Correcting the coverage of credible sets in Bayesian genetic fine-mapping

Anna Hutchinson[1*], Hope Watson[1], Chris Wallace[1,2**]

1 MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK

2 Cambridge Institute for Therapeutic Immunology and Infectious Disease, University of Cambridge, Cambridge, UK

* anna.hutchinson@mrc-bsu.cam.ac.uk

** cew54@cam.ac.uk

Abstract

Genome Wide Association Studies (GWAS) have successfully identified thousands of loci associated with human diseases. Bayesian genetic fine-mapping studies aim to identify the specific causal variants within GWAS loci responsible for each association, reporting credible sets of plausible causal variants, which are interpreted as containing the causal variant with some "coverage probability".

Here, we investigate the coverage probabilities of credible sets through simulations and find that these are systematically biased. We present a method to re-estimate the coverage of credible sets using rapid simulations based on the observed, or estimated, SNP correlation structure, we call this the "corrected coverage estimate". This is extended to find "corrected credible sets", which are the smallest set of variants such that their corrected coverage estimate meets the target coverage.

We use our method to improve the resolution of a fine-mapping study of type 1 diabetes. We found that in 27 out of 39 associated genomic regions our method could reduce the number of potentially causal variants to consider for follow-up, and found that none of the 95% or 99% credible sets required the inclusion of more variants – a pattern matched in simulations of well powered GWAS.

Crucially, our correction method requires only GWAS summary statistics and remains accurate when SNP correlations are estimated from a large reference panel. Using our method to improve the resolution of fine-mapping studies will enable more efficient expenditure of resources in the follow-up process of annotating the variants in the credible set to determine the implicated genes and pathways in human diseases.

Author summary

Pinpointing specific genetic variants within the genome that are causal for human diseases is difficult due to complex correlation patterns existing between variants. Consequently, researchers typically prioritise a set of plausible causal variants for functional validation - these sets of putative causal variants are called "credible sets". We find that the probabilistic interpretation that these credible sets do indeed contain the true causal variant are systematically biased, in that the reported probabilities consistently underestimate the true coverage of the causal variant in the credible set. We have developed a method to provide researchers with a "corrected coverage estimate" that the true causal variant appears in the credible set, and this has been extended to find "corrected credible sets", allowing for more efficient allocation of resources in the expensive follow-up laboratory experiments. We used our method to reduce the number of genetic variants to consider as causal candidates for follow-up in 27 genomic regions that are associated with type 1 diabetes.

Introduction

Genome-Wide Association Studies (GWAS) have identified thousands of disease-associated regions in the human genome, but the resolution of these regions is limited due to linkage disequilibrium (LD) between variants [1]. Consequently, GWAS identifies multiple statistical, but often non-causal, associations at common genetic variants (typically SNPs) that are in LD with the true causal variants. Follow-up studies are therefore required for the prioritisation of the causal variants within these regions, which is an inherently difficult problem due to convoluted LD patterns between hundreds or thousands of SNPs. Consequently, fine-mapping studies prioritise a set of variants most likely to be causal in each risk loci. Laboratory functional studies or large-scale replication studies may be used to identify the true causal variants within these sets, which can then be linked to their target genes to better understand the genetic basis of many human diseases [2, 3].

Early statistical approaches for fine-mapping tended to focus on the SNP in the region with the smallest P value, called the lead-SNP. However, it is generally acknowledged that this SNP may not be the causal variant in a given region due to correlations with the true causal variants [1, 4]. Studies may therefore prioritise the lead-SNP before extending the analysis to include either variants in high LD with this SNP or the top k variants with the highest evidence of association [5].

Fine-mapping is analogous to a variable selection problem with many highly correlated variables (the SNPs) [8]. As such, methods such as penalised regression have also been adopted for fine-mapping, with the aim of choosing the variables representing the variants most likely to be causal for inclusion in the final model [9]. Yet these methods ultimately select one final model and lack probabilistic quantification for this

selected model.

Bayesian approaches for fine-mapping [10–14] use posterior probabilities of causality (PPs) to quantify the evidence that a variant is causal for a given disease, and these can be meaningfully compared both within and across studies. The standard Bayesian approach for fine-mapping was developed by Maller et al. (2012) and assumes a single causal variant per genetic region to prioritise an "$(\alpha \times 100)\%$ credible set" of putative causal variants. This is derived by ranking variants based on their PPs and summing these until a threshold, α is exceeded - with the variants required to exceed this threshold comprising the credible set.

These credible sets are interpreted as having good frequentist coverage of the causal variant [10, 15, 16], although there is no mathematical basis for this [8]. For example, researchers often state that an $(\alpha \times 100)\%$ credible set contains the causal variant with $(\alpha \times 100)\%$ probability [17–21] or with probability $\geq (\alpha \times 100)\%$ [8, 22, 23]. More specifically, they may be interpreted as containing the causal variant with probability equal to the sum of the PPs of the variants in the credible set [24], for which the threshold forms a lower bound. A simulation study found that the coverage of the causal variant in 95% and 99% credible sets varied with the power to detect the signal (S1 Fig in [1]), implying that inferring the frequentist coverage estimate of these Bayesian credible sets may not be as straightforward as the literature suggests.

In this work, we investigate the accuracy of standard coverage estimates reported in the literature and find that these are systematically biased. We develop a new method to re-estimate the frequentist coverage of these credible sets, deriving a "corrected coverage estimate" and extending this to construct a "corrected credible set". We validate our method through simulations and demonstrate it's improved performance relative to standard coverage estimates reported in the literature.

Our method is available as a CRAN R package, `corrcoverage` (https://github.com/annahutch/corrcoverage, https://cran.r-project.org/web/packages/corrcoverage/index.html), which was used to decrease the size of 95% credible sets for 27 out of 39 genomic regions that are associated with type 1 diabetes. Crucially, our method requires only summary-level data and remains accurate when SNP correlations are estimated from a reference panel (such as the UK10K project [25]).

Results

Claimed coverage estimates for credible sets of genetic variants are biased

The standard results from single causal variant fine-mapping are credible sets of putative causal variants that are interpreted as containing the true causal variant with some specified probability [10]. To investigate the true coverage of the causal variant in these Bayesian credible sets, we simulated a variety of single

causal variant association studies using 1006 European haplotypes or 1322 African haplotypes from the 1000 Genomes Phase 3 data set [26]. Sets of SNPs were sampled from genomic regions with various LD patterns and here we present results from two regions; one of low LD (Fig 1D) and one of high LD (Fig 1H), although the results are similar for other LD patterns.

In each region, causality was randomly allocated to one of the variants (with minor allele frequency (MAF) > 0.05) with an additive phenotypic effect (OR; 1.05, 1.1 or 1.2). Sample sizes (NN; number of cases = number of controls = 5000, 10000 or 50000) were also varied across simulations. We calculated the frequentist empirical estimate of the true coverage for each simulated credible set as the proportion of 5000 replicate credible sets that contained the simulated causal variant.

We found that the threshold coverage estimate consistently under-estimated the true coverage (Fig 1A,E), implying that this value could be used as a lower bound for the coverage of the credible set [8, 22, 23]. As the minimum P value across all the SNPs in the region (P_{min}) decreases, the power of the study increases and the ability to detect the true causal variant (and therefore include it in the credible set) increases, while the threshold coverage estimate remains fixed.

We define the "claimed coverage" of credible sets as the sum of the PPs of the variants in the set [24]. In very high powered studies ($P_{min} < 10^{-12}$), the claimed coverage estimates were unbiased but with relatively large variability between estimates, especially in high LD regions. However, we found that these claimed coverage estimates are also systematically biased in representatively powered simulations where fine-mapping is usually performed, $P_{min} > 10^{-12}$ (Fig 1B,F).

Thus, the probabilities that the causal variant is contained within the credible set that are reported in the literature are typically too low, and researchers can afford to be "more confident" that they have captured the true causal variant in their credible set.

Our results imply that the accuracy of the coverage estimates depend on the background LD for the genomic region. We repeated the analysis, varying LD patterns over a much larger population (7562 European UK10K haplotypes) and averaging the results over a range of LD patterns (see Methods). We found that the results were similar to those of the high LD region in Fig 1 (S1 Fig).

We next investigated potential causes of this systematic bias. We found that the PPs themselves are empirically well calibrated (S2 Fig), implying that it is the procedure of forming the credible sets that is causing the bias. Sorting the variants into descending order of PPs prior to the assembly of the credible set ensures that the sets contain as few variants as possible. However, it also confers additional information which is not utilised in the procedure, specifically the ranking of each SNP's PP relative to all the other PPs for the SNPs in the region. We found that removing this ordering step from the algorithm, such that the variants were added to the set in a random order had only a minor effect on the bias (S3 Fig), implying that

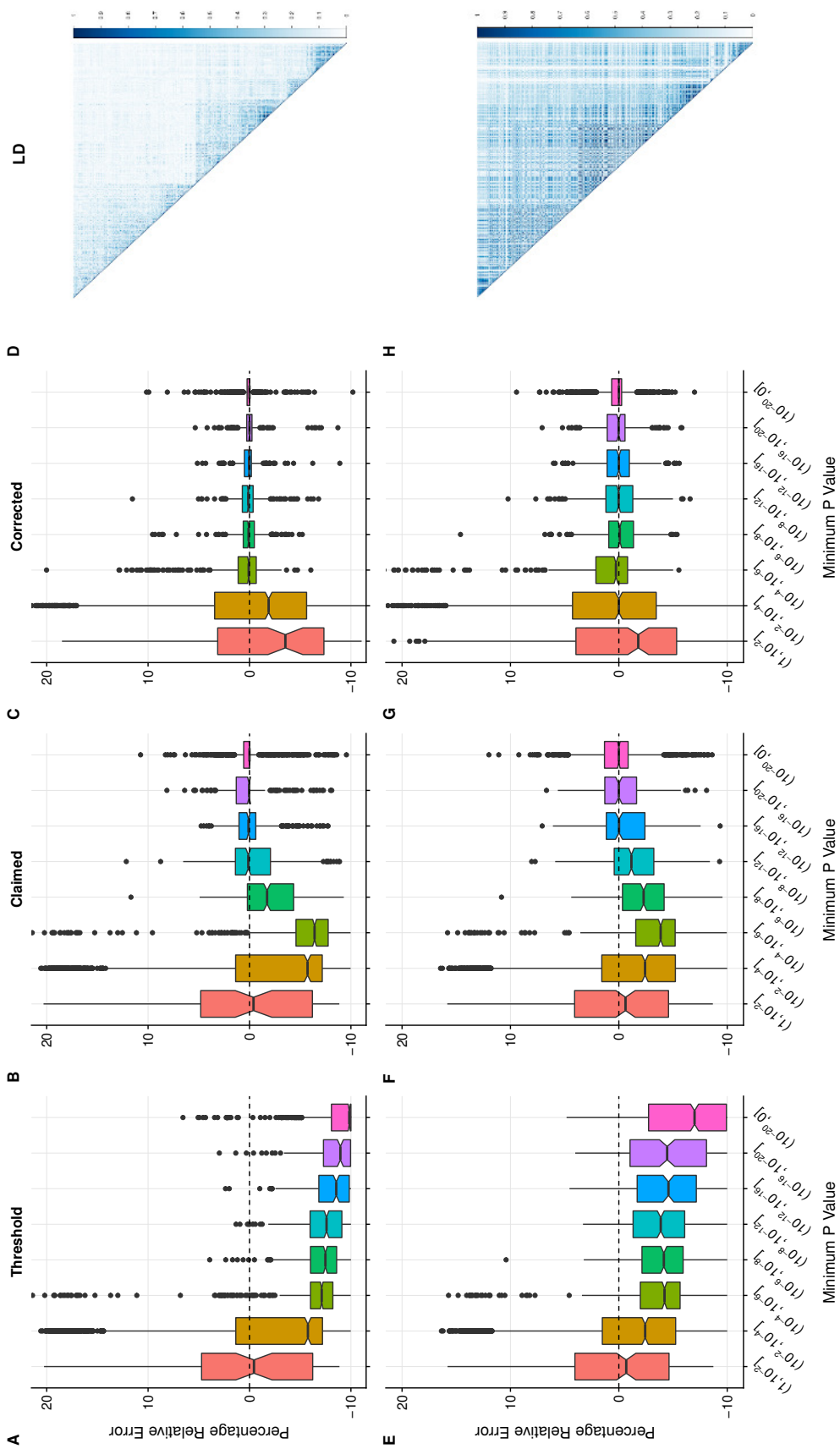

Fig 1. Percentage relative error of coverage estimates for 90% credible sets. Percentage relative error is calculated as [(estimated coverage − empirical coverage)/empirical coverage] × 100 where empirical coverage is the proportion of 5000 replicate credible sets that contain the causal variant. Boxplots showing coverage estimates for 5000 (A-D) low and (E-H) high LD simulations. (A,E) Coverage estimate is the threshold (0.9) (B,F) Coverage estimate is the claimed coverage (the sum of the posterior probabilities of the variants in the credible set) (C,G) Coverage estimate is the corrected coverage. (D,H) Graphical display of SNP correlation matrix.

the step of summing PPs until a target threshold is reached has an effect itself.

The standard Bayesian fine-mapping approach does not incorporate the null model of no genetic effect into the method. In low power, there may not be enough evidence to deduce that there actually is a causal variant in the region, such that if the null model was included in the analysis, then it may hold a substantial proportion of the posterior probability (S4 Fig). This means that omitting the null model from the calculations may contribute to the systematic bias we see in coverage estimates in low powered scenarios.

Corrected coverage estimate improves empirical calibration of credible sets

We developed a new estimator for the true coverage of the causal variant in credible sets, the "corrected coverage estimate", which is based on learning the bias in the system by repeatedly simulating summary GWAS data from the same MAF and LD structure as the observed data. We derive an estimate of effect size at the causal variant from the distribution of Z scores (S5 Fig) and use the observed PPs as weights to tailor the correction as closely as possible to the observed data (see Methods for detailed derivation).

For each of the simulated credible sets, we found that the corrected coverage estimates were better empirically calibrated than the claimed coverage estimates in simulations that are representative of those considered for fine-mapping ($P_{min} < 0.01$) (Fig 1C,G). Particularly, the median accuracy of the corrected coverage estimates improve for $10^{-12} < P_{min} < 10^{-2}$, and their variability also decreases where the claimed coverage estimates are unbiased but with large variability ($P_{min} < 10^{-12}$).

Corrected coverage robust to departures from single causal variant assumption

The Bayesian approach for fine-mapping described by Maller et al. assumes a single causal variant per genomic region, which may be unrealistic [28]. Using simulated data with 2 causal variants, and defining coverage as the frequency with which a credible set contained at least 1 causal variant, we found that the corrected coverage estimates tend to be more accurate than the claimed coverage estimates for causal variants in low LD ($r^2 < 0.01$, Fig 2A). When the 2 causal variants are in high LD ($r^2 > 0.7$), the corrected coverage estimates are still generally more accurate than the claimed coverage estimates, although both tend to underestimate the true coverage (and are thus conservative) (Fig 2B). These results imply that even when the key assumption underlying Bayesian fine-mapping is violated, the corrected coverage estimates are often still better empirically calibrated than the claimed coverage estimates.

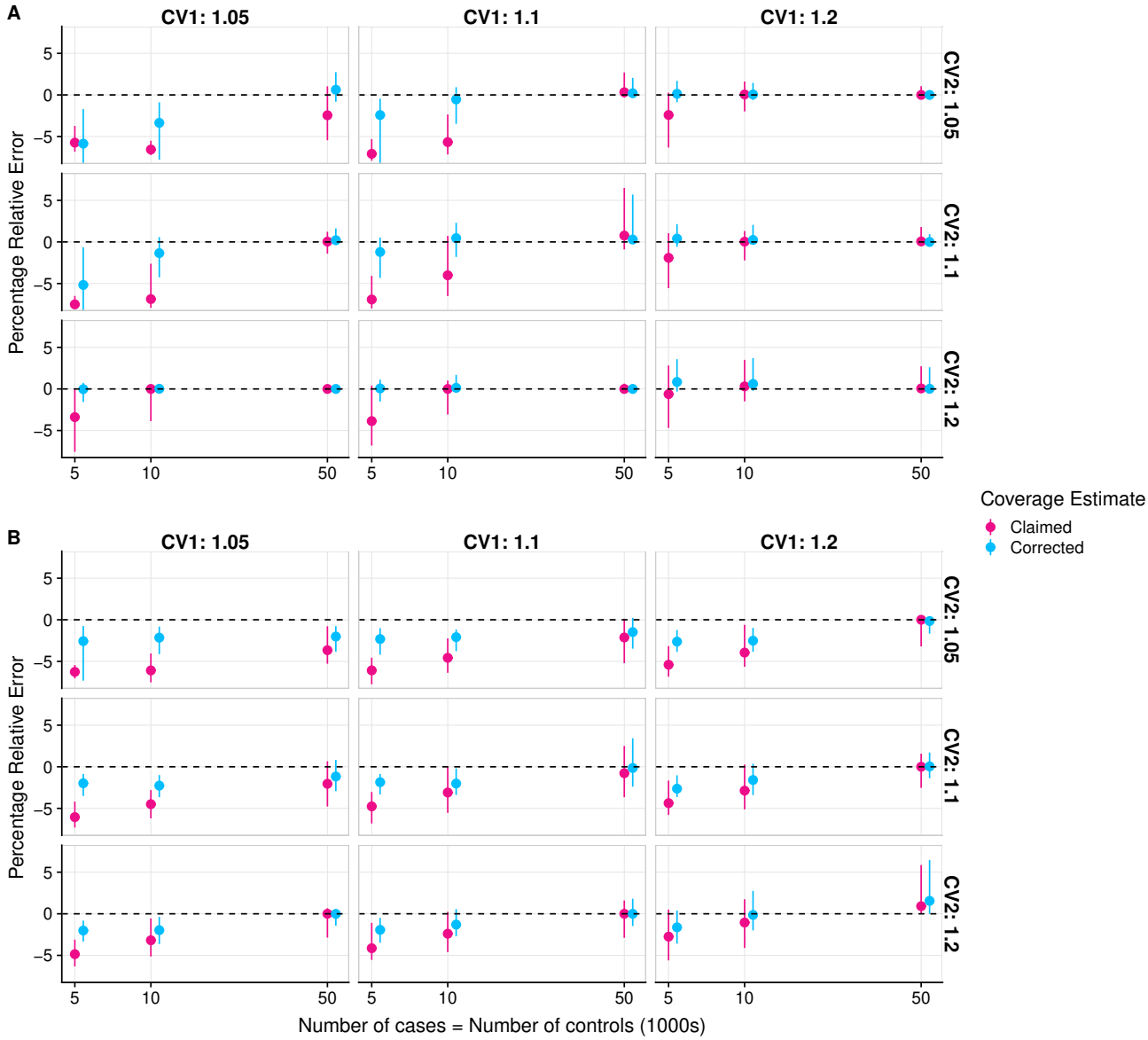

Fig 2. Percentage relative error of coverage estimates for 90% credible sets in regions with 2 causal variants. Percentage relative error is calculated as [(estimated coverage − empirical coverage)/empirical coverage] × 100 where empirical coverage is the proportion of the 5000 simulated 90% credible sets that contain at least one of the 2 causal variants and estimated coverage is the claimed or corrected coverage estimate as defined in the text. The median percentage relative error and interquartile range of claimed and corrected coverage estimates of 90% credible sets from 5000 simulated regions with 2 causal variants that are (A) in low LD ($r^2 < 0.01$) (B) in high LD ($r^2 > 0.7$). Faceted by odds ratio values at the causal variants.

Corrected coverage robust to reference panel estimated MAF and LD

Our method relies on MAF and SNP correlation data to simulate GWAS summary statistics representative of the GWAS data. So far we have assumed that this information is available from the GWAS samples, but this is not generally the case. We therefore evaluated the performance of our correction when using independent

reference data to estimate MAFs and SNP correlations. We applied our correction to sets simulated from the 1000 Genomes data using either 1000 Genomes or UK10K MAF and SNP correlation estimates. We found that the corrected coverage estimates remained accurate (Fig 3).

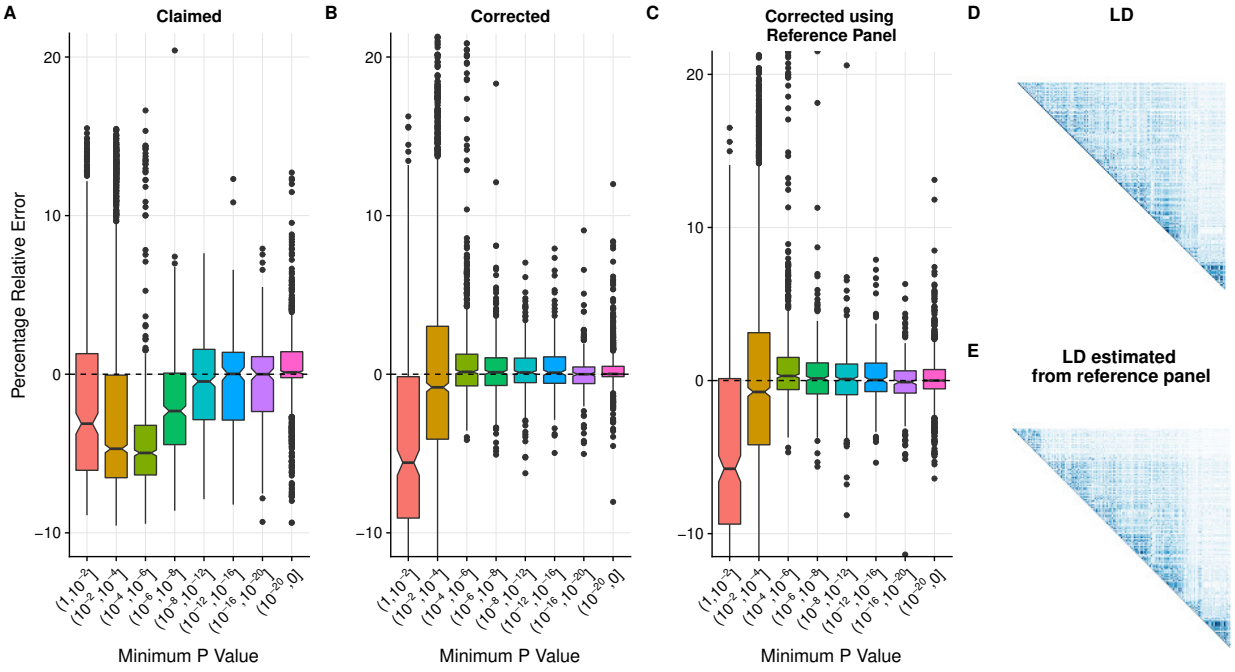

Fig 3. Percentage relative error of coverage estimates for 90% credible sets when using a reference panel to approximate MAFs and SNP correlations. Percentage relative error is calculated as [(estimated coverage − empirical coverage)/empirical coverage] × 100. Coverage estimates from 5000 simulated 90% credible sets. (A) Claimed coverage estimate (the sum of the posterior probabilities of causality for the variants in the credible set) (B) Corrected coverage estimate (C) Corrected coverage estimate using UK10K data to approximate MAFs and SNP correlations (D) Graphical display of SNP correlations in 1000 Genomes (E) Graphical display of the estimated SNP correlations in UK10K.

Corrected credible sets

Obtaining an accurate coverage estimate that the causal variant appears in the credible set is useful in its own right, but it is also beneficial to obtain a "corrected credible set" - that is, the smallest set of variants required such that the corrected coverage estimate of the resultant credible set achieves some desired coverage. For example, discovering that a 90% credible set actually has 99% coverage of the causal variant is useful, but an obvious follow-up question is "What variants do I need such that the coverage is actually 90%?".

We explored this using an example simulated GWAS across 200 SNPs. The 90% credible set, constructed using the standard Bayesian approach, contained 8 variants and had a claimed coverage value of 0.903. The corrected coverage estimate of this credible set was 0.969, which is close to the empirical coverage of the credible set, which was 0.972.

We used the root bisection method [29] to iteratively search for the threshold value required that yields a

credible set with accurate coverage of the causal variant. We found a corrected 90% credible set could be constructed using a threshold value of 0.781. This corrected credible set had a coverage estimate of 0.905 (empirical estimated coverage of 0.907) and reduced in size from 8 to 4 variants (Fig 4).

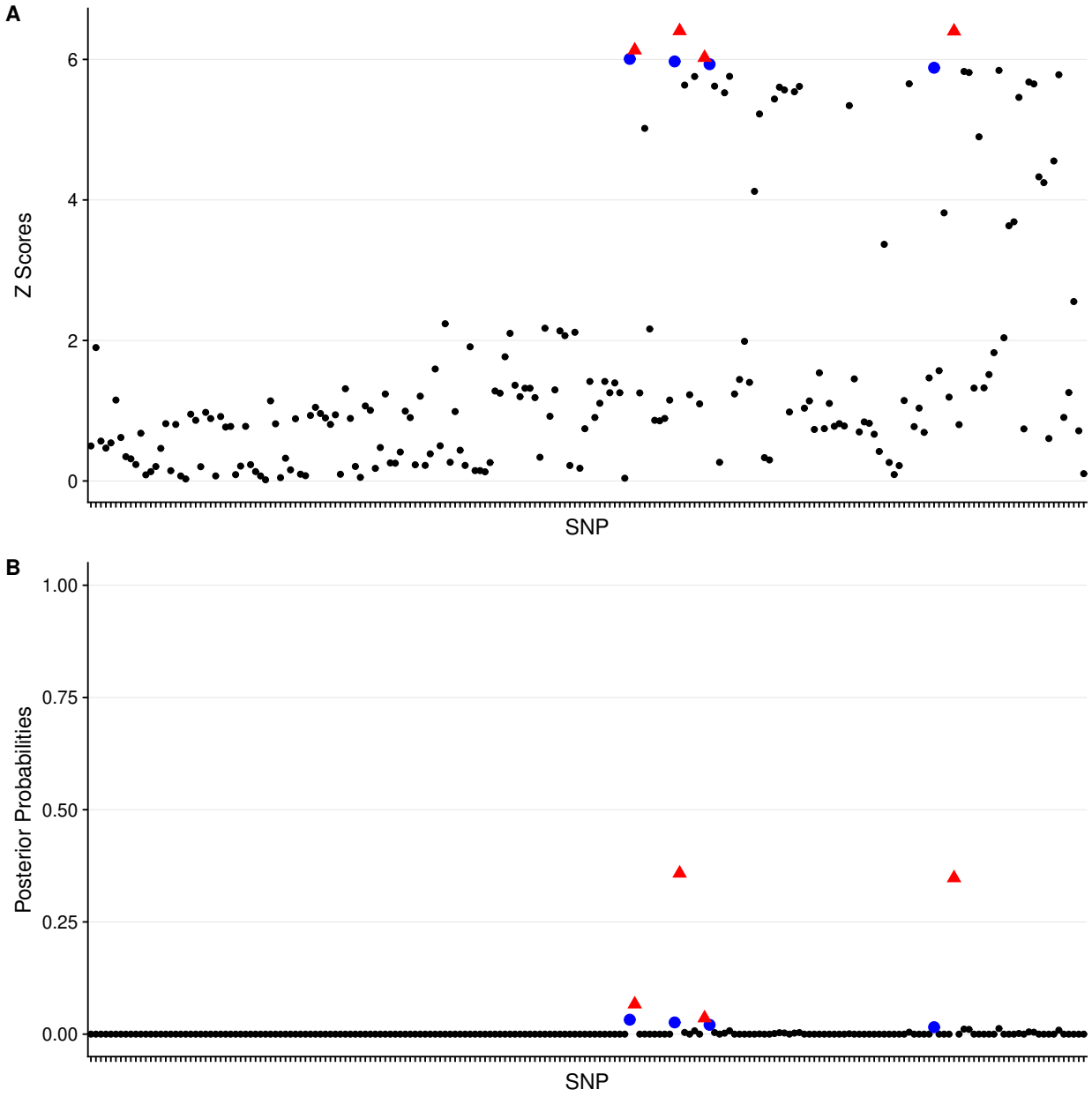

Fig 4. A simple example to illustrate the results of our correction method. (A) The absolute Z scores of the SNPs. (B) The PPs of the SNPs. Red SNPs are those in the corrected 90% credible set and blue SNPs are those that only appear in the standard 90% credible set. The credible set formed of the red SNPs has a corrected coverage estimate of 0.905 and the credible set formed of both the blue and red SNPs has a corrected coverage estimate of 0.969.

Simulations confirmed that the empirical coverage probabilities of corrected credible sets created in this way are accurate (S7 Fig), such that on average the empirical estimate of the true coverage of a corrected

90% credible set is indeed 90%.

`corrcoverage` R package

We created a CRAN R package, `corrcoverage` (https://annahutch.github.io/corrcoverage/, https://cran.r-project.org/web/packages/corrcoverage/index.html), that uses marginal summary statistics to derive corrected coverage estimates and corrected credible sets. The functions to calculate corrected coverage estimates are computationally efficient, taking approximately 1 minute for a 1000 SNP region (using one core of an Intel Xeon E5-2670 processor running at 2.6GHz, S6 Fig).

The functions to derive corrected credible sets require only the summary statistics needed to derive the corrected coverage estimate (Z scores, MAFs, sample sizes and SNP correlation matrix) plus some user-specified desired coverage. Users are able to customise the optional arguments to suit both their accuracy requirements and computational constraints. The algorithm then works iteratively such that the threshold and the corrected coverage estimate of each tested credible set is displayed, until the smallest set of variants with the desired coverage is established, offering researchers an easy tool to improve the resolution of their credible set.

Impact of correcting credible sets in a GWAS

We applied our corrected coverage method to association data from a large type 1 diabetes (T1D) genetic association study consisting of 6,670 cases and 12,262 controls [30]. In the original study, 99% credible sets are found for 40 genomic regions. Here we focus on 95% credible sets as these best illustrate the utility of our method due to the greater margin for error, and we exclude the *INS* region with lead SNP rs689 which failed QC in the ImmunoChip (and for which additional genotyping data was used in the original study).

The results match our previous findings - that the claimed coverage estimates are often too low (Fig 5). We found that the size of the 95% credible set could be reduced in 27 out of the 39 regions, without the use of any additional data (S1 Table S2 Table). Similarly, we found that the size of the 99% credible set could be reduced in 26 out of the 39 regions (S8 Fig, S2 File, S3 Table, S4 Table).

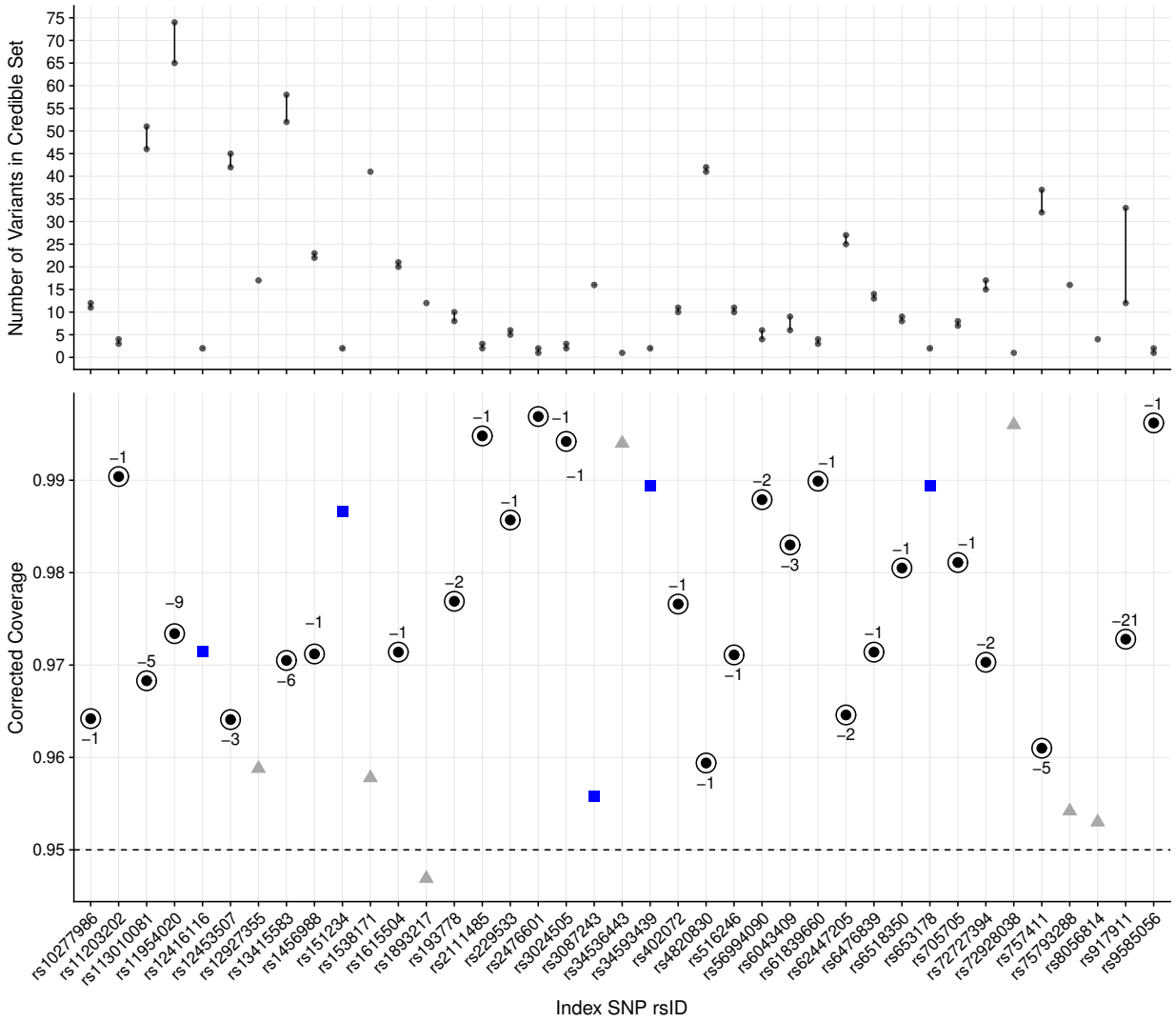

Fig 5. Summary of corrected coverage estimates and corrected credible sets in T1D data set. Top panel: The decrease in size of the credible set after correction. Bottom panel: The corrected coverage estimates of 95% Bayesian credible sets for T1D-associated genomic regions. Black points represents regions where the credible set changed after the correction and the "-" values for the circled points represent the decrease in the number of variants from the standard to the corrected 95% credible set. Blue points represent regions where the credible set did not change after the correction and grey points represent regions where the credible set did not need to be corrected since the threshold was contained in the 95% confidence interval of the coverage estimate, or because the credible set already contained only a single variant.

Fine mapping to single base resolution has been used as a measure of GWAS resolution [23]. Two of the original 95% credible sets only contained a single variant: rs34536443 (missense in *TYK2*) and rs72928038 (intronic in *BACH2*). After applying our correction, two additional 95% credible sets were narrowed down from two variants to a single variant. First, rs2476601 (missense variant R620W in *PTPN22*) was selected, dropping rs6679677 which is in high LD with rs2476601 ($r^2 = 0.996$). These SNPs have high PPs (0.856185774 and 0.143814226, respectively) and the corrected credible set containing only rs2476601 has a corrected

coverage estimate of 0.9501 with 95% confidence interval of (0.9392, 0.9613).

Second, rs9585056 was selected, dropping rs9517719 ($r^2 = 0.483$). rs9517719 is intergenic, while rs9585056 is in the 3' UTR of the lncRNA *AL136961.1*, but has been shown to regulate expression of *GPR183* which in turn regulates an IRF7-driven inflammatory network [31]. While it is likely that R620W is indeed the causal variant at *PTPN22*, there is no conclusive data to evaluate whether rs9585056 is more likely to be causal compared to rs9517719. Nonetheless, the enrichment for missense variants (from 1/2 to 2/4 single variant corrected credible sets) is encouraging. In total, the number of putative causal variants for T1D in credible sets reduced from 658 to 582 upon correction.

Discussion

Bayesian methods for fine-mapping genetic variants in genomic risk loci typically prioritise a credible set of putative causal variants. In this work, we have shown that the inferred probabilities that these credible sets do indeed contain the causal variant are systematically biased. Specifically, whilst credible sets do have good frequentist coverage in very high powered scenarios, this is not the case in the more modestly powered scenarios where fine-mapping is usually performed.

We could not pinpoint the exact source of the bias. We found that the posterior probabilities of causality themselves are well empirically calibrated, which implies that it is the procedure of forming the credible sets that is responsible for the bias. Investigating the cause of this bias and whether this problem arises in other variable selection problems is an interesting direction for future research.

Our method is limited in that it does not model multiple causal variants. Fine-mapping approaches that jointly model SNPs have been developed, such as GUESSFM [4] which uses genotype data and FINEMAP [11] and JAM [14] which attempt to reconstruct multivariate SNP associations from univariate GWAS summary statistics, differing both in the form they use for the likelihood and the method used to stochastically search the model space. The output from these methods are posterior probabilities for various configurations of causal variants, and therefore the grouping of SNPs to distinct association signals must be performed post-hoc to obtain similar inferences to that of single causal variant fine-mapping (e.g. to obtain credible sets).

The sum of single effects (SuSiE) method [8] removes the single causal variant assumption and groups SNPs to distinct association signals in the analysis, such that it aims to find as many credible sets of variants that are required so that each set captures an effect variant, whilst also containing as few variants as possible (similar to "signal clusters" in Lee et al.'s DAP-G method [32]). This sophisticated approach has great potential but the simulated 95% credible sets formed using both the SuSiE and DAP-G methods "typically had coverage slightly below 0.95, but usually > 0.9" (Fig 3 and Fig S3 in [8]). Our method could potentially

be extended to improve on the coverage of credible sets obtained using SuSiE and DAP-G fine-mapping methods.

Our method does not address all the limitations of single causal variant fine-mapping, but it improves on the common inferences that are reported in the literature by researchers. We recommend that our correction is used as an extra step in the single causal variant fine-mapping pipeline, to obtain a corrected coverage estimate that the causal variant is contained within the credible set and if required, to derive a corrected credible set.

Methods

Design of simulation pipeline

We simulated a variety of genetic association studies using African and European haplotypes present in the 1000 Genomes Phase 3 data set [26]. Regions were selected that contained approximately 700 SNPs in low LD (Chr10:6030194-6219451) or high LD (Chr10:60969-431161). Causality was randomly allocated to one of these variants (with $MAF > 0.05$) with an additive phenotypic effect (OR; 1.05, 1.1 or 1.2).

The values for the OR and MAF of the causal variant (CV) were selected such that the simulations reflect the common disease-common variant (CDCV) hypothesis, which asserts that common diseases are caused by common variants with small to modest effects [33, 34]. Sample sizes (NN; number of cases = number of controls = 5000, 10000 or 50000) were also varied across simulations.

The haplotype frequencies and sampled parameter values were then used in the `simGWAS` R package [27] to: (i) simulate the results of a case-control GWAS (the study to "correct") (ii) simulate results from 5000 case-control GWASs (to evaluate the accuracy of our method). These simulated GWAS results are marginal Z scores, which were then converted to PPs using the `corrcoverage::ppfunc` or `corrcoverage::ppfunc.mat` functions.

The variants are then sorted into descending order of their PPs and these are summed until the credible set threshold (0.9) is exceeded. The variants required to exceed this threshold comprise the 90% credible set. The frequentist empirical estimate of the true coverage is calculated as the proportion of the 5000 simulated credible sets that contain the CV. The claimed coverage is defined as the size of credible set that we wish to correct – that is, the sum of the PPs of the variants in the credible set [24], which must be greater than or equal to the threshold by virtue of the method. The corrected coverage estimate is also calculated for each credible set using the `corrcoverage::corrcov` function. The simulation procedure is repeated many times to obtain a final simulation data set, consisting of the sampled parameter values and the empirical, claimed

and corrected coverage estimates of the simulated credible sets.

For the evaluation of averaging results over a range of LD patterns, we used haplotypes from the UK10K data. In each simulation, genomic regions were randomly selected (bounded by recombination hotspots defined using the LD detect method [35]) on chromosome 22 and two non-overlapping sets of 100 adjacent variants were selected, so that the simulated region consisted of 200 correlated and non-correlated variants. The simulation pipeline described above was then followed to obtain a final simulation data set for various LD patterns.

For investigating the effect of the ordering step in the credible set algorithm, we re-ran the simulations whereby the variants were not sorted into descending order of PP prior to assembly of the credible set. This means that they were included into the credible set in a random order until the threshold was exceeded.

For investigating the effect of violating the single causal variant assumption, 2 CVs were simulated in each genomic region, which were in high LD ($r^2 > 0.7$) or low LD ($r^2 < 0.01$), and coverage was defined as the frequency with which a credible set contained at least one of the CVs. The odds ratio quantities of the simulated CVs were sampled independently and sample sizes were varied so that the power of the simulated systems varied (S10 Fig).

Corrected coverage estimate

Associations between a SNP and a trait are usually tested for using single-SNP models, such that marginal Z scores are derived. In contrast, if the SNPs in the region are jointly modelled, then joint Z scores can be derived. Under the assumption of a single CV per region, the expected joint Z scores can be derived and used to write down the joint Z score vector,

$$Z_J = (0, \ldots, 0, \mu, 0, \ldots, 0)^T, \quad (1)$$

where Z_J has length equal to the number of SNPs in the region, and all elements equal to 0 except at the causal SNP's position which takes the value μ.

The value of μ is unknown in genetic association studies and it must therefore be estimated in our method to derive the Z_J vector. We consider using the absolute Z score at the lead-SNP as an estimate for μ, but find this to be too high in low powered scenarios (S5 Fig). This is because $E(|Z|) > 0$ even when $E(Z) = 0$, and thus $E(|Z|) > E(Z)$ when $E(Z)$ is close to zero. Instead, we consider a weighted average of the absolute

Z scores, so that for a region comprising of k SNPs,

$$\hat{\mu} = \sum_{i=1}^{k} |Z_i| \times PP_i. \tag{2}$$

and find this estimate to have small relative error, even at small μ (S5 Fig).

The joint Z vector can now be estimated by

$$\hat{Z}_J = (0, \ldots, 0, \hat{\mu}, 0, \ldots, 0)^T, \tag{3}$$

and the expected marginal Z scores can be written as

$$E(Z) = \Sigma \times \hat{Z}_J, \tag{4}$$

where Σ is the SNP correlation matrix [36].

The asymptotic distribution of these test statistics is multi-variate normal (MVN) with variance equal to the SNP correlation matrix [36],

$$Z \sim MVN(E(Z), \Sigma). \tag{5}$$

We simulate multiple replicates of marginal Z scores, convert these into PPs and derive credible sets. Thus, the proportion of these credible sets that contain the assumed CV can be empirically calculated. For each SNP i considered as the CV, the joint Z vector is constructed as

$$\hat{Z}_J[j] = \begin{cases} 0 & j \neq i \\ \hat{\mu} & j = i \end{cases} \tag{6}$$

and we simulate $N = 1000$ marginal Z score vectors,

$$\mathcal{Z}^*_{N=1000} = \{Z^*_1, \ldots, Z^*_{1000}\} \overset{iid}{\sim} MVN(\Sigma \times \hat{Z}_J, \Sigma). \tag{7}$$

Each simulated Z^* vector is then converted to PPs and credible sets are formed using the standard Bayesian method (sort and sum). The proportion of the $N = 1000$ simulated credible sets that contain SNP i, $prop_i$, is calculated.

This procedure is implemented for each SNP in the genomic region with $PP > 0.001$ (this value can be altered using the 'pp0min' parameter in the software) considered as causal. The final corrected coverage estimate is then calculated by weighting each of these proportions by the PP of the SNP considered causal,

so that for a region containing p SNPs with $PP > 0.001$,

$$\text{Corrected Coverage Estimate} = \sum_{i=1}^{p} PP_i \times prop_i. \quad (8)$$

Intuitively, proportions obtained from realistic scenarios (SNPs with high posterior probabilities of causality considered as causal) are up-weighted and proportions obtained from more unrealistic scenarios (SNPs with low posterior probabilities of causality considered as causal) are down-weighted.

A value of $N = 1000$, so that 1000 credible sets are simulated for each SNP that is considered causal was found to be a robust choice, but is included as an optional parameter in the software. This allows users to increase or decrease the value as desired, for example in the interest of computational time for small or large numbers of SNPs in a genomic region, respectively.

Using a reference panel

We evaluated the performance of corrected coverage estimates when using a reference population to approximate MAFs and SNP correlations, that is, to derive the 'f' and 'Σ' parameters in the relevant functions in the `corrcoverage` R package. In this analysis, we selected an LD block (chr10:6030194-6219451) and chose only the SNPs in this region that could be matched by their position between the 1000 Genomes data and the UK10K data (578 SNPs) for our simulations. European haplotype data for these SNPs was collected from both the 1000 Genomes data (consisting of 503 individuals) and the UK10K data (consisting of 3781 individuals).

As in our standard simulation pipeline, causality was randomly allocated to one of these variants (with $MAF > 0.05$) with an additive phenotypic effect (OR; 1.05, 1.1 or 1.2). Sample sizes (NN; number of cases = number of controls = 5000, 10000 or 50000) were also varied across simulations. These sampled parameter values were then used with MAFs and haplotype data from the 1000 Genomes data to simulate marginal Z scores from various genetic association studies. The standard claimed and corrected coverage estimates (Fig 3A,B respectively) were then derived as usual and the corrected coverage estimates were also calculated when using the reference data to estimate MAFs and SNP correlations (Fig 3C).

For comparison, we also investigated the effect of using a reference panel for the correction in the high LD region previously discussed (we omitted the low LD region as this used African haplotypes, for which we do not have a large representative reference panel). The results were similar, indicating that there is minimal loss of accuracy in corrected coverage estimates when approximating SNP correlations using a reference panel (S9 Fig).

T1D data set

For the T1D data analysis, we used the index SNPs for the genomic regions reported in the original study [30]. For each of these index SNPs, we found the relevant 1000 Genomes build 37 genomic region and used ImmunoChip data to find the other SNPs in each of these regions. We then used the `corrcoverage` R package with default parameters to find 95% (and 99%) credible sets of variants, along with the claimed and corrected coverage estimates for each of these. 95% confidence intervals for the corrected coverage estimates were derived by calculating 100 corrected coverage estimates and taking the 2.5th and 97.5th percentile of these. If 0.95 (or 0.99 for 99% credible sets) did not fall within this confidence interval, then the `corrcoverage::corrected_cs` function with the following optional parameter values: 'acc = 0.0001, max.iter = 70' was used to find a corrected credible set; that is, the smallest set of variants required such that the corrected coverage of the resultant credible set is close to the threshold value (within 0.0001 or as close as possible within 70 iterations).

Supporting information

S1 Fig. Percentage relative error of coverage estimates for 90% credible sets using UK10K data. Percentage relative error is calculated as [(estimated coverage − empirical coverage)/empirical coverage] × 100. Boxplots showing percentage relative error of coverage estimates where (A) coverage estimate equals threshold ($\alpha = 0.9$) (B) coverage estimate equals claimed coverage (sum of the posterior probabilities of the variants in the set) (C) coverage estimate is corrected coverage. Results from 5000 simulations have been averaged over many genomic regions that vary in LD patterns.

S2 Fig. Empirical calibration of PPs. Estimated probability of causality against claimed posterior probability where claimed posterior probabilities are calculated using the ABF approach. Points are the prediction from a $logit(y) \sim logit(x)$ model fitted to 10000 simulations where y is a binary indicator of SNP causality and x is the claimed posterior probability. Grey ribbon shows the 95% confidence interval.

S3 Fig. Analysis of removing the ordering step in the credible set algorithm. The variants were not sorted into descending order of PP prior to assembly of the credible set, and were therefore included into the set in a random order ("unordered"). (A) Percentage relative error of coverage estimates (threshold and claimed) of 90% credible sets formed from unordered variants, calculated as [(estimated coverage − empirical coverage)/empirical coverage] × 100 for low and high LD regions. (B) The number of variants in the credible sets formed using unordered variants against using ordered variants. (C)

Empirical estimate of the true coverage of credible sets formed using ordered and unordered variants.

S4 Fig. Evaluation of including the null model of no genetic effect. (A) Posterior probability of the null model (calculated using 1e-04 as the prior for causality at each variant) for 5000 simulated GWAS. (B) Schematic of how the exclusion of the null model may affect the variant posterior probabilities in low power scenarios. While all variants and the null model are required to reach the target threshold of 0.9, ignoring the null model and rescaling, so that the PPs over the variants sum to 1, implies that some variants will be inappropriately dropped, causing the empirical coverage to be lower than the target.

S5 Fig. Estimating μ. Relative error of μ estimates calculated as $\hat{\mu}_X - \mu$. The x axis is the joint Z score at the CV. Line is fitted using a GAM as the smoothing function (geom_smooth() in `ggplot2`). (A) $\hat{\mu} = \underset{i \in \{1,\ldots,k\}}{max}(|Z_i|)$ (B) $\hat{\mu} = \sum_{i=1}^{k} |Z_i| \times PP_i$.

S6 Fig. R package timings. Curve showing the timings of the `corrcoverage::corrcov` function for different sized genomic regions. For each size of genomic region analysed, 50 replicates of the `corrcoverage::corrcov` function were ran and the mean time taken is plotted. Curve drawn using geom_smooth() function in `ggplot2`. Simulations ran using one core of an Intel Xeon Gold 6142 processor running at 2.6GHz.

S7 Fig. Empirical estimate of the true coverage of corrected 90% credible sets. 5000 simulated 90% credible sets were "corrected" using the `corrcoverage::corrected_cs` function (with default parameters and 'desired.cov=0.9'), and the "required threshold" value obtained from each simulation was used to form 5000 replicate credible sets to estimate the empirical coverage of these corrected 90% credible sets.

S8 Fig. Summary of corrected coverage estimates and corrected 99% credible sets in T1D data set. Top panel: The decrease in size of the credible set after correction. Bottom panel: The corrected coverage estimates of 99% Bayesian credible sets for T1D-associated genomic regions. Black points represents regions where the credible set changed after the correction and the "-" values for the circled points represent the decrease in the number of variants from the standard to the corrected 99% credible set. Blue points represent regions where the credible set did not change after the correction and grey points represent regions where the credible set did not need to be corrected since the threshold was contained in the 99% confidence interval of the coverage estimate, or because the credible set already contained only a single variant.

S9 Fig. Percentage relative error of coverage estimates for 90% credible sets using a reference panel to approximate MAFs and SNP correlations in a high LD region. Percentage relative error

is calculated as [(estimated coverage − empirical coverage)/empirical coverage] × 100. Coverage estimates from 5000 simulations. (A) Claimed coverage estimate (the sum of the posterior probabilities of causality for the variants in the credible set) (B) Corrected coverage estimate (C) Corrected coverage estimate using UK10K data to approximate MAFs and SNP correlations (D) Graphical display of SNP correlations in 1000 Genomes data (E) Graphical display of the estimated SNP correlations in UK10K data.

S10 Fig. Distribution of the minimum P value for 2 CV simulations (Fig 2). 2 CVs are (A) in low LD ($r^2 < 0.01$) (B) in high LD ($r^2 > 0.7$). Faceted by odds ratio values at the causal variants.

S1 File. Individual plots for 95% credible set T1D analysis. Zip file containing Z-score plots, PP plots and Manhattan plots for the 39 T1D association regions analysed.

S2 File. Individual plots for 99% credible set T1D analysis. Zip file containing Z-score plots, PP plots and Manhattan plots for the 39 T1D association regions analysed.

S1 Table. T1D corrected 95% credible set results.

S2 Table. List of 95% credible sets before and after correction.

S3 Table. T1D corrected 99% credible set results.

S4 Table. List of 99% credible sets before and after correction.

Funding

AH is supported by the the Engineering and Physical Sciences Research Council (EP/R511870/1) and GlaxoSmithKline (GSK). CW is supported by the Wellcome Trust (WT107881) and the Medical Research Council (MC UU 00002/4). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank Paul Newcombe and Rob Goudie for helpful discussions, and Kevin Kunzmann for advice on creating R packages.

Author Contributions

Conceived and designed the experiments: AH CW. Performed the experiments: AH HW. Software: AH CW. Wrote the paper: AH CW. Proofed the paper: HW.

References

1. van de Bunt M, Cortes A, Brown MA, Morris AP, McCarthy MI. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. PLoS Genetics. 2015;11(9):1–14. doi:10.1371/journal.pgen.1005535.

2. Ghosh S, Collins FS. The geneticist's approach to complex disease. Annual review of medicine. 1996;47:333–53. doi:10.1146/annurev.med.47.1.333.

3. 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. doi:10.1038/s41576-018-0016-z.

4. 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):1–22. doi:10.1371/journal.pgen.1005272.

5. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551:92.

6. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009;10:681.

7. Wakefield J. Bayes factors for Genome-wide association studies: Comparison with P-values. Genetic Epidemiology. 2009;33(1):79–86. doi:10.1002/gepi.20359.

8. Wang G, Sarkar AK, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. bioRxiv. 2018; p. 501114. doi:10.1101/501114.

9. Valdar W, Sabourin J, Nobel A, Holmes CC. Reprioritizing Genetic Associations in Hit Regions Using LASSO-Based Resample Model Averaging. Genetic Epidemiology. 2012;36(5):451–462. doi:10.1002/gepi.21639.

10. 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. doi:10.1038/ng.2435.

11. 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. doi:10.1093/bioinformatics/btw018.

12. 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. doi:10.1016/j.ajhg.2016.03.029.

13. 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. doi:10.1534/genetics.114.167908.

14. 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. doi:10.1002/gepi.21953.

15. Morris AP. Transethnic meta-analysis of genomewide association studies. Genetic Epidemiology. 2011;35(8):809–822. doi:10.1002/gepi.20630.

16. Kato N, Loh M, Takeuchi F, Verweij N, Wang X, Zhang W, et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. Nature Genetics. 2015;47(11):1282–1293. doi:10.1038/ng.3405.

17. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D), Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature Genetics. 2014;46(3):234–244. doi:10.1038/ng.2897.

18. Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nature Genetics. 2015;47(12):1449–1456. doi:10.1038/ng.3424.

19. Gormley P, Anttila V, Winsvold BS, Palta P, Esko T, Pers TH, et al. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. Nature Genetics. 2016;48(8):856–866. doi:10.1038/ng.3598.

20. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature Genetics. 2016;48(2):134–143. doi:10.1038/ng.3448.

21. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nature Genetics. 2019;51(1):63–75. doi:10.1038/s41588-018-0269-7.

22. Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Mägi R, Reschen ME, et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nature Genetics. 2015;47(12):1415–1425. doi:10.1038/ng.3437.

23. Huang H, Fang M, Jostins L, Umićević Mirkov M, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;547(7662):173–178. doi:10.1038/nature22969.

24. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human Molecular Genetics. 2015;24(R1):R111–R119. doi:10.1093/hmg/ddv260.

25. The UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82.

26. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68.

27. Fortune MD, Wallace C. simGWAS: a fast method for simulation of large scale case–control GWAS summary statistics. Bioinformatics. 2018;35(October 2018):1901–1906. doi:10.1093/bioinformatics/bty898.

28. Asimit JL, Rainbow DB, Fortune MD, Grinberg NF, Wicker LS, Wallace C. Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. Nature Communications. 2019;10(1):3216. doi:10.1038/s41467-019-11271-0.

29. Greene JM. Locating three-dimensional roots by a bisection method. Journal of Computational Physics. 1992;98(2):194–198. doi:10.1016/0021-9991(92)90137-N.

30. Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nature Genetics. 2015;47:381.

31. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature. 2010;467(7314):460–464. doi:10.1038/nature09386.

32. Lee Y, Francesca L, Pique-Regi R, Wen X. Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. bioRxiv. 2018; p. 316471. doi:10.1101/316471.

33. Reich DE, Lander ES, Reich DE, Lander ES. On the allelic spectrum of human disease. TRENDS in Genetics Vol17. 2001;17(9):502–510.

34. Pritchard JK. The allelic architecture of human disease genes: common disease-common variant... or not? Human Molecular Genetics. 2002;11(20):2417–2423. doi:10.1093/hmg/11.20.2417.

35. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2016;32(2):283–285. doi:10.1093/bioinformatics/btv546.

36. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics. 2017;18(2):117–127. doi:10.1038/nrg.2016.142.