

Controlled Discovery and Localization of Signals via Bayesian Linear Programming

Asher Spector¹ and Lucas Janson²

¹Department of Statistics, Stanford University

²Department of Statistics, Harvard University

Abstract

Scientists often must simultaneously discover signals and localize them as precisely as possible. For instance, in genetic fine-mapping, high correlations between nearby genetic variants make it hard to identify the exact locations of causal variants. So the statistical task is to output as many disjoint regions containing a signal as possible, each as small as possible, while controlling false positives. Similar problems arise in any application where signals cannot be perfectly localized, such as locating stars in astronomical surveys and change point detection in time series data. Our first contribution is to propose a notion of resolution-adjusted power for such problems. Second, we introduce Bayesian Linear Programming (BLiP), a Bayesian method for jointly detecting and localizing signals. BLiP overcomes an extremely high-dimensional and non-convex problem to verifiably nearly maximize expected power while provably controlling false positives. BLiP is very computationally efficient and can wrap around nearly any Bayesian model and algorithm. Applying BLiP to existing state-of-the-art analyses of UK Biobank data (for genetic fine-mapping) and the Sloan Digital Sky Survey (for astronomical point source detection) increased resolution-adjusted power by 30-120% in just a few minutes of computation. BLiP is implemented in the new packages `pyblip` (Python) and `blipr` (R).

1 Introduction

1.1 Motivation

Localizing signals is an important statistical task across disciplines. For example, consider the problem of controlled variable selection: given a set of variables X_1, \dots, X_p , analysts seek to identify a few key variables which impact an outcome Y . Here, X_1, \dots, X_p could represent genetic mutations or demographic data, and Y could represent a disease status or economic outcome. However, when variables are highly correlated, it can be challenging or impossible to say with confidence that any individual variable is important. For example, if X_1 and X_2 are nearly perfectly correlated variables, analysts may not be able to distinguish between them even if the data make clear that $\{X_1, X_2\}$ contains at least one important variable. Yet selecting different variables may lead to qualitatively different scientific conclusions, such as in genetic studies, where intervening on a causal variant

(e.g., X_1) might help cure a disease, whereas intervening on its highly correlated neighbor (e.g., X_2) could have no impact at all. As a result, analysts often must contend with significant uncertainty in which variables ought to be selected. Unfortunately, as observed by Wang et al. (2020), many modern variable selection methods cannot accomplish this task.

More generally, analysts in many settings may be able to tell that a signal exists without perfectly localizing it. Indeed, astronomers often can guarantee that a light source exists somewhere in a region of space without knowing its exact location. Similarly, economists may be able to tell that a time series has changed without knowing precisely when it did so. In these settings, analysts often want to localize signals as precisely as possible, i.e., to make statements like “there is a signal in this region, even if we do not know exactly where it is.” Indeed, even in variable selection problems, correlations often have localized or spatial structure, so it is scientifically interpretable to claim that at least one variable in a (highly correlated) region is relevant. (Correlations themselves can also implicitly define spatial structure.) In all cases, making the discovered regions as small as possible yields more precise scientific insights. However, practitioners may also want to control false positives to ensure their results are replicable, interpretable, and do not waste future resources.

The problem of localizing signals is particularly important in *genetic fine-mapping*, where researchers attempt to identify specific genetic variants which cause outcomes of interest, such as cardiovascular disease. However, genetic variants are highly locally correlated—for example, in this paper, we analyze a UK Biobank dataset of 337,000 individuals from Weissbrod et al. (2019), where over 40% of genetic variants are at least 99% correlated with at least one other genetic variant. In this context, it is particularly important that variable selection methods can select clusters of correlated variables, since there may be no way to guarantee that any single variant is individually important. At the same time, these clusters should be as small as possible to yield maximally precise information.

With this motivation, our paper introduces a novel procedure, called *Bayesian Linear Programming* (BLiP), which localizes signals as precisely as possible while controlling false positives. Before describing our contributions, however, we pause to formally define this problem.

1.2 Problem Statement

In this section, we define the problem of *resolution-adaptive signal detection*. To start, we assume that we have a space of locations \mathcal{L} at which there might be signals. For example, \mathcal{L} might represent locations on a genome, positions in the night sky, or specific points in time—indeed, we analyze each of these examples later. For each location $\ell \in \mathcal{L}$, let I_ℓ be the indicator of whether there is a signal at ℓ . For example, in genetic fine-mapping, we might say that $I_\ell = 1$ if the genetic variant at location ℓ on the genome has a causal relationship with a trait of interest, e.g., heart disease. Intuitively, one can think of I_ℓ as either an unobserved latent variable or the indicator that the null hypothesis that there is no signal at ℓ is false. To aid intuition, we pause to give a few concrete examples of this setup.

Example 1 (Variable selection in regression). Suppose we observe a response Y and variables $X = (X_1, \dots, X_p)$, and we seek to discover “important” variables. Here, the locations $\mathcal{L} = \{1, \dots, p\}$ correspond to X_1, \dots, X_p , and there is a signal at location $\ell \in \mathcal{L}$ if X_ℓ is “important.” For example, if we assume Y depends on X through linear coefficients $\beta \in \mathbb{R}^p$, we might set $I_\ell = \mathbb{I}(\beta_\ell \neq 0)$. More generally, we could set $I_\ell = 1$ if the conditional distribution $Y | X$ depends on X_ℓ , or equivalently, if $X_\ell \not\perp\!\!\!\perp Y | X_{-\ell}$. Note when (X_1, \dots, X_p) are highly correlated, we may not be able to discover

individual signal variables with confidence. However, we often do have power to discover that at least one variable in a group $G \subset \mathcal{L}$ is important.

Example 2 (Point source detection). Astronomers often seek to locate point sources (e.g., stars) in the night sky. Here, we define the set of locations \mathcal{L} to equal the region of the sky we are interested in, so $\mathcal{L} \subset \mathbb{R}^2$. For each coordinate $(x, y) \in \mathcal{L}$, $I_{(x,y)} = 1$ if and only if a source exists at location (x, y) . Our data consists of a set of photon counts $\{D_{ij}\}_{1 \leq i \leq d_1, 1 \leq j \leq d_2}$, where D_{ij} counts the photons observed by a telescope in the (i, j) th pixel of an image. Most images have blur, making it difficult to distinguish whether a source is located in one pixel versus its neighbors. This motivates a resolution-adaptive approach, where instead of estimating signal locations $\{(x_i, y_i)\}_{i=1}^R$, we output regions G_1, \dots, G_R which each contain a source with high confidence and are as small as possible.

Example 3 (Change point detection). Consider a time series Y_1, \dots, Y_T which is stationary except at a few change points, where the process changes. Often, we aim to identify change points. If time is measured on a discrete grid $\{1, \dots, T\}$, then the locations $\mathcal{L} = \{1, \dots, T\}$ represent the T times, and we say there is a signal at time t if there is a change point at time t . For example, one simple model might be $Y_t \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_t, \sigma^2)$, so $I_t = \mathbb{I}(\mu_t \neq \mu_{t-1})$. Resolution-adaptivity is important here because often, we cannot identify exactly *when* a time series has changed (perhaps because each observation Y_t is noisy), even if it is clear that $\{Y_t\}$ has change points. Thus, we opt to make statements like “there was at least one change point between time t_0 and t_1 .” Of course, whenever it is possible to perfectly localize a change point, a good resolution-adaptive method should automatically do so.

Our goal is to discover signals while controlling false positives, but in each of these examples, it is often difficult to perfectly localize signals. Thus, we allow ourselves to discover any group or region $G \subset \mathcal{L}$, which asserts that at least one signal exists in G . (Note that throughout the paper, we use the words “group” and “region” interchangeably.) This is a true detection if there actually is a signal in G , or formally, if $I_G := \mathbb{I}(\exists \ell \in G : I_\ell = 1) = 1$. Intuitively, detecting signals in larger regions corresponds to making discoveries at a coarser resolution, which is “easier” statistically than detecting signals in smaller regions but yields less specific information. For example, in genetic fine-mapping, discovering a large region G tells us that a causal genetic variant exists somewhere in G , whereas discovering $G = \{\ell\}$ precisely isolates the location of a single causal variant. Thus, we prefer to detect signals at as fine a resolution as possible, meaning that we discover regions G that are as small as possible.

With this motivation, the problem of *resolution-adaptive signal detection* is to select a disjoint set of regions G_1, \dots, G_R which maximizes expected power while controlling a false positive rate. Admittedly, the definition of statistical power is somewhat subtle, and one contribution of our paper will be to precisely define it later. For now, higher power intuitively corresponds to making more discoveries at a finer resolution.¹

Definition 1.1 (Resolution-adaptive signal detection). Suppose we seek to discover signals in a set of locations \mathcal{L} . Let the number of discoveries $R \geq 0$ and the discovered regions G_1, \dots, G_R be our optimization variables. Given a notion of statistical power for a set of discovered regions, denoted

¹We use the word “power” slightly loosely in this paper. Although “power” often refers to an expected quantity, we use it to refer to the *realized* (resolution-adjusted) number of true discoveries made using a single dataset.

by $\text{Power}(G_1, \dots, G_R)$, which we will say more about shortly, we seek to maximize

$$\max \quad \mathbb{E} [\text{Power}(G_1, \dots, G_R)] \quad (1.1)$$

$$\text{s.t.} \quad \text{FDR} := \mathbb{E} \left[\frac{\#\{1 \leq r \leq R : I_{G_r} = 0\}}{\max(1, R)} \right] \leq q, \quad (1.2)$$

$$G_1, \dots, G_R \subset \mathcal{L} \text{ are disjoint.} \quad (1.3)$$

Note that we constrain G_1, \dots, G_R to be disjoint to improve interpretability and prevent double-counting. For example, discovering $\{\ell_1\} \subset \mathcal{L}$ makes discovering $\{\ell_1, \ell_2\} \subset \mathcal{L}$ logically redundant, so this should not count as two separate discoveries when calculating power or FDR. We focus on the FDR because it is a popular and appealing error rate, but this problem is still well-defined if we replace the FDR with another error rate, such as the family-wise error rate (FWER) or the local FDR. Indeed, BLiP will be able to solve any of these problems.

A simplification of this problem would be to fix a prespecified partition $G_1, \dots, G_m \subset \mathcal{L}$ and test whether a signal exists in each of G_1, \dots, G_m . However, this non-adaptive approach will not optimally localize signals, because the best choice of partition G_1, \dots, G_m depends on the unknown data-generating process. Informally, when the “signal size” is large, we may be able to perfectly localize individual signals, whereas we may only be able to detect that a weak signal exists somewhere in a relatively large region. Indeed, in Appendix A, we give a concrete example of a regression problem where the best partition depends on the unknown relationship between Y and X , and thus a non-adaptive method will always make the wrong choice in some settings. In contrast, resolution-adaptive methods can use the data to discover regions G_1, \dots, G_R which are as small as possible. Of course, for computational reasons, it may not be possible to consider *every* region $G \subset \mathcal{L}$ as a potential discovery. However, we expect that methods which are more adaptive, meaning they can use the data to choose from among a larger set of candidate regions, to have higher power.

The problem as stated raises two important questions. First, it is not clear what statistical power means in Equation (1.1). Maximizing the number of discoveries is not the right goal, since we would prefer to make discoveries at finer resolutions to gain more specific information. For example, in variable selection problems, is the statistical power higher if we select the group $\{1\}$ and conclude that X_1 is important, or if we select groups $\{1, 2\}, \{3, 4\}$ and conclude that at least one of each pair $\{X_1, X_2\}, \{X_3, X_4\}$ is important? This is a fundamental question when performing resolution-adaptive signal detection, yet to our knowledge, prior works in this area have not considered it. Second, any resolution-adaptive method must search over combinatorially many possible groupings of the locations. How can we construct such a method which is maximally powerful, computationally efficient, and still controls some notion of the false positive rate? The goal of our work is to answer these questions.

1.3 Contribution

Our work makes two key contributions. First, we define a notion of *resolution-adjusted power*, which measures the statistical power of resolution-adaptive methods. Crucially, our definition is highly flexible and allows analysts to adapt our metric to fit various applied situations. Although a few previous works have introduced resolution-adaptive methods for variable selection, to our knowledge, our work introduces the first metric which actually allows for general comparisons between different resolution-adaptive methods.

Second, we introduce *Bayesian Linear Programming* (BLiP), which is a method for performing resolution-adaptive signal detection. As an input, BLiP takes the posterior probability that G contains a signal, also known as a posterior inclusion probability (PIP), for each region G in consideration. For example, if $Y | X$ follows a generalized linear model (GLM), one can use a Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution of the model coefficients. Alternatively, analysts can use fast variational methods like SuSiE (Wang et al., 2020).

Given this input, BLiP will output a set of disjoint regions, each containing a signal, which nearly maximizes expected power while provably controlling false positives. For example, in variable selection problems, BLiP will return a set of disjoint groups of variables so that (1) nearly all groups contain at least one signal variable, (2) we discover as many groups as possible, and (3) the groups are as small as possible. Indeed, in genetic fine-mapping, this corresponds to identifying as many important genetic variants as possible and simultaneously localizing them as precisely as possible.

At the outset, we highlight a few attractive features of BLiP:

Power: Where competitors exist, BLiP is often much more powerful than other methods, as we demonstrate in extensive simulations and two real data analyses. This is both because BLiP can take advantage of arbitrary Bayesian methods to compute PIPs, and also because BLiP finds nearly the optimal set of discoveries, a claim we make precise in Section 2.2. For this reason, when competitors exist, BLiP can wrap on top of those competitor methods to increase their power, often by wide margins.

Provable error control: Given correct PIPs, BLiP can provably control one of several error rates, including the FDR, FWER, and local FDR.

Flexibility: Since BLiP acts directly on PIPs, it can be applied on top of any Bayesian model. Furthermore, analysts can use nearly any algorithm to compute the PIPs, allowing them to leveraging recent advances in Bayesian MCMC or variational inference. This allows analysts to flexibly trade-off computational efficiency and statistical power.

Computational Efficiency: Given input PIPs, BLiP is extremely computationally efficient, allowing it to search over hundreds of millions of candidate regions to find a near-optimal set of discoveries. This allows BLiP to make discoveries at resolutions which are as precise as possible even in very large-scale problems without obvious hierarchical structure.

To quote Wang et al. (2020), “the output from Bayesian Variable Selection methods is typically a complex posterior distribution, and this can be difficult to distill into results that are easily interpretable.” BLiP is designed to solve exactly this problem (in the more general signal detection setting), and thus it offers a highly general framework to extract interpretable and replicable information from complex high-dimensional posterior distributions.

1.4 Related Literature

One natural approach when searching for signals is to partition the locations into groups in an unsupervised manner and then test whether there is a signal in each group. Here, note that in the frequentist literature, detecting a signal in a group $G \subset \mathcal{L}$ is often framed as rejecting a group null hypothesis of the form $H_G = \bigcap_{\ell \in G} H_\ell$, where H_ℓ is the null hypothesis that no signal exists at location ℓ . For example, in variable selection problems, many methods use X to cluster the variables

into groups and then test if there exists a signal (e.g., a nonzero coefficient) in each group using both X and Y (Hastie et al., 2001). The key difference between BLiP and these methods is that BLiP simultaneously groups the locations *and* discovers signals at the same time. For example, in variable selection, if X_1 and X_2 are moderately correlated, unsupervised clustering methods have to guess whether or not to group X_1 and X_2 together without looking at Y : in contrast, BLiP can use the full data to determine if there is enough signal to distinguish between X_1 and X_2 (see Appendix A for a concrete example where the best grouping is discernible from $Y | X$ but not X alone). As a result, BLiP can use the full data to localize signals as precisely as possible: this is the advantage of taking a resolution-adaptive approach.

A few recent works have developed methods to search over hierarchical sets of groups to detect signals at as fine a resolution as possible (Yekutieli, 2008; Renaux et al., 2018; Katsevich and Sabatti, 2019; Sesia et al., 2019; Katsevich et al., 2021). In the language of this paper, these methods perform resolution-adaptive signal detection or something quite similar to it, and indeed, these can be powerful methods in many settings. That said, they have a few disadvantages. First, each of these methods can only search over a somewhat limited set of candidate groups, often corresponding to a hierarchical tree. As a result, their adaptivity is limited. Although BLiP also can only make discoveries from a candidate set of groupings, BLiP places no restrictions on this candidate set and is efficient enough to search over, e.g., thousands of possible hierarchical groupings of the variables. As Wang et al. (2020) have observed, this is important in practical applications like genetic fine-mapping where the variables do not follow a single hierarchical structure. Second, these methods either assume complete independence of p-values (Yekutieli, 2008) or make conservative assumptions to control the FDR under dependence (Katsevich et al., 2021). As a result, we show in Section 3 that each of these methods is either unable to reliably control the FDR or loses statistical power. BLiP, in contrast, provably controls the Bayesian FDR at nearly the nominal level and has high power, since it is highly adaptive. Lastly, there is a much larger literature on hierarchical testing more broadly, including methods which control, e.g., the FWER (Meinshausen, 2008; Goeman and Solari, 2012) or weighted versions of the FDR (Benjamini and Hochberg, 1997; Bogomolov et al., 2020). (Note that like us, Benjamini and Hochberg (1997) introduce a weighted notion of power, but their definition of power is less general and has a fundamentally different motivation than ours since they do not consider resolution-adaptive inference or hierarchical testing problems.) Furthermore, there are resolution-adaptive methods in change point detection which control the FWER (see, e.g., Frick et al. (2014); Fang et al. (2016); Fang and Siegmund (2020); Fryzlewicz (2020)) or a slightly modified FDR (Li et al., 2016). It would be impossible to exhaustively compare to each method from this vast literature, however, and thus we only compare BLiP to methods which return a disjoint set of discoveries and target the FDR as defined in Definition 1.1, which is sometimes known as the “outer nodes FDR.” This choice is also motivated by our real data applications, where this FDR is a standard choice of error rate.

Our work is perhaps closest in spirit to that of Wang et al. (2020), who introduced “SuSiE,” a method for sparse Bayesian linear regression.² SuSiE uses a novel variational approximation (which is accurate when the number of signals is small) to approximate the posterior and then processes that approximate posterior to perform resolution-adaptive inference. In contrast, BLiP performs only the latter task, but it can do so on any posterior. For example, BLiP can apply directly to the approximate posterior from SuSiE, as we show in Section 3, where BLiP uniformly improves SuSiE’s power. Even when the SuSiE variational approximation is inaccurate, BLiP can often

²As is common in the literature, we will use the name “SuSiE” to refer to both the SuSiE model and the Iterative Bayesian stepwise selection (IBSS) algorithm.

partially correct this issue, leading to improved power and FDR control. Note that BLiP can also apply on top of the refined SuSiE procedure suggested in Zou et al. (2021), although the refined procedure was too computationally expensive for us to use in our simulations. To our knowledge, the only other comparable Bayesian method is DAP-G (Lee et al., 2018), which also requires a specific approximation to the posterior to perform resolution-adaptive inference. In principle, BLiP can also wrap on top of DAP-G to improve its power, although we focused only on combining SuSiE with BLiP in our simulations because SuSiE outperformed DAP-G. Alternatively, BLiP can be combined with any other method to approximate the posterior, e.g., MCMC, and thus BLiP offers an attractive alternative to perform resolution-adaptive inference without (necessarily) making any variational approximation.

1.5 Notation

For any $k \in \mathbb{N}$, let $[k]$ denote the set $\{1, \dots, k\}$. For any subset $J \subset [k]$ and vector $v \in \mathbb{R}^k$, let v_J denote the entries of v corresponding to the set J . Similarly, for any matrix $M \in \mathbb{R}^{m \times k}$, let M_J denote the columns of M corresponding to J , and let M_j denote the j th column for $j \in [k]$. To ease notation, let M_{-j} denote the set of all columns of M except column j , and similarly v_{-j} denotes all but the j th element of v . Much of this paper focuses on regression problems, and in these settings, we let $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ denote n independent samples of an outcome of interest, and we let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the design matrix for p variables. For convenience, we will let the non-bold notation $Y \in \mathbb{R}$ and $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ refer to an arbitrary single data-point of the outcome and covariates. We let \mathcal{L} denote the set of locations, and for any $G \subset \mathcal{L}$, I_G is the indicator of whether a signal exists in G . Note that we use the bold notation \mathbf{I}_k to specify the $k \times k$ identity matrix. In settings where we obtain samples from a posterior distribution, we let N denote the number of posterior samples. For any point $x \in \mathbb{R}^d$, let $S_r(x) \subset \mathbb{R}^d$ denote the closed d -sphere of radius r around x , and let $C_r(x) \subset \mathbb{R}^d$ denote the closed d -cube of side-length $2r$ around x . In all problems, we let \mathcal{D} denote all of the observed data.

1.6 Outline

The outline of the rest of the paper is as follows. In Section 2, we formally define our notion of resolution-adjusted power and then introduce BLiP. In Section 3, we demonstrate the advantages of BLiP over competitors via simulations. In Sections 4 and 5, we apply BLiP to a genetic fine-mapping analysis of UK Biobank data and to detect point sources in the Sloan Digital Sky Survey, respectively. Section 6 concludes with a discussion of future directions. For convenience, Section 7 includes a list of links to aid replication of the analyses in this paper.

2 Bayesian Linear Programming for Resolution-Adaptive Signal Detection

2.1 Resolution-Adjusted Power

In this section, we formally define *resolution-adjusted power*. Given locations \mathcal{L} , intuitively, we would like to make as many true discoveries as possible while controlling, e.g., the false discovery rate. However, as discussed in Section 1.2, there is some ambiguity in how to count the number of discoveries, since discovering a smaller group gives us more specific information than discovering a larger group. For example, it is not obvious whether it is preferable to perfectly localize a single signal or to instead discover two groups of size two.

To resolve this ambiguity, we suggest weighting discoveries by the size of the group, so that, e.g., discovering a group of size 3 counts as “fewer” discoveries than discovering a group of size 2. Indeed, one simple proposal might be to assign a group of size m weight $\frac{1}{m}$, so that discovering a group of size 1 and a group of size 2 would count as 1.5 discoveries total. Intuitively, it makes sense that if we can only say that a signal exists in one of two locations, this gives us (by some measure) half as much information as discovering the true location of the signal. However, this definition may not be appropriate for every problem: some analysts might prefer to use a different weighting scheme. Indeed, in variable selection problems, some analysts may be more willing to discover the group $\{1, 2\}$ if X_1 and X_2 are highly correlated. For this reason, the rest of this paper allows for analysts to use an arbitrary *weight function* w which maps each possible group G to a weight. After discovering a set of groups G_1, \dots, G_R , we define power to be the sum of the weights associated with each true discovery.

Definition 2.1 (Resolution-Adjusted Power). Given a set of locations \mathcal{L} and data \mathcal{D} , suppose a method discovers a disjoint set of groups $G_1, \dots, G_R \subset \mathcal{L}$. Let $w : \mathcal{P}(\mathcal{L}) \rightarrow \mathbb{R}$ be a *weighting function* on $\mathcal{P}(\mathcal{L})$, the set of all subsets of \mathcal{L} . Define the resolution-adjusted power as follows:

$$\text{Power}(G_1, \dots, G_R) = \sum_{r=1}^R I_{G_r} w(G_r), \quad (2.1)$$

where it may be helpful to recall that I_{G_r} is the indicator of whether a signal truly exists in $G_r \subset \mathcal{L}$.

From an economic perspective, $w(G)$ might represent the utility of discovering group G . Alternatively, in variable selection problems, w could account for the overall dependence in region G , or it could even account for a preference to discover signals with larger signal sizes, since w is allowed to depend on the data-generating process. The important point, however, is that BLiP can optimize for *any* possible weight function.

That said, unless otherwise specified, in this paper we will weight groups by the inverse of their size, so $w(G) = |G|^{-1}$. As discussed previously, this weighting corresponds to the intuition that discovering a region of size k gives us roughly k times less information than discovering an individual signal. The reason we use this weighting function is mainly because it is simple, interpretable, and does not depend on the specifics of the problem, making it easier to keep track of what power means across many different settings. For the rest of the paper, we use the word “power” to mean resolution-adjusted power unless we explicitly specify otherwise.

2.2 Bayesian Linear Programming for FDR Control

Having formally defined resolution-adjusted power, we now introduce a Bayesian Linear Program (BLiP) to perform resolution-adaptive signal detection. In this paper, we take a Bayesian approach, meaning that we aim to maximize expected power and control the false positive rate conditional on the data \mathcal{D} . Although BLiP can control other error rates as well, we start by discussing how to control the FDR, as it is the de facto error rate of choice in many applications including genetic fine-mapping.

This problem initially seems intractable, but it turns out that extremely high quality solutions can be found via a linear programming relaxation. Before we discuss this relaxation, however, note that it still may be too challenging computationally to search over all possible groupings, since there are an enormous (often infinite) number of possible subsets of \mathcal{L} . To narrow down the search space, we require that G_1, \dots, G_R are members of a set of *candidate groupings*, \mathcal{G} . This is not a particularly restrictive requirement, since one can make \mathcal{G} as large as is computationally feasible, and our algorithm can handle hundreds of millions of candidate groupings. For example, in a variable selection problem, one could cluster the variables X_1, \dots, X_p using (literally) a thousand different clustering algorithms and then let \mathcal{G} denote the union of the groups created by the clustering algorithms. We offer more suggestions for constructing \mathcal{G} in Section 2.4.

Given candidate groupings \mathcal{G} , the first key observation is that maximizing power corresponds to maximizing a linear function. To see this, let $p_G = \mathbb{E}[I_G | \mathcal{D}]$ be the posterior probability that there is a signal in group G , also known as a posterior inclusion probability (PIP). Then, let $x_G \in \{0, 1\}$ be the indicator of whether our procedure discovers group G : here, $\{x_G\}_{G \in \mathcal{G}}$ are our optimization variables which completely determine our discoveries G_1, \dots, G_R . Using this notation plus the definition of resolution-adjusted power, we obtain that

$$\mathbb{E}[\text{Power}(G_1, \dots, G_R) | \mathcal{D}] := \mathbb{E} \left[\sum_{r=1}^R I_{G_r} w(G_r) | \mathcal{D} \right] = \sum_{G \in \mathcal{G}} p_G w(G) x_G, \quad (2.2)$$

where the right-most equality sums over \mathcal{G} because $x_G = 1$ if and only if $G \in \{G_1, \dots, G_R\}$. Note that for simplicity, Equation (2.2) assumes that the weights $w(G)$ are constants which do not depend on any unknown parameters. However, one can also allow $w(G)$ to depend on unknown parameters by simply replacing $p_G w(G)$ with $\mathbb{E}[I_G w(G) | \mathcal{D}]$ (which will be no harder to approximate than $p_G w(G)$ given a set of samples from the posterior distribution) in Equation (2.2). Then, the rest of the analysis in the paper will carry through without further modification.

Equation (2.2) tells us that if we have access to $\{p_G\}_{G \in \mathcal{G}}$, the objective function is a linear function of $\{x_G\}_{G \in \mathcal{G}}$. Of course, computing the PIPs is often non-trivial and may require sophisticated MCMC methods or variational inference. However, this problem is *very* well studied in the Bayesian literature (Mitchell and Beauchamp, 1988; George and McCulloch, 1997; Shin and Liu, 2021; Albert and Chib, 1993; Polson et al., 2013; Jacob et al., 2020; Biswas et al., 2022), so for now, we assume that we do have access to the PIPs—we discuss practical ways to compute them throughout Section 3 and Appendix C.

Notably, the FDR constraint can also be formulated as a linear constraint. In particular, let

$V = R - \sum_{r=1}^R I_{G_r}$ be the number of false discoveries. Controlling the FDR at level q requires that

$$\text{FDR} := \mathbb{E} \left[\frac{V}{R} \mid \mathcal{D} \right] = \frac{\sum_{G \in \mathcal{G}} (1 - p_G) x_G}{\sum_{G \in \mathcal{G}} x_G} \leq q, \quad (2.3)$$

where in the above equation we use the convention that $0/0 = 0$. Multiplying by $\sum_{G \in \mathcal{G}} x_G$ on both sides yields the linear constraint $\sum_{G \in \mathcal{G}} (1 - p_G - q) x_G \leq 0$. Thus, as the following proposition makes clear, the resolution-adaptive signal detection problem can be formulated as a mixed-integer linear program (LP). A proof is given in Appendix B.1.

Proposition 2.1. *The solution to the resolution-adaptive signal detection problem in Definition 1.1 is the same as the solution to the following mixed-integer LP:*

$$\max_{\{x_G\}_{G \in \mathcal{G}}: x_G \in [0,1]} \quad \sum_{G \in \mathcal{G}} p_G w(G) x_G \quad (2.4)$$

$$\text{s.t.} \quad \sum_{G \in \mathcal{G}} (1 - p_G - q) x_G \leq 0, \quad (2.5)$$

$$\sum_{G \in \mathcal{G}: \ell \in G} x_G \leq 1 \quad \forall \ell \in \mathcal{L}, \quad (2.6)$$

$$x_G \in \{0, 1\} \quad \forall G \in \mathcal{G}. \quad (2.7)$$

Note that when \mathcal{G} or \mathcal{L} are infinite sets, it is possible to reduce this integer LP to an equivalent finite-dimensional problem assuming (i) the expected number of signals is finite and (ii) a very mild regularity condition on \mathcal{G} . In Appendix D, we introduce an efficient algorithm based on edge clique covers (Erdős et al., 1966) to perform this reduction.

Mixed-integer LPs are NP complete but nonetheless are very well studied (Jünger et al., 2010), so it is often possible to get approximate or even exact solutions fairly efficiently. That said, we would like to be able to search over millions or even billions of candidate groupings, so we would still like to improve efficiency. To this end, in large problems, we instead suggest solving the relaxed problem (2.4)-(2.6), which drops the integer constraint that $x_G \in \{0, 1\}$ and only requires that $x_G \in [0, 1]$. The relaxed problem is a vanilla linear program with highly sparse constraints, so it can be solved extremely efficiently (see, e.g., Huangfu and Hall (2015)). Relaxations of this form are a common technique in optimization problems which often yield good results in practice (Boyd and Vandenberghe, 2004). Furthermore, it turns out that this problem is sufficiently similar to a *knapsack* problem that the relaxed version returns solutions which are almost entirely composed of integers. Usually, any non-integer values roughly correspond to the need to randomize the discoveries to achieve an FDR of exactly level q . Indeed, in Appendix B.2, we show that the solution to the relaxed LP typically has a single-digit number of non-integer values when searching over tens of thousands of groups. Although we cannot prove that this will happen in general, Appendix B.2 gives intuition to explain this phenomenon, and we will present further empirical evidence for this claim shortly.

This motivates Algorithm 1, which defines BLiP for FDR control. To start, BLiP solves the relaxed LP corresponding to (2.4)-(2.6), which typically yields a solution $\{x_G^*\}_{G \in \mathcal{G}}$ with only a few non-integer values of x_G^* , denoted $\mathcal{H} = \{G \in \mathcal{G} : x_G^* \notin \{0, 1\}\}$. To obtain a deterministic rejection set, BLiP then runs the integer linear program (2.4)-(2.7) while holding the values of $\{x_G^* : G \in \mathcal{G} \setminus \mathcal{H}\}$ constant and only optimizing over $\{x_G : G \in \mathcal{H}\}$. Typically, this integer LP has only a few variables, usually in the single digits, so it runs practically instantly, yielding the output of BLiP. Note that

technically speaking, up to two things could go wrong in this step. First, if the integer LP is very large, it may be challenging to solve it efficiently, although in this case the user will know this in advance and can use polynomial-time heuristic methods instead (see Appendix B.4). Second, it is technically possible for the integer LP to be infeasible, in which case we propose a backtracking algorithm that iteratively finds the group G_{\min} in $\{G : x_G^* = 1\}$ with the smallest PIP and adds G_{\min} to \mathcal{H} . This guarantees that BLiP finds a feasible solution in a finite number of backtracking steps. That said, neither of these two phenomenon ever occurred in any of the simulations or real data applications in this paper, despite our applying BLiP tens of thousands of times in very large scale settings. This suggests that these modifications are only required in pathological examples. In all cases, however, the output of BLiP is a feasible solution to (2.4)-(2.7) and thus provably controls the FDR by Proposition 2.1, even in pathological examples.

Algorithm 1 BLiP for FDR control.

Input: Candidate groups \mathcal{G} , PIPs $\{p_G\}_{G \in \mathcal{G}}$, a weighting function w , a nominal level q .

- 1: Perform adaptive preprocessing on \mathcal{G} using the PIPs (Appendix B.3).
 - 2: Solve the relaxed LP corresponding to (2.4)-(2.6) to obtain a solution set $\{x_G^*\}_{G \in \mathcal{G}}$. Let $\mathcal{H} = \{G \in \mathcal{G} : x_G^* \notin \{0, 1\}\}$ denote the non-integer solutions.
 - 3: Fix $x_G = x_G^*$ for $G \in \mathcal{G} \setminus \mathcal{H}$ and run the integer LP (2.4)-(2.7) on $\{x_G : G \in \mathcal{H}\}$ and v , yielding (integer) solutions $\{x_G^{**} : G \in \mathcal{G}\}$.
 - 4: If the integer LP in Step 3 is feasible, detect signals in $\{G : x_G^{**} = 1\}$ and terminate.
Else, define $G_{\min} := \arg \min_{\{G : x_G^* = 1\}} p_G$, set $\mathcal{H} = \mathcal{H} \cup G_{\min}$, and return to Step 3.
-

Lastly, we can further speed up computation by *adaptively preprocessing* the candidate groups \mathcal{G} . The main idea is that after looking at the data, we can tell that certain groups and locations have a very low chance of containing a signal, so we exclude these groups from consideration before running the LP. We develop three algorithms to accomplish this, although for brevity, we defer details to Appendix B.3. Often, these heuristics improve computation by multiple orders of magnitude, since when the signals are sparse, most groups and locations have a low posterior probability of containing a signal and can be discarded. That said, it is important to note that this step in no way compromises BLiP’s provable Type I error guarantees, and furthermore, these common sense heuristics have almost no impact on power empirically. Intuitively, this makes sense: if a group G has only a 0.001% chance of containing a signal, it is extremely unlikely we would have discovered it anyway, so there is no need to consider it in the LP. Note that unlike methods which by default restrict \mathcal{G} to be very small, adaptive preprocessing does not sacrifice the adaptability of BLiP, since it uses the full data to prune the candidate groups \mathcal{G} . Overall, these heuristics substantially improve computation and we have never observed them to meaningfully reduce power, so we apply them by default for the rest of the paper.

Before discussing other error rates, we briefly discuss the heuristic claim that BLiP finds “nearly” the optimal grouping of the locations in the candidate set \mathcal{G} . As discussed in Appendix B.2, approximate optimality follows from the fact that the relaxed LP usually admits a solution which is almost entirely composed of integers. As a result, the expected power (objective function value) achieved by BLiP is extremely close to the expected power obtained by the non-integer LP solution, which is an upper-bound on the maximum achievable power. Of course, even when this is not true, one can compute and compare the expected power achieved by the relaxed LP and BLiP, so BLiP also comes with “warning lights” which signal when it is not optimal. However, Figure 1 confirms empirically that in three challenging high-dimensional variable selection problems with tens of

thousands of candidate groups, the nominal expected power achieved by BLiP is indistinguishable from the upper bound provided by the relaxed LP. This suggests BLiP is finding effectively the optimal solution. Lastly, even if there are rare cases when BLiP is not optimal, our simulations and data applications suggest that BLiP is more powerful than its competitors in practical settings.

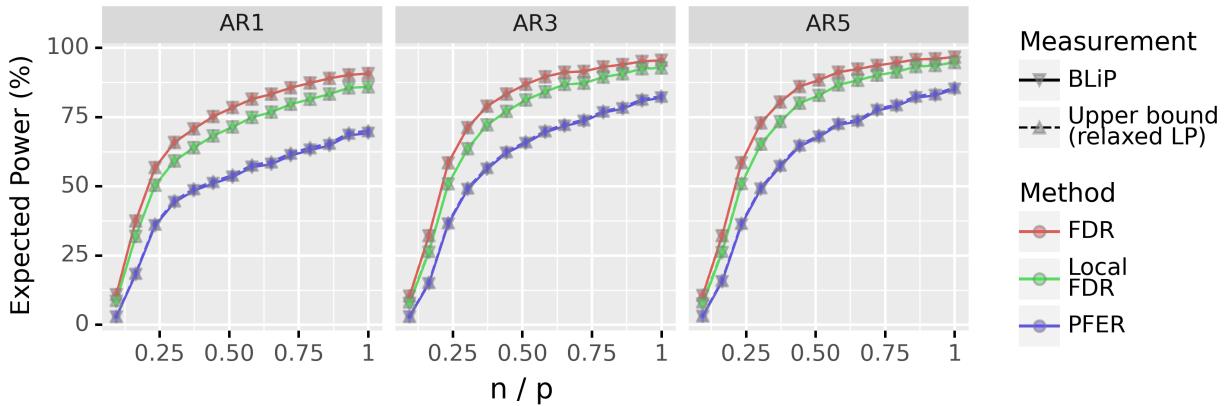


Figure 1: We sample $X \sim \mathcal{N}(0, \Sigma)$ for various Σ and $Y | X \sim \mathcal{N}(X\beta, 1)$, where β has 50 nonzero components sampled from $\mathcal{N}(0, 1)$ and $p = 1000$. Here, the facets correspond to different Σ : see Appendix E.2 for precise details on the setting. We run Algorithm 1 on top of a generic spike-and-slab MCMC algorithm with greater than 50,000 candidate groups, and we plot the objective function (expected power) achieved by BLiP and the relaxed LP. We also run BLiP to control the PFER and local FDR as discussed in Section 2.3. In all cases, the power of the relaxed LP is an upper bound on the expected power of any valid resolution-adaptive method. For all error rates, the upper bound completely overlaps with BLiP’s expected power, which suggests BLiP is effectively optimal in this setting. Note that we can create an analogous plot for the FWER, but its interpretation is slightly more complicated, so we defer it to Appendix B.2. Error bars denote two standard errors, but they are so small they are often not visible.

2.3 Other error rates

In this section, we introduce variants of BLiP that control the local FDR and FWER. We start with the local FDR, which requires that the PIP for each discovered group must be at least $1 - q$. To control the local FDR, we can start by letting $\mathcal{J} = \{G \in \mathcal{G} : p_G \geq 1 - q\}$ denote the set of groups which have a PIP of at least $1 - q$. Then, we can solve the following relaxed linear program to optimize for power while controlling the local FDR:

$$\begin{aligned} & \max_{\{x_G\}_{G \in \mathcal{J}}: x_G \in [0, 1]} \quad \sum_{G \in \mathcal{J}} p_G w(G) x_G \\ \text{s.t.} \quad & \sum_{G \in \mathcal{J}: \ell \in G} x_G \leq 1 \quad \forall \ell \in \mathcal{L}. \end{aligned}$$

As before, once we solve this LP, most of the solutions will be integers, and we can apply an integer LP or polynomial-time heuristic to the non-integer values to obtain discoveries G_1, \dots, G_R , just as in Algorithm 1. These discoveries trivially control the local FDR, since we have already filtered out groups G which have a PIP below $1 - q$.

The last error rate we will explicitly address is the family-wise error rate (FWER), which requires that there is at most probability q that we make any false discoveries, or equivalently, that $\mathbb{P}(V >$

$0 \mid \mathcal{D}) \leq q$. Since $\mathbb{P}(V > 0 \mid \mathcal{D}) \leq \mathbb{E}[V \mid \mathcal{D}]$, one valid way to control the FWER at level q is to ensure $\mathbb{E}[V \mid \mathcal{D}] \leq q$, which is also known as controlling the per-family error rate (PFER)—indeed, this is the approximation used by the standard Bonferroni correction. This is slightly conservative but works well empirically when q is fairly small, e.g., if $q \lesssim 0.2$. The advantage of making this approximation is that $\mathbb{E}[V \mid \mathcal{D}] = \sum_{G \in \mathcal{G}} (1 - p_G) x_G$ is linear in $\{x_G\}$. As a result, to control the FWER, we can run Algorithm 1 but replace the FDR constraint (2.5) with the linear constraint $\mathbb{E}[V \mid \mathcal{D}] \leq q$. Another alternative is to control the PFER at level v and perform a binary search over v to find the largest value v^* such that $\mathbb{P}(V > 0 \mid \mathcal{D}) \leq q$. Either method is guaranteed to control the FWER, though the latter will in general have slightly increased power at the cost of some extra computation.

By adding extra constraints or modifying the weights in existing constraints, one can easily modify BLiP to control various other error rates, such as a weighted FDR (Benjamini and Hochberg, 1997; Bogomolov et al., 2020), or even to control the FDR across multiple hierarchies simultaneously (Barber and Ramdas, 2017; Katsevich and Sabatti, 2019). However, we leave this possibility for future work.

2.4 Choosing the candidate groups and computing the PIPs

In this section, we discuss how to compute the primary inputs to BLiP: the candidate groups \mathcal{G} and the PIPs $\{p_G\}_{G \in \mathcal{G}}$. To begin with, how can one obtain good candidate groups \mathcal{G} ? Our overall advice is as follows: up to computational limits, we suggest adding every conceivably useful group to \mathcal{G} , since adding more groups to BLiP should only improve its power without affecting validity. Furthermore, BLiP is usually efficient enough to handle hundreds of millions of candidate groups. That said, we recommend two general approaches to find potentially useful groups.

First, we can let \mathcal{G} include the set of all *contiguous* regions below some maximum size m . This approach makes sense when the locations have spatial or temporal structure that we can exploit. For example, suppose $\mathcal{L} = \{\ell_1, \dots, \ell_p\}$ consists of p ordered locations of genetic variants on the genome. In this context, a contiguous group is of the form $\{\ell_i, \ell_{i+1}, \dots, \ell_{i+k}\}$ for some $i, k \in \mathbb{N}$. Since genetic variants exhibit local biological similarities and mostly local correlations, considering contiguous groups is arguably more interpretable and likely more useful than considering groups of far-flung genetic variants. This option is also attractive in change point detection (Example 3), where the set of locations $\mathcal{L} = \{1, \dots, T\}$ is a set of ordered times. Notably, there are roughly $m \cdot p$ contiguous groups of length m or less when considering p locations, so the number of candidate groups scales linearly with p . Lastly, when there are more dimensions of space and/or time, we suggest using spherical subsets of \mathcal{L} as candidate regions. For example, in our application to astronomical point source detection in Section 5, $\mathcal{L} = [0, 1]^2$, and we let \mathcal{G} include the set of circles of radius ϵ centered at one of a few million equidistant lattice points in $[0, 1]^2$, for many possible values of ϵ . We introduce efficient algorithms for this setting in Appendix D.

The second main approach we recommend is tailored to regression problems, where we seek to discover important variables among $X = (X_1, \dots, X_p)$. In this setting, we recommend applying many different clustering algorithms to X and letting \mathcal{G} denote the union of those groups. For example, one could generate candidate groupings by hierarchically clustering X based off its correlation matrix, or its partial correlation matrix (Bühlmann et al., 2012), or any combination thereof. Furthermore, we re-run these algorithms many times using different tuning parameters to add more candidate groups to \mathcal{G} . Finally, there is no need to choose between multiple approaches: when

(X_1, \dots, X_p) exhibit spatial structure, we can combine this approach with that of the previous paragraph.

The second important input to BLiP is the set of PIPs $\{p_G\}_{G \in \mathcal{G}}$. Since one can compute the PIPs based on the output of nearly any Bayesian MCMC algorithm or variational method, the question of how to compute the PIPs is essentially orthogonal to our contribution (BLiP), so we largely defer discussion of computing $\{p_G\}_{G \in \mathcal{G}}$ to Appendix C, where we give detailed suggestions. However, in this section, we will touch upon one important point, which is that of robustness. By construction, BLiP provably controls various Bayesian error rates, such as the FDR, under the assumption that the PIPs $\{p_G\}_{G \in \mathcal{G}}$ are correct. However, in many practical settings of interest, (a) Bayesian MCMC algorithms may not fully converge or (b) one may not have an accurate Bayesian prior. This is indeed an important concern: BLiP is only as good as its inputs, and like most Bayesian selective inference procedures, it will violate error control if the PIPs are inaccurate.

With this motivation, in Appendix C, we offer guidelines on how to compute PIPs in a way that is robust to these concerns. In brief, our advice is as follows. To address (a), we generally recommend sampling from multiple MCMC chains (or variational algorithms, when appropriate) with uniformly random initialization. Even when each individual chain fails to converge, we expect this approach to overstate the uncertainty in the location of a signal, leading to conservative but valid inference. To address (b), we recommend using *hierarchical priors* with conservative or estimated hyperparameters to account for uncertainty in the Bayesian model, such as modeling the sparsity level in regression problems. Since these are fairly standard techniques in Bayesian analysis, we defer details to Appendix C. Furthermore, in case there is non-negligible uncertainty in the PIPs, we note the possibility of replacing p_G in Equation (2.5) with p_G^{lower} , a high-confidence (or uniform) lower bound on p_G . For example, p_G^{lower} could account for uncertainty in the choice of prior or uncertainty due to MCMC error, although empirically this did not seem to be necessary in our experiments. This is equivalent to running a robust linear program (Ben-Tal et al., 2009), and it is easy to see that if $p_G \geq p_G^{\text{lower}}$ does hold for all G , then BLiP will still provably control the FDR at level q . Overall, in Section 3, we often use BLiP in combination with fairly uninformative priors for fairer comparison with frequentist methods. Despite this, BLiP is very powerful and consistently controls the FDR, even in high-dimensional settings where we have no reason to believe MCMC algorithms will fully converge.

3 Simulations

In this section, we demonstrate that BLiP is powerful, robust, and efficient compared to its competitors. We focus on the case of variable selection in Gaussian linear models and probit regression, although Appendix E.3 also includes simulations which show that BLiP holds promise as a method for change point detection. Note that Section 4 and Appendix C.1 contain further simulations related to genetic fine-mapping and the robustness of BLiP, respectively. All simulations were run on cores with 4 gigabytes of memory.

3.1 Variable selection for Gaussian linear models

We begin by applying BLiP to perform variable selection in (sparse) Gaussian linear models. We simulate $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$ where β has $[sp]$ randomly chosen nonzero coefficients, for sparsity

$s \in (0, 1)$. The nonzero coefficients are i.i.d. $\mathcal{N}(0, \tau^2)$ random variables. The locations $\mathcal{L} = [p]$ correspond to (X_1, \dots, X_p) , and we say $I_j = 1$ if $\beta_j \neq 0$. To capture a challenging setting where the variables are correlated, we sample \mathbf{X} from a nonstationary AR(k) model, meaning \mathbf{X} exhibits high local correlations with a complex structure—see Appendix E for details. Unless otherwise specified, we set $k = 5$ and $p = 1000$, and we control the FDR at level $q = 0.1$. We plot resolution-adjusted power as a percent, defined as the resolution-adjusted power divided by the total number of true signals.

We run BLiP on top of two types of models. First, we apply BLiP on top of a standard Gibbs sampler for linear spike-and-slab (LSS) models, and we consider both a nearly well-specified case, where the sampler uses the true values of s, τ^2 and σ^2 , and a misspecified case, where the sampler uses uninformative priors on s, τ^2 , and σ^2 ; see Appendix C.2 for detailed model descriptions.³ Second, we apply BLiP on top of SuSiE (Wang et al., 2020). We consider four main competitors: the focused Benjamini Hochberg (FBH) procedure (Katsevich et al., 2021), the “Yekutieli” procedure (Yekutieli, 2008), SuSiE, and DAP-G (Lee et al., 2018). The FBH and Yekutieli act on p-values, which we obtain using a lasso-based distilled conditional randomization test (dCRT) (Candès et al., 2018; Liu et al., 2021a). We used the dCRT because it was powerful in our simulations and because it allows us to obtain frequentist p-values in high-dimensional settings. We allow SuSiE/DAP-G to use the true sparsity level s . See Appendix E for more details.

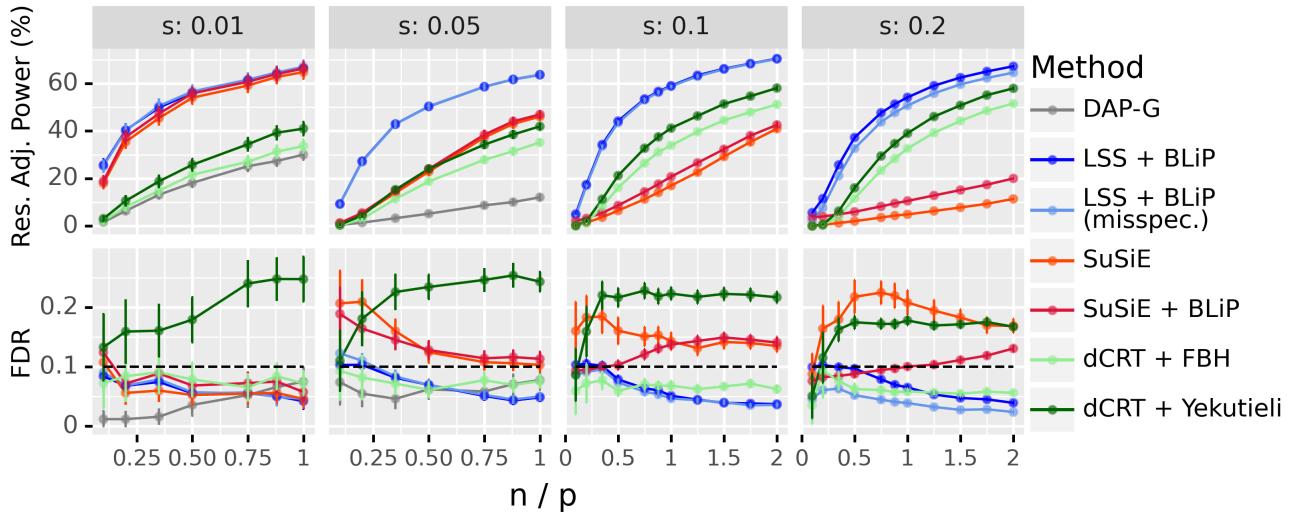


Figure 2: Resolution-adaptive variable selection for Gaussian linear models. Note that DAP-G is prohibitively expensive for $s > 0.05$, so we do not run DAP-G except in the left-most panels.

Figure 2 shows the (resolution-adjusted) power and FDR for each method while varying the number of data-points n and the sparsity s . It shows that LSS + BLiP is uniformly the most powerful method and reliably controls the FDR, even in high-dimensional settings where $p = 10n$ and the prior is misspecified. Remarkably, LSS + BLiP (misspec.) always achieves the same power as the (nearly) well-specified LSS model with oracle knowledge of the hyperparameters. Besides LSS + BLiP, only FBH and DAP-G reliably control the FDR, and LSS + BLiP often has double their power (or more). Similarly, LSS + BLiP is often many times more powerful than SuSiE and the Yekutieli procedure, even though these methods violate FDR control. Furthermore, SuSiE + BLiP has uniformly higher power than SuSiE and often simultaneously improves upon SuSiE’s FDR control. Overall, Figure

³The “nearly well-specified” case is not perfectly well-specified because the prior assumes there are $\text{Bin}(p, s)$ signals, whereas in our simulations there are always exactly $\lceil sp \rceil$ signals.

2 confirms the intuition from Section 1.4 about why BLiP outperforms its competitors, which we discuss in the next two paragraphs.

First, SuSiE behaves as suggested in Section 1. Recall that (roughly speaking) each iteration of SuSiE makes a variational approximation assuming that there is exactly one additional signal. When $s = 0.01$ and there are 10 signals total, this approximation is closest to the truth and SuSiE nearly matches the power of LSS + BLiP. Yet as s increases, the approximation becomes less accurate. As a result, SuSiE has much lower power than LSS + BLiP and violates FDR control in each of the other panels. Note that SuSiE’s variational approximation is inaccurate when the absolute number of signals is large, so SuSiE may perform poorly even in very sparse problems with large p (see Appendix E.2 for a simulation confirming this). However, applying BLiP on top of SuSiE can remedy this problem to some extent. Indeed, any disjoint output from SuSiE is a feasible output for BLiP, so we expect BLiP to uniformly improve SuSiE’s power, which is supported by all of our simulations. Furthermore, BLiP can simultaneously improve FDR control, as shown in the panels where $s \geq 0.1$. See Appendix C.4 for details on how SuSiE’s approximation breaks down and how BLiP can partially correct this problem.

LSS + BLiP also has uniformly higher power than the FBH/Yekutieli procedures, often by a factor of two or more. We suspect that BLiP gains power because BLiP can search over hundreds of times more candidate groups than the FBH/Yekutieli, which are restricted to search over a single hierarchical tree, and also because BLiP explicitly maximizes power when searching over candidate groups. In contrast, the FBH/Yekutieli only search heuristically over their input p-values and may not find a maximally powerful rejection set. (Note that without a hierarchical structure, the FBH is not guaranteed to return a disjoint rejection set while provably controlling the FDR.) That said, there are many differences between BLiP and the FBH/Yekutieli, so we cannot perfectly account for this power differential. For example, we implemented the most powerful versions of the FBH/Yekutieli we could subject to computational constraints, but perhaps there are more powerful p-values that could be used with the frequentist methods, although we were unable to find such p-values. Overall, we hope that the results speak for themselves, since either way it is not clear how one could apply the FBH/Yekutieli procedures in a way that is more powerful than LSS + BLiP. Note lastly that the Yekutieli procedure fails to control the FDR in any setting, presumably because it only has FDR guarantees when its input p-values are fully independent. The p-values also do not provably satisfy the conditions under which the FBH controls the FDR, but the FBH seems to be more robust.

Lastly, Figure 3 analyzes the runtime of BLiP in large-scale settings, with $s = 0.05$, $n = 0.5p$ and p varied from 500 to 10,000. It shows that the cost of applying BLiP is trivial compared to the cost of running SuSiE or the LSS sampler: when $p = 10,000$, BLiP runs in a few minutes, whereas fitting LSS and SuSiE requires 2 and 9 hours, respectively.

3.2 Variable selection for probit regression

We now apply BLiP to variable selection for probit regression. In particular, consider the same setting as Section 3.1, except instead of observing Y , we observe $Z = \mathbb{I}(Y \geq 0)$. We apply BLiP on top of a Gibbs sampler for this probit spike-and-slab (PSS) model, using the data augmentation scheme from Albert and Chib (1993). See Appendix C.3 for more details on the Gibbs sampler. We compare to the FBH/Yekutieli procedures, which we apply to p-values from a logistic-lasso-based dCRT. Although the PSS sampler has the same asymptotic runtime as the LSS sampler ($O(n_{\text{iter}}np)$),

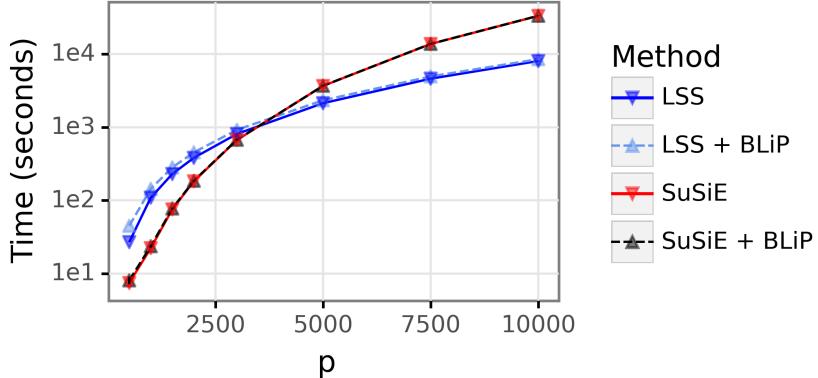


Figure 3: This figure shows computation time while varying p with $n = p/2$ and $s = 0.05$. It distinguishes between the computation required to fit the underlying model and the total time to both fit the model and run BLiP. It shows that the cost of running BLiP is effectively negligible compared to fitting a variational method (SuSiE) or Gibbs sampling in a spike-and-slab model. All methods controlled the FDR (see Appendix E.2).

in practice it is roughly an order of magnitude slower due to the data augmentation step. Indeed, inference is generally slower and more challenging for binary data—for this reason, many applied analyses in, e.g., genetic fine-mapping simply apply methods designed for linear regression to binary outcomes (Weissbrod et al., 2019). For this reason, we also compare to SuSiE and LSS + BLiP, as in Section 3.1.

Figure 4 shows the results in three settings with $p = 400$, with varied s and n . The results show that PSS + BLiP is uniformly the most powerful, often by a wide margin. SuSiE + BLiP has (uniformly) slightly higher power than SuSiE alone, with a slightly higher realized FDR (although occasionally the realized FDR is lower). This suggests that BLiP can enhance power in binary regression problems as well as linear regression.

4 Application to genetic fine-mapping

As discussed in Section 1, resolution-adaptive methods are particularly attractive in fine-mapping problems, where genetic variants are often very highly correlated with each other. As a result, it can be very challenging to detect individually important genetic variants. Resolution-adaptive methods instead allow the analyst to localize causal variants as precisely as possible given the data at hand, and for this reason, a few recent works (Weissbrod et al., 2019; Wang et al., 2020; Wallace, 2021) have used resolution-adaptive methods in fine-mapping problems. Furthermore, Bayesian variable selection methods are already commonly used in genetic fine-mapping problems (Guan and Stephens, 2011; Carbonetto and Stephens, 2012; Benner et al., 2015; Lee et al., 2018; Weissbrod et al., 2019). All this suggests that BLiP can help solve an important problem in the domain of fine-mapping.

To test BLiP’s effectiveness in fine-mapping problems, we apply BLiP to a dataset of $n \approx 337,000$ individuals from the UK Biobank with $p \approx 19,000,000$ genetic variants. In particular, we seek to identify causal genetic variants for four traits of interest: cardiovascular disease, height, low-density lipoprotein (LDL) cholesterol, and high-density lipoprotein (HDL) cholesterol. This dataset was previously analyzed by Weissbrod et al. (2019), and indeed, our work explicitly builds upon theirs.

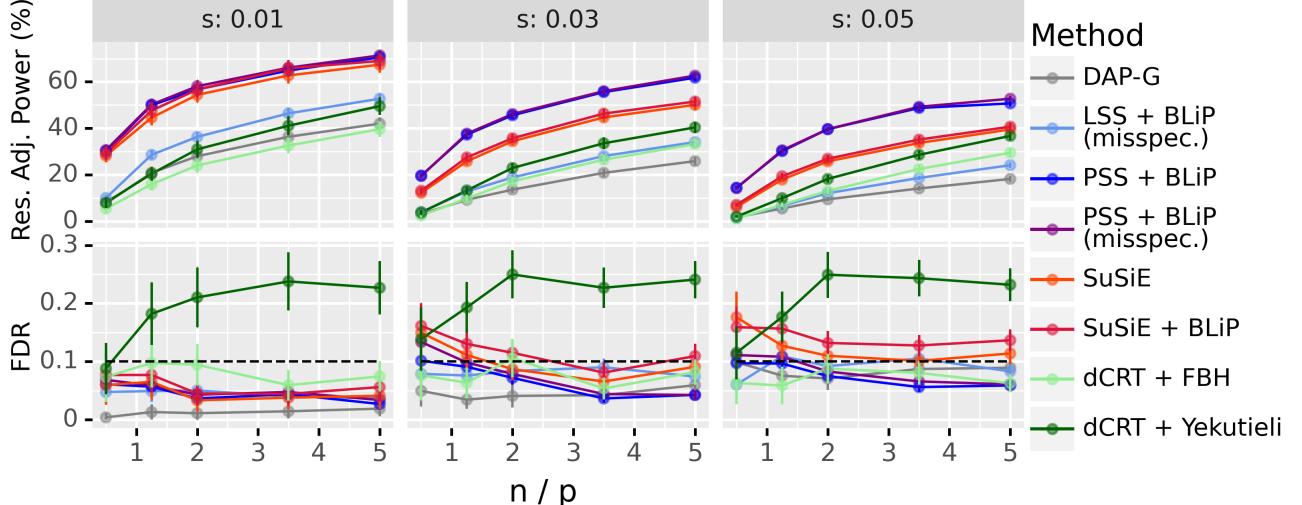


Figure 4: Power and FDR for resolution-adaptive variable selection in the probit model. Note that both the well-specified and misspecified versions of PSS + BLiP are present in all panels, although usually their power curves entirely overlap. See Appendix E.2 for more details.

SuSiE is an attractive model in this setting because we expect a priori that each genetic locus has a small number of causal variants, and our simulations suggest that SuSiE performs nearly as well as full Bayesian inference when the number of signals is small. Thus, we run BLiP directly on top of the SuSiE model that Weissbrod et al. (2019) fit on this dataset to identify causal genetic variants. Running BLiP requires less than 1 minute of computation per trait, but as shown by Figure 5, SuSiE + BLiP is consistently 30–50% more powerful than SuSiE alone. Crucially, this result is not at all sensitive to our definition of resolution-adjusted power: as shown by Figure 6, for every k , SuSiE + BLiP discovers more groups of size k or less than SuSiE alone, meaning that SuSiE + BLiP is making more discoveries at finer resolutions by nearly any metric. Indeed, for every group G discovered by SuSiE, SuSiE + BLiP discovers a group G' which overlaps with G . This suggests that BLiP is successfully optimally localizing signals based on the information available in the SuSiE model—see Appendix C.4 for more intuition on why SuSiE + BLiP can substantially outperform SuSiE alone. Notably, SuSiE + BLiP makes more singleton discoveries than SuSiE alone, in part because SuSiE + BLiP can use PIPs which are provably more powerful than the default SuSiE algorithm—however, we caution that discovering singleton groups is not the primary purpose of BLiP and the interpretation of this result is subtle, so we discuss this further in Appendix F.4. Lastly, note that each group we discover is roughly (but not perfectly) contiguous, meaning that each group only contains nearby genetic variants. As a result, detecting a signal in a group of size k is fairly interpretable: it tells us that one of k neighboring genetic variants has a causal effect on a trait. See Appendix F.1 for methodological details.

To validate our findings, we confirm in simulations using the real genotype data that SuSiE + BLiP controls the FDR in this context, as shown in Appendix F.2. Furthermore, we compare our findings on the real data to previous work: in particular, our findings replicate four of the eight discoveries made in Farh et al. (2015), which roughly matches the replication rate in previous studies (Weissbrod et al., 2019; Lareau et al., 2018). Note that we analyzed a much larger dataset than Farh et al. (2015), which explains why we made far more discoveries than they did. Lastly, as a sanity check, we also compare our findings with the SuSiE model from Weissbrod et al. (2019)

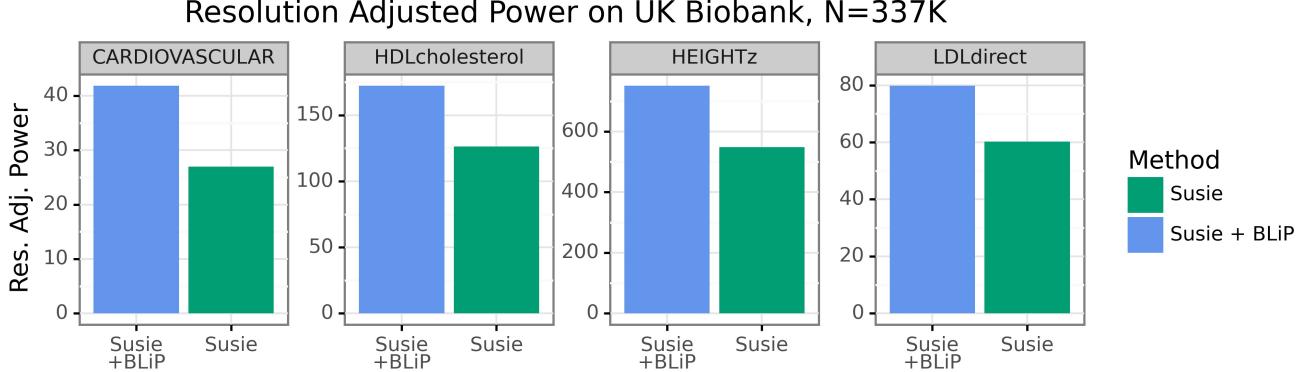


Figure 5: This figure shows the resolution-adjusted number of discoveries made by each method in our application to UK Biobank data. It shows that BLiP increases SuSiE’s power by 30–50%.

(this is the model represented by the green bars in Figure 5).⁴ Appendix F.3 shows that SuSiE + BLiP replicates every finding from the SuSiE model but makes roughly 15 – 20% more discoveries (note this number is not resolution-adjusted). Since SuSiE + BLiP has 30 – 50% higher resolution-adjusted power than SuSiE, this indicates that the power gain comes both from more precisely localizing existing discoveries and from making some entirely new discoveries. Overall, these results suggest that SuSiE + BLiP is working as intended, and they give no indication that the increased resolution-adjusted power of SuSiE + BLiP can be explained away as false discoveries. All code, data, and results are publicly available at https://github.com/amspector100/ukbb_blip.

Lastly, we remind the reader that BLiP can be applied on top of any Bayesian model, yielding higher power with very little additional computational cost. For example, in this section, SuSiE uses a relatively uninformative prior for simplicity, but several recent works have incorporated information from other sources into the prior, including functional annotations, other complex traits, and prior knowledge about the distribution of genetic effect sizes (Weissbrod et al., 2019; O’Connor, 2020; Weeks et al., 2020; Trippe et al., 2021). Similarly, the fine-mapping literature contains many inferential algorithms besides SuSiE (Carbonetto and Stephens, 2012; Hormozdiari et al., 2014; Benner et al., 2015; Kichaev et al., 2016), and some works have even used full Bayesian MCMC (Zhao et al., 2019). BLiP can wrap on top of any of these methods, and, we hope, enhance their power to make meaningful scientific discoveries.

5 Application to astronomical point source detection

In this section, we apply BLiP to the problem of astronomical point source detection. To formalize this problem, suppose that $\mathcal{L} = [0, 1]^2$ corresponds to some patch of the night sky, and we want to detect point sources (e.g., stars) in \mathcal{L} . We assume we observe data $\{D_{i,j}^{(b)}\}_{i,j \in [d], b \in [B]}$. Here, b denotes the *band* of an image (i.e., color), and $D_{i,j}^{(b)}$ (informally) counts the photons observed by a telescope in the (i, j) th pixel of an image, i.e., the subregion $[\frac{i-1}{d}, \frac{i}{d}] \times [\frac{j-1}{d}, \frac{j}{d}]$. If a source exists at $(x, y) \in \mathcal{L}$, we expect the pixels close to (x, y) to be brighter, allowing us to detect the star. However, the exact location of the source is always uncertain, both because of the discrete nature of the pixels and also because there are often multiple sources in the same pixel—indeed, due to the increasing sensitivity of telescopes, Feder et al. (2020) observed that astronomical images

⁴Note that this is not a replication analysis, since both analyses use the same dataset and model.

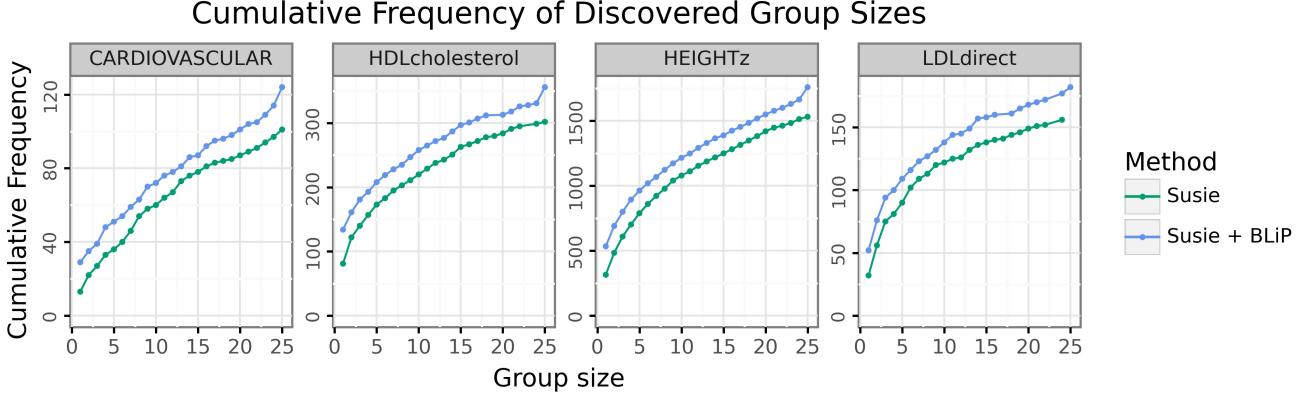


Figure 6: This figure plots the cumulative frequency of the discovered group sizes. To be precise, the point with x -value k on the green curve (resp. blue curve) counts the number of groups of size k or less discovered by SuSiE (resp. SuSiE + BLiP). This figure shows that SuSiE + BLiP makes more discoveries at finer resolutions than SuSiE alone, i.e., BLiP detects more causal variants and localizes them more precisely.

have become increasingly crowded over time. This phenomenon makes it difficult to discover as many point sources as possible, and it also causes ambiguity when deciding whether a bright pixel represents a single point source or two nearby point sources. The latter problem is particularly important since many downstream applications will yield misleading results unless sources are properly separated (Portillo et al., 2017), such as analyses which attempt to classify point sources based on, e.g., their intensity. These challenges have motivated the development of Bayesian models which can simultaneously capture uncertainty about both the number and locations of sources, including approaches which can scale to catalogue massive astronomical surveys (Regier et al., 2019; Liu et al., 2021b). However, extracting interpretable results from these models can be challenging (Feder et al., 2020). We now show that BLiP can solve exactly these types of problems.

In this section, we apply BLiP to a 100×100 pixel sub-image of the Messier 2 (M2) star cluster, taken by the Sloan Digital Sky Survey (SDSS). Several previous works (Portillo et al., 2017; Feder et al., 2020; Liu et al., 2021b) used this sub-image to validate point source detection methods, and indeed, our analysis directly builds on that of Liu et al. (2021b), who introduced StarNet, a Bayesian model for point source detection. The M2 region is a good testing ground because it was also imaged by the Hubble Space Telescope (HST) during the ACS Globular Cluster Survey (Sarajedini et al., 2007). Since the HST has a much greater resolution than the Sloan telescope, previous works used the HST detections as the ground truth. In particular, Liu et al. (2021b) compared StarNet’s MAP estimates of source locations to the HST locations. To account for both measurement differences between telescopes and also the statistical uncertainty in the positions of sources, each of these previous works classified an estimated source position as a “true source” if it was within 0.5 pixels of an HST source. Equivalently, let \mathcal{M} be the set of MAP estimates: for $(x, y) \in \mathcal{M}$, we can think of StarNet as detecting a signal in $C_\epsilon(x, y)$, where $C_\epsilon(x, y)$ is the square of radius ϵ centered at (x, y) . This procedure chooses ϵ before seeing the data, so it performs inference at a fixed resolution: we refer to it as the “MAP baseline at resolution ϵ ,” where Liu et al. (2021b) picked $\epsilon = \frac{0.5}{100}$. Since BLiP automatically accounts for statistical uncertainty in the location of sources, we use a more stringent standard: to account only for measurement differences, we classify each detected region from BLiP as a true detection if an HST source lies within 0.01 pixels of the region. (For fair comparison, we also give the baseline an additional 0.01 pixels of slack.) Despite this more stringent standard, we will see later that BLiP outperforms the MAP baseline for any ϵ .

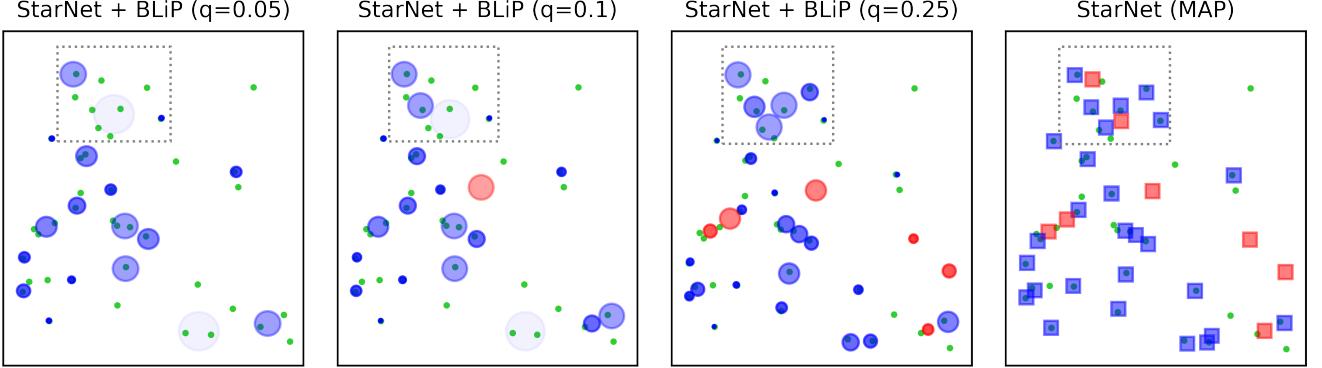


Figure 7: This figure shows the detections made by BLiP for a 20×20 pixel sub-image of the 100×100 M2 image using the “inverse radius” weight function. True and false positives are shown in blue and red, respectively, and the opacity of each outputted region is proportional to the inverse radius. Sources detected by the HST are treated as the ground truth and shown in green. Note that BLiP outputs circles of varying sizes, which reflects the varying degrees of uncertainty we have about the position of different sources. Furthermore, the right-most plot shows the output from Liu et al. (2021b), using their criterion for whether these are false positives. These plots show that at low nominal levels, BLiP successfully filters out false positives from the set of MAP estimators. Furthermore, as the nominal level increases, BLiP can better separate nearby sources while maintaining FDR control, as shown in the region highlighted by the dashed black box.

In particular, we propose applying BLiP directly on top of posterior samples from Liu et al. (2021b)’s pretrained StarNet model. Although in principle BLiP can output regions of arbitrary shapes, for simplicity we use a set of roughly 280 million overlapping circles of various radii as candidate groups. We applied BLiP with two weight functions to proxy two possible astronomical motivations. First, we used an “inverse radius” weight function, which assigns the weight $\frac{1}{\epsilon}$ to any candidate group with radius ϵ . This reflects a preference for discovering smaller regions. However, some practitioners may care less about perfectly localizing signals as long as the signals are correctly *separated*. In our second analysis, BLiP outputs both a region $G \subset \mathcal{L}$ and a contiguous credible interval $J \subset \mathbb{N}$ for the number of point sources in the region. For example, outputting $J = \{1, 2\}$ asserts that there are either one or two point sources in G . We use the “separation-based” weight function $w(G, J) = \frac{1}{|J|}$, which gives BLiP the incentive to precisely separate point sources whenever possible, since, e.g., discovering two regions containing exactly one star is four times as valuable as discovering a single region containing either one or two stars. For this second analysis, we count (G, J) as a true discovery if and only if the number of Hubble stars within 0.01 pixels of the region G is an element of J . Of course, there are myriad other potential ways to apply BLiP, including using different weight functions or even applying BLiP in higher dimensions to simultaneously capture uncertainty about the flux of discovered sources. That said, we hope our two analyses demonstrate the flexibility of BLiP, since it shows that BLiP can be directly applied to pretrained models and also can accommodate varying scientific objectives. See Appendix G for further methodological details, and see Section 6 for a discussion of more extensions of BLiP.

In Figure 7, we plot the detections from BLiP at various nominal levels against the true HST sources and the MAP estimates from Liu et al. (2021b). The figure shows that BLiP is able to adaptively quantify uncertainty in the locations of the sources and simultaneously filter out clear false positives from the set of MAP estimates. To formally test error control, we run BLiP to control the FDR at level q for various values of q and compare its discoveries to the HST sources. As depicted

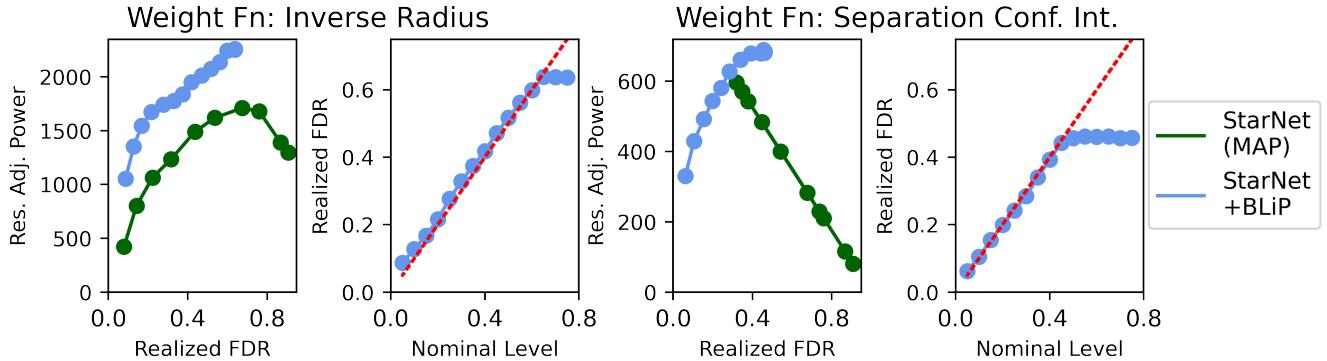


Figure 8: This figure shows the power and FDR of StarNet + BLiP at various nominal levels compared to the power of the fixed-resolution baseline based on MAP estimation. It shows that StarNet + BLiP successfully controls the FDR and simultaneously outperforms the baseline. The different dots on the green curves correspond to varying the fixed-resolution ϵ . Note that for the right-hand power plot, the StarNet (MAP) points are co-linear because according to the weight-function, each region discovered by the baseline gets the (maximal) weight of one, no matter the resolution. As a result, the realized power is a linear function of the realized FDR. Furthermore, the green line does not extend any further (the power of the baseline never goes higher for any fixed resolution).

in Figure 8, BLiP controls the FDR close to the exact nominal level for both weight functions, although for large q it becomes conservative due to the adaptive preprocessing steps (Appendix B.3). We interpret the remarkable alignment between realized and nominal FDR as a testament both to BLiP’s capacity to control the FDR and to the pretrained StarNet model, which appears to be very well calibrated.

Lastly, to analyze the power of BLiP, we compare StarNet + BLiP to the MAP baseline from Liu et al. (2021b), applied at various fixed resolutions ϵ . Figure 8 shows that BLiP achieves a much better power-FDR trade-off than the baseline. For example, when controlling the FDR at $q = 0.05$, StarNet + BLiP is more than twice as powerful than the MAP-based approach using the inverse-radius weight function. (Note that in addition to providing higher power for any realized FDR, BLiP’s output also comes with an FDR guarantee, unlike the MAP approach.) Similarly, for both weight functions, StarNet + BLiP run with $q = 0.30$ is more powerful than the baseline for *any* fixed resolution, even when the realized FDR for the baseline is as high as 90%. This is the benefit of resolution-adaptivity: whereas the baseline must discover each source at some fixed resolution ϵ , BLiP can use the data to account for the fact that we may be more uncertain of the location of some sources than others. And for the same reason, BLiP is better able to quantify uncertainty about whether two potential point sources have been properly separated for the “separation-based” weight function. In contrast, StarNet (MAP) always has a realized FDR above 30%. Overall, BLiP simultaneously controls the FDR and achieves much higher power than the MAP-based approach for both weight functions. Taken together, these results suggest that BLiP holds promise as a method for uncertainty quantification in point source detection. All code and data are publicly available at <https://github.com/amspector100/DeblendingStarfields>. Note that the whole analysis ran in under ten minutes on a laptop and required 11 gigabytes of memory.

6 Discussion

This paper introduces BLiP, a powerful method for performing resolution-adaptive signal detection. Our simulations and two applications show that BLiP is computationally efficient and very powerful while simultaneously providing provable error control. Another notable benefit of BLiP, however, is that it is a highly flexible method, and it also solves a very general problem. As a result, there are many possible extensions of BLiP which may be of interest.

First, in this paper, we use BLiP to maximize expected power while controlling the FDR or other error rates, but in principle, BLiP-like algorithms could optimize for different objective functions under different constraints. For example, it is fairly straightforward to tweak BLiP to maximize the harmonic mean of the expected true positive rate and expected power, also known as the F1 score (see Appendix B.5 for details). This objective function balances maximizing power against minimizing FDR, which may be useful in settings where strict FDR control is not necessary. Indeed, some previous work in astronomical point source detection attempted to maximize the F1 score (Liu et al., 2021b), suggesting this may be of interest. Similarly, as mentioned in Section 2.2, one can add or modify the linear constraints in the LP formulation of BLiP to control different notions of the error rate, such as controlling a weighted FDR (Benjamini and Hochberg, 1997; Bogomolov et al., 2020) or achieving multilayer FDR control (Barber and Ramdas, 2017; Katsevich and Sabatti, 2019).

Second, in Sections 3-4, we apply BLiP on top of Bayesian models with fairly uninformative priors. However, using more sophisticated models or algorithms could improve BLiP’s performance. To start, recent advances in Bayesian inference for spike-and-slab models or change point detection could boost power and computational efficiency (Cappello et al., 2021; Biswas et al., 2022), and generic advances in MCMC could ensure that the inputs to BLiP are at least unbiased (Jacob et al., 2020). More specifically, recent works in fine-mapping have developed informative priors based on, e.g., functional annotation data (Weissbrod et al., 2019; Weeks et al., 2020; Trippe et al., 2021). Indeed, BLiP can apply directly on top of nearly any Bayesian fine-mapping model or algorithm, e.g., those developed in Hormozdiari et al. (2014); Benner et al. (2015); Kichaev et al. (2016). Of course, in some cases, it might be worthwhile to develop sensible heuristics to optimize performance. For example, methods like FINEMAP (Benner et al., 2015) partially enumerate over configurations of causal variants to compute approximate PIPs, and it might make sense to account for this approximation before applying BLiP. Such questions may point to promising directions for future research.

Lastly, we hope BLiP may prove useful in many domains not considered in this paper. For example, BLiP might apply in problems with spatially structured signals, even if the signals are not isolated. One such application is the analysis of fMRI data, where researchers often study questions like “in which regions of the brain did an intervention change brain activity?” In this case, for some regions G , it might make sense to define the PIP p_G as the posterior probability that *every* point in G is a signal point (Maullin-Sapey et al., 2022). However, having made this adjustment to $\{p_G\}$, one could otherwise run BLiP exactly as written to perform resolution-adaptive signal detection. More generally, it can be challenging to extract interpretable results from complex posterior distributions in high-dimensional problems, such as in the Bayesian analysis of phylogenetic trees (Huelsenbeck et al., 2001; Kim et al., 2020) or the Bayesian estimation of graphical models (Li et al., 2018). By carefully defining “signals” of interest, could BLiP help extract intelligible results in these settings? We did not consider this question in this paper, but perhaps future work will address it.

7 URLs

For convenience, below are a collection of links which include all the code and data necessary to replicate our analyses.

`pyblip`, a Python package implementing BLiP: <https://github.com/amspector100/pyblip>.

`blipr`, an R package implementing BLiP: <https://github.com/amspector100/blipr>.

The code for the paper’s simulations: https://github.com/amspector100/blip_sims.

The code for the astronomical application: <https://github.com/amspector100/DeblendingStarfields>.

The code for the fine-mapping analysis: https://github.com/amspector100/ukbb_blip.

Open-access data for the fine-mapping analysis: https://data.broadinstitute.org/alkesgroup/UKBB_LD/ and https://data.broadinstitute.org/alkesgroup/polyfun_results/.

Acknowledgements

The authors would like to thank Niloy Biswas, Jun Liu, Minsuk Shin, and Benjamin Spector for suggestions related to Bayesian regression; Vinay Kashyap and Douglas Finkbeiner for their insight on the astronomical application; and Hilary Finucane and Luke O’Connor for valuable comments on the fine-mapping application. A.S. was partially supported by the Two Sigma Graduate Fellowship Fund and a Graduate Research Fellowship from the National Science Foundation. L.J. was partially supported by the William F. Milton Fund and a CAREER grant from the National Science Foundation (Grant #DMS2045981).

References

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.
- Barber, R. F. and Ramdas, A. (2017). The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4):1247–1268.
- Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press.
- Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24:407–418.
- Benner, C., Spencer, C. C., Ripatti, S., and Pirinen, M. (2015). Finemap: Efficient variable selection using summary data from genome-wide association studies. *bioRxiv*.
- Biswas, N., Mackey, L., and Meng, X.-L. (2022). Scalable spike-and-slab.

- Bogomolov, M., Peterson, C. B., Benjamini, Y., and Sabatti, C. (2020). Hypotheses on a tree: new error rates and testing strategies. *Biometrika*, 108(3):575–590.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, USA.
- Brinkhoff, T., Kriegel, H.-P., and Seeger, B. (1993). Efficient processing of spatial joins using r-trees. *SIGMOD Rec.*, 22(2):237–246.
- Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2012). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference* 2013, pages 1835–1858.
- Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577.
- Cappello, L., Padilla, O. H. M., and Palacios, J. A. (2021). Scalable bayesian change point detection with spike and slab priors.
- Carbonetto, P. and Stephens, M. (2012). Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis*, 7(1):73 – 108.
- Erdős, P., Goodman, A. W., and Pósa, L. (1966). The representation of a graph by set intersections. *Canadian Journal of Mathematics*, 18:106–112.
- Fang, X., Li, J., and Siegmund, D. (2016). Segmentation and estimation of change-point models: False positive control and confidence regions.
- Fang, X. and Siegmund, D. (2020). Detection and estimation of local signals.
- Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shores, N., Whitton, H., Ryan, R. J., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, J. C., Patsopoulos, N. A., Jager, P. L. D., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., and Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518:337–343.
- Feder, R. M., Portillo, S. K. N., Daylan, T., and Finkbeiner, D. (2020). Multiband probabilistic cataloging: A joint fitting approach to point-source detection and deblending. *The Astronomical Journal*, 159(4):163.
- Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.
- Fryzlewicz, P. (2020). Narrowest significance pursuit: inference for multiple change-points in linear models.
- George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373.
- Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities.

- Goeman, J. J. and Solari, A. (2012). The sequential rejection principle of familywise error control. *The Annals of Statistics*.
- Gramm, J., Guo, J., Hüffner, F., and Niedermeier, R. (2006). Data reduction, exact, and heuristic algorithms for clique cover. In *Proceedings of the Meeting on Algorithm Engineering & Experiments*, page 86–94, USA. Society for Industrial and Applied Mathematics.
- Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780 – 1815.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
- Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics*, 198(2):497–508.
- Huangfu, Q. and Hall, J. A. J. (2015). Parallelizing the dual revised simplex method.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.
- Jacob, P. E., O’Leary, J., and Atchadé, Y. F. (2020). Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.
- Jünger, M., Liebling, T., Naddef, D., Nemhauser, G., Pulleyblank, W., Reinelt, G., Rinaldi, G., and Wosley, L. (2010). *50 Years of Integer Programming 1958-2008*. Springer.
- Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1 – 33.
- Katsevich, E., Sabatti, C., and Bogomolov, M. (2021). Filtering the rejection set while preserving false discovery rate control. *Journal of the American Statistical Association*, 0(0):1–12.
- Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindström, S., Kraft, P., and Pasaniuc, B. (2016). Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2):248–255.
- Kim, J., Rosenberg, N. A., and Palacios, J. A. (2020). Distance metrics for ranked evolutionary trees. *Proceedings of the National Academy of Sciences*, 117(46):28876–28886.
- Kou, L. T., Stockmeyer, L. J., and Wong, C. K. (1978). Covering edges by cliques with regard to keyword conflicts and intersection graphs. *Commun. ACM*, 21(2):135–139.
- Lareau, C. A., Ulirsch, J. C., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., Satpathy, A. T., Salem, R., Hirschhorn, J. N., Finucane, H. K., Aryee, M. J., Buenrostro, J. D., and Sankaran, V. G. (2018). Interrogation of human hematopoiesis at single-cell and single-variant resolution. *bioRxiv*.
- Lee, Y., Luca, F., Pique-Regi, R., and Wen, X. (2018). Bayesian multi-snp genetic association analysis: Control of fdr and use of summary statistics. *bioRxiv*.

- Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918 – 959.
- Li, Z. R., McCormick, T. H., and Clark, S. J. (2018). Bayesian joint spike-and-slab graphical lasso.
- Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2021+a). Fast and powerful conditional randomization testing via distillation. *Biometrika*. To Appear.
- Liu, R., McAuliffe, J. D., and Regier, J. (2021b). Variational inference for deblending crowded starfields.
- Maullin-Sapey, T., Schwartzman, A., and Nichols, T. E. (2022). Spatial confidence regions for combinations of excursion sets in image analysis.
- Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278.
- Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.
- O'Connor, L. J. (2020). The distribution of common-variant effect sizes. *bioRxiv*.
- Orlin, J. (1977). Contentment in graph theory: Covering graphs with cliques. *Indagationes Mathematicae (Proceedings)*, 80(5):406–424.
- Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.
- Portillo, S. K. N., Lee, B. C. G., Daylan, T., and Finkbeiner, D. P. (2017). Improved point-source detection in crowded fields using probabilistic cataloging. *The Astronomical Journal*, 154(4):132.
- Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., McAuliffe, J., Thomas, R., and Prabhat (2019). Cataloging the visible universe through bayesian inference in julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.
- Renaux, C., Buzdugan, L., Kalisch, M., and Bühlmann, P. (2018). Hierarchical inference for genome-wide association studies: a view on methodology with software. *Computational Statistics*.
- Sarajedini, A., Bedin, L. R., Chaboyer, B., Dotter, A., Siegel, M., Anderson, J., Aparicio, A., King, I., Majewski, S., Marín-Franch, A., Piotto, G., Reid, I. N., and Rosenberg, A. (2007). The ACS Survey of Galactic Globular Clusters. I. Overview and Clusters without Previous Hubble Space Telescope Photometry. *The Astronomical Journal*, 133(4):1658–1672.
- Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2019). Multi-resolution localization of causal variants across the genome. *bioRxiv*.
- Shin, M. and Liu, J. S. (2021). Neuronized priors for bayesian sparse linear regression. *Journal of the American Statistical Association*, 0(0):1–16.
- Trippe, B. L., Finucane, H. K., and Broderick, T. (2021). For high-dimensional hierarchical models, consider exchangeability of effects across covariates instead of across datasets.

- Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genetics*, 17(9):1–11.
- Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300.
- Wang, X. and Emerson, J. W. (2015). Bayesian change point analysis of linear models on graphs.
- Weeks, E. M., Ulirsch, J. C., Cheng, N. Y., Trippe, B. L., Fine, R. S., Miao, J., Patwardhan, T. A., Kanai, M., Nasser, J., Fulco, C. P., Tashman, K. C., Aguet, F., Li, T., Ordovas-Montanes, J., Smillie, C. S., Biton, M., Shalek, A. K., Ananthakrishnan, A. N., Xavier, R. J., Regev, A., Gupta, R. M., Lage, K., Ardlie, K. G., Hirschhorn, J. N., Lander, E. S., Engreitz, J. M., and Finucane, H. K. (2020). Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv*.
- Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A., van de Geijn, B., Reshef, Y., Márquez-Luna, C., O'Connor, L., Pirinen, M., Finucane, H. K., and Price, A. L. (2019). Functionally-informed fine-mapping and polygenic localization of complex trait heritability. *bioRxiv*.
- Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316.
- Zhao, Y., Zhu, H., Lu, Z., Knickmeyer, R. C., and Zou, F. (2019). Structured genome-wide association studies with bayesian hierarchical variable selection. *Genetics*, pages 397–415.
- Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2021). Fine-mapping from summary data with the “sum of single effects” model. *bioRxiv*.

A A toy example motivating resolution-adaptivity

In this section, we give a toy example of a variable selection problem where the optimal way to group the variables depends on the unknown relationship between X and Y .

Example 4. Suppose we observe regression data $\mathbf{X} \in \mathbb{R}^{n \times 2}$, $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$ and $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{I}_n)$. X_ℓ is a signal variable if $\beta_\ell \neq 0$, for $\ell \in \{1, 2\}$. Here, we could either try to detect individual signals, or we could try to detect whether *either* X_1 or X_2 is a signal variable. A non-adaptive method must choose one option *before* observing \mathbf{y} . Which is the correct choice?

As we will prove in a moment, the answer to this question depends on the unknown coefficients $\beta \in \mathbb{R}^2$. Suppose in truth that $\beta_1 \neq 0, \beta_2 = 0$, although the analyst does not know this. In Figure 9, we plot the power of the standard F-tests corresponding to the two previous resolutions, with appropriate multiplicity corrections. As shown by Figure 9, detecting X_1 individually requires a much larger signal size than detecting a signal in (X_1, X_2) , simply because X_1 and X_2 are correlated. Indeed, for $\beta_1 \approx 5$, we have nearly a 100% chance of detecting a signal in (X_1, X_2) but less than

a 5% chance of detecting X_1 . However, when β_1 is large, we can detect X_1 individually with high probability, which yields more specific information. For this reason, the resolution-adjusted power of the finer-resolution method is higher for large β_1 and lower for small β_1 in Figure 9. This means that every non-adaptive method will make the wrong choice in some scenarios, whereas adaptive methods like BLiP can use the data to choose a resolution.

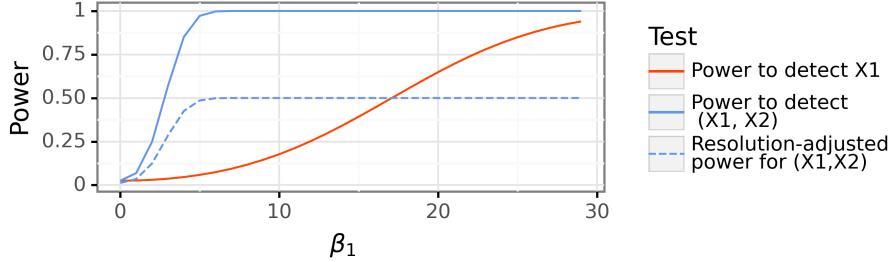


Figure 9: Power of F-tests described in Example 4 while varying β_1 . This figure shows that the correct choice of resolution in Example 4 depends on the unknown coefficient β_1 . Note resolution-adjusted power is formally defined in Section 2.1: the main idea is that discoveries at coarser resolutions count as fewer discoveries since they yield less specific information.

To derive the power curves in Figure 9, we review classical results for the power of the F-tests testing (1) whether $\beta_1 = 0$ and (2) whether $(\beta_1, \beta_2) = 0$. As notation, let $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ be the projection matrix onto the column space of \mathbf{X} and let H_2 be the projection matrix onto the column space of \mathbf{X}_2 . Let $\chi_{k,\mu}^2$ denote a non-central χ^2 distribution with non-centrality parameter μ . Then Cochran's theorem (see Agresti (2015) Chapter 3 for a review) tells us that for $n \geq 3$,

$$\mathbf{y}^T (H - H_2) \mathbf{y} \sim \chi_{1,\mu_1}^2, \quad \mathbf{y}^T H \mathbf{y} \sim \chi_{2,\mu_2}^2, \quad \mathbf{y}^T (\mathbf{I}_n - H) \mathbf{y} \sim \chi_{n-2}^2,$$

where in our setting each of the first two quantities is marginally independent of the third quantity, and furthermore,

$$\begin{aligned} \mu_1 &= \beta^T \mathbf{X}^T (H - H_2) \mathbf{X} \beta = (1 - \rho)^2 \beta_1^2 \\ \mu_2 &= \beta^T \mathbf{X}^T H \mathbf{X} \beta = \beta^T \mathbf{X}^T \mathbf{X} \beta = \beta_1^2. \end{aligned}$$

Note that in the last step of the previous two equations, we use the fact that $\beta_2 = 0$ by assumption as well as the form of $\mathbf{X}^T \mathbf{X}$. As a result, we can calculate the distributions of the two standard F-statistics:

$$T_1 = \frac{\mathbf{y}^T (H - H_2) \mathbf{y}}{\mathbf{y}^T (\mathbf{I}_n - H) \mathbf{y} / (n - 2)} \sim F_{1,n-2,(1-\rho^2)\beta_1^2},$$

$$T_2 = \frac{\mathbf{y}^T H \mathbf{y} / 2}{\mathbf{y}^T (\mathbf{I}_n - H) \mathbf{y} / (n - 2)} \sim F_{2,n-2,\beta_1^2},$$

where $F_{n_1, n_2, \mu}$ denotes a noncentral F distribution with n_1, n_2 degrees of freedom and non-centrality parameter μ . At a significance level of 5%, the standard F-tests for the two resolutions reject when $T_1 \geq t_1$ and $T_2 \geq t_2$, respectively, where t_1, t_2 , are the 0.95th quantiles of the $F_{1,n-2}, F_{2,n-2}$ distributions. For the finer resolution, the analyst without oracle knowledge would have to test two hypotheses (one for each variable): thus we apply a Bonferroni correction and let t_1 denote the 0.975th quantile of $F_{1,n-2}$. Then, the power of these tests can be calculated as the CDF of the relevant non-central F-distributions evaluated at t_1, t_2 .

B Further details on BLiP

B.1 Proof of Proposition 2.1

In this section, we prove Proposition 2.1.

Proposition 2.1. *The solution to the resolution-adaptive signal detection problem in Definition 1.1 is the same as the solution to the following mixed-integer LP:*

$$\max_{\{x_G\}_{G \in \mathcal{G}}: x_G \in \{0,1\}} \sum_{G \in \mathcal{G}} p_G w(G) x_G \quad (2.4)$$

$$\text{s.t.} \quad \sum_{G \in \mathcal{G}} (1 - p_G - q) x_G \leq 0, \quad (2.5)$$

$$\sum_{G \in \mathcal{G}: \ell \in G} x_G \leq 1 \quad \forall \ell \in \mathcal{L}, \quad (2.6)$$

$$x_G \in \{0, 1\} \quad \forall G \in \mathcal{G}. \quad (2.7)$$

Proof. Note that throughout this proof, we use the convention that $\frac{0}{0} = 0$. Recall that Problem 1.1 maximizes expected power subject to the constraint that the selected groups G_1, \dots, G_R are disjoint and control the FDR at level q . As argued in Section 2.2, the first line (2.4) is equivalent to maximizing expected power. Additionally, (2.6) is equivalent to the disjointness constraint. Indeed, consider two candidate groups G, G' which are not disjoint. This means G, G' contain a common location $\ell \in \mathcal{L}$, and therefore (2.6) requires that $x_G + x_{G'} \leq 1$. Thus, (2.6) permits at most one of G, G' to be discovered. Conversely, if G_1, \dots, G_R are all disjoint, $x_{G_1} = 1, \dots, x_{G_R} = 1$ does not violate (2.6) since no two groups in G_1, \dots, G_R share a location ℓ .

It remains to show that (2.5) is equivalent to controlling the FDR. For the first direction, suppose G_1, \dots, G_R control the FDR at level q . By Equation (2.3), this implies $\frac{\sum_{G \in \mathcal{G}} (1 - p_G) x_G}{\sum_{G \in \mathcal{G}} x_G} \leq q$, and multiplying by $\sum_{G \in \mathcal{G}} x_G$ on both sides implies (2.5) holds. For the other direction, note that if $\{x_G\}_{G \in \mathcal{G}}$ is feasible, Equation (2.5) implies that $\sum_{G \in \mathcal{G}} (1 - p_G) x_G \leq q \sum_{G \in \mathcal{G}} x_G$. Dividing both sides of this inequality by $\sum_{G \in \mathcal{G}} x_G$ implies that we control the FDR: the only subtlety is that $\sum_{G \in \mathcal{G}} x_G$ could be zero. However, $\sum_{G \in \mathcal{G}} (1 - p_G) x_G = 0$ whenever $\sum_{G \in \mathcal{G}} x_G = 0$. Since we use the convention that $0/0 = 0$, we can divide by $\sum_{G \in \mathcal{G}} x_G$ in all cases to yield the result. \square

B.2 Integer solutions to the relaxed LP from Section 2

In this section, we give some intuition as to why the relaxed LPs described in Section 2 tend to return (largely) integer solutions. To see this, we briefly review a canonical relaxed knapsack problem. Note that for simplicity, in this section we consider the version of BLiP which controls the PFER, but this intuition extends to other error rates as well. At the end of this section, we also present the corresponding plot of Figure 1 for the FWER.

Suppose we have m liquids with weights $w = (w_1, \dots, w_m) \in \mathbb{R}_{\geq 0}^m$ and volumes $v = (v_1, \dots, v_m) \in \mathbb{R}_{\geq 0}^m$. The “knapsack problem” is to select a proportion of each liquid $x_1, \dots, x_m \in [0, 1]$ which maximizes the total weight $w^T x$ subject to the volume constraint $v^T x \leq v^*$ for $v^* \geq 0$ —here, the volume constraint corresponds to ensuring the liquids fit in a “knapsack.” This is a simple LP, and it is well known that optimal solution has at most one non-integer value, in particular because the

greedy solution is optimal. In other words, the optimal solution is to order the liquids in decreasing order of their “density” w_j/v_j and select all of the first k liquids where k is the largest number such that $\sum_{j=1}^k v_j \leq v^*$. Then, we select some fractional proportion of the $k+1$ th liquid to use up the rest of our volume budget. In other words, the solution is of the form $x_1, \dots, x_k = 1$, $x_{k+1} \in [0, 1]$, $x_{k+2}, \dots, x_\ell = 0$ assuming that $\frac{w_1}{v_1} \geq \frac{w_2}{v_2} \geq \dots \geq \frac{w_m}{v_m}$.

When controlling the PFER, the relaxed version of the resolution-adaptive signal detection problem is simply a knapsack problem with extra linear constraints that prevent some variables (or “liquids”) from being selected simultaneously. In many cases, despite the extra constraints, the optimal solution to the relaxed LP in BLiP consists of only one non-integer value, just like in a vanilla knapsack problem. In this case, let G^* be the unique group such that $x_{G^*} \notin \{0, 1\}$. Given such a solution, to exactly achieve PFER control and maximize expected power, one could simply detect signals in all the groups G such that $x_G = 1$ and detect a signal in group G^* with probability x_{G^*} . Alternatively, to avoid a randomized procedure, one could only discover $\{G : x_G = 1\}$ at the cost of making up to one fewer discovery and being slightly conservative.

Of course, the extra constraints do often affect the solution to the LP. To build intuition for this, consider a simple example where there are only two groups, $G = \{1\}$ and $G' = \{1, 2\}$, with $w(G) = 1$, $w(G') = 0.5$ and PIPs $p_G = 0.8$, $p_{G'} = 0.9$. Thinking in terms of a knapsack problem, G and G' have “weights” $p_G w(G) = 0.9$, $p_{G'} w(G') = 0.45$ and “volumes” $1 - p_G = 0.2$, $1 - p_{G'} = 0.1$, respectively. Thus, the relaxed signal detection problem is of the form

$$\max \quad 0.8x_G + 0.45x_{G'} \tag{B.1}$$

$$\text{s.t.} \quad 0.2x_G + 0.1x_{G'} \leq q, \tag{B.2}$$

$$x_G + x_{G'} \leq 1, \tag{B.3}$$

$$x_G, x_{G'} \in [0, 1].$$

As usual, the objective corresponds to maximizing power, (B.2) corresponds to controlling the PFER, and (B.3) is the disjointness constraint. Without the disjointness constraint (B.3), the optimal (greedy) solution would be to first make $x_{G'}$ as large as possible under the PFER constraint (B.2). Then, if it is possible to set $x_{G'} = 1$ and there is additional slack in the PFER constraint, we would make x_G as large as possible while controlling the PFER. However, this behavior changes when we once again consider the disjointness constraint. For example, when $q = 0.15$, after setting $x_{G'} = 1$, there is extra slack in the PFER budget, but one cannot naively set $x_{G'} = 1$, $x_G > 0$ without violating the disjointness constraint (B.3). So instead, the optimal solution is to *simultaneously* make $x_{G'}$ smaller and x_G larger until there is no slack in the PFER constraint. In this case, the optimal solution is to set $x_G = 0.5$ and $x_{G'} = 0.5$. Intuitively, what is happening is that we would like to make discoveries at a finer resolution and discover a signal in the group $\{1\}$: however, we cannot do that and control the PFER at level $q = 0.15$. Instead, the most powerful procedure is a randomized one, which discovers $\{1\}$ or $\{1, 2\}$ each with probability 0.5. And the most powerful non-randomized procedure is the slightly conservative one, which discovers $\{1, 2\}$ with probability 1. In the general formulation of the LP, we say that G, G' are a *randomized pair* if $x_G + x_{G'} = 1$ with $x_G, x_{G'}$ non-integers. Notably, this same intuition extends to each of the other error rates, where to achieve a Type I error of exactly q in expectation, we would have to randomly choose between two or more discovery sets. Of course, BLiP yields a deterministic solution, so the output of BLiP will usually be (very slightly) conservative for this reason.

The main takeaway from all this is as follows: in general, we should expect the solution to the relaxed LP to be nearly all integers, with a few randomized pairs and a few other non-integer values

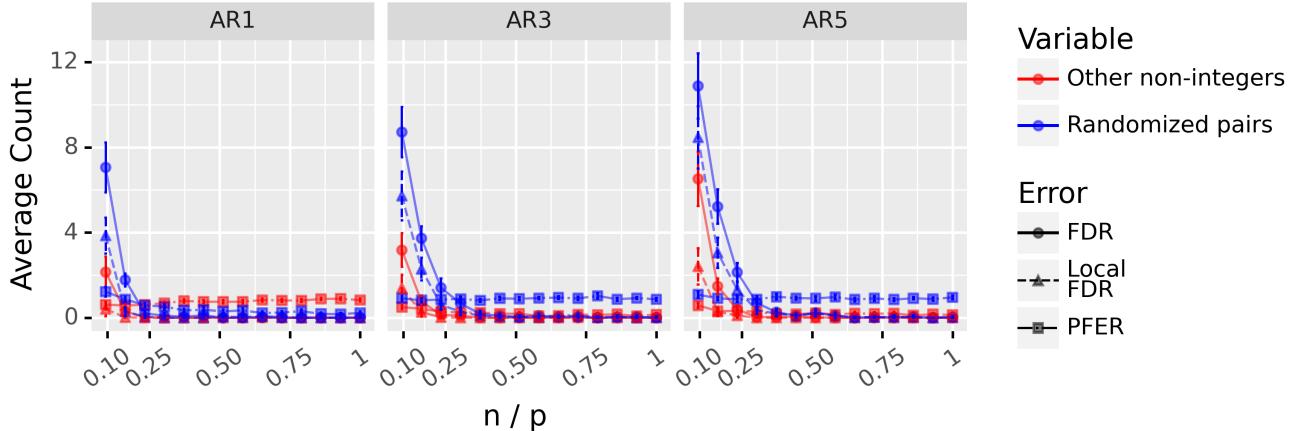


Figure 10: In the same setting as Figure 1, this figure shows the average number of randomized pairs and non-integer solutions to the relaxed LP (2.4)-(2.6), and the corresponding relaxed LPs for other error rates. The total number of candidate groups exceeds 50,000 for all methods, so this plot shows that only a tiny proportion of variables in the LP have nonzero values. See Appendix E.2 for the precise simulation setting.

to help us achieve a PFER equal to q . Indeed, Figure 10 confirms that in the same setting as Figure 1, solving the relaxed LP yields integer solutions for all but a few decision variables out of many thousands. Notably, as n/p gets smaller, the number of non-integer values increases, which corresponds to the fact that in low-power settings, more randomization may be necessary to achieve exact optimal power.

Lastly, we now present Figure 11, the corresponding plot of Figure 1 for the FWER. Recall that to control the FWER, we use a more heuristic method which controls the PFER at the largest nominal level q^* such that the FWER is controlled. In particular, $q^* \geq q$ because the PFER is a stricter notion of error control than the FWER, although usually q^* is only slightly larger than q . Figure 1 shows that this binary search yields very slightly higher power than choosing $q^* = q$. Since the FWER constraint cannot be represented as a linear constraint on the optimization variables $\{x_G^*\}_{G \in \mathcal{G}}$, unlike in Figure 1, the LP bound in Figure 11 is not an upper bound on the power of any method which controls the FWER. Instead, it is an upper bound on any method that controls the PFER at level q^* , with q^* identified via binary search. This is not a principled upper bound, but at least it shows that our relaxation of the integer LP (2.4)-(2.7) yields a nearly optimal solution to the integer LP. And of course, the method still provably controls the FWER—we just cannot give any provable guarantees about its optimality.

B.3 Adaptive preprocessing

In this section, we detail the adaptive preprocessing steps discussed in Section 2.2.

First, after we have computed PIPs, we can *pre-filter the candidate groups* \mathcal{G} before adding them into the linear program in BLiP. Indeed, whenever we seek to control the FWER, local FDR, or PFER below some level q , we cannot detect a signal in a group $G \in \mathcal{G}$ unless $p_G \geq 1 - q$: otherwise we would definitionally violate error control. Thus, we can throw out all groups G where $p_G < 1 - q$. We also suggest performing a similar procedure when controlling the FDR, where we throw out all groups such that $p_G < \kappa$ for some $\kappa \leq 1 - q$. Although this does not provably preserve optimality, in practice we do not expect this to reduce power for an appropriate κ . For example, suppose $q = 0.05$

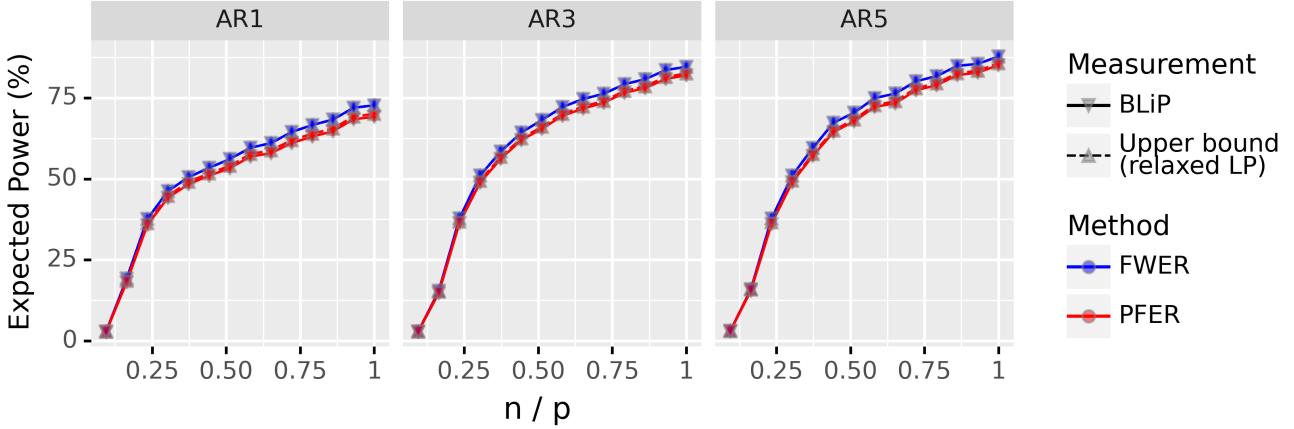


Figure 11: This plot is identical to Figure 1 except we apply BLiP to control the FWER. Note that for the FWER, the upper bound from the relaxed LP is not a provable upper bound on the power of any method controlling the FWER at level q . Instead, as discussed in Appendix B.2, it upper bounds the power of any method controlling the PFER at some level q^* , where $q^* \geq q$ is chosen to be as large as possible while still controlling the FWER at level q . The plot also shows that choosing $q^* \geq q$ yields a slight boost in power for the FWER method compared to the PFER method. See Appendix E.2 for simulation details.

and $p_G \leq 0.1$, which means that group G has at most a 10% chance of containing a signal. It is unlikely that a FDR controlling procedure would select G . Even if selecting G is optimal for power, it is arguably desirable to *not* select G anyway, since G would be “free-riding” on the other selected groups with high PIPs. In practice, we suggest setting $\kappa = 0.5$.

Second, when \mathcal{L} is discrete, we suggest *pre-filtering the locations*. Suppose for example that for some location $\ell \in \mathcal{L}$, the posterior probability that there is a signal exactly at ℓ is extremely low, so $p_{\{\ell\}} \approx 0$. In this case, if $G_1 \subset \mathcal{L}$ contains ℓ , we could just as easily discover $G_2 = G_1 \setminus \{\ell\}$, which yields more precise information. Formally, a union bound yields that $p_{G_1} \leq p_{G_2} + p_{\{\ell\}} \approx p_{G_2}$, so G_1 has roughly the same chance of containing a signal as G_2 , even though $G_2 \subset G_1$. As a result, there is effectively no reason to include ℓ in any group in the candidate groups. This suggests the following procedure: after computing marginal PIPs $\{p_{\{\ell\}} : \ell \in \mathcal{L}\}$, we only construct candidate groups based on locations with a sufficiently high chance of harboring a signal, i.e., $\mathcal{L}_\kappa = \{\ell : p_{\{\ell\}} \geq \kappa\}$. Typically, we choose κ to be quite small, e.g., $\kappa \approx 0.001$. This approach often actually increases power in practice, since it prunes out irrelevant locations from our candidate groups, yielding a better set of candidate groups and allowing us to make discoveries at finer resolutions. Note that this approach is not appropriate when the set of locations is continuous and $p_{\{\ell\}} = 0$ for all $\ell \in \mathcal{L}$.

Lastly, when controlling the FDR or local FDR, we suggest *pre-narrowing \mathcal{G}* . To define this, suppose $G_1, G_2 \in \mathcal{G}$ such that $G_1 \subset G_2$, which means detecting a signal in G_1 yields more precise information than detecting a signal in G_2 . As a result, it is usually a safe assumption that $w(G_1) > w(G_2)$ (this is true in all of our applications). Suppose also that $p_{G_1}w(G_1) \geq p_{G_2}w(G_2)$, meaning that detecting a signal in G_1 leads to higher expected power than detecting a signal in G_2 . If this is true and $p_{G_1} \leq \alpha$ for some $\alpha < q$, this means that we can already discover G_1 at below the nominal level, so there is not much reason to discover G_2 : as a result, we exclude G_2 from consideration. Like the previous approach, this is a heuristic method, although it provably preserves validity. We suggest setting $\alpha = q/2$.

B.4 Polynomial-time alternatives to the integer LP

Our algorithm for BLiP (Algorithm 1) involves solving a relaxed LP and then running an integer LP on the non-integer solutions from the relaxed LP. In practice, the number of non-integer solutions from the relaxed LP is typically in the single-digits, as in Figure 10. As a result, the integer LP runs extremely quickly. That said, it is mathematically possible for the relaxed LP to admit solutions which include a very large number of non-integer values, in which case the integer LP may be too expensive. We emphasize that this never occurred in any of our simulations or real data applications, including in the fine-mapping application in Section 4 with 19 million covariates and hundreds of millions of candidate groups. For the sake of completeness, however, we will detail a heuristic polynomial time alternative to the integer LP. Note that this section follows the notation from Section 2.2.

Given optimal solutions $\{x_G^*\}_{G \in \mathcal{G}}$ to the relaxed LP (2.4)-(2.6) with $x_G^* \in [0, 1]$, we seek to obtain integer solutions $\{x_G^{**}\}_{G \in \mathcal{G}}$ which are feasible for the integer LP (2.4)-(2.7) and have as high expected power as possible. The key idea is as follows: we will use $\{x_G^*\}$ as guidelines and sample random feasible integer solutions $\{z_G^{(i)}\}_{G \in \mathcal{G}}$ for $1 \leq i \leq n_{\text{sample}}$, and then pick the solution $\{z_G^{(i)}\}_{G \in \mathcal{G}}$ which maximizes expected power. Intuitively, since $\{x_G^*\}_{G \in \mathcal{G}}$ are the optimal solution to the relaxed problem, we might expect that there is an integer solution “close to” $\{x_G^*\}$ which has similar expected power. One simple way to convert $\{x_G^*\}$ into a reasonably good solution would be to sample $z_G^{(i)}$ such that $\mathbb{P}(z_G^{(i)} = 1) = x_G^*$. Indeed, one can think of this as a directed search algorithm, where we search over solutions “close to” x_G^* by sampling $z_G^{(i)} \sim \text{Bern}(x_G^*)$. Although it is not always possible to ensure $\mathbb{P}(z_G^{(i)} = 1) = x_G^*$ while simultaneously satisfying the disjointness constraint (2.6), Algorithm 2 iteratively samples $\{z_G^{(i)}\}$ in a way that guarantees $\{z_G^{(i)}\}$ are a feasible solution to the integer LP while ensuring $\mathbb{P}(z_G^{(i)} = 1) \approx x_G^*$ at each step.

Since each $\{z_G^{(i)}\}_{G \in \mathcal{G}}$ is feasible by construction, Algorithm 2 is guaranteed to return a feasible solution to the integer LP (2.4)-(2.7), and furthermore, its runtime is polynomial in $|\mathcal{G}||\mathcal{L}|$. In our simulations, however, we found that in practice Algorithm 2 does not run faster than the integer LP in the default version of BLiP, and both algorithms achieve similar power. Since Algorithm 2 returns a randomized solution, for the rest of the paper, we always use the default version of BLiP specified in Algorithm 1 (which is deterministic).

B.5 Balancing the false positive and false negative rates

In this section, we discuss extending BLiP to balance the expected false positive and false negative rates. Throughout this section, we make one very mild assumption, which is that perfectly localizing a signal attains the maximum weight of 1. Formally, this means $w(G) \leq 1$ for all $G \in \mathcal{G}$, and $w(\{\ell\}) = 1$ for all $\ell \in \mathcal{L}$. We also assume that $\mathcal{L} = \{1, \dots, p\}$ is discrete and finite, although we assume this mainly for convenience (the analysis extends to the continuous case). We use the notation defined in Section 2.

For any set of detections defined by $\{x_G\}_{G \in \mathcal{G}}$, we define the Bayesian expected (resolution-adjusted) false negative rate (FNR) as

$$\mathbb{E}[\text{FNR} \mid \mathcal{D}] := 1 - \frac{\sum_{G \in \mathcal{G}} w(G)p_G x_G}{\sum_{\ell \in \mathcal{L}} p_{\{\ell\}}}.$$

Algorithm 2 Polynomial time alternative to the integer LP in BLiP.

Input: Locations \mathcal{L} , PIPs $\{p_G\}_{G \in \mathcal{G}}$ and solutions $\{x_G^*\}$ to the relaxed LP (2.4)-(2.6).

- 1: For each $\ell \in \mathcal{L}$, compute $s_\ell = \sum_{G:\ell \in G} x_G^*$, the target probability of discovering a group containing ℓ .
 - 2: Sort \mathcal{L} in decreasing order of $\{s_\ell\}$.
 - 3: **for** $i = 1, 2, \dots, n_{\text{sample}}$ **do**
 - 4: Initialize $\mathcal{D} = \{\}$ (the discovered groups). Later, we will set $z_G^{(i)} = 1$ iff $G \in \mathcal{D}$.
 - 5: Initialize $\mathcal{F} = \mathcal{G}$. Here, \mathcal{F} represents the set of “feasible” groups we can add to \mathcal{D} without violating the disjointness constraint.
 - 6: **for** $\ell \in \mathcal{L}$ **do**
 - 7: Let $\mathcal{F}_\ell = \{G \in \mathcal{F} : \ell \in G\}$ denote the feasible groups containing location ℓ .
 - 8: Sample $U \sim \text{Unif}(0, 1)$.
 - 9: **if** $U \leq s_\ell$ **then**
 - 10: Sample a group G_ℓ from \mathcal{F}_ℓ , with $\mathbb{P}(G_\ell = G) \propto x_G^*$ for $G \in \mathcal{F}_\ell$.
 - 11: Set $\mathcal{D} = \mathcal{D} \cup \{G_\ell\}$ and $\mathcal{F} = \mathcal{F} \setminus \{G' : G_\ell \cap G' \neq \emptyset\}$.
 - 12: **else**
 - 13: Set $\mathcal{F} = \mathcal{F} \setminus \mathcal{F}_\ell$ and continue to the next location.
 - 14: Sort \mathcal{D} in increasing order of $\{p_G\}_{G \in \mathcal{D}}$.
 - 15: Iteratively remove elements of \mathcal{D} until the discoveries \mathcal{D} satisfy the FDR constraint (2.3).
 - 16: For $G \in \mathcal{G}$, set $z_G^{(i)} = \mathbb{I}(G \in \mathcal{D})$.
 - 17: Return $\{z_G^{(i*)}\}_{G \in \mathcal{G}}$ for $i* = \arg \max_i \sum_{G \in \mathcal{G}} (1 - p_G) w(G) z_G^{(i)}$.
-

In the equation above, note that the numerator $\sum_{G \in \mathcal{G}} w(G) p_G x_G$ is simply the expected resolution-adjusted power, and the denominator is the expected number of signals, which is also the maximum expected resolution-adjusted power by the aforementioned assumptions. Thus, the expected FNR is zero if we discover signals at every location (which corresponds to maximal power), and it is one if we make no discoveries. It is important to note that $\mathbb{E}[\text{FNR} | \mathcal{D}]$ is a linear function of the optimization variables $\{x_G\}_{G \in \mathcal{G}}$, since the PIPs $\{p_G\}_{G \in \mathcal{G}}$ are constants.

There are many possible objective functions which balance the FDR against the expected FNR. For brevity, we will focus on the F1 score, which is defined as the harmonic mean of the precision and recall. In particular, define the resolution-adjusted F1 score as

$$\text{F1} := \frac{2}{(1 - \text{FDR})^{-1} + (1 - \mathbb{E}[\text{FNR} | \mathcal{D}])^{-1}}.$$

We are interested in optimizing $\{x_G\}_{G \in \mathcal{G}}$ to maximize the F1 score subject to the disjointness constraint (2.6), without necessarily controlling the FDR at some fixed level $q \in (0, 1)$. However, note that the F1 score is decreasing in both the FDR and expected FNR. Since BLiP minimizes the expected FNR given a constraint on the FDR, this implies that the solution $\{x_G^*\}_{G \in \mathcal{G}}$ which maximizes the F1 score is *also* a solution to the BLiP integer LP (2.4)-(2.7) for some FDR level q^* . To see this, suppose that $\{x_G^*\}_{G \in \mathcal{G}}$ maximizes the F1 score. Then, $\{x_G^*\}_{G \in \mathcal{G}}$ must be an optimal solution for the integer LP (2.4)-(2.7) with q^* equal to the FDR level of $\{x_G^*\}$, e.g., $q^* = \frac{\sum_{G \in \mathcal{G}} (1 - p_G) x_G^*}{\sum_{G \in \mathcal{G}} x_G^*}$. Otherwise, it would be possible to strictly improve the F1 score by decreasing the FNR without increasing the FDR. Thus, finding the F1-maximal solution simply corresponds to finding q^* and solving BLiP to control the FDR at level q^* .

This observation suggests that to maximize the F1 score, we can first run BLiP while controlling the

FDR at various levels $q \in (0, 1)$. By performing a line search or binary search over q , we can pick the value $q^* \in (0, 1)$ which attains the highest value of the F1 score. This algorithm is heuristic, but we suspect it will yield a reasonably good solution, since finding q^* only requires solving a univariate optimization over q . Overall, this shows that by slightly changing the objective function for BLiP, analysts could optimize for many different objectives, including balancing power with false positive control.

C Guidelines for computing the PIPs

In this section, we detail the methods we used to compute PIPs as an input to BLiP. We start by offering a few general guidelines in Section C.1. Next, Sections C.2 and C.3 detail the Gibbs samplers we used in variable selection problems for sparse linear regression and sparse probit regression, respectively. We picked these samplers because some other existing software yielded inaccurate PIPs which caused BLiP to substantially violate error control. In contrast, these (very simple) samplers performed well empirically and are quite efficient. Finally, Section C.4 describes how to compute PIPs based on SuSiE (Wang et al., 2020) and explains why SuSiE + BLiP can outperform SuSiE alone by wide margins.

C.1 General guidelines

In this section, we offer a few general guidelines on computing the PIPs $\{p_G\}_{G \in \mathcal{G}}$. To begin with, it is straightforward to estimate p_G if we can sample from the posterior distribution of the locations of true signals. Indeed, let $\gamma_1, \dots, \gamma_N \subset \mathcal{L}$ be such samples, which are straightforward to obtain from standard Bayesian methods. For example, in sparse linear regression problems where $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$ and the signals correspond to nonzero coefficients in β , one can use any standard method to obtain samples $\tilde{\beta}^{(1)}, \dots, \tilde{\beta}^{(N)} \in \mathbb{R}^p$ from the posterior of β and then set $\gamma_i = \{j : \tilde{\beta}_j^{(i)} \neq 0\}$. Having obtained $\gamma_1, \dots, \gamma_N$, we can estimate p_G as the average number of posterior samples γ_i which include a signal in G , i.e., $p_G \approx \frac{1}{N} \sum_{i=1}^N \mathbb{I}(G \cap \gamma_i \neq \emptyset)$. That said, obtaining good posterior samples can be challenging in high dimensional problems where (a) the prior may be misspecified and (b) standard MCMC algorithms may not converge quickly. Below, we describe a few heuristics to ensure the posterior samples and BLiP are robust to these issues.

To address (a), we recommend using *hierarchical priors* which account for some uncertainty in the Bayesian model. For example, when performing variable selection, many Bayesian models require some knowledge of p_1 , the proportion of signals. Unless one has very precise prior knowledge of these parameters, we suggest picking a fairly uninformative prior for them, such as letting $p_1 \sim \text{Unif}(0, p_{\max})$ for some $p_{\max} \leq 1$. To determine the hyper-parameters for these priors (e.g., p_{\max}), we suggest either using a conservative choice or taking an empirical Bayesian approach. (Here, “conservative” choices mean choices that may yield an error rate slightly below the nominal level.) The simulations in Section 3 show that even fairly conservative choices, such as when $p_{\max} = \frac{1}{2}p_1$, do not lose much power and reliably control the FDR even when the hyperprior is quite different from the true sparsity level.

To address (b), when using MCMC algorithms to sample from the posterior, we generally recommend sampling from *multiple MCMC chains with random initialization*. Even if each chain does not converge, we expect that aggregating results across chains will usually overestimate the uncertainty

in the location of a signal. This allows BLiP to control the error rate, even if it is somewhat conservative. For example, suppose that in a bivariate regression problem, $\{X_1, X_2\}$ clearly contains a signal variable, but X_1 and X_2 are highly correlated, so it is not clear which one is the signal variable. Consider a worst-case scenario where the MCMC algorithm randomly initializes (e.g.) X_1 to be a signal variable, but then keeps X_1 as a signal variable at every iteration. If we compute $p_{\{1\}}$ just using this chain, we will falsely conclude that $p_{\{1\}} \approx 1$. However, if we run 5 – 10 MCMC chains which each have a 50% chance of initializing X_1 or X_2 as a signal variable, then we will conclude $p_{\{1\}}, p_{\{2\}} \approx 50\%$, or equivalently, that we know X_1 or X_2 is a signal but we are maximally uncertain about which one is the signal. This will yield (conservative) error rate control in this toy example, even though the MCMC algorithm did not converge whatsoever. This intuition also extends to variational approaches which use random initialization or MCMC sampling.

To empirically demonstrate the effect of using multiple MCMC chains, in Figure 12, we rerun simulations in the setting of Figure 2 but with only 200 MCMC samples per chain and a burn-in of 20 samples. (In contrast, in Figure 2, we used 5000 samples per chain with a burn-in of 500 samples.) This is a very high-dimensional setting where BLiP searches over tens of thousands of candidate groups over $p = 1000$ covariates, so we should not expect the first 220 MCMC samples to converge. Indeed, when using only one chain, LSS + BLiP dramatically violates FDR control. When using 10 chains, however, LSS + BLiP successfully controls the FDR while achieving nearly the same power.

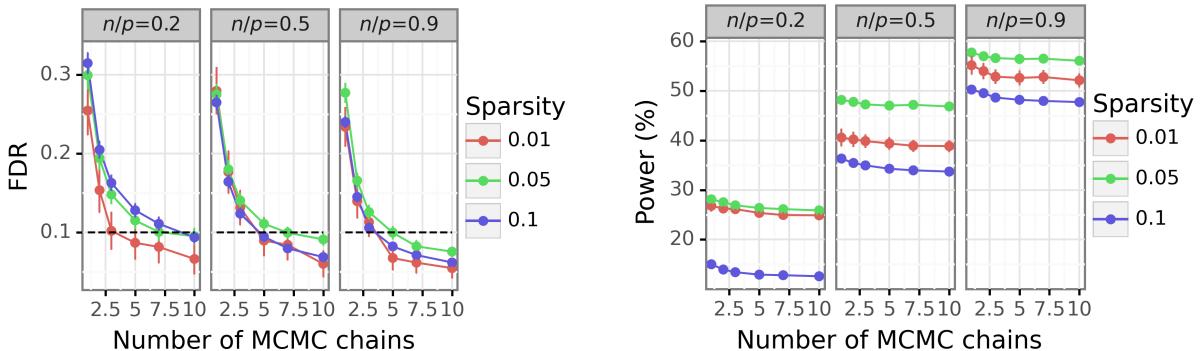


Figure 12: This figure replicates the “LSS + BLiP (misspec.)” method from Figure 2 but uses only 200 samples per MCMC chain. The individual chains clearly do not converge, as shown by the left-hand plot, where the realized FDR is up to three times the nominal level when using just one chain. Despite this, aggregating results from 5-10 chains leads to FDR control without substantially losing power (as shown in the right-hand figure). The simulation details are otherwise identical to Figure 2 and are detailed in Appendix E.2

C.2 Details on the linear spike-and-slab sampler

In this section, we review details on the linear spike and slab (LSS) Gibbs sampler we use in Sections 3-4. We assume \mathbf{y} follows a Gaussian linear model, so $\mathbf{y} | \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$. The prior specification is as follows:

$$\begin{aligned}\sigma^2 &\sim \text{invGamma}(a_\sigma, b_\sigma) \\ \tau^2 &\sim \text{invGamma}(a_\tau, b_\tau) \\ p_0 &\sim \text{TruncBeta}(a_0, b_0, p_{\min})\end{aligned}$$

where $\text{TruncBeta}(a_0, b_0, p_{\min})$ denotes a $\text{Beta}(a_0, b_0)$ distribution truncated to the interval $[p_{\min}, 1]$. We assume that p_0, σ^2, τ^2 are jointly independent a priori. The exception to this prior scheme is that in the “well-specified” case, we assume that p_0, σ^2, τ^2 are known and use their true values. Finally, we assume the following prior for β :

$$\beta_j \mid \tau^2, p_0 \stackrel{i.i.d.}{\sim} \begin{cases} 0 & \text{w.p. } p_0 \\ \mathcal{N}(0, \tau^2) & \text{w.p. } 1 - p_0. \end{cases}$$

To sample from the posterior of the parameters, we employ Gibbs sampling. However, because \mathbf{X} may exhibit high correlations, instead of sampling individual parameters from their conditional distributions, at each iteration we partition the covariates into blocks J_1, \dots, J_m of size k or less and sample from the joint distribution of β_{J_1} , then β_{J_2} , and so on. In our implementation, we choose J_1, \dots, J_m to be contiguous groups of size k , but in principle one could apply any clustering algorithm. Although setting $k = 1$ and sampling from the distribution of individual parameters works fairly well, setting (e.g.) $k \approx 3$ can yield a substantial boost in power by helping the chain mix faster. Algorithm 3 gives an outline of the general sampling scheme.

Algorithm 3 Linear spike-and-slab Gibbs sampler.

Input: \mathbf{y}, \mathbf{X} , hyperpriors, and the maximum block-size k .

- 1: Initialize $\beta^{(1)} = 0$ and $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta^{(1)}$.
 - 2: **for** $\ell = 1, 2, \dots, n_{\text{iter}}$ **do**
 - 3: Partition $[p]$ into blocks J_1, \dots, J_m , e.g., by clustering \mathbf{X} .
 - 4: **for** $J = J_1, \dots, J_m$ **do**
 - 5: Update $\mathbf{r} = \mathbf{r} + \mathbf{X}_J \beta_J^{(\ell)}$.
 - 6: Sample $\beta_J \mid \mathbf{y}, \beta_{-J} = \beta_{-J}^{(\ell)}, \tau^2, \sigma^2, p_0$ using \mathbf{r} for efficiency, as in Equations (C.1), (C.2).
 - 7: Set $\beta_J^{(\ell)} = \beta_J$ and update $\mathbf{r} = \mathbf{r} - \mathbf{X}_J \beta_J^{(\ell)}$
 - 8: Sample $\tau^2, \sigma^2 \mid \mathbf{y}, \beta = \beta^{(\ell)}$ using inverse-Gamma conjugacy as in Equations (C.4), (C.3).
 - 9: Sample $p_0 \mid \mathbf{y}, \beta = \beta^{(\ell)}$ using rejection sampling as in Equation (C.5).
 - 10: Return $\beta^{(1)}, \dots, \beta^{(n_{\text{iter}})}$.
-

The most computationally intensive step in this algorithm is to efficiently sample $\beta_J \mid \beta_{-J}, \sigma^2, p_0, \tau^2, \mathbf{y}$ for any $J \subset [p]$. We describe how to do this in two steps. First, define $A = \{j \in J : \beta_j \neq 0\}$ to be the active set (set of nonzero coefficients) of J . Note that A is a random latent variable. The first step is to sample $A \mid \beta_{-J}, \sigma^2, p_0, \tau^2, \mathbf{y}$. To lighten notation, let $H = (\sigma^2, p_0, \tau^2)$ denote the hyper-parameters. Note that for any $J_0 \subset J$,

$$\mathbb{P}(A = J_0 \mid \beta_{-J}, H, \mathbf{y}) \propto \mathbb{P}(A = J_0 \mid H, \beta_{-J}) p(\mathbf{y} \mid A = J_0, \beta_{-J}, H)$$

where we use $p(\cdot \mid \cdot)$ to denote an arbitrary conditional density. Now, the form of the prior yields that $\mathbb{P}(A = J_0 \mid H, \beta_{-J}) = p_0^{|J|-|J_0|} (1 - p_0)^{|J_0|}$. To analyze the second term in the product, note that since $\beta_{J_0} \mid A = J_0$ is multivariate Gaussian (and $\beta_{J \setminus J_0} = 0$),

$$\mathbf{y} \mid A = J_0, \beta_{-J}, H \sim \mathcal{N}(\mathbf{X}_{-J} \beta_{-J}, \sigma^2 \mathbf{I}_n + \tau^2 X_{J_0} X_{J_0}^T).$$

As a result, we conclude

$$p(\mathbf{y} \mid A = J_0, \beta_{-J}, H) \propto \det(\sigma^2 \mathbf{I}_n + \tau^2 X_{J_0} X_{J_0}^T)^{-1/2} \exp\left(-\frac{1}{2} r^T (\sigma^2 \mathbf{I}_n + \tau^2 X_{J_0} X_{J_0}^T)^{-1} r\right).$$

To further simplify this, define $\mathbf{r} = \mathbf{y} - X_{-J}\beta_{-J}$ to be the residuals without X_J and let $Q_{J_0} = I_{|J_0|} + \frac{\tau^2}{\sigma^2} X_{J_0}^T X_{J_0}$. Using the Woodbury identity and the matrix determinant lemma, we conclude

$$\mathbb{P}(A = J_0 \mid \beta_{-J}, H, \mathbf{y}) \propto p_0^{|J|-|J_0|} (1 - p_0)^{|J_0|} \det(Q_{J_0})^{-1/2} \exp\left(\frac{\tau^2}{2\sigma^4} \mathbf{r}^T X_{J_0} Q_{J_0}^{-1} X_{J_0}^T \mathbf{r}\right). \quad (\text{C.1})$$

Since $Q_{J_0} \in \mathbb{R}^{|J_0| \times |J_0|}$ and $|J_0| \leq |J|$, we can compute this expression in $O(n|J| + |J|^3)$. We repeat this for each of the $2^{|J|}$ possible values for A , yielding a computational complexity of $O(2^k(k^3 + nk))$ to sample A . Since typically k is a constant which does not scale with n or p (we suggest taking $k \leq 5$), we will write this as $O(2^k nk)$. Repeating this for each of the m blocks and n_{iter} iterations yields a time complexity of $O(2^k nkmn_{\text{iter}}) = O(2^k n_{\text{iter}} np)$ since $km = O(p)$. Since k is a constant, this reduces to $O(n_{\text{iter}} np)$, although the constant 2^k naturally affects the performance of the algorithm.

Once we have sampled A , the second step is to sample $\beta_J \mid \beta_{-J}, A, \mathbf{y}, H$. Of course, $\beta_{J \setminus A} = 0$, so we need only sample β_A . Observe that conditional on A, β_{-J} , we have that

$$(\mathbf{r}, \beta_A) \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2 \mathbf{I}_n + \tau^2 X_A X_A^T & \tau^2 X_A \\ \tau^2 X_A^T & \tau^2 \mathbf{I}_{|A|} \end{bmatrix}\right).$$

Using standard Gaussian update rules, we can sample from the desired conditional distribution. To do this efficiently, we also apply the Woodbury identity to $\sigma^2 \mathbf{I}_n + \tau^2 X_A X_A^T$. Thus, the second step is to sample from

$$\beta_A \mid \beta_{-J}, A, H, \mathbf{y} \sim \mathcal{N}\left(\frac{\tau^2}{\sigma^2} X_A^T \mathbf{r} - \frac{\tau^4}{\sigma^4} X_A^T X_A Q_A^{-1} X_A^T \mathbf{r}, \tau^2 \mathbf{I}_{|A|} - \frac{\tau^4}{\sigma^2} X_A^T X_A + \frac{\tau^6}{\sigma^4} X_A^T X_A Q_A^{-1} X_A^T X_A\right) \quad (\text{C.2})$$

which can be done in $O(n|A|^2 + |A|^3) = O(nk^2) = O(n)$. Lastly, for completeness, we now specify how to sample the hyper-parameters $H = (\sigma^2, p_0, \tau^2)$ given β . Let $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$ and let $n_{\text{active}} = \#\{j : \beta_j \neq 0\}$ denote the number of active (nonzero) features. Then

$$\tau^2 \mid \mathbf{y}, \beta, \sigma^2, p_0 \sim \text{invGamma}\left(n_{\text{active}}/2 + a_\sigma, \frac{1}{2} \sum_{j:\beta_j \neq 0} \beta_j^2 + b_\sigma\right), \quad (\text{C.3})$$

$$\sigma^2 \mid \mathbf{y}, \beta, \tau^2, p_0 \sim \text{invGamma}\left(n/2 + a_\sigma, \mathbf{r}^T \mathbf{r}/2 + b_\sigma\right). \quad (\text{C.4})$$

In the case where $p_{\min} = 0$, one can use beta-binomial conjugacy to sample

$$p_0 \mid \mathbf{y}, \beta, \sigma^2, \tau^2 \sim \text{Beta}(a_0 + p - n_{\text{active}}, b_0 + n_{\text{active}}). \quad (\text{C.5})$$

If $p_{\min} \neq 0$, one can repeatedly rejection sample using (C.5) until the result is greater than p_{\min} . Since our algorithm maintains a running buffer of \mathbf{r} , sampling the hyperparameters requires either constant time or $O(n)$ computations, which is not a bottleneck. The overall time complexity is thus $O(n_{\text{iter}} np)$ as discussed earlier.

C.3 Details on the probit spike-and-slab sampler

In the case of probit regression (Section 3.2), we assume the same model as in Section C.2 except that we observe $\mathbf{z} = \mathbb{I}(\mathbf{y} \geq 0) \in \{0, 1\}^n$ instead of the continuous outcome \mathbf{y} . To sample from the posterior $\beta, \sigma^2, \tau^2, p_0 \mid \mathbf{z}, \mathbf{X}$, we follow Albert and Chib (1993) and employ a data-augmentation strategy.

In particular, we perform exactly the same steps as Algorithm 3 except we add an intermediate step where we sample the latent variables \mathbf{y} from the posterior distribution $\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2, \tau^2, p_0, \mathbf{z}$. This is simple, since if we let $\mu = \mathbf{X}\beta$, then conditional on $\mu, \mathbf{z}, \sigma^2$, the coordinates of \mathbf{y} are independent truncated Gaussians. In particular,

$$Y_j \mid \mathbf{X}, \beta, \sigma^2, \tau^2, p_0 \stackrel{\text{ind}}{\sim} \begin{cases} \text{truncNorm}(\mu_j, \sigma^2, 0, \infty) & z_j = 1 \\ \text{truncNorm}(\mu_j, \sigma^2, -\infty, 0) & z_j = 0, \end{cases}$$

where $\text{truncNorm}(\mu, \sigma^2, a, b)$ represents the Gaussian distribution with mean μ and variance σ^2 truncated to (a, b) . Note that we resample the whole vector \mathbf{y} using this formula every time we update a block of coordinates β_J in Algorithm 3: otherwise, the algorithm for the probit case is identical to Algorithm 3. It requires $O(n)$ operations to resample \mathbf{y} , but it also requires $\Omega(n)$ operations to update a block β_J , so the computational complexity of the probit sampler is still $O(n_{\text{iter}}np)$. Practically speaking, resampling \mathbf{y} is a bottleneck, so we implemented the algorithm from Geweke (1991) to do this as quickly as possible.

C.4 Further discussion of BLiP and SuSiE

In this section, we describe how to run BLiP on top of the outputs from SuSiE (Wang et al., 2020) and also explain why SuSiE + BLiP can have much higher power than SuSiE alone. To begin with, a short review of SuSiE is in order.

C.4.1 Review of SuSiE

The building block of SuSiE is the *single effect regression* (SER) model, which is essentially a spike-and-slab Gaussian regression model with exactly one nonzero coefficient. Formally, this model assumes that $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$ and decomposes $\beta = b\gamma$, with priors $b \sim \mathcal{N}(0, \tau^2) \in \mathbb{R}$ and $\gamma \sim \text{Mult}(1, \pi) \in \{0, 1\}^p$ for $\pi \in [0, 1]^n$ satisfying $\sum_{i=1}^p \pi_i = 1$. Also, b and γ are assumed to be independent a priori. Since the random vector $\gamma \in \{0, 1\}^p$ has exactly one nonzero entry, only one entry of β is nonzero. As notation, let $\ell^* \in [p]$ denote the (random) index such that $\gamma_{\ell^*} = 1$. For simplicity, we will assume that σ^2, τ^2 and π are known and fixed in advance, although this need not be the case.

Notably, it is easy to compute the posterior distribution of b and γ as

$$\begin{aligned} \gamma \mid \mathbf{y}, \mathbf{X} &\sim \text{Mult}(1, \alpha) \\ b \mid \mathbf{y}, \mathbf{X}, \gamma_j = 1 &\sim \mathcal{N}(\mu_j, v_j^2) \end{aligned}$$

where there are analytical formulas for $\alpha \in [0, 1]^n$, $\mu = (\mu_1, \dots, \mu_p) \in \mathbb{R}^n$ and $v^2 = (v_1^2, \dots, v_p^2) \in \mathbb{R}_+^n$ given in Wang et al. (2020). Here, $\alpha_1, \dots, \alpha_p$ are the PIPs for coefficients β_1, \dots, β_p . A key advantage of this model is that one can easily construct a minimal-width credible set for the location of the single nonzero coefficient. Indeed, let $\varphi : [p] \rightarrow [p]$ be the permutation which sorts α in decreasing order, so that $\alpha_{\varphi(1)} \geq \alpha_{\varphi(2)} \geq \dots \geq \alpha_{\varphi(p)}$. Then for each k , the set $\{\varphi(1), \dots, \varphi(k)\}$ contains ℓ^* (the index of the nonzero coefficient) with probability $\sum_{j=1}^k \alpha_{\varphi(j)}$. The minimal-width $1 - q$ credible set simply chooses k to be as small as possible such that $\sum_{j=1}^k \alpha_{\varphi(j)} \geq 1 - q$. As notation, let $\text{CS}(\alpha) = \{\varphi(1), \dots, \varphi(k)\}$ for this choice of k .

The Sum of Single Effects (SuSiE) model stacks L SER models on top of each other by assuming that $\beta = \sum_{i=1}^L b^{(i)}\gamma^{(i)}$, where $b^{(i)} \stackrel{\text{ind}}{\sim} \mathcal{N}(0, \tau^2)$ and $\gamma^{(i)} \stackrel{\text{ind}}{\sim} \text{Mult}(1, \pi)$. To fit a SuSiE model, Wang et al. (2020) proposed sequentially fitting L SER models, and this algorithm is known as Iterative Bayesian stepwise selection (IBSS). To be precise, at the first iteration, the IBSS algorithm fits an SER model to obtain outputs $\alpha^{(1)}, \mu^{(1)}, v^{(1)}$ as well as a credible set $\text{CS}(\alpha^{(1)})$ for the first signal. Let $B^{(1)} = \alpha^{(1)} \odot \mu^{(1)}$ be the estimated coefficients, where \odot denotes elementwise multiplication. At this point, one computes the residuals $\mathbf{r} = \mathbf{y} - \mathbf{X}B^{(1)}$ and then fits a second SER model on \mathbf{X} and the residuals \mathbf{r} to obtain outputs $\alpha^{(2)}, \mu^{(2)}, v^{(2)}$. The algorithm repeats for L steps, at each step fitting an SER model and then recomputing the residuals based on the estimated coefficients. For our purposes, the main outputs of SuSiE are the PIPs $\alpha^{(1)}, \dots, \alpha^{(L)}$ and the SuSiE credible sets $\text{CS}(\alpha^{(1)}), \dots, \text{CS}(\alpha^{(L)})$. (It is worth noting that the original SuSiE algorithm only returns $\text{CS}(\alpha^{(i)})$ as a discovery if $\text{CS}(\alpha^{(i)})$ passes a heuristic post-processing check ensuring that the minimum pairwise correlation among variables in $\text{CS}(\alpha^{(i)})$ is greater than 50%.)

C.4.2 Computing the inputs to BLiP from the SuSiE output

Given these outputs, it is straightforward to compute the inputs for BLiP, namely the PIPs for a set of candidate groups \mathcal{G} . Recall that in a SER model with PIPs α , for any group $G \subset [p]$, the PIP $p_G = \mathbb{P}(I_G = 1) = \sum_{j \in G} \alpha_j$ by definition of α . The only question is how to combine the PIPs $\alpha^{(1)}, \dots, \alpha^{(L)}$ which correspond to fitting L SER models. However, Wang et al. (2020) showed that the IBSS algorithm is a variational approximation to the posterior which assumes that the posterior of the SuSiE model factors into L independent SER models. Under this variational approximation, by independence, we conclude that

$$p_G = \mathbb{P}(I_G = 1) = 1 - \prod_{i=1}^L \left(1 - \sum_{j \in G} \alpha_j^{(i)} \right), \quad (\text{C.6})$$

which allows us to compute PIPs for arbitrary candidate groups based on the SuSiE outputs $\alpha^{(1)}, \dots, \alpha^{(L)}$. Note that this equation is a trivial extension of the formula for PIPs of single variables in the original paper. However, it has a useful consequence: if SuSiE discovers disjoint groups G_1, \dots, G_m , then for each G_k there must exist some $i' \in \{1, \dots, L\}$ such that $\sum_{j \in G_k} \alpha_j^{(i')} \geq 1 - q$. This plus the previous equation implies $p_{G_k} \geq 1 - q$, so the set of discoveries made by SuSiE are feasible for BLiP when controlling the FDR. Since BLiP optimizes to maximize the expected resolution-adjusted power, we should expect SuSiE + BLiP to make discoveries with higher expected power than SuSiE for any input.

C.4.3 Why SuSiE + BLiP outperforms SuSiE alone

The previous paragraph only suggests that SuSiE + BLiP weakly outperforms SuSiE. Now, we discuss why SuSiE + BLiP can outperform SuSiE by wide margins. At first, this might seem a bit surprising, since at each iteration i SuSiE obtains a minimal credible interval based on $\alpha^{(i)}$. However, SuSiE is not able to combine information across iterations, which can dramatically reduce power when there are multiple non-nulls. To see this, consider a simple example where $\sigma^2 = 1, \pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}), p = L = 4, \mathbf{X}^T \mathbf{X} = \mathbf{I}_4$ and $\mathbf{X}^T \mathbf{y} = (100, 100, 0, 0)$. In this case, many methods (such as a frequentist t-test) can detect that $\beta_1 \neq 0$ and $\beta_2 \neq 0$. However, by symmetry, the SuSiE outputs will be $\alpha^{(1)} = \alpha^{(2)} = \alpha^{(3)} = \alpha^{(4)} \approx (0.5, 0.5, 0, 0)$, suggesting that at each iteration, there is

a 50% chance that either X_1 or X_2 is a signal variable. As a result, when providing level $q = 0.1$ error control, SuSiE will output the same discovery set $\{1, 2\}$ four times. In contrast, BLiP can use PIPs which combine information across iterations, allowing it to discover the groups $\{1\}, \{2\}$, since $p_{\{1\}} = p_{\{2\}} = 1 - \left(\frac{1}{2}\right)^4 = 0.9375 \geq 1 - q$.

Of course, this example is a bit contrived, since in this case anyone inspecting the marginal PIPs from SuSiE can tell that X_1, X_2 are signal variables, and furthermore it is unlikely the equality $\mathbf{X}_1^T \mathbf{y} = \mathbf{X}_2^T \mathbf{y}$ would hold exactly in real data. That said, this general phenomenon is not uncommon: if at any iteration of SuSiE $\mathbf{X}_j^T \mathbf{r} \approx \mathbf{X}_k^T \mathbf{r}$, SuSiE will assign PIPs satisfying $\alpha_j^{(i)} \approx \alpha_k^{(i)} \leq 0.5$, even if it is obvious that both X_j and X_k are signal variables. The consequence is that SuSiE will only partially regress out the effects of X_j and X_k , so it is also likely that $\alpha_j^{(i+1)}, \alpha_k^{(i+1)}$ will be fairly large in the next iteration as well. Admittedly, this does not always happen when there are a small number of signal variables, but when there are a large number of signal variables, the maximum signal size may concentrate around some value. As a result, $\mathbf{X}_j^T \mathbf{r} \approx \mathbf{X}_k^T \mathbf{r}$ may occur fairly frequently, and the SuSiE algorithm will have artificially low power. In our view, the best way to recover this power is by applying BLiP, since BLiP can leverage information from $\alpha^{(1)}, \dots, \alpha^{(L)}$ to compute more accurate PIPs. As exemplified by Figure 2, this can simultaneously lead to higher power and lower FDR.

To illustrate this further, we consider a simple setting where $p = 200$ and there are between 0 and 10 signal variables, although in each case we run SuSiE with $L = 10$ iterations. For each of the detected groups G returned by SuSiE + BLiP, the left plot in Figure 13 shows the mean difference between $p_G = 1 - \prod_{i=1}^L \left(1 - \sum_{j \in G} \alpha_j^{(i)}\right)$, the aggregate PIP for group G , and $p_G^{\max} = \max_{1 \leq i \leq L} \sum_{j \in G} \alpha_j^{(i)}$, the maximum PIP for group G at any iteration i of SuSiE. This difference is relevant because SuSiE + BLiP makes detections based on p_G and SuSiE (roughly speaking) makes detections based on p_G^{\max} . Figure 13 shows that for true discoveries, the aggregate PIPs p_G are often as much as five percentage points larger than p_G^{\max} on average, which is a substantial difference since we are controlling the FDR at level $q = 0.1$. In contrast, p_G is usually only half a percentage point larger than p_G^{\max} for null variables. This suggests that aggregating over the iterations of SuSiE yields more accurate PIPs, which is remarkable since this appears to be true even when SuSiE runs for $L = 10$ iterations and there are only 2 signal variables. Indeed, the right plot in Figure 13 shows that SuSiE + BLiP has higher power than SuSiE and successfully controls the FDR. See Appendix E for more details.

That said, we caution that if L is *much* larger than the number of signal variables, for example if $L = 100$ and there are only one or two signals, we expect that SuSiE + BLiP as currently defined will violate FDR control. This motivates our discussion in the next subsection. That said, $L = 10$ is the default in genetic fine-mapping, and we emphasize that when $L = 10$ we have seen no evidence that SuSiE + BLiP violates FDR control even when there are as few as zero signals.

C.4.4 Further possibilities for SuSiE + BLiP

We conclude this section by discussing two additional ways to apply BLiP on top of SuSiE. First, one might worry that aggregating evidence across all L SuSiE iterations in Equation (C.6) could artificially increase the PIPs, since L is often larger than the true number of signals. This did not appear to be a problem in any of our simulations, e.g., in Figures 13 and 21, where $L = 10$ and SuSiE + BLiP controls the FDR even when there are zero signals. However, one additional

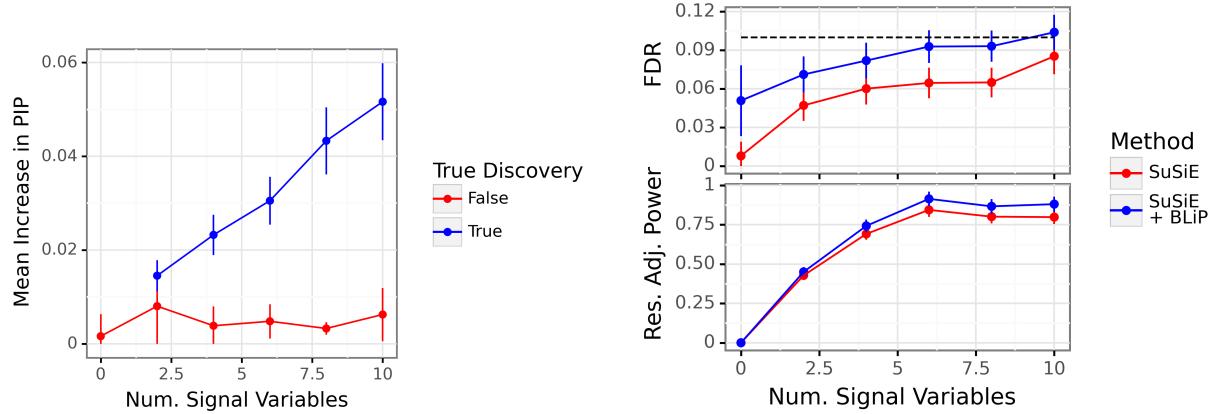


Figure 13: This figure illustrates why SuSiE + BLiP outperforms SuSiE alone in a high-dimensional linear regression setting where $p = 200$, $k = 1$, $n/p = 0.2$, and we run SuSiE with $L = 10$ iterations. For each group G detected by SuSiE + BLiP, the left plot shows the mean difference between the full PIPs p_G , which are used by SuSiE + BLiP, and the iteration-specific PIPs p_G^{\max} used by SuSiE. It shows that for true discoveries, p_G is substantially larger than p_G^{\max} , whereas for false discoveries, p_G and p_G^{\max} are quite similar. This confirms the intuition in the preceding section and explains why SuSiE + BLiP can have higher power than SuSiE alone but good FDR control even when L is larger than the true number of signals. The right plot shows that SuSiE + BLiP has higher (unnormalized) power than SuSiE and controls the FDR. See Appendix E for details.

way to address this concern would be to only aggregate evidence across iterations of SuSiE which pass the heuristic post-processing steps from Wang et al. (2020). In other words, before computing the PIPs in Equation (C.6), we could completely discard $\alpha^{(i)}$ if $\alpha^{(i)}$ does not yield a sufficiently “pure” credible set (see Wang et al. (2020) for details). Alternatively, one could compute the PIPs as normal but discard all candidate groups which are not sufficiently “pure.” The main point is that any post-processing heuristic which can be applied to SuSiE can likely be applied to SuSiE + BLiP as well, although in this paper, we never found it necessary to use postprocessing heuristics when applying BLiP on top of SuSiE.

Second, Zou et al. (2021) recently introduced a refinement of the IBSS algorithm which is guaranteed to produce a better model fit. Unfortunately, we could not use this new algorithm in our simulations, since it is computationally prohibitive when SuSiE makes many discoveries. For example, in Figure 2, the refined SuSiE algorithm would involve refitting the IBSS procedure hundreds or thousands of times with different initializations. That said, in principle, BLiP can apply on top of this refined SuSiE procedure in exactly the same manner as described above. Note that although the refined procedure produces a better fit, its confidence sets still do not aggregate evidence across multiple iterations. Thus, we expect that BLiP will improve the power of the refined IBSS procedure for the same reasons outlined in this section.

D Applying BLiP when the locations are continuous

In this section, we discuss how to apply BLiP when \mathcal{L} is a “continuous” set of locations. For example, in our application to astronomical point source detection in Section 5, $\mathcal{L} = [0, 1]^2$. The continuous case can be challenging when the candidate groups \mathcal{G} and locations \mathcal{L} are infinite sets. This poses issues both for solving the relaxed LP in the BLiP algorithm and for computing the PIPs

$\{p_G\}_{G \in \mathcal{G}}$ efficiently. The following two sections address these problems separately. First, Appendix D.1 introduces efficient algorithms to compute $\{p_G\}_{G \in \mathcal{G}}$ even when \mathcal{G} is extremely large or infinite. Second, Appendix D.2 shows that even when $|\mathcal{L}| = \infty$ and applying BLiP seems to require solving an infinite-dimensional LP, there exist efficient algorithms to reduce the infinite-dimensional LP to an equivalent finite-dimensional LP of (approximately) minimal dimension. Overall, as demonstrated by our astronomical application in Section 5, BLiP can be very efficient and powerful even when \mathcal{L} is continuous and infinite.

Throughout this section, we will assume for simplicity that $\mathcal{L} \subset \mathbb{R}^d$, although the ideas in this section extend beyond \mathbb{R}^d . We make three further assumptions. First, we assume that M is finite, where M denotes the expected number of signals under the posterior. This assumption is satisfied in any setting where the analyst does not expect to make an infinite number of discoveries. Second, we assume there exists some constant $R \in \mathbb{N}$ such that for each $x \in \mathbb{R}^d$, there are at most R sets $G \in \mathcal{G}$ containing x . This intuitively means that \mathcal{G} contains groups at roughly R resolutions. For example, in our astronomical application, the candidate groups are a collection of circles of radius $r \in \{r_1, \dots, r_m\}$ centered at lattice points. It turns out that each point x is contained by up to 5 circles of radius r for each r , so therefore each point x is contained by at most $R = 5m$ candidate groups. Overall, this is not a particularly restrictive assumption because R can be arbitrarily large, and this assumption still permits \mathcal{G} to be infinite, e.g., if $\mathcal{L} = \mathbb{R}^d$ and \mathcal{G} covers \mathcal{L} with finite rectangles. Lastly, we assume that given any $x \in \mathbb{R}^d$, we can find all groups $G_1, \dots, G_m \in \mathcal{G}$ containing x in a computational complexity of $O(\varphi(d, R))$, where $m \leq R$ and φ is some function not depending on x . Throughout this paper, $\varphi(R, d) = Rd$, because \mathcal{G} is a collection of shapes (spheres, cubes) on a collection of grids of lattice points (see Appendix G for a precise description). Thus, to find the groups containing x , we just round x to the R nearest lattice points, which requires $O(dR)$ time. However, we use the notation for $\varphi(d, R)$ to allow for greater generality.

As discussed in Section 6, there are possible extensions to BLiP which apply when M is infinite, although we do not consider them here.

D.1 Computing PIPs for a continuous set of locations

Suppose we have access to N samples $\{\gamma^{(1)}, \dots, \gamma^{(N)}\}$ from the posterior of the locations of the signals, which means that the i th posterior sample is a set of the form $\gamma^{(i)} = \{\ell_1^{(i)}, \dots, \ell_{m_i}^{(i)}\} \subset \mathcal{L} \subset \mathbb{R}^d$. Note that the i th posterior sample asserts that there are m_i signals at locations $\ell_1^{(i)}, \dots, \ell_{m_i}^{(i)}$, which implies that the expected number of signals is approximately $M \approx \frac{1}{N} \sum_{i=1}^N m_i$. Given these inputs and a set of candidate groups \mathcal{G} , in this section we consider the following question: how can we compute the PIPs $\{p_G\}_{G \in \mathcal{G}}$ efficiently?

Recall that for any candidate group $G \in \mathcal{G}$, the PIP p_G is simply the proportion of posterior samples $\{\gamma^{(i)}\}$ which include a signal inside G . Thus, a naive algorithm would loop through all candidate groups $G \in \mathcal{G}$ and compute this proportion. Unfortunately, the number of candidate groups is potentially infinite, and even when it is finite, $|\mathcal{G}|$ generally grows exponentially in the dimension d , so the naive algorithm is too slow. However, we assume the number of signals is finite, and each signal $\ell_i^{(j)}$ in our posterior samples is contained by at most R candidate groups. Therefore, at most RNM candidate groups contain a signal, and thus when \mathcal{G} is large, the vast majority of candidate groups $G \in \mathcal{G}$ have PIPs of zero and are irrelevant. This suggests the following algorithm: instead of looping through the candidate groups, we can loop through the signals in the posterior samples $\{\ell_i^{(j)}\}$. By the assumptions at the start of Appendix D, for each $1 \leq i \leq N$, we can compute

the groups $G_1, \dots, G_{M_i} \in \mathcal{G}$ which contain a signal in $\gamma^{(i)}$, and we are guaranteed $M_i \leq Rm_i$. After eliminating duplicates in G_1, \dots, G_{M_i} , we can then increment the PIPs corresponding to these candidate groups by $\frac{1}{N}$ and proceed to the next posterior sample. This is formally stated in Algorithm 4.

Algorithm 4 Computing PIPs when $\mathcal{L} \subset \mathbb{R}^d$.

Input: Candidate groups \mathcal{G} and posterior samples $\{\gamma^{(i)}\}_{i=1}^N$ for $\gamma^{(i)} \subset \mathcal{L}$.

- 1: Initialize a hash map PIP, which will map candidate groups to PIPs.
 - 2: **for** $i \in \{1, \dots, N\}$ **do**
 - 3: Initialize $\mathcal{A} = \{\}$, the set of candidate groups which contain signals according to $\gamma^{(i)}$.
 - 4: **for** $\ell \in \gamma^{(i)}$ **do**
 - 5: Let C_1, \dots, C_R denote the candidate groups in \mathcal{G} containing ℓ .
 - 6: Set $A = A \cup \{C_1, \dots, C_R\}$.
 - 7: **for** $C \in \mathcal{A}$ **do**
 - 8: If $\text{PIP}(C)$ is uninitialized, initialize $\text{PIP}(C) = 1/N$. Else, set $\text{PIP}(C) = \text{PIP}(C) + 1/N$.
 - 9: Return PIP, whose keys are the set of candidate groups with nonzero PIPs.
-

As an illustrative example, suppose $d = 2$ and consider the problem of computing PIPs for $\mathcal{G} = \{S_r(rz) : z \in \mathbb{Z}^d, r \in \{r_1, \dots, r_m\}\}$, a collection of circles of varying radii centered at lattice points. As noted previously, for any $\ell \in \mathbb{R}^2$, there are at most $5m$ circles in \mathcal{G} containing ℓ , and we can find these $R = 5m$ circles in $O(R)$ by checking the distances between ℓ and its surrounding lattice points. Furthermore, by assumption, the cost of finding the candidate groups G_1, \dots, G_{M_i} containing ℓ is $\varphi(d, R)$, where d is the dimension. Thus, Algorithm 4 runs in $O(NM\varphi(R, d))$. In this paper, $\varphi(d, R) = O(Rd)$, so Algorithm 4 runs in $O(NMRd)$ for our choices of \mathcal{G} .

Algorithm 4 is implemented in `pyblip` and `blipr` when \mathcal{G} is composed of rectangles and/or spheres, although it applies (in principle) to arbitrary shapes.

D.2 Reducing the locations to a finite set

Given a set of candidate groups \mathcal{G} and PIPs $\{p_G\}_{G \in \mathcal{G}}$, it is not immediately obvious how to run BLiP when \mathcal{L} has infinite cardinality. To start, note that the size of \mathcal{G} is not a problem, per se, for two reasons. First, when the estimated PIPs $\{\hat{p}_G\}_{G \in \mathcal{G}}$ are estimated via sample means from a finite number of samples from the posterior, then in Appendix D.1 we showed that $\{G \in \mathcal{G} : \hat{p}_G > 0\}$ has finite cardinality. Second, as discussed in Appendix B.3, we recommend adaptively preprocessing \mathcal{G} and only considering the sets $\{G \in \mathcal{G} : p_G > \epsilon\}$ for some $\epsilon > 0$. In practice, in sparse problems, choosing a reasonable ϵ (e.g., $\epsilon = 0.5$ when controlling the FDR at $q = 0.1$) ensures \mathcal{G} is not too large. And formally, the assumptions at the beginning of Appendix D imply that $\{G \in \mathcal{G} : p_G > \epsilon\}$ has finite cardinality for any $\epsilon > 0$. To see this, let ν be the measure which maps $G \subset \mathcal{L}$ to the expected number of signals in G under the posterior, so for example $M = \nu(\mathcal{L})$ is the total number of expected signals. Let f be the density of ν with respect to a base measure μ on \mathbb{R}^d . Then note that

$$\sum_{G \in \mathcal{G}} p_G \leq \sum_{G \in \mathcal{G}} \int_G f d\mu \leq R \int_{\mathcal{L}} f d\mu = RM < \infty,$$

where the second inequality follows because each point $x \in \mathcal{L}$ is contained in at most R groups in \mathcal{G} . This shows that $\{G \in \mathcal{G} : p_G > \epsilon\}$ has finite cardinality. Thus, under our assumptions, \mathcal{G} can be

infinite without causing too many problems.

However, when $|\mathcal{L}| = \infty$, it is impossible to apply BLiP as written. To see this and to ease readability, we reproduce the BLiP integer LP below:

$$\max_{\{x_G\}_{G \in \mathcal{G}}: x_G \in \{0,1\}} \sum_{G \in \mathcal{G}} p_G w(G) x_G \quad (2.4)$$

$$\text{s.t.} \quad \sum_{G \in \mathcal{G}} (1 - p_G - q) x_G \leq 0, \quad (2.5)$$

$$\sum_{G \in \mathcal{G}: \ell \in G} x_G \leq 1 \quad \forall \ell \in \mathcal{L}, \quad (2.6)$$

$$x_G \in \{0, 1\} \quad \forall G \in \mathcal{G}. \quad (2.7)$$

The penultimate line represents an infinite number of linear constraints when $|\mathcal{L}| = \infty$. Since there are only a finite number of optimization variables, clearly all but finitely many of these linear constraints are redundant! The question is how to compute a (minimal) subset of the constraints in (2.6) which are equivalent to the original constraints. Since each constraint is associated with a location $\ell \in \mathcal{L}$, this is equivalent to finding a finite subset $\mathcal{L}_0 \subset \mathcal{L}$. As the following proposition makes clear, it turns out this can be reduced to the problem of finding a (minimal) clique cover of the edges of a graph. Before proving this, we first formally define an edge clique cover.

Definition D.1 (Edge clique cover). Let (V, E) be a simple undirected graph with vertices $V = \{v_1, \dots, v_n\}$ and edges $E \subset V \times V$. A subset of the vertices $C \subset V$ is a *clique* if it is fully connected, i.e., for all $v_1 \neq v_2 \in C$, $(v_1, v_2) \in E$. A set of cliques C_1, \dots, C_k is an *edge clique cover* (ECC) of the edges if for all edges $(v_1, v_2) \in E$, there exists some C_i such that $v_1, v_2 \in C_i$.

The next proposition tells us the following: form the graph with \mathcal{G} as the vertices and an edge between $G_1, G_2 \in \mathcal{G}$ iff G_1 and G_2 intersect. Given an edge clique cover C_1, \dots, C_k of this graph, we can reduce the integer LP (2.4)-(2.7) to an equivalent integer LP with only k constraints. Proposition D.1 makes this claim precise.

Proposition D.1. *Consider the integer linear program (2.4)-(2.7). Consider the graph (V, E) where the vertices are the candidate groups, so $V = \mathcal{G}$, and $(G_1, G_2) \in E$ if and only if $G_1 \neq G_2$ and $G_1 \cap G_2 \neq \emptyset$. Let C_1, \dots, C_k be an edge clique cover of E . Then the integer linear program (2.4)-(2.7) is equivalent to the same integer LP but with Equation (2.6) replaced by the following constraints:*

$$\sum_{G \in C_i} x_G \leq 1 \quad \forall i \in [k]. \quad (D.1)$$

Proof. It suffices to show that the constraints in Equation (2.6) are equivalent to the constraints in Equation (D.1). As notation, for any $\ell \in \mathcal{L}$, let $\mathcal{G}_\ell = \{G \in \mathcal{G} : \ell \in G\}$.

For the first direction, suppose that $\{x_G\}_{G \in \mathcal{G}}$ are feasible for Equation (2.6): under this assumption, we will show $\{x_G\}_{G \in \mathcal{G}}$ satisfy Equation (D.1). In particular, fix any clique C_i . It suffices to show that $\sum_{G \in C_i} x_G \leq 1$. There are two cases here. First, if $x_G = 0$ for all $G \in C_i$, this condition is satisfied trivially. Second, suppose there exists at least one $G \in C_i$ such that $x_G = 1$. We must then show that for every other $G' \in C_i$, $x_{G'} = 0$. This is true because if $G, G' \in C_i$, there is an edge between G and G' , and thus there exists some common $\ell \in G \cap G'$. Therefore, the constraint

corresponding to ℓ in (2.6) guarantees that $x_G + x_{G'} \leq 1$, so therefore $x_{G'} = 0$. This completes the proof of the first direction.

For the second direction, suppose that $\{x_G\}_{G \in \mathcal{G}}$ are feasible for Equation (D.1). To show $\{x_G\}_{G \in \mathcal{G}}$ are also feasible for Equation (2.6), it suffices to show that for any $\ell \in \mathcal{L}$, $\sum_{G \in \mathcal{G}: \ell \in G} x_G = \sum_{G \in \mathcal{G}_\ell} x_G \leq 1$. (The equality is by definition of \mathcal{G}_ℓ .) To see this, once again, there are two cases. The first case is that $x_G = 0$ for all $G \in \mathcal{G}_\ell$, in which case the inequality holds trivially. The second case is that there exists at least one $G \in \mathcal{G}_\ell$ such that $x_G = 1$. In this case, it suffices to show that for any other $G' \in \mathcal{G}_\ell$, $x_{G'} = 0$. However, because G, G' both contain ℓ , they form an edge in the graph (V, E) . Therefore, there exists a clique C_i such that $G, G' \in C_i$. This guarantees that $x_G + x_{G'} \leq 1$, and since $x_G = 1$, this implies $x_{G'} = 0$. This completes the proof. \square

This proposition motivates the following strategy for applying BLiP. First, we form the graph (V, E) by checking whether each pair of candidate groups $G_1, G_2 \in \mathcal{G}$ intersect. A naive algorithm which loops over all pairs G_1, G_2 already runs in polynomial time of the size of \mathcal{G} , specifically $O(|\mathcal{G}|^2)$. The naive algorithm proved to run quickly in our applications, but in large-scale applications, it is likely possible to use spatial tree-structures such as (for example) R-trees (Brinkhoff et al., 1993) to compute (V, E) substantially more efficiently, although we did not explore that possibility. Second, given the graph (V, E) , we find an edge clique cover of (approximately) minimal size and use it to form a finite-dimensional version of BLiP, following Proposition D.1.

Unfortunately, for any k_0 , the decision problem asking whether or not there exists an ECC C_1, \dots, C_k with $k \leq k_0$ is known to be NP complete (Orlin, 1977). This means that there is no known polynomial-time algorithm which can find a *minimal* ECC. However, we do not need to find an optimal solution: we only need to find a “good” solution where k is small. This is certainly possible in polynomial time. At worst, a trivial edge clique cover of E is simply the set of cliques corresponding to all the edges in E , e.g., $\{(v_1, v_2) : (v_1, v_2) \in E\}$, since these are cliques of size 2. This shows that it is always possible to find *an* edge clique cover of size $|E|$ in polynomial time, although in many problems, we can do much better. For example, in our astronomical application, after preprocessing, $|\mathcal{G}| \approx 26,000$ and the graph formed by \mathcal{G} has over 1,100,000 edges. Despite this, we are able to find an edge clique cover of size $\approx 12,000$, a 100x improvement.

There are known polynomial time heuristics for the minimal ECC problem (Kou et al., 1978; Gramm et al., 2006), but we found it easier to implement a simple iterative algorithm (Algorithm 5), and we found it worked well. Algorithm 5 does the following. Given (V, E) , we iterate through the edges $e \in E$ and maintain a running list of cliques C_1, \dots, C_k as well as a set of “uncovered” edges E' . For each $e \in E$, if e is uncovered ($e \in E'$), then we pick a maximal clique C_{k+1} containing e and add it to our list C_1, \dots, C_k and remove all edges in C_{k+1} from E' . Note that we try to pick C_{k+1} to contain as many uncovered edges as possible using a heuristic based on the degrees of the graph (V, E') . Otherwise, if e has already been covered, we move on to the next edge.

The worst-case complexity of Algorithm 5 is $O(|V|^2|E|)$, since the cost of adding any clique is at most $O(|V|^2)$, and we add at most one clique per edge. Of course, when the degree-based heuristic performs well, adding a clique of size k requires $O(|V|k)$ operations and will cover $k - 1$ uncovered edges, yielding a complexity of $O(|V|^2)$. In our astronomical application, Algorithm 5 ran in only a few seconds when $|V| = 26,000$, $|E| = 1,100,000$.

Algorithm 5 Polynomial-time heuristic to find minimal edge clique covers.

Input: A graph (V, E) with $V = \{1, \dots, n\}$.

- 1: Check that (V, E) has one connected component. Else, apply the algorithm separately to each connected component.
 - 2: Initialize $E' = E$, $\mathcal{C} = \{\}$.
 - 3: Initialize $D[1], \dots, D[n]$ to be the degrees of vertices $1, \dots, n$.
 - 4: **for** $e = (v_1, v_2) \in E$ **do**
 - 5: If $e \notin E'$, proceed to the next edge e .
 - 6: Else, initialize a new clique $C = \{v_1, v_2\}$.
 - 7: Let $N_C = \{v \in V : D[v] > 0, (v_1, v) \in E, (v_2, v) \in E\}$ denote the common neighbors of C with nonzero degree in the uncovered graph (V, E') .
 - 8: **while** $|N_C| > 0$ **do**
 - 9: Let $v^* = \arg \max_{v \in N_C} D[v]$.
 - 10: Set $C = C \cup \{v^*\}$ and update $N_C = N_C \cap \{v : (v, v^*) \in E, D[v] > 0\}$.
 - 11: **for** $(v, v') \in C$ **do**
 - 12: Set $E' = E' \setminus (v, v')$
 - 13: Set $D[v] = D[v] - 1$, $D[v'] = D[v'] - 1$.
 - 14: Set $\mathcal{C} = \mathcal{C} \cup \{C\}$.
 - 15: Return \mathcal{C} , an edge clique cover of (V, E) .
-

E Further simulations and simulation details

E.1 Our choice of candidate groups

In this section, we describe our method for choosing the candidate groups \mathcal{G} , which we use in all simulations in the paper except those in Appendix F. To introduce notation, recall that all of our simulation settings can be represented as a sparse (Bayesian) GLM. Let \mathbf{y}, \mathbf{X} be the data and let $\beta \in \mathbb{R}^p$ be the linear coefficients, where nonzero coefficients correspond to signals. The set of locations is $\mathcal{L} = \{1, \dots, p\}$. Furthermore, in each setting, we use various methods to compute samples $\tilde{\beta}^{(1)}, \dots, \tilde{\beta}^{(N)} \in \mathbb{R}^p$ approximating the posterior of β . (The exception to this is when we run BLiP on top of SuSiE, which we will discuss at the end of this section.) Given these samples, let $\epsilon_{ij} = \mathbb{I}(\tilde{\beta}_j^{(i)} \neq 0)$ be the indicator of whether there is a signal at location j in posterior sample i , forming the binary matrix $\epsilon \in \{0, 1\}^{N \times p}$. As a reminder, given any group $G \subset \mathcal{L}$, the estimated PIP p_G is

$$p_G \approx \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\exists j \in G : \epsilon_{ij} = 1).$$

Finally, let $\hat{\Sigma} = \frac{1}{n} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ be the empirical correlation matrix of \mathbf{X} , where $\tilde{\mathbf{X}}$ is \mathbf{X} but shifted and scaled so the columns of $\tilde{\mathbf{X}}$ sum to zero and have unit variance. Similarly, let $\hat{\Sigma}_\epsilon$ denote the empirical correlation matrix of ϵ .

Below, we specify how we choose the candidate groups.

1. First, for some choice of $\kappa > 0$ (we will specify κ in step 4), we adaptively pre-filter the locations to obtain $\mathcal{L}_\kappa = \{\ell \in \mathcal{L} : p_{\{\ell\}} \geq \kappa\} \subset \mathcal{L}$. Note that we can represent $\mathcal{L}_\kappa = \{\ell_1, \dots, \ell_k\}$ for some k such that $\ell_1 < \ell_2 < \dots < \ell_k$.

2. Second, we consider all *contiguous* sub-groups of \mathcal{L}_κ of maximum size less than m : recall that a contiguous group is a group of the form $\{\ell_i, \ell_{i+2}, \ell_{i+3}, \dots, \ell_{i+m_0}\}$ where we restrict $m_0 \leq m$, so there are $O(mk)$ contiguous subgroups. By default, we pick $m = 25$ unless specified otherwise. Let $\mathcal{G}_\kappa^{\text{seq}}$ denote these subgroups.
3. Third, we hierarchically cluster \mathcal{L}_κ based on a dissimilarity metric $d : \mathcal{L}_\kappa \times \mathcal{L}_\kappa \rightarrow \mathbb{R}_+$. For each dissimilarity metric, we apply single-linkage, average-linkage, and complete-linkage clustering, yielding three tree-structured sets of candidate groups $\mathcal{G}_{\kappa,d}^{\text{single}}, \mathcal{G}_{\kappa,d}^{\text{average}}, \mathcal{G}_{\kappa,d}^{\text{complete}}$. We also restrict these sets to have cardinalities of m or less. We repeat this for two dissimilarity metrics, one based on \mathbf{X} where $d_1(i,j) = |1 - \hat{\Sigma}_{i,j}|$ and a second based on ϵ where $d_2(i,j) = 1 + (\hat{\Sigma}_\epsilon)_{i,j}$.
4. We repeat this process for $\kappa \in \mathcal{K}$ where we pick $\mathcal{K} = \{0, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2\}$. Then, the final set of candidate groups is

$$\mathcal{G} = \bigcup_{\kappa \in \mathcal{K}} \mathcal{G}_\kappa^{\text{seq}} \cup \left(\bigcup_{d \in \{d_1, d_2\}} \mathcal{G}_{\kappa,d}^{\text{single}} \cup \mathcal{G}_{\kappa,d}^{\text{average}} \cup \mathcal{G}_{\kappa,d}^{\text{complete}} \right).$$

After creating \mathcal{G} , we adaptively preprocess \mathcal{G} and only include candidate groups G such that $p_G \geq 0.5$. Since we control the FDR at level $q = 0.1$, it is unlikely that including groups with $p_G < 0.5$ would lead to higher power (Appendix B.3).

Lastly, if we are applying BLiP on top of SuSiE, we use the same procedure, with two exceptions. First, we add the groups discovered by the default SuSiE algorithm to the candidate groups. Second, naturally, we do not include the hierarchical groups generated using the covariance matrix $\hat{\Sigma}_\epsilon$, since SuSiE does not produce posterior samples ϵ . See Appendix C.4 for more discussion of how to compute PIPs when working with outputs from SuSiE.

E.2 Details for variable selection simulations

In this section, we describe the data-generating process and the methods used for all variable selection simulations in the paper. To be clear, all of the descriptions in this section apply to Figures 1, 2, 3, 10, 11, 12, 13 and 15. We also include the corresponding power and FDR plots for Figure 3, which show that the power of SuSiE decreases compared to other methods as the absolute number of signals grows. All code for these simulations is available at https://github.com/amspector100/blip_sims.

To start, we describe the data-generating processes for all figures involving variable selection.

1. Throughout, we sample $X \in \mathbb{R}^p$ to follow a non-stationary AR(k) process. In particular, to sample X , we sample $Z_1, \dots, Z_p \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and initialize $X_1 = Z_1$. We then iteratively set

$$X_j = \rho_0^{(j)} Z_j + \sum_{\ell=1}^{\min(j-1,k)} \rho_\ell^{(j)} X_{j-\ell}.$$

For each j , we sample $(\rho_0^{(j)}, \dots, \rho_k^{(j)}) \stackrel{i.i.d.}{\sim} \text{Dirichlet}((0.2, \frac{0.8}{k-1}, \dots, \frac{0.8}{k-1}))$, and then we rescale $(\rho_0^{(j)}, \dots, \rho_k^{(j)})$ so that $\text{Var}(X_j) = 1$. This sampling structure was motivated by fine-mapping

Example of AR5 covariance matrix ($p=10$)										
0	1	0.7	0.88	0.84	0.33	0.91	0.75	0.81	0.49	0.42
1	-0.7	1	0.95	0.97	0.38	0.92	0.85	0.89	0.56	0.47
2	-0.88	0.95	1	0.99	0.39	0.99	0.88	0.93	0.58	0.49
3	-0.84	0.97	0.99	1	0.39	0.98	0.88	0.93	0.58	0.49
4	-0.33	0.38	0.39	0.39	1	0.39	0.76	0.68	0.65	0.66
5	-0.91	0.92	0.99	0.98	0.39	1	0.87	0.92	0.57	0.49
6	-0.75	0.85	0.88	0.88	0.76	0.87	1	0.98	0.73	0.66
7	-0.81	0.89	0.93	0.93	0.68	0.92	0.98	1	0.7	0.63
8	-0.49	0.56	0.58	0.58	0.65	0.57	0.73	0.7	1	0.6
9	-0.42	0.47	0.49	0.49	0.66	0.49	0.66	0.63	0.6	1
	0	1	2	3	4	5	6	7	8	9

Figure 14: An example of the AR5 covariance matrix we used in variable selection simulations.

studies, since it creates very strong local dependencies. Figure 14 gives an example of what the covariance matrix looks like with $k = 5$ and $p = 10$. In all simulations, we choose $k = 5$ unless otherwise specified in the plot.

2. In all regression settings, for simulations with sparsity $s \in [0, 1]$, we set $\beta \in \mathbb{R}^p$ to equal all zeroes except for $\lceil sp \rceil$ randomly chosen coefficients. The nonzero coefficients are sampled as i.i.d. $\mathcal{N}(0, \tau^2)$ random variables, except that we ensure all nonzero coefficients have absolute values greater than 0.1τ . We do this so that the resolution-adjusted power of each method approaches $\lceil sp \rceil$ with a reasonable sample size (otherwise, there will be a few coefficients which are extremely close to zero, and no method will detect them without $n \gg p$). We sample $\mathbf{y} | \mathbf{X}, \beta \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ with $\sigma^2 = 1$. In the linear regression simulations, we observe (\mathbf{X}, \mathbf{y}) : in the probit simulations, we only observe (\mathbf{X}, \mathbf{z}) where $\mathbf{z} = \mathbb{I}(\mathbf{y} \geq 0)$. The only parameter which changes between figures is τ^2 . We picked $\tau^2 = 0.5$ in Figures 1, 4, and 11, and $\tau^2 = 1$ otherwise.

Next, we describe the methods we compare between in more detail.

1. **LSS + BLiP:** For the method LSS + BLiP, we use the LSS sampler described in Section C.2 with the true values of σ^2, τ^2, p_0 . This means that we do not use hyperpriors. The LSS + BLiP (misspec.) method is identical except that we use the hyperpriors specified in Section C.2 with $a_\sigma = 2, b_\sigma = 1, a_\tau = 2, b_\tau = 1, a_0 = 1, b_0 = 1$ and $p_{\min} = 0.9$. Note that these are fairly uninformative priors. For the Gibbs update, we use a block size of 5 for all simulations except Figures 1 and 3, where we use a block size of 1. We run $N = 10000$ samples per chain with 10 chains, unless otherwise specified. After fitting these samplers, we apply BLiP to the candidate groups described in Appendix E.1.
2. **PSS + BLiP:** The methods PSS + BLiP and PSS + BLiP (misspec.) are identical to the corresponding LSS-based methods except that we use the probit sampler with the extra data-augmentation step described in Appendix C.3.
3. **SuSiE-based methods:** In all simulations, we apply SuSiE with the default parameters in the susieR package (Wang et al., 2020) except that we provide SuSiE the oracle value for the number of signals. For SuSiE + BLiP, we use the same SuSiE model and extract PIPs from SuSiE as described in Appendix C.4.

4. **DAP-G**: In all instances, we apply DAP-G to Z-statistics based on \mathbf{X}, \mathbf{y} . We run DAP-G with the default parameters in the package from Lee et al. (2018) except that we give DAP-G oracle knowledge of the sparsity s . For the sake of computation, we tell DAP-G the exact number of non-nulls, so that it does not search over models with more than $\lceil sp \rceil$ parameters.
5. **dCRT+FBH and dCRT+Yekutieli**: For both of these methods, we apply the d_0 CRT based on a cross-validated lasso with analytical p-values as described in Liu et al. (2021a). The FBH and Yekutieli methods apply to tree-structured p-values, so to create the tree, we hierarchically cluster the covariates with average linkage based on the empirical correlation matrix of \mathbf{X} . We cut the correlation tree at 10 evenly spaced values between 0 (which corresponds to individual covariates) and the first correlation at which the maximum group size exceeds $m = 25$, the same maximum groupsize that we apply with BLiP. We cut the correlation tree at 10 levels for two reasons. First, we chose the number of levels to approximately balance the benefit of more levels (high adaptivity) against the drawback, namely that it makes the FBH more conservative by increasing the multiplicity correction. Note we were only able to do this because we could simulate from the true data-generating process, and in general it is not clear how one would choose the number of levels. Second, computing many d CRT p-values is very expensive, and cutting the tree at (e.g.) 100 different levels would have made these procedures computationally intractable. As is, the dCRT-based methods were 100x more expensive than the Bayesian samplers in Figure 2.

It may also be helpful to review exactly how we extended the d_0 CRT of Liu et al. (2021a) to test group hypotheses. As a short review, the dCRT is used to test the conditional independence hypotheses $H_j : X_j \perp\!\!\!\perp Y | X_{-j}$, where in a generalized linear model this is equivalent to testing $H_j : \beta_j = 0$ under mild conditions on the distribution of X . The dCRT yields valid p-values under the model-X assumption (Candès et al., 2018), which assumes knowledge of the distribution of X . (In contrast, the Bayesian procedures, including BLiP, assume a full Bayesian model of $Y | X$ but no knowledge of X .) For fair comparison, we give the dCRT oracle knowledge of the distribution of X .

Consider the problem of obtaining a d_0 CRT-based p-value for the group null hypothesis $H_G : \bigcap_{j \in G} H_j$, for any $G \subset [p]$. We will describe this in the linear regression case first and then generalize to the probit case. The first step is identical to the normal d_0 CRT, which is to compute \mathbf{d}_y , the *distillation* of \mathbf{y} . Intuitively, \mathbf{d}_y is a proxy of $\mathbb{E}[\mathbf{y} | \mathbf{X}_{-G}]$, although it can be any function of $\mathbf{X}_{-G}, \mathbf{y}$ while maintaining Type I error control. We choose \mathbf{d}_y to be predictions from a cross-validated lasso fit on $\mathbf{y}, \mathbf{X}_{-G}$. The next step is to compute analytical p-values based on $\mathbf{y}, \mathbf{d}_y, \mathbf{X}_G$. In particular, in our setting, $X \sim \mathcal{N}(0, \Sigma)$ is multivariate Gaussian. As a result, $X_G | X_{-G} \sim \mathcal{N}(A_{-G}X_{-G}, \Sigma_G)$ where $A_{-G} \in \mathbb{R}^{|G| \times (p-|G|)}$ and $\Sigma_G \in \mathbb{R}^{|G| \times |G|}$ are known quantities depending on Σ . Under the null H_G , this is also true conditional on \mathbf{y}, \mathbf{d}_y . Let $\mathbf{x}_G = A_{-G}\mathbf{X}_{-G}$. Then, under the null, we have that

$$(\mathbf{X}_G - \mathbf{x}_G)^T (\mathbf{y} - \mathbf{d}_y) | \mathbf{y}, \mathbf{X}_{-G} \sim \mathcal{N}(0, \|\mathbf{y} - \mathbf{d}_y\|_2^2 \Sigma_G)$$

and therefore if we apply the whitening transformation $\Sigma_G^{-1/2}$, we can use the test statistic

$$T_G := \|((\mathbf{X}_G - \mathbf{x}_G)\Sigma_G^{-1/2})^T (\mathbf{y} - \mathbf{d}_y)\|_2^2$$

where $T_G | \mathbf{X}_{-G}, \mathbf{y} \sim \chi_{|G|}^2$ under the null. Therefore, the p-value for group G is simply $1 - F_{\chi_{|G|}^2}(T_G)$, where $F_{\chi_{|G|}^2}$ is the CDF of a $\chi_{|G|}^2$ random variable. The p-value for the probit

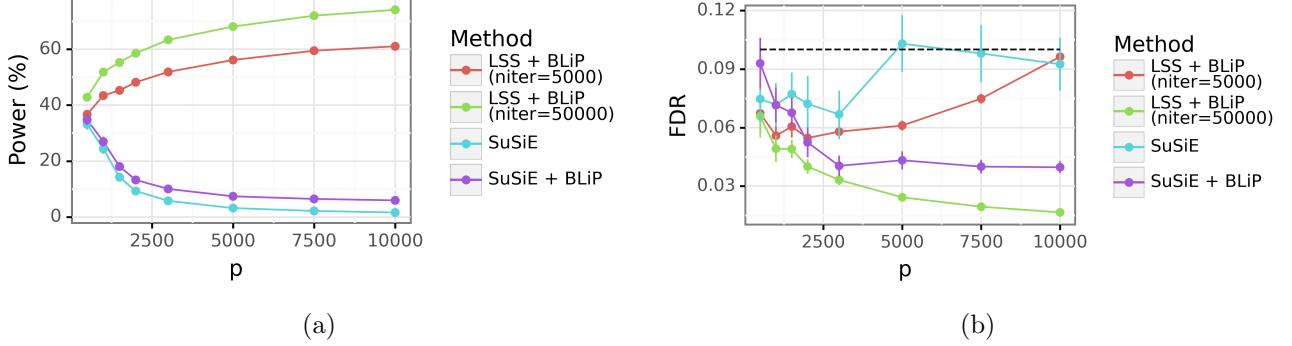


Figure 15: The corresponding power and FDR plots of Figure 3. Panel (a) shows SuSiE loses power compared to the other methods as p grows, even as n/p and the signal size are held constant. Panel (b) shows that all methods controlled the FDR up to Monte Carlo error.

case is identical except we replace \mathbf{y} with \mathbf{z} and apply a cross-validated logistic lasso instead of a lasso during the distillation.

Note that we do not use the pre-screening procedure from Liu et al. (2021a) in either the linear or probit case. We made this choice because pre-screening reduced power empirically, perhaps because the very high correlations in \mathbf{X} make it very easy to accidentally screen out signal variables.

Lastly, Figure 15 is the corresponding power and FDR plot for Figure 3.

E.3 Simulations for change point detection

In this section, we show that BLiP has potential as a method for detecting change points in a time series $\{Y_1, \dots, Y_T\}$. We consider a model where $Y_t \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_t, \sigma^2)$ and $\mu_t = \mu_{t-1} + \beta_t$ where β_t is a mixture of a point mass at zero and a $\mathcal{N}(0, \tau^2)$ distribution. We chose this model for sake of simplicity and because it has been frequently discussed in prior work, such as in Wang et al. (2020); Fang and Siegmund (2020); Cappello et al. (2021), although BLiP could easily wrap on top more complex models, for example those which account for temporal dependence. As noted by Wang et al. (2020), this problem can be described as a sparse regression problem where $\mathbf{X} \in \mathbb{R}^{T \times T}$ is defined so that $\mathbf{X}_j = (\mathbf{1}_{T-j+1}, \mathbf{0}_{T-1}) \in \mathbb{R}^T$. As illustrated by Figure 16, BLiP will output regions where it detects at least one change point, and it makes these regions as narrow as possible while controlling the FDR. As discussed in Section 1.4, there are several frequentist methods which can accomplish this while controlling the FWER (Frick et al., 2014; Fang et al., 2016; Fang and Siegmund, 2020; Fryzlewicz, 2020). Similarly, Li et al. (2016) control a slightly modified FDR, although they do not aim to precisely localize signals. Since these methods are not directly comparable to BLiP while controlling the FDR, our simulations only compare against SuSiE.

To illustrate one possible advantage of BLiP, we first consider a toy example from Wang et al. (2020), where there is one very large change point at $T = 50$ which immediately is cancelled out by another change point at $T = 52$. Since SuSiE is a greedy method and adding a single change point never improves the model fit, SuSiE discovers no change points. In contrast, LSS + BLiP overcomes this optimization hurdle and correctly identifies the two change points, as in Figure 16. We emphasize that this toy example is motivated by real applications, where (for example) the mean of an economic indicator might temporarily jump before reverting to its previous value (Fang

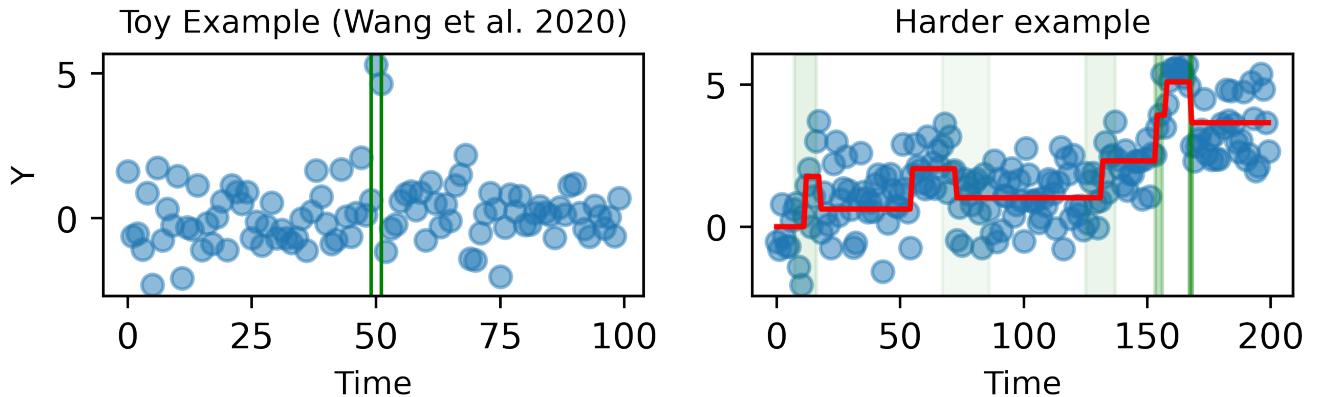


Figure 16: Two examples of LSS + BLiP applied to change point detection. In both cases, the blue points represent $\{Y_t\}$ and the green regions represent detections from LSS + BLiP. The left example is a setting from Wang et al. (2020) where SuSiE cannot escape a local minimum and makes no discoveries, but LSS + BLiP discovers both change points. The right example shows BLiP’s output in a more challenging setting. Note in the right example, the discontinuities in the red line denote true change points.

and Siegmund, 2020). Of course, as discussed in Wang et al. (2020), SuSiE can perform very well as a change point detection algorithm in other settings, and in these cases, SuSiE + BLiP can slightly improve SuSiE’s power.

For more comprehensive simulations, we set $T = 300$ with $\lceil sT \rceil$ change points and vary the signal size τ^2 and the sparsity s . To capture this “cancellation” effect, we chose the change points so that the distances between consecutive change points are distributed as i.i.d. $\text{Unif}([2d])$ random variables. Intuitively, d represents the “spacing” or average distance between change points. We consider six methods. To start, as discussed in Section E.3, this problem can be represented as a sparse linear regression problem, so we apply the methods SuSiE, SuSiE + BLiP, LSS + BLiP, and LSS + BLiP (misspec.) from the linear regression setting in Section 3.1. We use the exact same parameters as described in Appendix E.2 for these methods. Second, we apply BLiP on top of posterior samples from the BCP method from Wang and Emerson (2015). We use the default samples from the BCP package except in the well-specified setting, where we use oracle values for the parameters governing the sparsity level and signal size. For all methods involving BLiP, we use the candidate groups described in Appendix E.1.

As shown in Figure 17, for change points which are close together, BCP + BLiP and LSS + BLiP are often much more powerful than the SuSiE-based methods. That said, when the change points are fairly spaced out, SuSiE performs quite well, although the BLiP-based methods do give a slight boost in power. All methods controlled the FDR up to Monte Carlo error, as shown by Figure 18.

Next, we consider the same setting as before, but with an additional twist: instead of sampling the nonzero coefficients as i.i.d. $\mathcal{N}(0, \tau^2)$ random variables, we specify that at each change point t where $\mu_t \neq 0$, with probability r we set $\mu_{t+1} = 0$, i.e., the process reverts to having mean zero. This is a challenging setting because consecutive change points may cancel out previous change points, causing change point detection algorithms to get stuck in local maxima. However, Figure 19 shows that BLiP-based methods perform comparatively well in this setting.

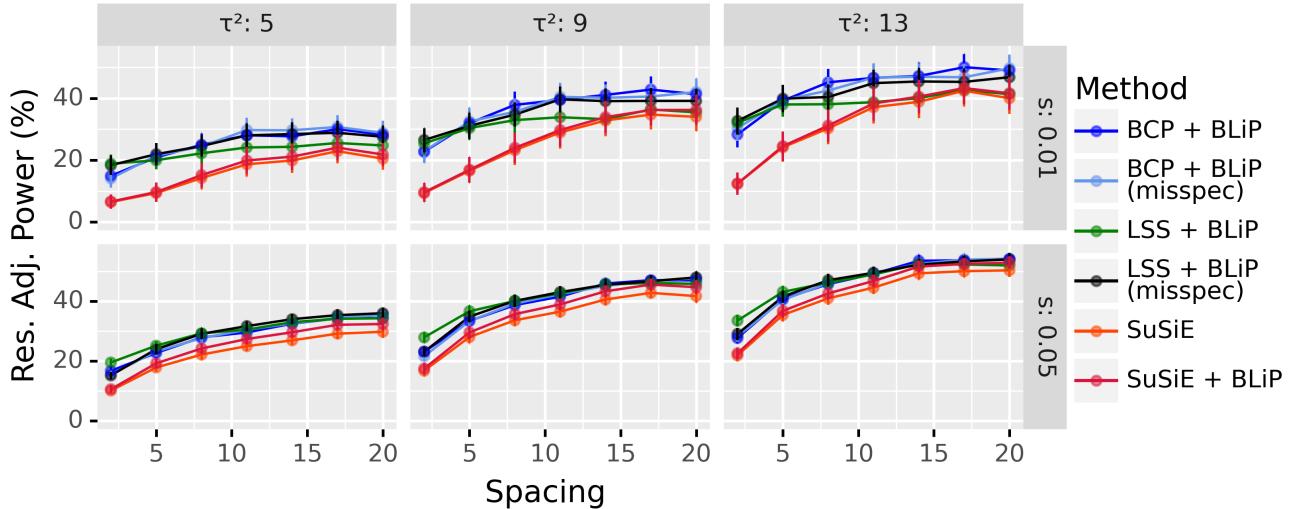


Figure 17: We apply LSS + BLiP, BCP + BLiP, SusiE, and SuSiE + BLiP to perform change point detection in a Gaussian spike and slab model, as described in Section E.3. The x-axis denotes the average spacing between consecutive change points.

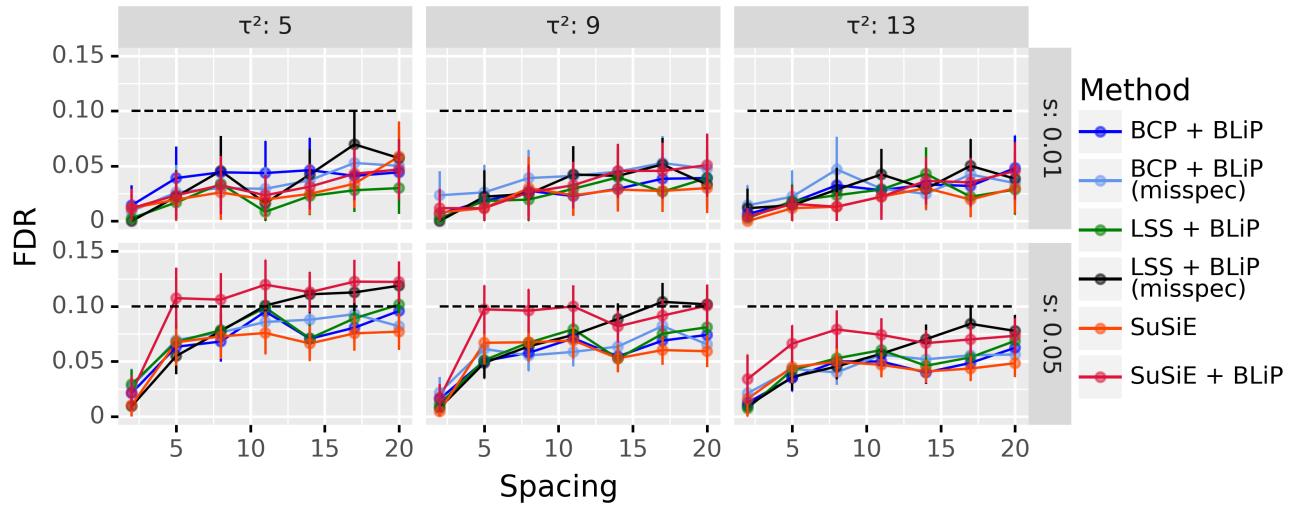


Figure 18: The corresponding FDR plot of Figure 17.

F Further details on Section 4

F.1 Methodological details

In this section, we describe our fine-mapping analysis in detail. As a review, Weissbrod et al. (2019) split the genome into $m = 2763$ loci spanning 3 mega-base-pairs (Mb) each. For each of these loci, Weissbrod et al. (2019) fit SuSiE models with $L = 10$ causal SNPs, yielding $m = 2763$ SuSiE outputs $\alpha^{(1)} \in \mathbb{R}^{p_1 \times 10}, \dots, \alpha^{(m)} \in \mathbb{R}^{p_m \times 10}$, where p_i denotes the number of SNPs in locus i and $\alpha^{(i)}$ denote the matrix of SuSiE PIPs as described in Appendix C.4. Please see Weissbrod et al. (2019) for further details on these SuSiE models and the generation of the loci. Note that Weissbrod et al. (2019) fit SuSiE models both with uninformative priors and with (novel) priors based on functional annotation data. For simplicity, we use the SuSiE models with uninformative priors, although more sophisticated priors could improve the power of SuSiE and SuSiE + BLiP.

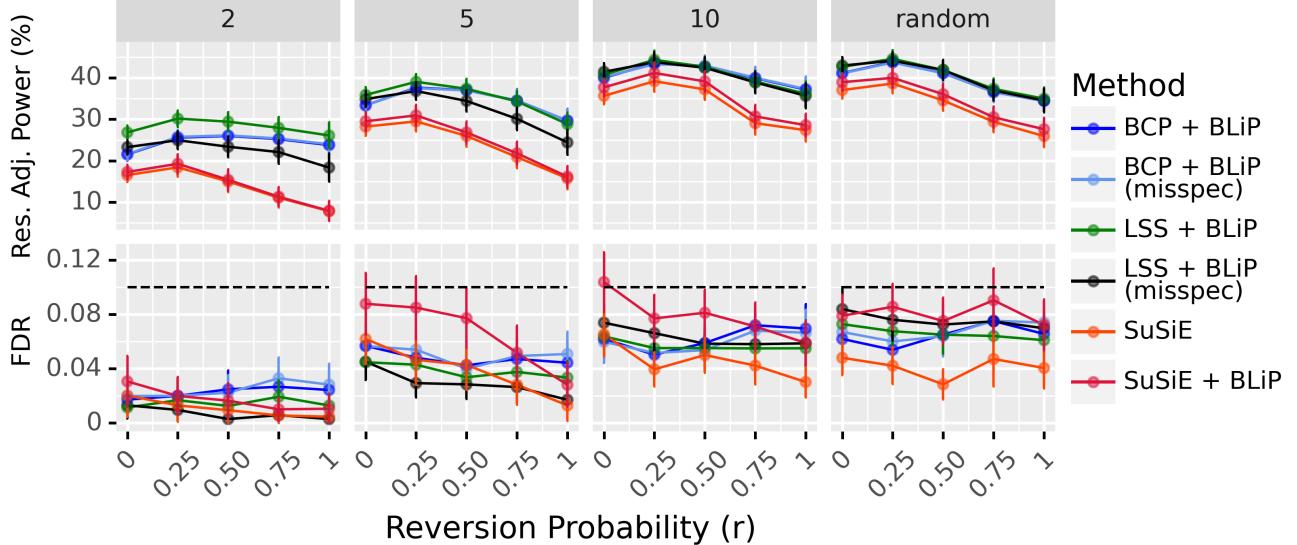


Figure 19: This plot shows the same setting as Figure 17 with $\tau^2 = 9$ and $s = 0.05$, with one exception. At each change point, with probability r , the time series reverts to a global mean of zero. When r is high, this causes many successive change points to cancel each other out, which can trap change point detection algorithms in local maxima. In particular, SuSiE has comparatively lower power than the other methods in this setting. The y-panel shows the average spacing between change points (“random” spacing means the locations of the change points are chosen uniformly at random).

For each $\alpha^{(i)}$, we generated the candidate groups as follows. To speed up computation, we pre-filter the SNPs to exclude SNPs from consideration unless they have a posterior inclusion probability of at least 1%, since it is quite unlikely we would discover any groups containing those SNPs. Formally, if $\mathcal{L}^{(i)}$ denotes the set of SNPs in locus i , let $\mathcal{L}_\kappa^{(i)} = \{\ell \in \mathcal{L}^{(i)} : p_{\{\ell\}} > 0.01\}$. Then, we generate the set of all contiguous subgroups of $\mathcal{L}_\kappa^{(i)}$ of maximum size less than 25, as described in Appendix E.1. For good measure, we also include the groups discovered by the default SuSiE model as candidate groups. Let $\mathcal{G}^{(i)}$ denote these candidate groups.

Given the candidate groups $\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(m)}$ for each locus, let $\mathcal{G} = \bigcup_{1 \leq i \leq m} \mathcal{G}^{(i)}$ denote the set of candidate groups for the whole genome. As discussed in Appendix B.3, we pre-filter \mathcal{G} to only include groups G where the PIP is greater than 75%, so $G \in \mathcal{G} \implies p_G \geq 0.75$. We then apply BLiP to \mathcal{G} to control the FDR at level $q = 0.05$. This yields the results described in Section 4.

F.2 Simulations on real genotype data

In this section, we describe the simulations we performed using the real LD matrix from the fine-mapping analysis.

We create simulated data using exactly the same set-up as Weissbrod et al. (2019). As notation, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the genetic covariate data for a fixed locus, so $n \approx 337,000$ and p ranges between 5,000 and 50,000. Suppose that \mathbf{X} has been scaled so its columns have mean zero and unit variance, let $\mathbf{y} \in \mathbb{R}^n$ be the phenotype data scaled so that $\mathbb{E}[\mathbf{y}^T \mathbf{y}] \approx n$, and let $R \approx \frac{1}{n} \mathbf{X}^T \mathbf{X}$ be the reported LD matrix from Weissbrod et al. (2019). As in Weissbrod et al. (2019), we assume a Gaussian linear model where $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ and $\sigma^2 = 1 - \beta^T \mathbf{X}^T \mathbf{X} \beta := 1 - h_g^2$, so h_g^2 is the locus-specific heritability. Note that R is not exactly equal to $\frac{1}{n} \mathbf{X}^T \mathbf{X}$, but in the fine-mapping

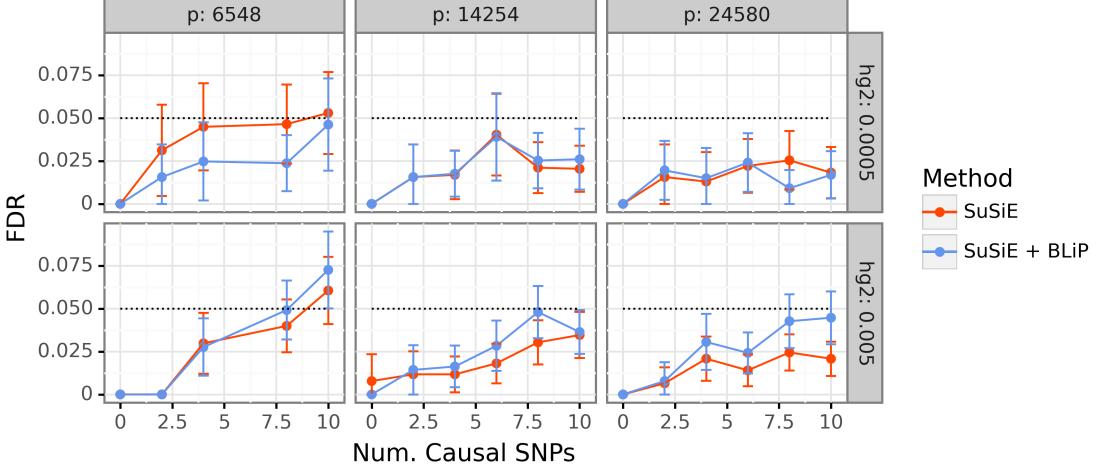


Figure 20: This figure shows the FDR of SuSiE and SuSiE + BLiP in the fine-mapping simulations described in Appendix F.2. It shows that both methods control the FDR (up to Monte Carlo error) in a variety of settings, even when the number of causal SNPs is smaller than 10.

literature, it is common to assume that $\frac{1}{n} \mathbf{X}^T \mathbf{y} \sim \mathcal{N}(R\beta, \sigma^2 R/n)$, and Weissbrod et al. (2019) make this assumption. Initially, the raw LD matrix R may not be positive semidefinite (PSD), but we use the same procedure as Weissbrod et al. (2019) to perturb R to be positive definite while changing R as little as possible.

For a fixed value of h_g^2 and a fixed number of causal SNPs L , we run simulations as follows. First, we randomly choose L nonzero coefficients of β and sample their values as $\mathcal{N}(0, 1)$ random variables. Then, we scale β such that $\beta R \beta = h_g^2$. Second, for this β , we sample the sufficient statistics $\frac{1}{n} \mathbf{X}^T \mathbf{y} \sim \mathcal{N}(R\beta, \sigma^2 R/n)$. Third, we fit a SuSiE model based on the sufficient statistics $(\frac{1}{n} \mathbf{X}^T \mathbf{y}, R)$ using the algorithm from Zou et al. (2021), and we run BLiP on top of SuSiE as specified in Appendix F.1. Note that when fitting the SuSiE model, we follow Weissbrod et al. (2019) and use a modified HESS estimator to estimate the prior variance of the non-null coefficients—see Weissbrod et al. (2019) for details. Then, we evaluate the resolution-adjusted power and FDR of SuSiE and SuSiE + BLiP. We repeated this process for 128 replications on three loci from chromosomes 1, 10, and 12, chosen because they represent various values of p . The results are shown in Figures 20 and 21 for various values of h_g^2 and L , where we adapted our values of h_g^2 and L from Weissbrod et al. (2019). Note that in all cases, we fit the SuSiE model assuming there are $L = 10$ causal variants, even when in truth there are fewer. As expected, the figures show that BLiP improves the power of SuSiE and that both methods successfully control the FDR.

That said, the power boost from BLiP in these simulations is smaller than in the real analysis. We suspect this is because these simulations are locus-specific, whereas in the real analysis, we apply BLiP to the whole genome. Since there are vastly more candidate groups when working on the whole genome, the optimization problem of which groups to discover is much harder, and it is much more likely that BLiP can substantially improve upon the heuristic solution from SuSiE. In contrast, when working with only one 3Mb locus, it seems that SuSiE comes close to finding the optimal solution. Admittedly, this difference suggests that these simulations are not perfectly representative of the real analysis, and we would prefer to have done simulations on the entire genome. Unfortunately, these locus-specific simulations were already quite computationally expensive, so a genome-wide simulation was not possible. (Note Weissbrod et al. (2019) did not perform a genome-wide simulation either.)

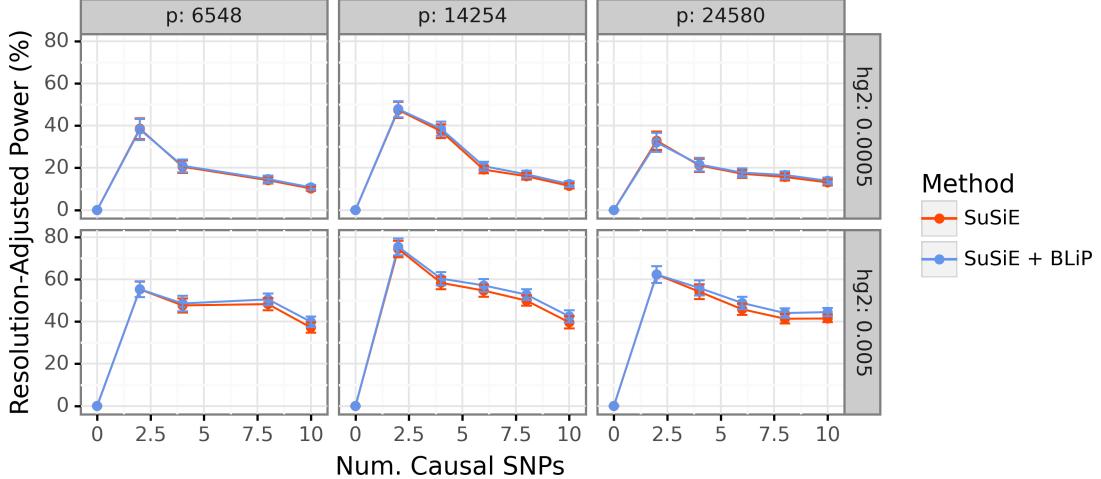


Figure 21: The corresponding power plot for Figure 20—see Appendix F.2 for a discussion.

Trait	SNP	PIP (Farh et al. (2015))	Replicated by SuSiE + BLiP?
HDL	rs581080	0.978	No
HDL	rs964184	1.000	Yes
HDL	rs12229654	1.000	No
HDL	rs2074356	1.000	No
HDL	rs1532085	0.986	No
HDL	rs1800961	1.000	Yes
LDL	rs11591147	1.000	Yes
LDL	rs964184	1.000	Yes

Table 1: This table shows the eight significant PIPs from Farh et al. (2015) for the traits HDL cholesterol and LDL direct. It also shows whether SuSiE + BLiP replicated these findings. Note that in each case where SuSiE + BLiP replicated one of these findings, SuSiE + BLiP discovered the individual SNP (by coincidence SuSiE + BLiP did not discover groups of size larger than 1 containing any of these SNPs).

F.3 Comparison to previous studies

In this section, we compare our findings to those of Farh et al. (2015) and Weissbrod et al. (2019). First, we replicate four of the eight findings from Farh et al. (2015), who analyzed the HDL cholesterol and LDL direct traits using a separate dataset. Note that SuSiE + BLiP makes far more than eight discoveries for these two traits, likely due to the fact that the SuSiE model we use from Weissbrod et al. (2019) was fit on a much larger dataset than the one used in Farh et al. (2015).

Second, we compare the results from SuSiE + BLiP to those of SuSiE alone, using the SuSiE model from Weissbrod et al. (2019). Note that this is certainly **not** a replication analysis, since the dataset is the same for both analyses—rather, it is a sanity check to ensure that SuSiE + BLiP does not completely change the findings from Weissbrod et al. (2019). For each of the four traits we analyze, Table 2 shows the degree of overlap between SuSiE and SuSiE + BLiP. To be precise, we say that the proportion of overlap for SuSiE is the proportion of discoveries G made by SuSiE which overlap with at least one other group G' discovered by SuSiE + BLiP, and analogously for SuSiE + BLiP. For example, for the cardiovascular trait, the first row of Table 2 shows that 100% of the discoveries made by SuSiE overlapped with at least one discovery made by SuSiE + BLiP, and 81.4% of the discoveries made by SuSiE + BLiP overlapped with at least one discovery made by

Trait	Prop. Overlap for SuSiE	Prop. Overlap for SuSiE + BLiP
Cardiovascular	100.0%	81.4%
Height	100.0%	87.5%
HDL	100.0%	84.8%
LDL	100.0%	86.2%

Table 2: For the four traits analyzed in Section 4, this figure shows the proportion of overlap (as defined in Appendix F.3) for SuSiE and SuSiE + BLiP. For example, the first row shows that for the cardiovascular trait, 100% of the discoveries made by SuSiE overlapped with at least one discovery made by SuSiE + BLiP, whereas 81.4% of the discoveries made by SuSiE + BLiP overlapped with at least one discovery made by SuSiE.

SuSiE. Holistically, Table 2 shows that SuSiE + BLiP confirms every discovery that SuSiE makes but also that SuSiE + BLiP makes roughly 15 – 20% more discoveries that do not overlap with any previous discoveries from SuSiE. Since BLiP increased SuSiE’s resolution-adjusted power by 30 – 50%, this suggests that the power gain comes both from more precisely localizing existing discoveries *and* from making some completely new discoveries.

F.4 Discussion on making discoveries at a single-variant resolution

One notable result from Section 4 is that in addition to having higher resolution-adjusted power, SuSiE + BLiP discovers more individual variants than SuSiE alone. This result raises the question: does BLiP help make more discoveries at a single-variant resolution? The answer to this question is subtle, so we pause to discuss it here.

First, if one aims to perform *resolution-adaptive inference*, then yes, we expect SuSiE + BLiP to make more single-variant discoveries than SuSiE alone. The reason for this, as explained in Section C.4, is that SuSiE + BLiP can simultaneously use full PIPs from the SuSiE model and also make resolution-adaptive discoveries. In contrast, prior to BLiP, analysts using SuSiE faced a trade-off: SuSiE can construct credible sets, but to do so, SuSiE must use PIPs which are provably more conservative than the full PIPs. This may cause SuSiE to make fewer single-variant discoveries, as seen in Section 4. Alternatively, one could use SuSiE’s full PIPs to make many discoveries at the single-variant resolution, but in this case it is not clear how to make any discoveries at coarser resolutions. Thus, SuSiE + BLiP gets the best of both worlds, since it can both use the full SuSiE PIPs *and* make resolution-adaptive discoveries.

Of course, if one’s *exclusive* goal is to discover single variants without discovering any larger groups, then BLiP will reduce to the following trivial (known) procedure: calculate PIPs $\{p_{\{\ell\}}\}$ for each variant ℓ , sort them in decreasing order, and sequentially make rejections until the average PIP of the rejections drops below $1 - q$. In short, BLiP is not a meaningful innovation in settings where one does not wish to take a resolution-adaptive approach. That said, there is basically no cost to being resolution-adaptive, since BLiP will discover single variants whenever this is possible, and when this is not possible, it will (likely) discover other, larger groups which may still be of scientific interest. Note that this same argument applies when applying BLiP on top of any Bayesian model.

A last note is that studies which primarily aim to make discoveries at a single-variant resolution may wish to consider a different notion of the FDR than the (standard) one we consider in this paper. In particular, controlling the FDR as defined in Equation (1.2) does not guarantee that one will control the FDR when only considering singleton discoveries—for example, the FDR of the

singleton discoveries in our application in Section 4 is $5.99\% > 5\%$. Of course, this is easy to fix, since BLiP can control a variety of error rates. For example, one might require that for each $k \in \mathbb{N}$, the FDR for each set of discoveries of size k is controlled at level q . Similarly, one could also define a weighted or “resolution-adjusted” FDR constraint. In general, we suspect that the right notion of the error rate depends on the study in question: the good news is that BLiP can be easily adapted to control any of these error rates.

G Further details on Section 5

In this section, we describe our method from Appendix 5. Note that this section follows the notation of Appendix D. We create the inputs to BLiP as follows:

1. First, we load the prefit StarNet “wake-sleep” encoder published by Liu et al. (2021b), and we sample $N = 1000$ samples from the posterior of this StarNet model. Let $\gamma^{(1)}, \dots, \gamma^{(N)} \subset \mathcal{L} = [0, 1]^2$ denote these posterior samples. Furthermore, let $\mathcal{M} = \{\ell^{(1)}, \dots, \ell^{(m)}\} \subset \mathcal{L}$ denote the MAP estimators of point source locations from the StarNet model.
2. As candidate groups, for a fixed radius r , let $\{S_r(rz) : z \in \mathbb{Z}^2\}$ denote the set of circles of radius r on the lattice points rz for $z \in \mathbb{Z}^2$. Furthermore, let $\{S_r(\ell) : \ell \in \mathcal{M}\}$ denote the circles of radius r around the MAP estimators. We set

$$\mathcal{G} = \bigcup_{r \in \mathcal{R}} \{S_r(rz) : z \in \mathbb{Z}^2\} \cup \{S_r(\ell) : \ell \in \mathcal{M}\},$$

where we take \mathcal{R} to be 25 evenly spaced values on the log scale between 0.01 and 0.0005. For intuition, recall that a radius of 0.01 corresponds to one pixel.

3. We compute PIPs $\{p_G\}_{G \in \mathcal{G}}$ for these candidate groups using Algorithm 4 from Appendix D. We then pre-filter \mathcal{G} to only include groups G such that $p_G \geq 0.25$ (so each candidate group must have at least a 25% chance of containing a signal). This takes roughly eight minutes of serial computation on a laptop. Note that this step is the bottleneck, but it is quite parallelizable, so in large-scale applications it should not pose a barrier to running BLiP.
4. In this case, there are an infinite number of locations, so we compute a finite-dimensional reduction of BLiP using the algorithms from Appendix D.2 (notably Algorithm 5). This takes only a few seconds.
5. Finally, we apply BLiP to control the FDR at levels $q \in \{0.05, 0.1, 0.15, \dots, 0.75\}$, which takes 5-10 seconds per value of q . We use the weight functions described in Section 5.

To evaluate both BLiP and the baseline, we use the Hubble locations as the true sources and calculate resolution-adjusted power using the weight functions from Section 5. We give all methods 0.01 pixels of slack when calculating false discoveries to account for measurement differences between the Hubble and SDSS surveys.