Controlled Discovery and Localization of Signals via Bayesian Linear Programming (BLiP)

Asher Spector

Stanford University, Department of Statistics

Coauthor

Lucas Janson (Harvard Statistics)

One-slide overview

Challenge: Collinearity/correlations make it challenging to perform controlled variable selection.

- Often, we can tell that *some* variables influence the outcome Y, but we don't know which ones.
- Even after fitting a model, it's unclear how to *localize* which variables may affect Y.

This talk: Given a model, how can we "localize" signal variables?

- This talk is *not* about fitting the model!
- It *is* about extracting useful information from a pre-fit model.
- For this talk, I will assume the model is Bayesian.

Empirically: Increases power 20-50% on a large-scale GWAS with ≤ 1 min of added computation!

Outline

Outline

Motivation I: genetic fine-mapping

UK Biobank dataset ($n \approx 377,000$):

- Y is disease status
- $(X_1, \ldots, X_p)$ are genetic variants ($p \approx 19,000,000$)
- Question: which features X_j influence Y? Which are "signals?"

For simplicity, let's temporarily assume a linear model:

$$Y = X^T \beta + \epsilon \text{ with } \mathbb{E}\left[\epsilon \mid X\right] = 0$$

Challenge:

- $(X_1, \ldots, X_p)$ exhibit strong local correlations, e.g., $\mathrm{Cor}(X_1, X_2) = 0.999$
- We may have no power to detect that $\beta_1 \neq 0$
- But, maybe we know (X_1, X_2) contains a signal, i.e., $\beta_{1,2} \neq 0$!

Motivation in a picture

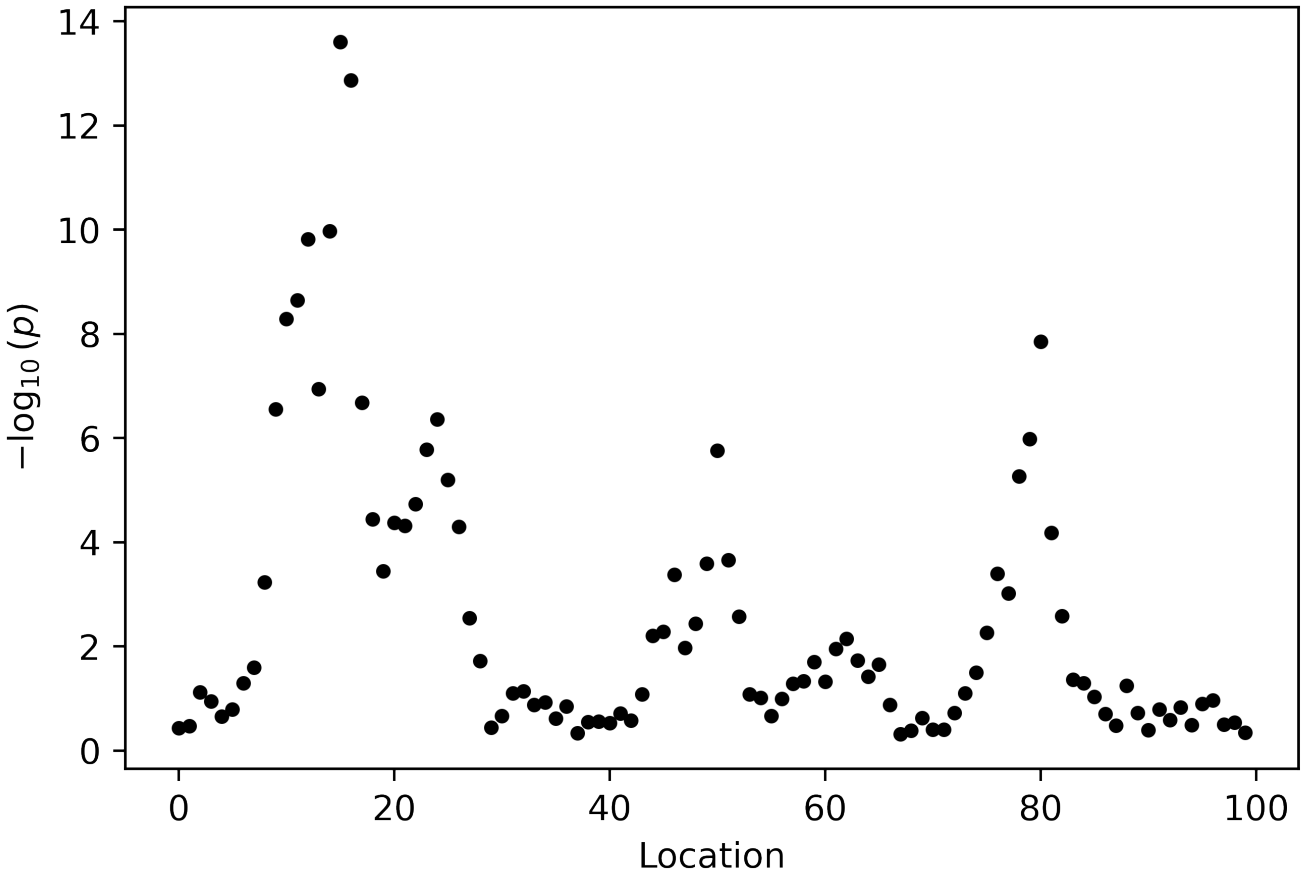


Figure: Cartoon Manhattan plot of genome; y-axis shows a measure of $\mathrm{Corr}(Y, X_j)$ for $j = 1, \ldots, 100$.

Motivation in a picture

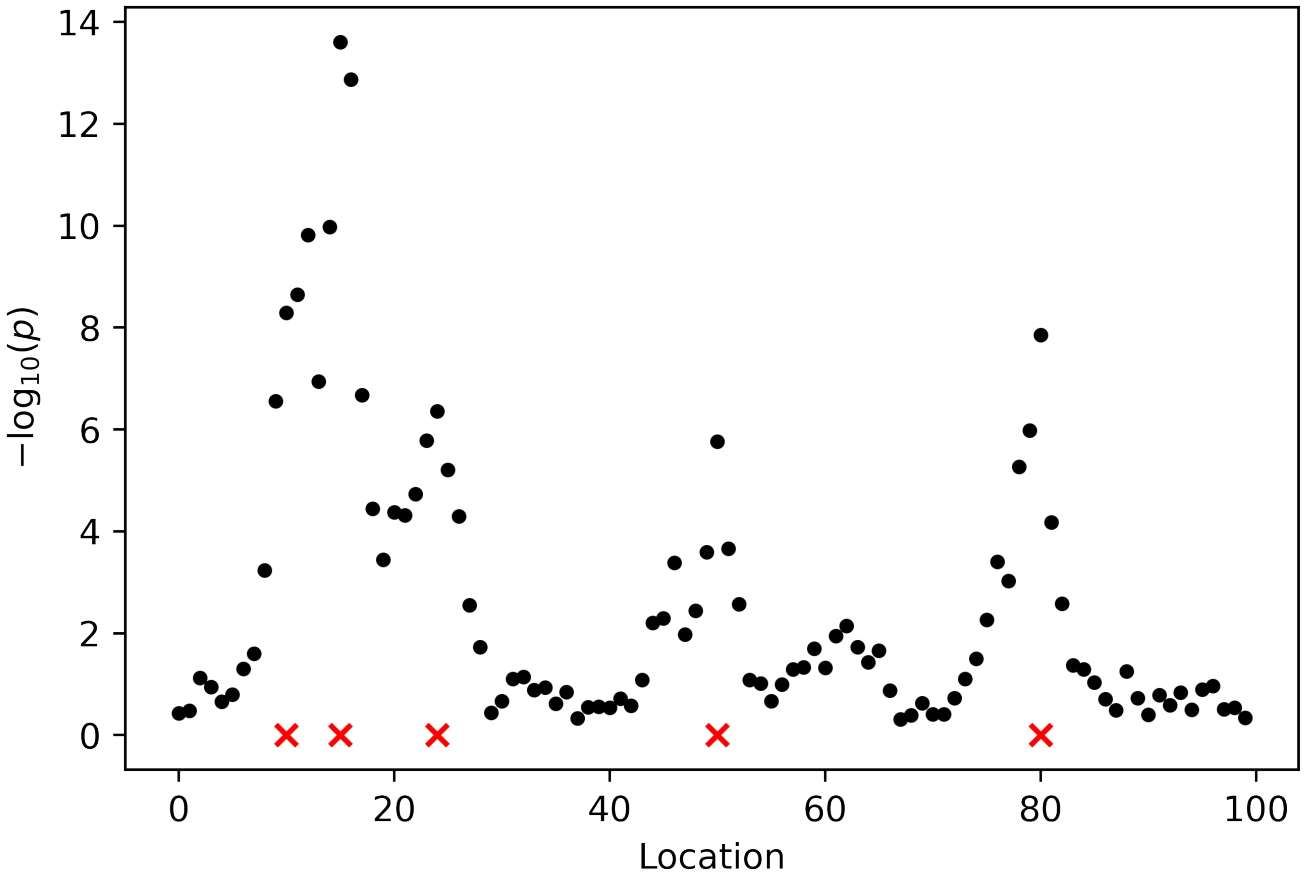


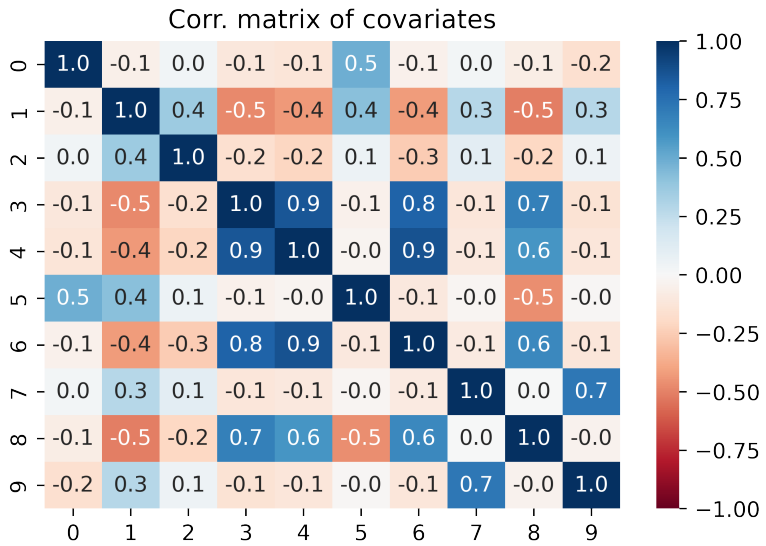
Figure: Cartoon Manhattan plot of genome; y-axis shows a measure of $\mathrm{Corr}(Y, X_j)$ for $j = 1, \ldots, 100$.

If you fit a linear model, you might find that nothing is significant!

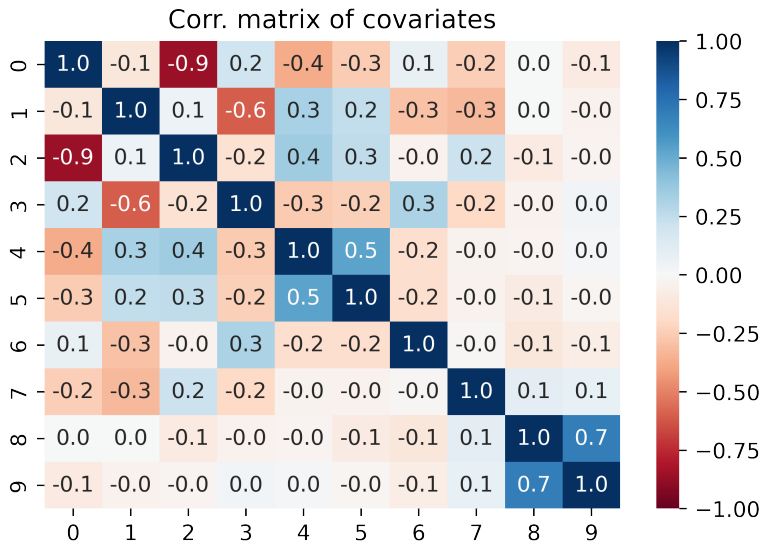
Motivation II: exploratory analysis of clinical trials

Correlated features appear in exploratory analysis of clinical trials:

- Proteomic/genomic data, even demographic data (age, baselines, etc)
- Throughout, think of X as a generic set of features (could include treatments/interaction terms)



(a) Blum et al. (2010)



(b) Siedner et al. (2017)

Moral: for small to medium n, moderate correlations make it harder to identify treatment effect moderators / prognostic variables.

Contribution

Goals:

- Discover disjoint groups $G_1, \ldots, G_R \subset \{1, \ldots, p\}$ which each contain a signal
- Make R large and $G_1, \ldots, G_R$ small—both matter a lot!
- Control (e.g.) the FDR

Method: Bayesian Linear Programming (BLiP).

Input: posterior samples from any Bayesian model (e.g. Bayesian GLM/GAM)

Output: groups $G_1, \ldots, G_R$ that maximize power subject to FDR control.*

* FDR control assumes the Bayesian model is well-specified.

Contribution in a picture

Input: Samples from sparse Bayesian linear model.
Output:

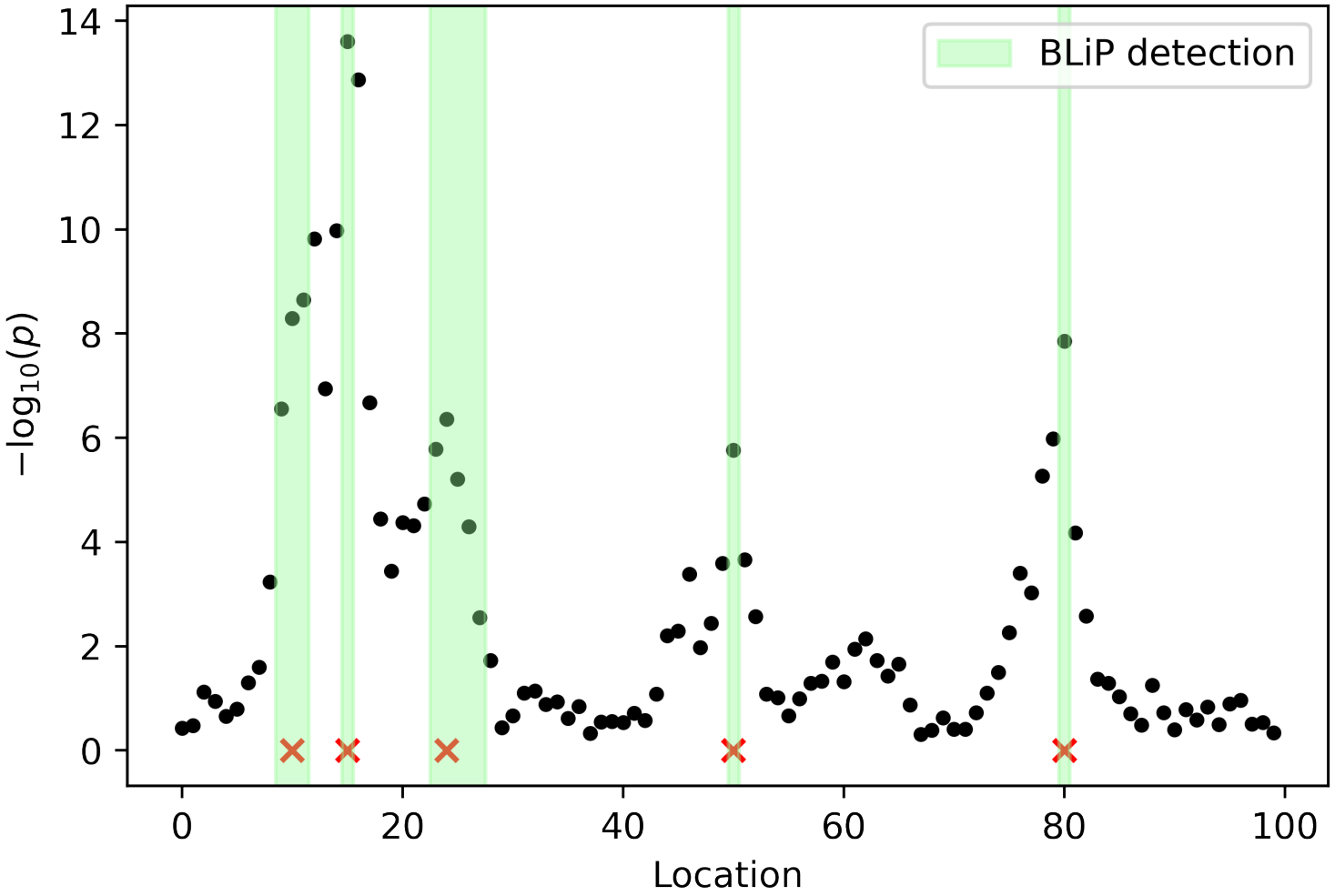


Figure: Cartoon of partial Manhattan plot of genome

Outline

Notation and assumptions

Notation: $X \in \mathbb{R}^p$ are features, $Y \in \mathbb{R}$ is an outcome, $\mathcal{D}$ is dataset.

Assumption 1: The analyst specifies a Bayesian model which implies

$$\mathbb{E}[Y \mid X] = f_\theta(X) \text{ for } \theta \in \Theta$$

with $\theta \sim \pi$ sampled from some prior distribution.

- $S = \{j : f_\theta(X) \text{ depends on } X_j\} \subset [p]$ is the set of *signal variables.*

Assumption 2: The analyst can sample from the law of $\theta \mid \mathcal{D}$.

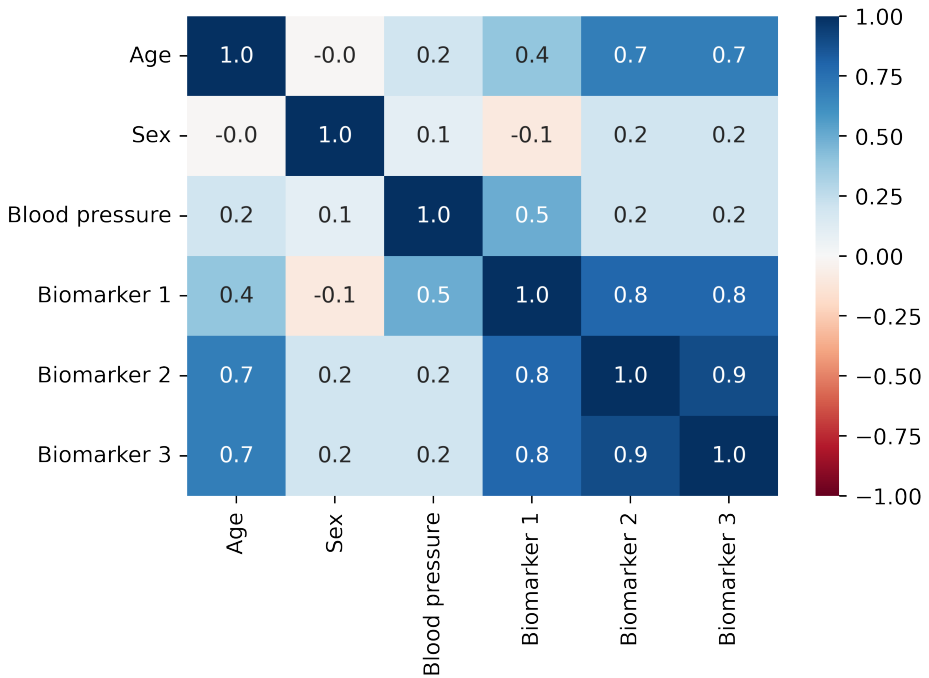
- These assumptions are not always reasonable! (see Section 5)
- But there is an enormous literature on sampling from these models
- Our question: how do we extract useful insights from these models after fitting them?

Problem statement (I)

Goal:

- Discover disjoint groups $G_1, \ldots, G_R \subset \{1, \ldots, p\}$ which each contain a signal
- Make R large and $G_1, \ldots, G_R$ small—both matter a lot!
- Control the (Bayesian) FDR

Emphasis: we want to pick $G_1, \ldots, G_R$ after seeing the data. E.g.:



How to group the features depends on the (unknown) signal size!

Problem statement (II)

Don't want to narrow potential discovery regions until *after* seeing data

Goal: look at the data and outputs regions $G_1, \ldots, G_R$ so as to:

$$\begin{aligned} \max \quad & \mathbb{E}\left[\text{Power}(G_1, \ldots, G_R) \mid \text{Data}\right] \\ \text{s.t.} \quad & \text{FDR} := \mathbb{E}\left[\frac{\#\{G_r \text{ containing no signal}\}}{\max(1, R)} \mid \text{Data}\right] \leq q, \\ & G_1, \ldots, G_R \subset [p] \text{ are disjoint.} \end{aligned}$$

What does high Power() look like?

- As many (true) discovered regions G_r as possible
- Discovered regions G_r should be as small as possible

Existing work: no formalization of what "power" means, so cannot optimize it.

Defining resolution-adjusted power

Define a weighting function $w(G)$ that measures value of discovering a group

- Should penalize larger groups
- Canonical choice is inverse-size weighting: $w(G) = 1/|G|$
- Sum weights of true rejections to get Power():

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

where $I_G = \mathbb{I}(G \cap S \neq \emptyset)$ is the indicator that G contains a signal (i.e., is a true discovery)

Remarks:

- Different w can accommodate very different scientific objections
- In practice, do we exactly know our "utility function"?
- We will see that the results are not too sensitive to precise specification

Method: Bayesian Linear Programming (I)

Method: directly solve the optimization problem:

$$\begin{aligned} \max \quad & \mathbb{E}\left[\text{Power}(G_1, \ldots, G_R) \mid \text{Data}\right] \\ \text{s.t.} \quad & \text{FDR} := \mathbb{E}\left[\frac{\#\{G_r \text{ containing no signal}\}}{\max(1, R)} \mid \text{Data}\right] \leq q, \\ & G_1, \ldots, G_R \subset [p] \text{ are disjoint.} \end{aligned}$$

Key observation: the power of a Bayesian method that discovers $G_1, \ldots, G_R$ is

$$\mathbb{E}[\text{Power}(G_1, \ldots, G_R) \mid \text{Data}] = \mathbb{E}\left[\sum_{r=1}^{R} I_{G_r} w(G_r) \,\middle|\, \text{Data}\right] = \sum_{G \subseteq [p]} p_G w(G) z_G,$$

- $p_G = P(G \text{ contains a signal} \mid \text{Data})$ can be computed Assumptions 1-2
- $z_G \in \{0, 1\}$ is indicator that we discover G

Method: Bayesian Linear Programming (II)

Theorem: the optimization problem is equivalent the following integer LP:

$$\max_{\{z_G\}_{G\subseteq[p]}} \quad \sum_G p_G w(G) z_G \qquad \text{(Power)}$$

$$\text{s.t.} \quad \sum_G (1 - p_G - q) z_G \leq 0 \qquad \text{(FDR)}$$

$$\sum_{G\subset[p]: j\in G} z_G \leq 1 \qquad \forall j = 1, \ldots, p \quad \text{(disjoint discoveries)}$$

for decision variables $z_G \in \{0, 1\}$.

This is progress. Yet we have 2^p integer decision variables.

Method: Bayesian Linear Programming (III)

Problem: 2^p integer decision variables.
Solutions:

1. Narrow the search space *after* looking at the data
   - Sublinear algorithm to discard $\{G : p_G \leq 0.001\}$
2. Narrow the search space by imposing desirable structural constraints
   - E.g., ensure $|G| \leq 25$
3. If needed, solve the continuous relaxed LP; then round to obtain integers.

Result: Provable FDR control, verifiable near-optimality.

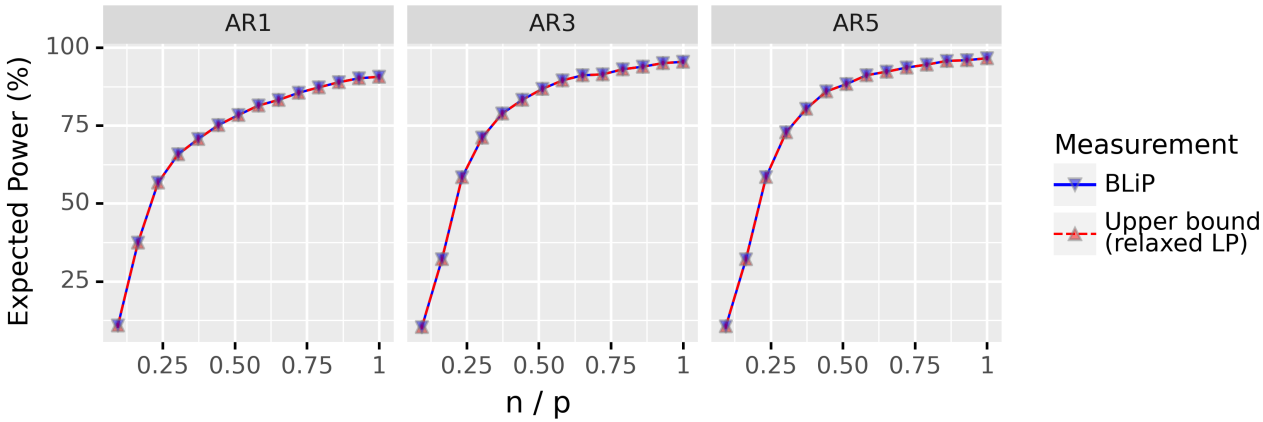


Figure: Expected power (BLiP) vs. upper bound

Putting it all together: BLiP

Input: Nearly any Bayesian model (via MCMC, variational inference) and any desired structural constraints on the discovery set

Output: disjoint discoveries which (1) verifiably nearly maximize power and (2) control the FDR.

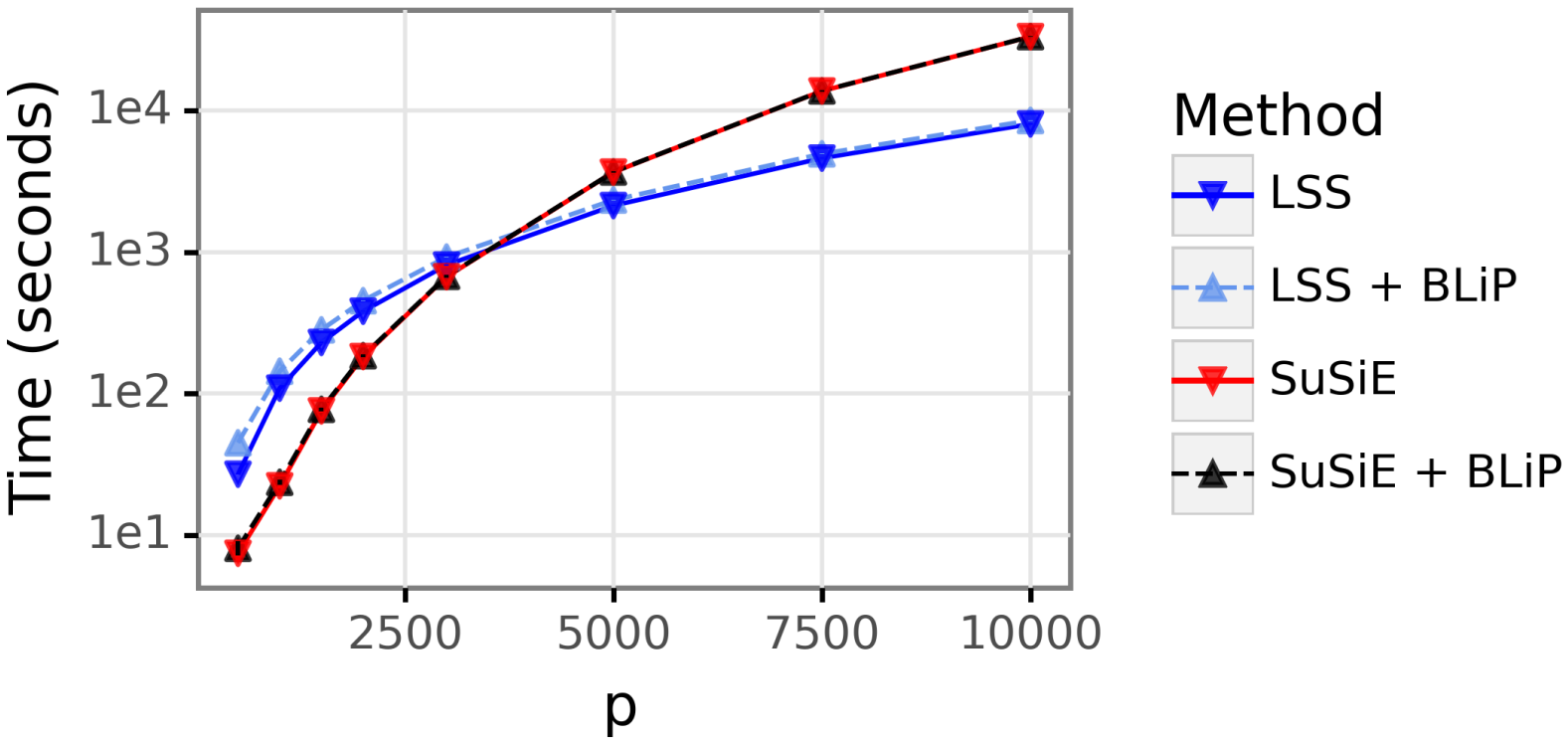


Figure: p denotes dimension of linear model being fit, with $n = p/2$

Outline

Fine-mapping setup

Dataset: $n \approx 337,000$, $p \approx 19,000,000$, four traits of interest.

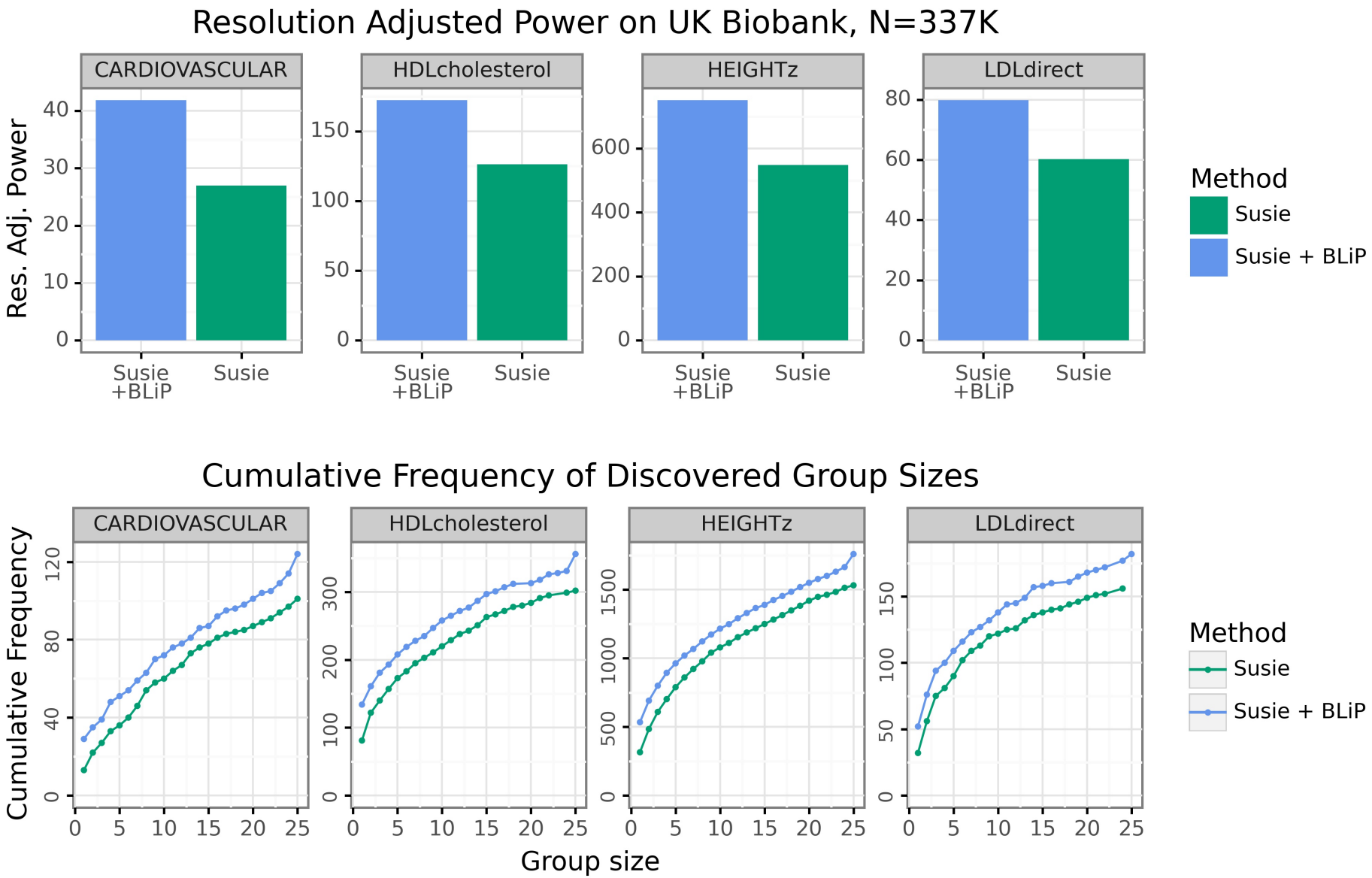
Bayesian model: SuSiE (Wang et al., 2020)

1. SuSiE is a sparse Bayesian linear model that can be fit highly efficiently.
2. Like BLiP, SuSiE returns regions $G_1^{\text{SuSiE}}, \ldots, G_R^{\text{SuSiE}}$ of the genome.
3. However, SuSiE's regions are constructed heuristically.
    - Can we do better using a principled approach (BLiP)?

We run BLiP on top of a pre-fit SuSiE model from Weissbrod et al. (2019).

Fine-mapping results

UK Biobank data: $n \approx 337,000$, $p \approx 19,000,000$; BLiP takes $\underline{< 1 \text{ min}}$ per trait



Are our results real?

Trait	Corroboration Rate (SuSiE)	Corroboration Rate (*new*)
Height	53.5%	45.0%
HDL	57.0%	50.0%
LDL	67.3%	60.0%
Cardiovascular	82.2%	65.2%

Table: The proportion of discoveries which can be corroborated by a separate study in the NHGRI-EBI GWAS Catalog (Buniello et al., 2018).

Note the right-hand column only contains entirely new discoveries made by BLiP.

Our interpretation: this is a positive result, since all of the "low-hanging fruit" should lie in the left-hand column. Nonetheless, the numbers are comparable.

Outline

BLiP with knockoffs?

A weakness: BLiP assumes the Bayesian model is well-specified.

Gablenz and Sabatti (2024) also solve the BLiP optimization problem...

- ...but obtain model-free frequentist FDR guarantees.

Insights:

- Use knockoffs for model-free error control (Candes et al., 2018)
- Technical insight: use e-values to account for multiplicity (Wang and Ramdas, 2022)

TL;DR: one can perform resolution-adaptive variable selection as a frequentist.

Outline

Conclusion

BLiP is a powerful and efficient method for resolution-adaptive variable selection

- Provable (Bayesian) error control and verifiable near-optimality
- Substantial power gains in minutes on fine-mapping
- Software packages `pyblip` (Python) and `blipr` (R)

More in the paper:

- Applications to astronomy, change-point detection
- Potential for other signal discovery problems with spatial structure?

Paper available at: `https://arxiv.org/abs/2203.17208`

All code posted at: `https://github.com/amspector100/blip_sims/`

Thank you!

References

Blum, J., Winikoff, B., Raghavan, S., Dabash, R., Ramadan, M. C., Dilbaz, B., Dao, B., Durocher, J., Yalvac, S., Diop, A., Dzuba, I. G., and Ngoc, N. T. N. (2010). Treatment of post-partum haemorrhage with sublingual misoprostol versus oxytocin in women receiving prophylactic oxytocin: a double-blind, randomised, non-inferiority trial. *Lancet*, 375(9710):217–223.

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

Candes, 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.

Gablenz, P. and Sabatti, C. (2024). Catch me if you can: signal localization with knockoff e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae042.

Siedner, M. J., Bwana, M. B., Moosa, M.-Y. S., Paul, M., Pillay, S., McCluskey,