Supplementary Note
Genome-wide fine-mapping improves identification of causal variants

Supplementary Note1. Derivation of estimated true positive rate using PIP

We show below that the genome-wide true positive rate given a posterior inclusion probability (PIP) threshold α can be estimated using the PIPs from genome-wide PIPs,

$$\begin{aligned}
\hat{\mathrm{TPR}}(\alpha) &= \Pr(\mathrm{PIP} \geq \alpha | \mathrm{H_1 is\ true}) && (1)\\
&= \frac{\Pr(\mathrm{PIP} \geq \alpha, \mathrm{H_1 is\ true})}{\Pr(\mathrm{H_1 is\ true})} && (2)\\
&= \frac{\Pr(\mathrm{H_1 is\ true} | \mathrm{PIP} \geq \alpha)\Pr(\mathrm{PIP} \geq \alpha)}{\pi} && (3)\\
&= \frac{\sum_i [\mathrm{PIP}_i | \mathrm{PIP}_i \geq \alpha]}{\#\{\mathrm{PIP}_i \geq \alpha\}} * \frac{\#\{\mathrm{PIP}_i \geq \alpha\}}{\mathrm{M}} * \frac{1}{\pi} && (4)\\
&= \frac{\sum_i [\mathrm{PIP}_i | \mathrm{PIP}_i \geq \alpha]}{\mathrm{M}\pi} && (5)
\end{aligned}$$

where M is the total number of SNPs, and π is the proportion of SNPs with nonzero effects on complex trait. One underlying assumption for our derivation above is $\Pr(\mathrm{H_1 is\ true} | \mathrm{PIP} \geq \alpha) = \mathrm{PIP}$, which has been shown is true in Figure 2 of our main text.

When considering each focal SNP j, we can also compute its corresponding true positive rate with the genome-wide PIPs by replacing the threshold α with an observed PIP_j,

$$\hat{\mathrm{TPR}}_j = \frac{\sum_i [\mathrm{PIP}_i | \mathrm{PIP}_i \geq \mathrm{PIP}_j]}{\mathrm{M}\pi} \quad (6)$$

Supplementary Note2. Prediction of power of genome-wide fine-mapping

The aim is to develop a method to predict the power of fine-mapping using a genome-wide Bayesian mixture model (GBMM). From a GBMM, we can estimate the SNP-based heritability (h^2_{SNP}) and the proportion of variants belong to each of the K components of the mixture distribution (π_k with $k \in \{1, ..., K\}$) from the data. Given these estimates, we can estimate the required sample size for achieving a desired power of identifying all causal variants or identifying a subset that explain a desired proportion of h^2_{SNP}, as described below.

Throughout the derivation below, we assume both genotypes and phenotypes are standardised with mean zero and variance one.

1 Sampling distribution of PIP

In GBMM, posterior inclusion probability (PIP) is computed for each SNP and used as the test statistic to test whether the SNP is a causal variant. For example, PIP > 0.9 is often regarded as evidence for detecting a causal variant, assuming all causal variants are observed.

To predict power, our first step is to derive the sampling distribution of PIP. For each SNP, we assume that its effect follows a mixture of a point mass at zero and a number of normal distributions. Here, we focus on a mixture model with 5 components (e.g., SBayesR or SBayesRC), but the theory applies to a arbitrary number of components (e.g., SBayesC with two components).

$$\beta_j \sim \sum_{k=1}^{5} \pi_k N(0, \gamma_k \sigma^2_g) \tag{7}$$

where $\gamma = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5]' = [0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]'$ are the prespecified coefficients to constrain the variance in each effect size distribution with respect to the total genetic variance σ^2_g, and π_k is the prior probability for the SNP effect belong to the k^{th} distribution. Let δ_j be the indicator variable for the distribution membership for SNP j. The PIP is calculated as

$$\text{PIP}_j = 1 - \Pr(\delta_j = 1|y) \tag{8}$$

For notation brevity, we will ignore the subscript j in the following derivation, as it is generic to any SNP. Expanding the posterior probability $\Pr(\delta_j = 1|y)$, we have

$$\text{PIP} = 1 - \frac{f(y|\delta = 1)\pi_1}{\sum_{k=1}^{5} f(y|\delta = k)\pi_k} \tag{9}$$

$$= 1 - \frac{1}{1 + \sum_{k=2}^{5} \frac{f(y|\delta=k)}{f(y|\delta=1)} \frac{\pi_k}{\pi_1}} \tag{10}$$

When $\delta = 1$, the likelihood function is

$$f(y|\delta = 1) \propto \exp\{-\frac{y_c' y_c}{2\sigma_e^2}\} \tag{11}$$

where y_c is the adjusted y for all other effects except β.

When $\delta = k$, the likelihood function is

$$f(y|\delta = k) \propto \int f(\beta|\delta) f(y|\delta = k, \beta) \mathrm{d}\beta \tag{12}$$

$$\propto \exp\{-\frac{y_c' y_c}{2\sigma_e^2}\} (\gamma_k \sigma_g^2)^{-\frac{1}{2}} (C_k^{-1} \sigma_e^2)^{\frac{1}{2}} \exp\{\frac{1}{2} r C_k^{-1} r\} \tag{13}$$

$$\propto \exp\{-\frac{y_c' y_c}{2\sigma_e^2}\} \lambda_k^{\frac{1}{2}} C_k^{-\frac{1}{2}} \exp\{\frac{r^2}{2C_k}\} \tag{14}$$

where

$$\lambda_k = \frac{\sigma_e^2}{\gamma_k \sigma_g^2} \tag{15}$$

$$C_k = n + \lambda_k \tag{16}$$

$$r = X' y_c = X'(e + X\beta) = X'e + n\beta \tag{17}$$

with n being GWAS sample size and $X'X = n$ given standardised genotypes.

Putting equations (11) and (14) together, the ratio of likelihood function can be written as

$$\frac{f(y|\delta = k)}{f(y|\delta = 1)} = \lambda_k^{\frac{1}{2}} C_k^{-\frac{1}{2}} \exp\{\frac{r^2}{2C_k}\} \tag{18}$$

As a result, the summation term in Eq (10) is

$$\sum_{k=2}^{5} \frac{f(y|\delta = k)}{f(y|\delta = 1)} \frac{\pi_k}{\pi_1} = \sum_{k=2}^{5} \frac{\pi_k}{\pi_1} \lambda_k^{\frac{1}{2}} C_k^{-\frac{1}{2}} \exp\{\frac{r^2}{2C_k}\} \tag{19}$$

In the equation above, only r is a random variable that depends on y. Given the true β value, following the normality assumption on y, we have the sampling distribution for r,

$$\mathrm{E}[r] = n\beta \tag{20}$$

$$\mathrm{Var}[r] = n\sigma_e^2 \tag{21}$$

$$r \sim N(n\beta, n\sigma_e^2) \tag{22}$$

Let $v = \beta^2$ be the variance explained by the SNP, then we have

$$r^2 \sim n\sigma_e^2 \chi_1^2(\frac{nv}{\sigma_e^2}) \tag{23}$$

To simplify the equation above, let

$$Z = \chi_1^2(\frac{nv}{\sigma_e^2}) \tag{24}$$

then equation (19) can be expressed as

$$\sum_{k=2}^{5} \frac{f(y|\delta = k)}{f(y|\delta = 1)} \frac{\pi_k}{\pi_1} = \sum_{k=2}^{5} \frac{\pi_k}{\pi_1} \lambda_k^{\frac{1}{2}} C_k^{-\frac{1}{2}} \exp\{\frac{n\sigma_e^2}{2C_k} Z\} \tag{25}$$

Finally, let

$$A_k = \frac{\pi_k}{\pi_1} \lambda_k^{\frac{1}{2}} C_k^{-\frac{1}{2}} \tag{26}$$

$$B_k = \frac{n\sigma_e^2}{2C_k} \tag{27}$$

the sampling distribution of PIP is

$$\text{PIP} = 1 - \frac{1}{1 + \sum_{k=2}^{5} A_k \exp\{B_k Z\}} \tag{28}$$

It shows that PIP is a function of a non-central Chi-square variable with NCP = $\frac{nv}{\sigma_e^2}$. Given the unit phenotypic variance, $\sigma_g^2 = h_{SNP}^2$ and $\sigma_e^2 = 1 - h_{SNP}^2$.

2 Analytic solution for two-component mixture

When the number of mixture components is 2 (i.e, $K = 2$), then a point-normal mixture (i.e., BayesC) is assumed,

$$\beta \sim \pi N(0, \sigma_\beta^2) + (1 - \pi)\phi_0 \tag{29}$$

In this case, the distribution of PIP is

$$\text{PIP} = 1 - \frac{1}{1 + A \exp\{BZ\}} \tag{30}$$

where

$$A = \frac{\pi}{1 - \pi} \lambda^{\frac{1}{2}} C^{-\frac{1}{2}} \tag{31}$$

$$B = \frac{n\sigma_e^2}{2C} \tag{32}$$

For notation brevity, we define P = PIP. Rearraging equation (30) gives

$$Z = \frac{1}{B}(\log P - \log(1 - P) - \log A) \tag{33}$$

$$= u(P) \tag{34}$$

The derivative of $u(P)$ with respect to P is

$$\frac{\partial u(P)}{\partial P} = \frac{1}{B}(\frac{1}{P} + \frac{1}{1-P}) = \frac{1}{B}\frac{1}{P(1-P)} \tag{35}$$

Therefore, the probability density function of PIP is

$$f(P) = f_Z(u(P))|\frac{\partial u(P)}{\partial P}| \tag{36}$$

$$= \text{dchisq}(u(P), 1, \frac{nv}{\sigma_e^2})\frac{1}{B}\frac{1}{P(1-P)} \tag{37}$$

However, it is not straightforward to obtain an analytical solution for a multi-component mixture model. Instead, we use numerical integration for the distribution of PIP to calculate the power, shown in Section 5.

3 Power calculation

Given a GWAS sample size, we are interested in the power for identifying the causal variants. This will, in turn, predict the required sample size to achieve a certain level of power.

Conditional on a per-SNP variance explained v, when a positive result is claimed at the PIP threshold of 0.9, the power can be calculated as

$$\text{Power}_v = \Pr(PIP > 0.9|v) \tag{38}$$

$$= \int_{0.9}^{1} f(P|v)\text{d}P \tag{39}$$

To compute the power for identifying any causal variant, we need to further integrate out v:

$$\text{Power} = \int_{0.9}^{1}\int_{0}^{\infty} f(P|v)f(v)\text{d}v\text{d}P \tag{40}$$

Since $v = \beta^2$, then $\beta = v^{\frac{1}{2}} = u(v)$. The distribution of v is

$$f(v) = f_\beta(u(v))2|u'(v)| \tag{41}$$

$$= f_\beta(v^{\frac{1}{2}})v^{-\frac{1}{2}} \tag{42}$$

where f_β is the distribution of β, which is a mixture of normal distributions.

4 Proportion of SNP-based heritability explained

Given a set of identified variants, we can estimate the expected genetic variance explained (GVE) by these variants

$$\text{E[GVE]} = \text{NCV} \times \text{E}[v|P > 0.9] \tag{43}$$

$$= \text{NCV} \times \int_{0}^{\infty} vf(v|P > 0.9)\text{d}v \tag{44}$$

where

$$f(v|P>0.9) = \frac{f(v, P>0.9)}{f(P>0.9)} \tag{45}$$

$$= \frac{f(P>0.9|v)f(v)}{f(P>0.9)} \tag{46}$$

$$= \frac{f(P>0.9|v)f(v)}{\int_0^\infty f(P>0.9|v)f(v)\mathrm{d}v} \tag{47}$$

$$= \frac{\text{Power}_v \times f(v)}{\text{Power}} \tag{48}$$

Therefore, Eq (52) can be wirtten as,

$$\text{E[GVE]} = m(1-\pi_1)\int_0^\infty \text{Power}_v \times vf(v)\mathrm{d}v \tag{49}$$

The expected proportion of SNP-based heritability explained (PHE) is

$$\text{E[PHE]} = \frac{m(1-\pi_1)\int_0^\infty \text{Power}_v \times vf(v)\mathrm{d}v}{m(1-\pi_1)\int_0^\infty vf(v)\mathrm{d}v} \tag{50}$$

$$= \frac{\int_0^\infty \text{Power}_v \times vf(v)\mathrm{d}v}{\int_0^\infty vf(v)\mathrm{d}v} \tag{51}$$

5 Numerical integration for power calculation

The double integral in equation (40) can be computed numerically after being rewritten to increase the stability of the numerical integration:

$$\int_{0.9}^1 \int_0^\infty f(P|v)\, f(v)\, dv\, dP = \int_{u(0.9)}^\infty \int_0^\infty f(Z|v)\, f(v)\, dv\, dZ \tag{52}$$

Note that $u(P)$ cannot be expressed in closed form but can be easily computed by dichotomy since u is monotonic.
Let λ be the non-centrality parameter: $\lambda = \frac{nv}{\sigma_e^2}$, then

$$f(Z|v) = \text{dchisq}(Z, 1, \lambda) \tag{53}$$

$$= \frac{1}{2}e^{-(Z+\lambda)/2}\left(\frac{Z}{\lambda}\right)^{-1/4} I_{-1/2}(\sqrt{\lambda Z}) \tag{54}$$

$$= \frac{1}{2}e^{-(Z+\lambda)/2}\left(\frac{Z}{\lambda}\right)^{-1/4} e^{\sqrt{\lambda Z}}e^{-\sqrt{\lambda Z}} I_{-1/2}(\sqrt{\lambda Z}) \tag{55}$$

$$= \frac{1}{2}e^{-(Z-2\sqrt{\lambda Z}+\lambda)/2}e^{-\sqrt{\lambda Z}} I_{-1/2}(\sqrt{\lambda Z}) \tag{56}$$

$$= \frac{1}{2}e^{-(\sqrt{Z}+\sqrt{\lambda})^2/2}e^{-\sqrt{\lambda Z}} I_{-1/2}(\sqrt{\lambda Z}) \tag{57}$$

where $I_\nu(w)$ is a modified Bessel function of the first kind and we compute $e^{-w}I_\nu(w)$ to avoid overflow when w is large.

Moreover, the distribution of v follows

$$f(v) = \sum_{k=2}^{5} \frac{\pi_k}{1-\pi_1} \text{dnorm}(\sqrt{v}, 0, \sigma_k^2) v^{-1/2} \tag{58}$$

and (52) becomes

$$\text{Power} = \sum_{k=2}^{5} \frac{\pi_k}{1-\pi_1} \int_{u(0.9)}^{\infty} \int_0^{\infty} \text{dchisq}(Z, 1, \lambda)\, \text{dnorm}(\sqrt{v}, 0, \sigma_k^2)\, v^{-1/2}\, dv\, dZ \tag{59}$$

$$= \sum_{k=2}^{5} \frac{\pi_k}{1-\pi_1} \int_0^{\infty} \text{dnorm}(\sqrt{v}, 0, \sigma_k^2) \int_{u(0.9)}^{\infty} \text{dchisq}(Z, 1, \lambda)\, v^{-1/2}\, dZ\, dv \tag{60}$$

Finally, we proceed to the change of variable $w = \ln(\lambda) = \ln(\frac{nv}{\sigma_e^2})$ to obtain:

$$\text{Power} = \sum_{k=2}^{5} \frac{\pi_k}{1-\pi_1} \int_{-\infty}^{\infty} \text{dnorm}(\sqrt{v}, 0, \sigma_k^2) \int_{u(0.9)}^{\infty} \text{dchisq}(Z, 1, e^w)\, \sqrt{v}\, dZ\, dw \tag{61}$$

with $v = \frac{\sigma_e^2}{n} e^w$

6 Monte Carlo integration

To check if our numerical integration is correct, we estimate the power and PHE using Monte Carlo integration.

First, we draw 10,000 samples of β_j based on the estimated mixture distribution of SNP effects (Eq (7)) and compute $v_j = \beta_j^2$. For each v_j, we draw 1,000 samples from the non-central chi-square distribution (Eq (24)) and compute $P_k|v_j$.

Then, the power to detect v_j in Eq (39) can be computed as

$$\text{Power}_{v_j} = \frac{1}{1000} \sum_{k=1}^{1000} I(P_{k|v_j} > 0.9) \tag{62}$$

where $I(\cdot)$ is an indicator function gives 1 if true, otherwise gives 0.

The power to detect any causal variant in Eq (40) can be computed as

$$\text{Power} = \frac{1}{10000} \sum_{j=1}^{10000} \text{Power}_{v_j} \tag{63}$$

The expected proportion of SNP-based heritability explained in Eq (48) is

$$\text{E[PHE]} = \frac{\sum_{j=1}^{10000} v_j \times \text{Power}_{v_j}}{\sum_{j=1}^{10000} v_j} \tag{64}$$

We found the results from numerical integration and Monte Carlo integration are highly consistent.