;82(5):1273–1300.

18. Sesia M, Katsevich E, Bates S, Candès E, Sabatti C. Multi-resolution localization of causal variants across the genome. Nature Communications. 2020;11:1093.

19. Wallace C, Cutler AJ, Pontikos N, Pekalski ML, Burren OS, Cooper JD, et al. Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping. PLoS Genetics. 2015;11(6):e1005272.

20. Hutchinson A, Watson H, Wallace C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. PLoS Computational Biology. 2020;16(4):e1007829.

21. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics. 2017;18(2):117–127.

22. Chipman H, George EI, McCulloch RE. The practical implementation of Bayesian model selection. In: Model Selection. vol. 38 of IMS Lecture Notes. Institute of Mathematical Statistics; 2001. p. 65–116.

23. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genetics. 2008;4(10):e1000214.

24. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. American Journal of Human Genetics. 2014;94(4):559–573.

25. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genetics. 2014;10(5):e1004383.

26. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–2191.

27. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

28. Benner C, Havulinna AS, Järvelin MR, Salomaa V, Ripatti S, Pirinen M. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. American Journal of Human Genetics. 2017;101(4):539–551.

29. Zhu X, Stephens M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Annals of Applied Statistics. 2017;11(3):1561–1592.

30. Lozano JA, Hormozdiari F, Joo JWJ, Han B, Eskin E. The multivariate normal distribution framework for analyzing association studies. bioRxiv. 2017;doi:10.1101/208199.

31. Park Y, Sarkar AK, He L, Davila-Velderrain J, De Jager PL, Kellis M. A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. bioRxiv. 2017;doi:10.1101/219428.

32. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 2004;88(2):365–411.

33. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics. 2014;30(20):2906–2914.

34. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genetic epidemiology. 2017;41(6):469–480.

35. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA. DIST: direct imputation of summary statistics for unmeasured SNPs. Bioinformatics. 2013;29(22):2925–2927.

36. Chen W, Wu Y, Zheng Z, Qi T, Visscher PM, Zhu Z, et al. Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. bioRxiv. 2020;doi:10.1101/2020.07.09.196535.

37. LaPierre N, Taraszka K, Huang H, He R, Hormozdiari F, Eskin E. Identifying causal variants by fine mapping across multiple studies. bioRxiv. 2020;doi:10.1101/2020.01.15.908517.

38. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209.

39. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Annals of Applied Statistics. 2011;5(3):1780–1815.

40. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics. 2016;48(5):481–487.

41. Han B, Kang HM, Eskin E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genetics. 2009;5(4):1–13. doi:10.1371/journal.pgen.1000456.

42. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics. 2012;91(6):1011–1021.

43. Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genetic Epidemiology. 2009;33(1):79–86.

44. Albert A. Regression and the Moore-Penrose pseudoinverse. New York, NY: Academic Press; 1972.

45. Trefethen LN, Bau D. Numerical linear algebra. Philadelphia, PA: SIAM; 1997.

46. Brent RP. Algorithms for minimization without derivatives. Mineola, NY: Dover; 2002.

47. Kim Y, Carbonetto P, Stephens M, Anitescu M. A fast algorithm for maximum likelihood estimation of mixture proportions using sequential quadratic programming. Journal of Computational and Graphical Statistics. 2020;29(2):261–273.

48. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine. 2015;12(3):e1001779.

49. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nature Genetics. 2018;50(11):1593–1599.

50. R: a language and environment for statistical computing; 2018. Available from: `https://www.R-project.org`.

51. Zou Y, Carbonetto P, Stephens M. stephenslab/dsc_susierss: release of dsc_susierss repository prior to journal submission; 2021. Available from: `https://doi.org/10.5281/zenodo.5611713`.