COMPUTATIONAL TOOLS

From genome-wide associations to candidate causal variants by statistical fine-mapping

Daniel J. Schaid¹*, Wenan Chen² and Nicholas B. Larson¹

Abstract | Advancing from statistical associations of complex traits with genetic markers to understanding the functional genetic variants that influence traits is often a complex process. Fine-mapping can select and prioritize genetic variants for further study, yet the multitude of analytical strategies and study designs makes it challenging to choose an optimal approach. We review the strengths and weaknesses of different fine-mapping approaches, emphasizing the main factors that affect performance. Topics include interpreting results from genome-wide association studies (GWAS), the role of linkage disequilibrium, statistical fine-mapping approaches, trans-ethnic studies, genomic annotation and data integration, and other analysis and design issues.

Genome-wide association studies

(GWAS). Scans of genetic markers, typically singlenucleotide polymorphisms (SNPs), across DNA of many subjects to find variants statistically associated with a complex trait.

Complex traits

Either quantitative traits (for example, blood pressure and height) or common diseases (for example, major cancers) that are caused by many genetic and environmental factors working together, each having a relatively small effect and few, if any, being absolutely required for disease to occur.

¹Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA.

²Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA. *e-mail: schaid@mayo.edu

https://doi.org/10.1038/ s41576-018-0016-z

Common complex human traits¹ — including quantitative traits and diseases - often result from multiple environmental and genetic causes. Genome-wide association studies (GWAS)¹ have been widely used to identify the genomic regions on chromosomes that harbour genetic determinants of complex traits²⁻⁸. To date, 47,681 statistical associations of single-nucleotide polymorphisms (SNPs) with complex traits are summarized in the US National Human Genome Research Institute (NHGRI)-European Bioinformatics Institute (EBI) GWAS Catalog9, representing 2,185 traits with association P values less than 10^{-5} . This success can be attributed to cost-efficient genotyping microarrays containing a large number of SNPs. However, the SNPs on microarrays typically do not cause¹⁰ the trait. Rather, the SNPs on microarrays, called tag SNPs, are chosen because they are highly correlated (that is, have a large amount of linkage disequilibrium (LD)) with neighbouring SNPs, thereby serving as surrogates for large genomic regions that contain unmeasured SNPs^{11,12}. The association between a tag SNP and a trait can be indirect, resulting from a tag SNP associated with a causal SNP, which in turn is associated with a trait. Because patterns of LD among SNPs can be complex, it can be challenging to determine the underlying causal variants. This is when fine-mapping can help¹³. The principles discussed here are also relevant for common genetic variants measured by whole-genome sequencing studies.

Fine-mapping seeks to determine the genetic variant (or variants) responsible for complex traits, given evidence of an association of a genomic region with a trait and assuming at least one causal variant exists. After an initial GWAS identifies at least one SNP strongly associated with a trait (for example, *P* value $< 5 \times 10^{-8}$), fine-mapping typically proceeds by the general steps outlined in FIG. 1. These steps are discussed in detail below, but the general strategy is to use the GWAS list of SNPs associated with a trait to identify regions of interest. Each region is then visually explored for its LD structure and for genes known to be mapped to the region. Because it is simpler to fine-map one causal variant at a time, each region is partitioned into subregions that have approximately independent effects on the trait. Next, a finemapping strategy is chosen. The strategies we discuss are illustrated in FIG. 2. The selected SNPs are further evaluated for their likely function based on publicly available genomic annotation in order to prioritize costly and time-consuming laboratory-based functional studies.

The main focus of this Review is the statistical methods used to fine-map each region of interest. The overarching goal is to determine which variants are most likely to be functional and to quantify the strength of evidence. This information can then be used for follow-up studies, such as large-scale replication studies of specific candidates, or laboratory functional studies. Although the initial GWAS can provide statistical evidence that a region is likely to harbour a causal variant, additional statistical methods are needed in order to discriminate likely functional variants from variants that are merely correlated with the functional variants. A variety of methods have been used, ranging from simple heuristic methods, to the use of penalized regression for high-dimensional data and to more refined Bayesian methods tailored for fine-mapping (illustrated in FIG. 2). Some of the methods can be applied to a single study or to multiple studies

Tag SNPs

Single-nucleotide polymorphisms (SNPs) that are sufficiently correlated with neighbouring SNPs such that the tag SNP serves as a surrogate for unmeasured SNPs.

Linkage disequilibrium

(LD). Nonrandom association of alleles at different loci on a haplotype in a given population. LD is key to fine-mapping because coinheritance without recombination of alleles from different variants implies that the variants are proximal on the same chromosome.

Causal variants

Genetic variants that mechanistically contribute to diseases or quantitative traits but are not fully penetrant in the sense that the variant may not be a sufficient cause in isolation.

Fine-mapping

To refine the genomic localization of causal variants by the use of statistical, bioinformatic or functional methods.

Penalized regression

A way to estimate regression coefficients by maximizing the log-likelihood of the data while placing a penalty that constrains the size of the regression coefficients, shrinking small coefficients towards zero, sometimes exactly to zero. Although this causes coefficient estimates to be biased, it improves the overall prediction of the model by decreasing the variance of the coefficient estimates.

Summary statistics

Measures of statistical association between a trait and one or more single-nucleotide polymorphisms (SNPs) that summarize the size of effects of the SNPs on the trait, the variances of the effect sizes and how the effect sizes are correlated among themselves. For case-control studies, summary statistics include the estimated log-odds ratios from logistic regression, the variances of the log-odds ratios and the correlations among the log-odds ratios

combined by meta-analysis techniques. When combining multiple studies, special techniques have been developed to simplify sharing of data based on summary statistics¹⁴. When the ethnic background of subjects differs across studies, trans-ethnic fine-mapping can sometimes improve the resolution of fine-mapping. Each of these topics is discussed, emphasizing their strengths, weaknesses and challenges. We also present the main factors that influence power and resolution of fine-mapping, which in turn offer guidance for study designs. We review the use of genomic annotation for fine-mapping, as well as the integration of gene expression data with GWAS data. Finally, we discuss future challenges as our understanding of the genetic basis of complex traits evolves.

Interpreting lead SNPs from GWAS

The decision to fine-map a region typically follows the discovery of genome-wide significant results from a GWAS. It is common to summarize GWAS results with a Manhattan plot of all *P* values that measure the marginal association of one SNP at a time with a trait, followed with LocusZoom plots for regions of interest¹⁵ (FIG. 1). This allows one to focus on the SNPs with the smallest (that is, most significant) *P* values in distinct regions, sometimes called the lead or index SNPs. GWAS results are most reliable when SNP associations achieve the accepted genome-wide statistical significance threshold of *P* value < 5×10^{-8} (REFS^{16,17}), a threshold that accounts for multiple testing correction, although some investigators use a weaker threshold of *P* value < 10^{-6} to highlight regions that are suggestive of harbouring causal variants.

A limitation of the lead SNP is that there is a reasonable chance that it is not the causal variant. This can occur because GWAS microarrays are based on tag SNPs, with the tag SNP merely correlated with the unmeasured causal SNP. Furthermore, even if the causal SNP is measured or imputed, there is a reasonable chance that the statistical association of the causal SNP with the trait is not the most significant association among all the associated SNPs when statistical power is not large. For example, based on simulations with 1,000 cases and 1,000 controls, there was a 79% chance that the lead SNP was found to be causal when the odds ratio (OR) for disease risk was 1.5 and the risk allele frequency was 50% but only a 2.4% chance when the OR was 1.1 and the risk allele frequency was 5%18. Zaykin et al.19 considered multiple causal variants and the impact of LD and drew similar conclusions that true associations are not likely to result in the smallest P values, in part owing to the small effect sizes of variants on complex traits. These findings emphasize the importance of caution when considering the lead SNP as likely causal and the importance of fine-mapping in order to identify the causal variant or variants.

LD for population fine-mapping

Fine-mapping in population-based studies leverages measures of nonrandom associations between pairs of loci. When loci are near each other and the frequency of recombination between them is low, the alleles from the different loci that occur on the same chromosome (called a haplotype) tend to be inherited as a unit. For alleles on a haplotype, occurrence together more than by chance is referred to as gametic association or more commonly as LD. Various measures of LD have been proposed²⁰, but they all depend on the difference between the observed joint frequency of alleles occurring on the same haplotype versus that expected by random chance. The most frequently used measure of LD is a standardized difference, which can be easily estimated by the Pearson correlation between the counts of the minor alleles (that is, less-common alleles) for two SNPs. This correlation coefficient is directly related to statistical power, and it is a reasonable measure for fine-mapping, although for case–control studies of rare diseases, measures such as attributable risk can perform better²¹.

Using LD to fine-map a complex trait is based on the premise that ancestral meiotic recombinations diminish LD, implying that the SNP with the strongest association with a trait is either the causal variant or close to the causal variant. However, evaluating one SNP at a time can be misleading owing to the complex patterns of LD in a genomic region. A classic example is the association of Alzheimer disease with multiple SNPs around the APOE locus on chromosome 19. As the distance from APOE increases, the P values for the association of SNPs with disease do not follow a monotonic pattern but rather increase and decrease²². This initiated a debate of whether there are genes in this region other than APOE that cause Alzheimer disease. However, the large amount of LD in this region has made it challenging to resolve this issue²³. Factors beyond recombination that influence LD are mutation rates of the genetic markers, natural selection, population migrations and admixture, population bottlenecks and demographic history (for example, population size and mating patterns)24. Because LD is influenced by factors other than recombination, it is too limiting to rely solely on patterns of pairwise LD, or even haplotype blocks, to provide reliable fine-mapping of complex traits.

Factors influencing fine-mapping

A number of factors influence the performance of finemapping, including the number of causal SNPs in a region and their effect sizes on the trait, the local LD structure, sample size, SNP density and whether the causal variants can be measured. Careful definition of the phenotype to enrich for genetic causes (for example, disease severity and strength of family history) might increase the genetic effect size. The local LD structure is difficult to control, but trans-ethnic meta-analyses can capitalize on differences in LD structure. Factors that can be controlled in study designs are the sample size and SNP density. Sample size can be increased by pooling different studies or performing meta-analyses, recognizing that disease heterogeneity might increase, thus diluting efforts. Because it is critical to have high SNP density to capture the causal variants, in the following section, we focus on strategies to increase density.

Increasing SNP density

The density of SNPs can be increased by DNA sequencing, but the costs for sequencing the large sample sizes required for fine-mapping can be prohibitive. Alternative cost-efficient strategies include genotype imputation and additional genotyping.



Fig. 1 | Flow of a typical process from initial GWAS to annotation of SNPs selected from fine-mapping analyses. Based on genome-wide association study (GWAS) *P* values summarized in a Manhattan plot¹¹², a list of single-nucleotide polymorphisms (SNPs) that achieve genome-wide statistical significance (that is, *P* value < 5 × 10⁻⁸) is used to determine regions of interest for fine-mapping. Each region is typically explored according to the structure of linkage disequilibrium (LD) among SNPs using Haploview plots. Statistical associations are viewed with LocusZoom plots that illustrate the patterns of association of each SNP with the lead SNP, as well as the annotation of genes in the region. The regions can then be partitioned into independent subregions to ease computational burden, based on statistical models that evaluate the simultaneous effects of multiple SNPs on a trait. Statistical fine-mapping is conducted in each region, using one of the methods illustrated in FIG. 2. The SNPs selected from fine-mapping are then annotated with genomic features to prioritize follow-up functional studies. eQTL, expression quantitative trait locus. Figure is reproduced with permission from REF.¹¹², Springer.

Trans-ethnic

A type of genetic association study that includes subjects from more than one ethnic background.

Multiple testing correction When testing more than one

statistical association, the probability of declaring at least one significant result increases as the number of statistical tests increases. If each of m independent statistical tests uses P value < α to declare significance, then the chance that at least one of the *m* tests is found to be significant is approximately ma. Multiple testing correction maintains the overall chance of declaring at least one significant result by using more stringent P value thresholds for each association tested. The Bonferroni correction uses P value $< \alpha/m$ to test each association

Genotype imputation. Imputation of SNPs can fill in sporadic missing genotypes, harmonize data from different GWAS genotyping arrays to perform a pooled or meta-analysis and increase the density of SNPs for fine-mapping^{25,26}. Key criteria for imputation success are high correlation of directly assayed SNPs with the untyped SNPs and appropriate reference panels that provide templates for LD patterns and allele frequencies that are representative of the study sample²⁵. Popular reference panels are the 1000 Genomes Project²⁷ and the Haplotype Reference Consortium²⁸. Although imputed SNPs tend to be robust to the choice of quality control filters²⁹, power to detect associations with a trait decreases as imputation accuracy decreases²⁵.

Additional genotyping. Because imputation accuracy depends on the local LD structure, regions with weak LD might require actual genotyping to accurately evaluate their association with a trait. Situations where additional genotyping helps are: validation of imputed SNPs, possibly improving fine-mapping by reducing genotype measurement error; discovery of low-frequency SNPs that are not in strong LD with a lead SNP; identification of SNPs that are not well represented in a reference panel; and disentanglement of

SNPs that appear to be in perfect LD because of a small reference sample³⁰.

Cost-efficient genotyping is becoming more accessible with the development of custom genotyping arrays that target certain diseases or traits, such as OncoArray for common cancers³¹; Metabochip for metabolic, cardiovascular and anthropometric traits³²; and Immunochip for major autoimmune and inflammatory diseases³³. These specialized arrays are primarily designed to evaluate associations of SNPs that are known to be associated with specific diseases or traits. Notable advantages of these arrays are increased density of SNPs in known gene regions and cost-efficiency to achieve large sample sizes for fine-mapping. However, there are limitations. First, array content is focused on SNPs or genes that are already known to be associated with specific diseases or traits, which might exclude the finding of novel genes. These arrays also contain a backbone of genome-wide tag SNPs that might overcome this limitation, albeit at a lower resolution than high-density arrays that are typically used for initial GWAS discovery. Second, array content can be based on reference data with incomplete coverage. Third, array content is based on SNPs rather than structural variation. Fourth, SNPs may be excluded from

















• Leverage ethnic differences in LD at a given locus

Ea Locus 1 Eb Locus 2 Eb Loc

E Multi-region fine-mapping with annotation



494 | AUGUST 2018 | VOLUME 19

Fig. 2 | Hypothetical examples of fine-mapping strategies. All subfigures are based on LocusZoom-style illustrations of marginal single-nucleotide polymorphism (SNP) associations. The $-\log_{10}(P)$ values are presented on the left y axis, and variant positions are on the x axis. The gold diamond for each locus represents the peak SNP. The results for other SNPs are coloured by descending degree of linkage disequilibrium (LD) with the peak SNP (ordered red, orange, green and blue dots). The purple bars represent additional variant-level statistics produced by fine-mapping (that is, β -values for penalized regression and posterior inclusion probabilities (PIPs) for Bayesian methods), and the corresponding scale is on the right y axis. The light grey boxes represent the regions selected by fine-mapping. A | The heuristic approach is based on LD patterns with the peak SNP (rs12345). All SNPs that meet the orange LD category threshold have been selected. **B** | The penalized regression approach selects all SNPs whose effects (β , right axis) are not shrunk to zero. **C** | Bayesian fine-mapping produces SNP-level PIPs (right axis), which can be summed to form credible sets on the basis of a specified coverage probability threshold (for example, 95%). This example illustrates that the peak SNP does not correspond to the SNP with the highest PIP, which can occur because of the correlation structure among all SNPs in the region. D Bayesian trans-ethnic fine-mapping. Results for the same locus are illustrated for two diverse study populations (Pop. 1 (part Da) and Pop. 2 (part Db)) with different local LD structures. The peak SNPs for the two analyses differ (rs12345 versus rs23456), and combining the results through meta-analysis yields a narrowed fine-mapping credible region (part Dc). E | Joint analysis of multiple loci by integrating annotation in Bayesian fine-mapping can improve fine-mapping by borrowing annotation information across loci. Presented are three example independent loci (parts Ea, Eb and Ec), along with two corresponding regional annotations, indicated by bands below each locus plot. For loci 1 and 2, the peak SNPs overlap with Annotation 1, indicating enrichment. This enrichment results in the SNP with the highest PIP in locus 3 to be different from the peak SNP in locus 3.

arrays owing to assay failure. Finally, array designs are primarily based on European ancestry.

Partition to independent regions

When analysing one SNP at a time, it is common to find multiple SNPs in a region to be marginally associated with a trait. This can occur when multiple non-causal SNPs are correlated with a single causal SNP. The LD among the SNPs will make each SNP appear to be associated with a trait when analysing one SNP at a time. By contrast, when analysing all SNPs jointly, only the causal SNP is expected to be associated with the trait when accounting for the LD among the SNPs, simplifying fine-mapping efforts. Alternatively, when there are in fact multiple causal SNPs and they are correlated with each other and with other non-causal SNPs, finemapping can be more challenging because joint analyses are not expected to highlight just one causal SNP.

To simplify fine-mapping, it can be advantageous to first partition the region of interest according to the independent effects of SNPs on the trait and then separately fine-map each partition. To determine if multiple marginally associated SNPs are driven by one or a few independent statistical associations, conditional forward stepwise regression is often used (for example, logistic regression for case-control studies or linear regression for quantitative traits). This entails conditioning on the lead SNP from a GWAS by treating it as an adjusting covariate in a regression model and testing the remaining SNPs in the region of interest. If a secondary SNP is found to be statistically significant after adjusting for the primary SNP, sequential testing proceeds by treating some SNPs as adjusting covariates while screening the remaining SNPs until no conditional tests are significant. A challenge is to decide on the threshold for significance. Some investigators use the stringent GWAS threshold

P value $< 5 \times 10^{-8}$, while others use more liberal thresholds, such as *P* value $< 10^{-4}$ or even *P* value < 0.05.

There are several limitations of forward stepwise conditional regression. First, as the number of steps increases, the number of statistical tests increases. If there are *m* SNPs, then after *k* sequential steps, approximately *km* statistical tests will be performed, increasing the chance of a false positive result. Second, when *m* is large, perhaps close to the number of subjects in the sample, and when a liberal threshold is used to include SNPs at each step, the forward selection procedure becomes unstable³⁴ and is overly optimistic regarding the trait variation explained by the selected SNPs³⁵. Third, the probability (that is, statistical power) of detecting secondary signals diminishes as the correlation among SNPs increases.

To illustrate diminishing power to detect a secondary SNP by conditional analysis, we derived simple formulas (Supplementary Box 1) to examine power. We assumed 90% power to detect the primary SNP when it explains 1% of a quantitative trait variation. We set the effect size for the secondary SNP to be a fraction of that for the primary SNP (fraction 50-100%), and we assumed the stringent GWAS threshold *P* value $< 5 \times 10^{-8}$ for testing the effect of the secondary SNP. FIGURE 3 shows a dramatic loss in power to detect secondary signals as the correlation of SNPs increases, even at levels of SNP correlation $\rho = 0.2$. Power also decreases as the effect size of the secondary SNP decreases. Hence, one should be cautious when using conditional testing to declare only a single SNP to be statistically significant because low power can cause missed secondary associations.

Types of fine-mapping approaches

A number of approaches have been used to perform fine-mapping. We present three main strategies that have been used in the literature: heuristic methods, penalized regression models and Bayesian methods (illustrated in FIG. 2). Heuristic approaches were the first to be used, having grown out of practical experience and educated guesses, but with loosely defined criteria. Penalized regression models were developed in other fields of statistics, with the aim to reduce highdimensional predictor variables (for example, SNP data for fine-mapping) to a much smaller set that are strongly associated with a trait. In recent years, Bayesian methods have been specifically tailored for fine-mapping. The main features of these three approaches are discussed, along with their benefits and limitations.

Heuristic fine-mapping approaches. Because the LD structure around the lead SNP from GWAS has a substantial role in fine-mapping, it is popular to first examine the correlation among the SNPs surrounding a lead SNP. One approach is to filter SNPs according to their pairwise correlation (r^2) with the lead SNP, retaining as potentially causal only those SNPs with an r^2 above a threshold. Alternatively, hierarchical clustering of all SNPs in a region based on their pairwise r^2 has been used to create clusters. However, both of these approaches depend on arbitrary thresholds to filter or form clusters. More rigorous penalized models and Bayesian methods

Statistical power

The probability of correctly rejecting a null hypothesis of no statistical association between a single-nucleotide polymorphism (SNP) and a trait when in truth a statistical association exists. Power depends on the magnitude of the SNP effect, the sample size and the *P* value threshold for deciding statistical significance.

Haplotype

A combination of alleles found on the same chromosome.

Haplotype block

A set of highly associated alleles on a chromosome that tend to be inherited together.

1.0 Effect size SNP, (% of SNP, effect) - 100% - - - 80% ----- 50% 0.8 0.6 Power 0.4 0.2 0.0 0.0 0.2 0.4 0.6 0.8 1.0 SNP correlation, ρ

Fig. 3 | **Power of conditional analysis.** This figure illustrates how conditional analyses have weaker power to detect secondary associated single-nucleotide polymorphisms (SNPs) than does an initial genome-wide association study (GWAS). Power of conditional analyses diminishes as the correlation of a primary SNP (indicated by SNP₁) and a secondary SNP (indicated by SNP₂) increases and when the effect size of a secondary SNP is weaker than that for a primary SNP. For this figure, the power for an initial GWAS to detect a primary SNP₁ is 90% for an effect size of $R^2 = 1\%$ of explained trait variation. The effect size of a secondary SNP₂ is varied from 100% to 50% of the effect size of primary SNP₁.

Genotype imputation

A method for estimating (imputing) the unobserved genotypes of study subjects, both for individuals with missing or unreliable genotypes at a genotyped single-nucleotide polymorphism (SNP) and for all individuals at an ungenotyped SNP.

Recombination hot spots

Genomic regions where the rate of recombination is much higher than the neutral expectation.

Cross-validation

A technique to build a prediction model by randomly partitioning the sample into a training set to train the model (for example, determining which single-nucleotide polymorphisms (SNPs) to include in a model) and a test set to measure its predictive performance (for example, average squared prediction error). It is common to split the original sample into ten equally sized subsamples, use nine to train and one to test, repeat this process ten times such that each of the ten subsamples is used as a test sample, and then average the predictive performance over the ten training subsamples.

that jointly model the simultaneous effects of multiple SNPs on a trait are more informative.

Another way to view correlation structure is by pairwise LD among SNPs within haplotypes, using software such as Haploview³⁶ (FIG. 1). This gives a visual impression of discrete haplotype blocks³⁷. Combining the GWAS lead SNP with SNPs in the same haplotype block is an alternative way to select potential causal SNPs. However, one should be cautious with this approach. Recombination hot spots strongly influence block features, yet block boundaries can be arbitrary owing to the choice of statistical model parameters and computational method, as well as genetic marker density and allele frequencies^{38,39}. Internal inconsistencies of blocks are troubling, such as when two markers in strong LD flank markers with little LD with one of the flanking markers³⁸. The above heuristic methods to choose SNPs for functional follow-up are too limiting because they do not account for the joint effects of the SNPs on the trait, and they do not give an objective measure of the confidence that a SNP is causal but rather rely on somewhat arbitrary thresholds and subjective interpretations of correlations among SNPs.

Penalized regression models. An alternative finemapping approach is to use a regression model to jointly analyse all the SNPs in a region. Traditional model building is based on forward selection (or alternative stepwise methods), using *P* values to determine whether a SNP should be included in a model. However, a large number of SNPs and high correlation among the SNPs can make traditional regression models unstable. A more robust approach is provided by penalized regression models. These models simultaneously perform estimation of SNP effect sizes and SNP selection into a model by shrinking small effect estimates towards zero. Popular penalized models are lasso⁴⁰, elastic net⁴¹, minimax concave penalty⁴² and a normal-exponential-y shrinkage prior43 implemented in the HyperLasso software. Simulation studies show that penalized models tend to perform better than forward selection. Forward selection can be too conservative when using a very stringent *P* value threshold to select SNPs, yet a liberal threshold increases the chance of falsely selecting SNPs⁴⁴. Penalized models use tuning parameters to select SNPs into a model, with tuning parameters chosen to encourage SNPs with small effect sizes to be removed from the model. Tuning parameters are often estimated by cross-validation to choose a model with minimum prediction error. Penalized models tend to result in sparse models, selecting only one or a few SNPs belonging to a group of correlated SNPs. This can result in a good prediction model that includes non-causal SNPs and excludes a causal SNP when they are highly correlated. In Supplementary Box 1, we provide R code for readers to further see how high correlation combined with sparse models reduces the chance of selecting the causal variant. By contrast, Bayesian methods have been specifically designed for fine-mapping, offering advantages over heuristic and penalized regression approaches.

Bayesian methods overview. The challenge of both penalized regression and Bayesian variable selection methods is determining which SNPs have non-zero effect sizes (regression β -values) on a trait. Although we refer to SNPs with non-zero effect sizes as causal, it is important to realize that statistical methods alone cannot determine causality. Penalized models choose SNPs based on cross-validation that minimizes the error of predicting a trait. By contrast, Bayesian inference focuses on the probability of a specific hypothesis or specific model, thus providing probabilistic interpretation of models of interest.

A model for fine-mapping can be represented by an indicator variable for each SNP, with values of 1 for causal and 0 for not, and by organizing these indicators for all SNPs of interest in vector c. For m SNPs, there are 2^m possible *c* vectors (hence, 2^m possible models), ranging from all values of c equal to 0 for no SNPs causal to all values equal to 1 for all SNPs causal. Using Bayes' formula, the prior probability of a model can be combined with the likelihood of the data D (trait and SNPs) to compute the posterior probability of a specified model M_c , which we denote $P(M_c \mid D)$. See BOX 1 for how the posterior probability is calculated. There are several ways to specify the prior probability for a model, such as assuming that variants are independent and equally likely to be causal45,46 or assuming a fixed number of causal variants out of the total variants^{46,47} (see Supplementary Box 1 for more details). The posterior probabilities for different models can be used to determine the posterior probability of including each SNP in any of the models (posterior inclusion probability (PIP)), as well as determining the minimum set of SNPs required to capture the likely causal SNPs (credible sets). A variety

Box 1 | Bayesian methods

The goal of Bayesian methods is to compute the probability of a specific model, conditional on the observed data, denoted D, which includes the trait and single-nucleotide polymorphisms (SNPs). For a model M_c specified by a vector c of indicator variables, with each indicator variable having a value of 1 if a SNP is causal and 0 if a SNP is non-causal, the posterior probability of model M_c is

$$P(M_c|D) = \frac{P(D|M_c)P(M_c)}{\sum_{M \in M} P(D|M)P(M)}$$

where the sum in the denominator is over all models from a specified model space. For example, if the model space were to allow exactly one causal SNP out of *k* SNPs, there would be *k* possible *c* vectors and *k* possible models in the space.

The term $P(D|M_c)$ is the probability of the data under an assumed model. If only a fixed set of β -values were considered, $P(D|M_c)$ would be the likelihood of the data, denoted $P(D|\beta)$. This is the same likelihood used in frequentist analyses. However, Bayesian inference treats the β -values as random in order to account for their uncertainty. This is accomplished by assuming a prior distribution for the β -values, which we denote $P(\beta|M_c)$, and averaging over their possible values. This leads to the marginal likelihood, computed by integrating over the distribution of β -values:

 $P(D|M_c) = \int P(D|\beta)P(\beta|M_c)\partial\beta$

The ingredients of Bayesian methods are the likelihood ($P(D|\beta)$) and the assumed prior distribution of the models and their parameters. Details regarding choices of likelihood models and prior probabilities for Bayesian fine-mapping, as well as approximate Bayes factors, are discussed in Supplementary Box 1.

of Bayesian fine-mapping methods have been developed and are summarized in TABLE 1.

Bayesian methods: posterior inclusion probability. Bayesian methods for fine-mapping have been specialized in order to focus on the SNPs that have the largest chance of being causal^{45,48}. The PIP for a SNP is the probability of including a SNP as causal in any of the models. For SNP *j*, the PIP is computed by the sum of the posteriors over all models that include SNP *j* as causal,

$$PIP_j = P(c_j = 1 | D) = \sum_{M, c_j = 1} P(M | D)$$

After the posterior probabilities for the different models are computed, the PIP is rapid to compute and is an output from many Bayesian fine-mapping software packages (TABLE 1). Ranking SNPs by their PIP is a convenient way to select putative causal SNPs⁴⁶. For example, the top *k* SNPs ranked by their PIP maximize the expected number of causal SNPs across all possible subsets of size k^{47} . However, caution is warranted when multiple SNPs in a region are highly correlated and all are approximately equally correlated with the phenotype. In this situation, none of the individual PIPs would be very large. It would be better to estimate the posterior expected number of causal SNPs by summing the estimated PIPs for all SNPs in the region⁴⁵.

Bayesian methods: credible sets. Bayesian methods can be used to determine the credible set⁴⁹, the minimum set of SNPs that contains all causal SNPs with probability α . When assuming only one causal SNP, α is the sum of PIPs for SNPs in a set. This means that an α credible set is equivalent to ranking SNPs from largest to smallest PIPs⁵⁰ and taking the cumulative sum of PIPs until it is at least α .

Bayesian methods can also allow for multiple causal SNPs in a region. Although computation time increases because of the larger number of models required for multiple causal SNPs, recent developments^{46,51-54}, notably the JAM software⁵⁵, make it feasible to stochastically search over a wide array of possible models. An excellent demonstration of the benefits of Bayesian fine-mapping is provided by a fine-mapping analysis of 84 prostate cancer GWAS loci using 143,804 subjects⁵⁶. Single signals of causal variants were found in 63 regions, and 12 regions showed evidence of multiple independent causal variants. Furthermore, only 15 (5.4%) of the 280 original GWAS tag SNPs remained as potential causal variants.

There are numerous advantages to Bayesian methods for fine-mapping. First, unlike *P* values, posterior probabilities for SNPs can be directly compared. Second, they tend to select fewer SNPs as potentially causative compared with selecting SNPs on the basis of their correlation with the lead SNP¹⁸. Third, simulation studies have shown Bayesian methods to perform better than both conditional stepwise regression^{46,55,57} and penalized regression models⁴⁷. Finally, because Bayesian models are based on the joint effects of SNPs, they control for SNPs with large effects, improving power to detect SNPs with lesser effects. See FIG. 2 for visual aspects of Bayesian fine-mapping.

Combining studies and meta-analyses

Combining data from multiple cohorts can increase fine-mapping resolution, and the fine-mapping strategies discussed above can be used when individual-level data are pooled. However, because it can be challenging to obtain individual-level data from multiple cohorts, summary statistics for the associations of a trait with SNPs can be used¹⁴. This strategy is gaining popularity because it simplifies data sharing and computational issues. When summary statistics are appropriately chosen, there is no loss of information compared with using individual-level data⁵⁸. Most summary statistics focus on marginal regression β -values and their variances. To obtain the joint effects of multiple SNPs using individual-level data, one must regress trait y on all SNPs simultaneously, $y = X\beta^{j} + e$, where β^{j} is a vector of the joint effects of SNPs, X is a matrix of the coded SNP effects (for example, counts of minor alleles) and e is random error. By contrast, in summary data when only the marginal β -values are available, it is still possible to determine the joint effects of SNPs needed for finemapping. When effect sizes are small, there is a simple relationship between the joint and marginal effects⁵⁹, determined by $\beta^{J} = R^{-1}\beta^{M}$, where β^{M} is a vector of the marginal effects and R^{-1} is the inverse of the matrix of pairwise SNP correlations (assuming the columns of X are standardized to have a mean of 0 and a variance of 1). The key aspect is the SNP correlation matrix, which summarizes the LD among the SNPs. When the original data are used to estimate the SNP correlations, there is no loss of information relative to analysing individuallevel data⁵⁸. In practice, an appropriate reference sample

Prior probability

In Bayesian probability theory,

prior knowledge of their values

In Bayesian probability theory

distribution of parameters of

(PIP). The marginal probability

included in any causal model, conditional on the observed

data, thereby providing weight

of evidence that a SNP should

be included as potentially

interest, conditional on the

the probability distribution

assigned to parameters of interest, specified to represent

before observing the data.

Posterior probability

the updated probability

observed data

probability

causative

Posterior inclusion

that a single-nucleotide

polymorphism (SNP) is

able 1 Commonly used Dayesian fine mapping software								
Software	Trait type ^a	Input covariates ^b	Uses summary statistics?	Maximum number of causal variants ^c	Input annotation?	Causal search	Main output	Refs
BIMBAM v1.0	qt and binary	No	No	Fixed	No	Exhaustive	Bayes factor	113,114
mvBIMBAM v1.0.0	mqt	No	Yes	1	No	Exhaustive	Bayes factor	115,116
SNPTEST v2.5.4-beta3	qt, binary, mqt and multinomial	No	No	1	No	Exhaustive	Bayes factor	117
piMASS v0.9	qt and binary	No	No	Computed	No	MCMC	Bayes factor and PIP	45
BVS v4.12.1	Binary	Yes	No	Computed	Yes	MCMC	Bayes factor and PIP	97,118,119
FM-QTL	qt	No	No	Computed	Yes	MCMC	Bayes factor and PIP	96
DAP v1.0.0	qt	Yes	Yes	1, fixed and computed	Yes	Exhaustive	Bayes factor and PIP	52
Fine-mapping	Multinomial	Yes	No	Computed	No	Greedy	PIP	30
Trinculo	Multinomial	Yes	No	Computed	No	Greedy	Bayes factor and PIP	30,120
BayesFM	Binary	Yes	No	20	No	MCMC	PIP	30
ABF	qt and binary ^d	Yes	Yes	1	No	Exhaustive	Bayes factor	121
fgwas v0.3.6	qt and binary ^d	No	Yes	1	Yes	Exhaustive	Bayes factor and PIP	95
CAVIAR/eCAVIAR	qt and binary ^d	No	Yes	Fixed	No	Exhaustive	ρ probability confidence set and PIP	46,101
PAINTOR v3.0	qt, binary ^d and mqt	No	Yes	Fixed and computed	Yes	Exhaustive and MCMC	Bayes factor and PIP	54,57,71
CAVIARBF v0.2.1	qt and binary ^d	No	Yes	Fixed	Yes	Exhaustive	Bayes factor and PIP	47,94
FINEMAP v1.1	qt and binary ^d	No	Yes	Fixed	No	Shotgun stochastic search	Bayes factor and PIP	53
JAM in R2BGLiMS v0.1	qt and binary $^{\rm d}$	No	Yes	Fixed and computed	No	Exhaustive and MCMC	Bayes factor and PIP	55

Table 1 | Commonly used Bayesian fine-mapping software

MCMC, Markov chain Monte Carlo; PIP, posterior inclusion probability. ^aTrait types are binary, single binary trait; mqt, multiple quantitative traits; multinomial, trait with more than two categories; and qt, single quantitative trait. ^bFor software that does not allow covariates to be input, the traits can be adjusted for covariates by first regressing out the covariates (that is, subtracting trait predicted by covariates from trait values). ^cA fixed number is specified by the user to reduce computational cost. It is usually small (for example, three) when the number of candidate variants is large. When computed, the number of causal variants is determined by the software. As indicated, some software allow different options for whether the maximum number of causal variants is fixed by the user or computed by the software. ^dApplication to binary traits is based on linear regression, an approximation assuming small effect sizes and large sample sizes.

is often used to estimate the SNP correlations, such as the 1000 Genomes Project, allowing marginal summary statistics from single SNP analyses to be combined for a joint analysis⁵⁹. However, if the LD patterns in the reference sample do not represent those in the analysed samples, the estimated joint effects can be biased, leading to errors in fine-mapping. Furthermore, the size of the reference sample should not be too small and should increase in size as the GWAS sample size increases⁶⁰.

Trans-ethnic fine-mapping

Comparisons of GWAS findings across ethnically diverse populations have revealed that associations of SNPs with complex traits are often consistent across populations, with similar direction of effects of alleles on traits^{61,62}. Trans-ethnic meta-analyses that combine GWAS results of the same trait across genetically diverse populations can aid fine-mapping by capitalizing on ethnic differences in LD patterns^{63,64} (FIG. 2).

A critical issue is the choice of ethnic groups. Analyses based on a mix of different European ancestries or a mix of European and Asian ancestries provide little gain for fine-mapping^{30,65}. Rather, simulations have shown that a substantial reduction in the size of Bayesian fine-mapping credible sets can be attained by including subjects of African ancestry⁶⁵ because they have much narrower LD. Including Hispanic ancestry might also improve fine-mapping over using only European or Asian ancestries⁶⁵. The balance of ethnic groups can also impact fine-mapping performance. Simulations have shown that equal proportions of African and European ancestry are optimal for fine-mapping¹⁸. Although current samples with African ancestry tend to be much smaller than those with European ancestry, including African ancestry can still improve fine-mapping resolution⁶⁵.

Trans-ethnic analyses are often conducted as metaanalyses of summary statistics using random-effects methods, recognizing that a SNP can have different effect sizes across different ethnic groups. This heterogeneity can result from differences in study designs, differences due to interactions with other SNPs or differences due to environmental or lifestyle factors that influence the effects of genes. METASOFT⁶⁶, which implements a modified form of the traditional random-effects approach, is popular among trans-ethnic studies⁶⁷⁻⁷⁰. Alternative methods that model the heterogeneity can improve fine-mapping resolution. One approach is based on Bayesian methods, as implemented in the PAINTOR software⁷¹. Other approaches that use GWAS SNP data to explicitly adjust for heterogeneity include MANTRA and MR-MEGA. MANTRA is based on a computationally intensive Bayesian partition method that creates discrete clusters of ethnically similar subjects, assuming that subjects within the same cluster have the same allelic effects for a variant, in contrast to subjects from different clusters having different allelic effects72. A good example of trans-ethnic fine-mapping with MANTRA, followed by use of PAINTOR and CAVIAR software to estimate PIPs and credible sets, as well as use of functional annotation, is provided in a large fine-mapping study of high-density lipoprotein cholesterol⁷³. In contrast to the discrete clusters created by MANTRA, MR-MEGA considers allelic heterogeneity on a continuum, uses principal components of genetic similarity among subjects to determine axes of genetic variation and then uses the principal components to adjust for heterogeneity of allelic effect sizes⁷⁴. Simulations show that MR-MEGA can offer improved fine-mapping over other methods while retaining computational efficiency74.

Factors affecting Bayesian PIPs for fine-mapping

A number of factors influence the resolution and power of fine-mapping. To quantify this influence, we derived a simple formula for the expected Bayes PIP of a single causal SNP when studying a quantitative trait (Supplementary Box 1). This expected posterior probability depends on the effect size of the causal SNP on a trait (measured by multiple regression R^2 , the per cent of trait variation explained by the causal SNP) and the sample size (N). Importantly, these two factors combine into the non-centrality parameter that determines power, $\lambda = NR^2 / (1 - R^2)$. Other factors that influence the expected posterior probability are the number of SNPs in the fine-mapping effort (we assume one causal SNP and *m* non-causal SNPs) and SNP correlation structure. To simplify the correlation structure, we assume that all SNPs are equally correlated with correlation ρ , as might occur when examining a small genomic region or filtering on SNPs to achieve a correlation threshold. Finally, the assumed prior probabilities that SNPs are causal also influence the posterior probability. Based on these assumptions, the expected posterior probability for a causal SNP can be expressed as,

$$post_{c} = \frac{pr_{c}}{pr_{c} + \sum_{i=1i\neq c}^{m} pr_{i}exp\{-(1-\rho)NR^{2}/(1-R^{2})\}}$$

Where pr_i is the prior probability that the *i*th SNP is causal, and subscript *c* is for the causal SNP.

Assuming $R^2 = 1\%$, a relatively small effect size, we illustrate in FIG. 4 that large sample size and small SNP correlations provide the ideal setting to achieve a high posterior probability for a causal SNP. Increasing the number of non-causal SNPs tends to decrease the posterior probability by spreading out the posterior probability among multiple non-causal SNPs, particularly when

the non-centrality parameter λ is not large. FIGURE 4 also emphasizes that it is difficult to achieve a large posterior probability for a causal SNP when SNPs are highly correlated. Hence, it might be premature to pre-filter SNPs to have a large correlation with a lead SNP before performing fine-mapping efforts because the SNPs without large correlations could still be viable candidates. Furthermore, when power is weak or SNP correlation is large, well-informed prior probabilities are most critical because in these situations they have a larger effect on the posterior probabilities. Of course, mis-specified prior probabilities can result in a non-causal SNP having a large posterior probability, misleading fine-mapping efforts.

Genomic annotation

Genomic annotation that assigns biological function to DNA sequences can be informative about the likely function of SNPs selected by fine-mapping analyses and can aid prioritization of follow-up functional studies. Large-scale initiatives have increased publicly available resources, including Gene Ontology⁷⁵, GENCODE⁷⁶, ENCODE⁷⁷, FANTOM5⁷⁸ and the Roadmap Epigenomics Project⁷⁹. By integrating multiple datatypes for a variety of tissues and cell types, current annotations provide functional context for approximately 80% of the human genome⁸⁰. Analyses of published GWAS findings have identified significant enrichment for functional annotations among complex trait associations⁸¹⁻⁸³, motivating the use of annotation to improve fine-mapping resolution. Annotations are generally categorized on the basis of whether they occur in protein-coding versus non-protein-coding sequences.

Protein-coding annotation. Annotation for SNPs in genes that code for proteins focuses on their impact on the resulting protein structure. Examples of annotation include whether a SNP occurs in an exon, an intron or a splice site or whether it is involved in alternative splicing. A large number of bioinformatics annotation methods are available to functionally characterize coding SNPs and provide impact scores that predict their deleterious effect^{84,85}. Although prediction accuracy can be low from individual methods, combining methods can lead to improved predictions, such as by the CADD⁸⁶ and REVEL⁸⁷ methods.

Non-protein-coding annotation. The Encyclopedia of DNA Elements (ENCODE) project has shown that the genome is pervasively transcribed and that the majority of bases are found in primary transcripts, including non-protein-coding transcripts⁸⁸. Genetic variation in non-coding regions is often involved in gene regulation. Some examples of non-coding annotation are promoters, enhancers, long non-coding RNA loci, transcription start sites, transcription factor binding sites, regulatory sequences, features of chromatin accessibility and histone modification patterns and DNaseI hypersensitive sites. Variant impacts on putative transcription factor binding site (TFBS) motifs can be estimated by position weight matrices from databases, such as TRANSFAC⁸⁹ and JASPAR⁹⁰. Annotation tools, such as



Fig. 4 | Posterior probability for a single causal SNP when 5–40 SNPs are in a region of interest. The prior probability that a single-nucleotide polymorphism (SNP) is causal is assumed to be equal for all SNPs. Sample size (*N*) ranges from 500–20,000, and the per cent of trait variation explained by the causal variant, R^2 , is 1%. SNPs are assumed to be equally correlated with magnitude ρ . The horizontal dotted line is for equal prior probabilities for SNPs, and the posterior probability approaches this line when the data have little information to distinguish causal from non-causal SNPs.

FIRE⁹¹, RegulomeDB⁹² and CADD⁸⁶, provide regulatory impact scores for non-coding SNPs by aggregating multiple evidence sources of functional importance.

Integrating annotation into fine-mapping. Ad hoc review of SNP annotations is often applied to SNPs selected by fine-mapping analyses in order to identify patterns of annotation enrichment and prioritize candidates for functional validation. This subjective approach can be cumbersome and biased. Alternative methods can improve fine-mapping, such as using functional annotation to weight SNPs93 in regression models or extending Bayesian models to allow the prior probability that a SNP is causal to depend on annotation^{57,94-97} (Supplementary Box 1). Advantages of Bayesian methods are that the weights given to functional annotation are inferred from the data, and that jointly mapping multiple regions simultaneously provides a way to share annotation information across different regions^{57,95}. Most methods allow for a relatively small number of annotation features, which requires screening each annotation one at a time in order to select a subset for a final model.

An alternative approach that incorporates automatic selection of annotations from a large number of features overcomes this limitation⁹⁴. Bayesian mapping software that allows use of annotation is highlighted in TABLE 1.

The broad impact of the use of annotation on fine-mapping is yet unknown. By simulations, Van de Brunt et al.¹⁸ found that incorporating annotation into Bayesian prior probabilities gave modest benefit for fine-mapping by increasing the frequency of small credible sets (for example, less than ten SNPs) from 27% without annotation to 36% with annotation. In real applications of fine-mapping four blood lipid traits, the 90% credible set size reduced from an average of 17.5 SNPs without use of annotation to 13.5 SNPs with use of annotation⁵⁷. When using a Bayesian method to fine-map three diseases in the Wellcome Trust Case Control Consortium, very few SNPs in the credible sets were found to have annotated functions⁵⁰. Therefore, use of annotation information can be limiting from two angles: first, incorporation of annotation into prior probabilities has limited impact on wellpowered studies, and second, current understanding of broad genomic function may be too limiting to accurately

Expression quantitative trait loci

(eQTLs). Genomic regions that harbour one or more nucleotide variants that influence the amount of expression of a gene. improve prior probabilities of causation. Conversely, annotation might help when association signals are at best moderate, in regions of high LD, when there are multiple causal SNPs in a region, or when different regions share enrichment for specific annotation features.

Integrating GWAS with gene expression. More than 90% of trait-associated alleles discovered by GWAS map to non-coding regions, with strong evidence of enrichment for regulatory elements, such as enhancers, promoters, insulators and silencers⁸¹. Furthermore, SNPs associated with complex traits are significantly more likely to be expression quantitative trait loci (eQTLs) than other SNPs on genotype arrays with the same allele frequencies⁹⁸. This suggests that SNPs discovered by GWAS influence the amount of expression of nearby genes, and this altered expression ultimately influences the trait.

Statistical methods have been developed to integrate eQTL data with GWAS data to quantify the evidence

Box 2 | Guidelines for evaluating fine-mapping

The goal of fine-mapping is to determine which variants in a genomic region are most likely to be causally related to a trait after accounting for how the variants in the region are correlated. One challenge with Bayesian fine-mapping can be that the output credible sets contain a large number of variants. A way to reduce the number of putative causal variants is the improved use of annotation, perhaps by careful selection of relevant cell types for gene regulation annotation. Another way is to increase sample size. With large samples, Bayesian fine-mapping can sometimes result in a credible set of only a few variants or perhaps a single variant. To gain confidence in fine-mapping results, it is worthwhile to evaluate the robustness of the results. Some guidelines for evaluating robustness are given below.

- What is the spread of the posterior inclusion probability (PIP) values for the variants in a credible set? If the PIPs are similar for multiple variants in a credible set, it can be worthwhile to view the correlation among the variants and the effect sizes of the variants with the effects of variants adjusted for each other. Large correlations among variants with similar effect sizes can suggest that it is not possible to distinguish among the variants with the available sample.
- When using summary statistics and a reference panel to estimate the SNP correlations, how well does the reference panel match the study sample? When possible, it is worthwhile to compare the correlations among the variants in the study sample with the correlations in the reference panel. Large differences between the study sample and reference panel warrant further fine-mapping analyses with the correlations obtained by the study sample. When correlations among SNPs from the study sample are not available, comparisons of allele frequencies between the reference panel and study sample can sometimes reveal problems, such as large differences in allele frequencies.
- Because highly correlated SNPs can make fine-mapping difficult, sometimes compromising numerical computations, it is worthwhile to examine the correlations among the SNPs chosen on the basis of fine-mapping. High correlations (for example, more than 0.98) might warrant more careful scrutiny.
- Different software for fine-mapping make different assumptions, and the computational methods can be implemented differently. To evaluate the robustness of a finding, it can sometimes be worthwhile to compare results across different software packages. Consistent results can improve confidence in the fine-mapping findings. However, inconsistent results can be caused by multiple factors, including different modelling assumptions, rounding errors or differences in stochastic search methods. Hence, although comparing results across different software packages can be useful, expertise in the software is required in order to adequately interpret the results.
- Trans-ethnic fine-mapping assumes that the set of causal SNPs are the same for all
 populations but accounts for differences in effect sizes across populations. For this
 reason, it can be worthwhile to evaluate the estimated effect sizes of the SNPs
 selected by fine-mapping to determine if just one, or a few, of several populations
 have effect sizes that dominate the fine-mapping.

of a causal pathway from SNP to gene expression to a complex trait. The intermediate variable, mRNA, is the mediator between a SNP and a trait. One approach to test a causal pathway is by a causal inference test⁹⁹, with small *P* values inferring causality. Alternative Bayesian approaches focus on the posterior probability that a single SNP is causally related to both mRNA level and the trait^{100,101}. Mendelian randomization is yet another approach that can be used to distinguish whether a single SNP influences both gene expression and the trait versus whether separate SNPs in LD influence gene expression and the trait¹⁰². Details of these various approaches are provided in Supplementary Box 1.

A critical issue for integrating eQTL and GWAS results is the type of tissue for which the expression was measured. Complex diseases often result from dysfunction of multiple tissues or cell types, and the expression of genes varies widely across different types of tissues. Selecting relevant tissue types for a given disease process or complex trait can be a substantial challenge. Numerous public resources facilitate the exploration of gene expression in different tissues, such as the Genotype-Tissue Expression (GTEx) project, which includes genotypes, gene expression, and histological and clinical data for 449 human donors across 42 distinct tissues¹⁰³.

Conclusions

Fine-mapping efforts have made considerable advancements to refine the most likely genetic variants discovered by large-scale genetic association studies of complex traits. We reviewed a variety of analytical methods, with the more sophisticated and relevant methods based on Bayesian fine-mapping. A common underlying basis of all methods is LD between measured SNPs and causal variants, which makes fine-mapping feasible, as well as challenging. Although the details of a particular study will determine the success of fine-mapping, general guidelines for evaluating the robustness of fine-mapping efforts are provided in BOX 2.

Most of this Review discusses SNPs. Other types of genetic variants can be analysed with similar strategies, such as small insertions or deletions (indels) analysed as binary variables. Currently, there are few general methods to incorporate large structural variants into GWAS fine-mapping. However, germline structural rearrangements have helped to discover simple Mendelian disease genes, such as partial deletion of chromosome 15 leading to discovery of the gene UBE3A that causes Angelman syndrome¹⁰⁴ or a 1.5 Mb inversion on chromosome 17q12 leading to renal cysts and diabetes syndrome (RCAD), an autosomal dominant disorder¹⁰⁵. Genetic variants with more than two alleles would require extension of current methods. For example, the human leukocyte antigen (HLA) locus on chromosome 6 is highly polymorphic and is involved with immune response and multiple diseases. This region has unique fine-mapping challenges, some of which can capitalize on GWAS data (Supplementary Box 1).

As fine-mapping efforts proceed, an important consideration is the expected level of resolution. Using simulations that closely align with genomic structure of

the UK 10,000 Genomes (UK10K) sample of 3,642 unrelated subjects of European ancestry, Wu et al.¹⁰⁶ found that at least 80% of common variants identified in published GWAS that used imputed data were within 33.5 kb of causal variants. This provides optimism for finemapping GWAS results when using the study designs and computational methods discussed in this Review. Sharing of data in large meta-analyses and improved genomic annotation will probably improve resolution.

The frequency of variants has a large effect on the success of fine-mapping by statistical methods. Wu et al.¹⁰⁶ found that rare untyped causal variants (minor allele frequency less than 0.01) are unlikely to map to a common variant in GWAS, whether using imputed data or whole-genome sequencing. Less-common variants tend to have lower levels of LD with other variants, making it challenging to impute less-common variants by statistical methods. However, low levels of LD can be advantageous for fine-mapping when the variants are accurately genotyped or imputed and a study has high power to detect associations with single variants. Wholeexome sequencing and whole-genome sequencing open new avenues to discover rare variants. A challenge of these types of sequencing studies is the very large sample size (for example, approximately 100,000 subjects) needed to discover regions that harbour rare causal variants for complex traits¹⁰⁷. Furthermore, the statistical methods for rare-variant associations with a trait do not focus on single rare variants but rather evaluate an entire region¹⁰⁸. Therefore, the initial discovery of a region provides little evidence of specific variants that are likely to be causal. Nonetheless, whole-genome sequencing can be informative for fine-mapping rare variants, but in this situation, various sources of annotation of variants will be key to success¹⁰⁹. Alternative approaches that combine functional assay results with computational prediction of function¹¹⁰ might prove useful for fine-mapping, albeit with the requirement of functional assays developed for each specific region of interest.

Our understanding of the genetic basis of human disease has evolved from single major genes influencing rare Mendelian disorders to multiple genes — polygenic influencing common complex traits. Future finemapping challenges will face complex traits driven by a large number of variants with small effects, possibly based on a genetic architecture of regulatory networks composed of a small number of core genes that directly affect a trait but with a large number of genes outside the core that indirectly affect the trait — an omnigenic model¹¹¹. Understanding the genetic architecture of regulatory networks offers new opportunities to integrate this information into future fine-mapping strategies.

Published online 29 May 2018

- Hardy, J. & Singleton, A. Genomewide association studies and human disease. N. Engl. J. Med. 360, 1759–1768 (2009).
- Wellcome Trust Case Control Consortium. Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678 (2007).
- Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569 (2010).
- Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283 (2013).
- Nikpay, M. et al. A comprehensive 1,000 Genomesbased genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47, 1121–1130 (2015).
- Al Olama, A. A. et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* 46, 1103–1109 (2014).
- 7. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
- Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427 (2014).
- MacArthur, J. et al. The new NHCRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901 (2017).
- 10. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
- Ding, K. & Kullo, I. J. Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance. *Eur. J. Hum. Genet.* 15, 228–236 (2007).
- Stram, D. Tag SNP selection for association studies Genet. Epidemiol. 27, 365–374 (2004).
 Spain, S. L. & Barrett, J. C. Strategies for fine-
- Spain, S. L. & Barrett, J. C. Strategies for finemapping complex traits. *Hum. Mol. Genet.* 24, R111–R119 (2015).
- Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18, 117–127 (2017).
 This paper reviews the developments and progress of using summary statistics from genetic association studies to perform joint analyses of genetic variants for use in fine-mapping and to perform transcription-wide association studies (TWAS).

- Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337 (2010).
- Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176 (2010).
- Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* 32, 381–385 (2008).
- Zaykin, D. V. & Zhivotovsky, L. A. Ranks of genuine associations in whole-genome scans. *Cenetics* 171, 813–823 (2005).
- Hedrick, P. W. Gametic disequilibrium measures: proceed with caution. *Genetics* 117, 331–341 (1987).
- Devlin, B. & Risch, N. A comparison of linkage diequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322 (1995).
- Martin, E. R. et al. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* 67, 383–394 (2000).
- Guerreiro, R. J. & Hardy, J. TOMM40 association with Alzheimer disease: tales of APOE and linkage disequilibrium. *Arch. Neurol.* 69, 1243–1244 (2012).
- Slatkin, M. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485 (2008).
- Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511 (2010).
- Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genom. Hum. Genet.* 10, 387–406 (2009).
- The Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
- McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48, 1279–1283 (2016).

- Southam, L. et al. The effect of genome-wide association scan quality control on imputation outcome for common variants. *Eur. J. Hum. Genet.* 19, 610–614 (2011).
- Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* 547, 173–178 (2017).
 This paper applies three complementary Bayesian fine-mapping methods to a large data set and nicely illustrates novel methods and their interpretations, along with strategies for using annotation to interpret fine-mapping results. The supplemental material is particularly informative for computational strategies for Bayesian
- fine-mapping.
 Amos, C. I. et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomarkers Prevent.* 26, 126–135 (2017).
- Voight, B. F. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8, e1002793 (2012).
- Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Cenet.* 14, 661–673 (2013).
- Hocking, R. A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics* 32, 1–49 (1976).
- Freedman, D. A note on screening regression equations. *Am. Statistician* **37**, 152–155 (1983).
 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J.
- Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265 (2005).
- Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232 (2001).
- Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 4, 587–597 (2003).
- Tibshirani, R. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B 58, 267–288 (1996).
- Cho, S., Kim, H., Oh, S., Kim, K. & Park, T. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc.* 3 (Suppl. 7), S25 (2009).

- Breheny, P. & Huang, J. Penalized methods for bi-level variable selection. *Statist. Interface* 2, 369–380 (2009).
- Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* 4, e1000130 (2008).
 Ayers, K. L. & Cordell, H. J. SNP selection in genome-
- Ayers, K. L. & Cordell, H. J. SNP selection in genomewide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* 34, 879–891 (2010).
- Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Ann. Appl. Statist.* 5, 1780–1815 (2011).
 This paper provides a Bayesian computational

This paper provides a Bayesian computational framework to consider a large number of causal variants.

- Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
- 47. Chen, W. et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* 200, 719–736 (2015). This paper links Bayesian fine-mapping using summary statistics and full data and describes an efficient computational approach using only relevant variables for each candidate model
- relevant variables for each candidate model.
 48. Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C. & Schildkraut, J. M. Bayesian model search and multilevel inference for SNP association studies. *Ann. Appl. Statist.* 4, 1342–1364 (2010).
- Carlin, B. & Louis, T. Bayesian Methods for Data Analysis 3rd edn, (Chapman and Hall/CRC, Boca Raton, FI, USA, 2008).
- Maller, J. B. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* 44, 1294–1301 (2012).
- Wallace, C. et al. Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *PLoS Genet.* 11, e1005272 (2015).
- Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* 98, 1114–1129 (2016).
- Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501 (2016).
 Kichaev, G. et al. Improved methods for multi-trait fine
- Kichaev, G. et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* 33, 248–255 (2017).
- Newcombe, P. J., Conti, D. V. & Richardson, S. JAM: a scalable bayesian framework for joint analysis of marginal SNP effects. *Cenet. Epidemiol.* 40, 188–201 (2016).

This paper builds on prior developments of Bayes methods for fine-mapping and develops a computationally efficient method to explore a wide range of models that can include multiple causal variants in regions of interest.

- 56. Dadaev, T. et al. Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nat. Commun.* https:// doi.org/10.1038/s41467-018-04109-8 (2018). This paper illustrates practical approaches to fine-mapping many genomic regions using Bayesian methods and illustrates the use of quantile regression to evaluate how genomic annotation is associated with SNPs that have a large Bayes posterior probability of being causally related to prostate cancer.
- 57. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* 10, e1004722 (2014). This is the first of a series of papers regarding PAINTOR software for fine-mapping, allowing multiple causal variants and summary statistics and integrating functional annotations.
- Lin, D. Y. & Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.* 34, 60–66 (2010).
- Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* 44, 369–375 (2012).
- Benner, C. et al. Prospects of fine-mapping traitassociated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
- 61. Ntzani, E. E., Liberopoulos, G., Manolio, T. A. & Ioannidis, J. P. Consistency of genome-wide

associations across major ancestral groups. *Hum. Genet.* **131**, 1057–1071 (2012). Marigorta, U. M. & Navarro, A. High trans-ethnic

- 62. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* 9, e1003566 (2013). This paper illustrates that common genetic associations of complex traits are highly conserved across diverse ethnic populations and motivates the application of trans-ethnic analysis.
- Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* 6, 91 (2014).
- Zaitlen, N., Pasaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* 86, 23–33 (2010).
- Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* 24, 1330–1336 (2016).

This paper demonstrates that reductions in fine-mapping credible sets are heavily dependent on ancestral composition of contributing studies and emphasizes the importance of trans-ethnic study design.

- Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* 88, 586–598 (2011).
- Wang, X. et al. Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 22, 2303–2311 (2013).
- van Rooij, F. J. et al. Genome-wide trans-ethnic meta-analysis identifies seven genetic loci influencing erythrocyte traits and a role for RBPMS in erythropoiesis. *Am. J. Hum. Genet.* **100**, 51–63 (2017).
- Franceschini, N. et al. Variant discovery and fine mapping of genetic loci associated with blood pressure traits in Hispanics and African Americans. *PLoS ONE* 11, e0164132 (2016).
- Larson, N. B. et al. Trans-ethnic meta-analysis identifies common and rare variants associated with hepatocyte growth factor levels in the Multi-Ethnic Study of Atherosclerosis (MESA). Ann. Hum. Genet. 79, 264–274 (2015).
- Kichaev, G. & Pasaniuc, B. Leveraging functionalannotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
- Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* 35, 809–822 (2011).
 This paper introduces a Bayesian partition model

framework for trans-ethnic fine-mapping by clustering study populations based on genetic similarity in order to account for heterogeneity of allelic effects on a trait.

- Cannon, M. E. et al. Trans-ancestry fine mapping and molecular assays identify regulatory variants at the ANGPTL8 HDL-C GWAS locus. G3 7, 3217–3227 (2017).
- Magi, R. et al. Trans-ethnic meta-regression of genomewide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* 26, 3639–3650 (2017).
- Yon Rhee, S., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509 (2008).
- Harrow, J. et al. CENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774 (2012).
- ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004).
- Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461 (2014).
- Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
- Pennisi, E. ENCODE project writes eulogy for junk DNA. Science 337, 1159–1161 (2012).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195 (2012).
 This paper leverages cell-line regulatory annotation to identify disease-relevant cell types and reveals that common genetic trait associations are enriched in functional DNA.
- Ma, M. et al. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics* 16 (Suppl. 8), S3 (2015).

- Trynka, C. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130 (2013).
- Mudge, J. M. & Harrow, J. The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* 17, 758–772 (2016).
- Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* 18, 599–612 (2017).
- Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 510–315 (2014).
- Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99, (877–885 (2016).
- Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816 (2007).
- Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* 24, 238–241 (1996).
- Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147 (2014).
 Ioannidis, N. et al. FIRE: functional inference of genetic
- Ioannidis, N. et al. FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics* 33, 3895–3901 (2017).
- Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797 (2012).
- Sveinbjornsson, G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* 48, 314–317 (2016).
- Chen, W., McDonnell, S., Thibodeau, S., Tillmans, L. & Schaid, D. Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics* 204, 933–958 (2016).
- Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
- Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* 11, e1005176 (2015).
- Quintana, M. A. et al. Incorporating prior biologic information for high-dimensional rare variant association studies. *Hum. Hered.* **74**, 184–195 (2012).
- Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888 (2010).
- Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling molecular relationships with a causal inference test. *BMC Genet.* **10**, 23 (2009).
- 100. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10, e1004383 (2014).
- Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99, 1245–1260 (2016).
- 102. Zhu, Z. H. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487 (2016).
- 103. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017).
- 104 204 213 (2017), 104 Magenis, R. E., Brown, M. G., Lacy, D. A., Budden, S. & LaFranchi, S. Is Angelman syndrome an alternate result of del(15)(q11q13)? *Am. J. Med. Genet.* 28, 829–838 (1987).
- Antonacci, F. et al. Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* 18, 2555–2566 (2009).
- 106. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18**, 86 (2017).
- Auer, P. L. et al. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI Exome Sequencing Project. *Am. J. Hum. Genet.* **99**, 791–801 (2016).
- Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

- Morrison, A. C. et al. Practical approaches for wholegenome sequence analysis of heart- and blood-related traits. *Am. J. Hum. Genet.* **100**, 205–215 (2017).
- Guidugli, L. et al. Assessment of the clinical relevance of BRCA2 missense variants by functional and computational approaches. *Am. J. Hum. Genet.* **102**, 233–248 (2018).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186 (2017).
- Haralambieva, I. H. et al. Genome-wide associations of CD46 and IFI44L genetic variants with neutralizing antibody response to measles vaccine. *Hum. Genet.* 136, 421–435 (2017).
- 113. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 3, e114 (2007).
- Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet.* 4, e1000279 (2008).
- Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PloS ONE* 8, e65245 (2013).
- 116. Shim, H. et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PIoS ONE* 10, e0120758 (2015).
- 117. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genomewide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

- 118. Quintana, M. A. & Conti, D. V. Integrative variable selection via Bayesian model uncertainty. *Stat. Med.* 32, 4938–4953 (2013).
- 119. Quintana, M. A., Berstein, J. L., Thomas, D. C. & Conti, D. V. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genet. Epidemiol.* **35**, 638–649 (2011).
- Jostins, L. & McVean, G. Trinculo: Bayesian and frequentist multinomial logistic regression for genomewide association studies of multi-category phenotypes. *Bioinformatics* 32, 1898–1900 (2016).
- Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* 33, 79–86 (2008).

Acknowledgements

This research was supported by the US Public Health Service and National Institutes of Health (contract grant number GM065450).

Author contributions

All authors contributed to researching content for the article, discussing content and writing. D.J.S. was responsible for reviewing and editing the manuscript before submission.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information

Nature Reviews thanks D. Conti and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s41576-018-0016-z.

RELATED LINKS

BayesFM: https://sourceforge.net/projects/bayesfmmcmc-v1-0/

BIMBAM v1.0: http://www.haplotype.org/bimbam.html BVS v4.12.1: https://cran.r-project.org/web/packages/BVS CAVIAR/eCAVIAR: http://genetics.cs.ucla.edu/caviar/ CAVIARBF v0.2.1: https://bitbucket.org/Wenan/caviarbf DAP v1.0.0: https://github.com/xqwen/dap

fgwas v0.3.6: https://github.com/joepickrell/fgwas FINEMAP v1.1: http://www.christianbenner.com/ Fine-mapping: https://github.com/hailianghuang/Fine-

apping

FM-QTL: https://github.com/xqwen/fmeqtl JAM in R2BGLiMS v0.1: https://github.com/pjnewcombe/ R2BGLiMS

mvBIMBAM v1.0.0: http://stephenslab.uchicago.edu/ software.html#mvbimbam

PAINTOR v3.0: https://github.com/gkichaev/PAINTOR_V3.0 piMASS v0.9: http://www.haplotype.org/pimass.html SNPTEST v2.5.4-beta3: https://mathgen.stats.ox.ac.uk/ genetics_software/snptest/snptest.html

Trinculo: https://sourceforge.net/projects/trinculo/