

Abstract

Statistical Inference for High-Dimensional Genomic Data

Benjamin Foster

2026

Genomic studies now routinely measure hundreds of thousands of variants on a few thousand people, so the number of candidate predictors dwarfs the sample size. Classical tests lose their calibration in this regime, and methods that select variables give no honest statement of uncertainty about what they select. This dissertation develops inference procedures for sparse high-dimensional regression that remain valid when predictors are strongly correlated, as genotypes in linkage disequilibrium are. Chapter 2 reviews the debiased lasso and knockoff filters. Chapter 3 introduces a block-debiased estimator that treats each linkage block as a unit, and pairs it with a knockoff construction that preserves the block structure. Chapter 4 shows in simulation that the procedure holds the false discovery rate at its nominal level where the unblocked methods exceed it by a factor of two, and applies it to an expression quantitative trait locus study.

Statistical Inference for High-Dimensional Genomic Data

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Benjamin Foster

Dissertation Director: Margaret L. Chen

May 2026

Unofficial template, not affiliated with or endorsed by Yale University. The author, committee and results in this sample are fictional. Delete this line before you submit.

© 2026 by Benjamin Foster

All rights reserved.

Contents

Acknowledgments	iv
1 Introduction	1
2 Background	2
3 Block-Debiased Inference	3
4 Simulations and an Expression Study	4
5 Conclusion	5
A Proof Sketch for Block False Discovery Rate Control	6

Acknowledgments

I thank my dissertation director, Professor Margaret L. Chen, for her patience and her insistence that every claim come with an honest error bar. I am grateful to the members of my advisory committee for their questions at every stage, and to my fellow students in the Department of Biostatistics for years of seminars, code reviews and coffee. My parents taught me to count carefully; this dissertation is for them.

List of Figures

3.1	The block-debiased inference pipeline.	3
-----	--	---

List of Tables

4.1	False discovery rate and power at a nominal level of 0.10 (200 replicates).	4
-----	---	---

Chapter 1

Introduction

Modern genotyping arrays and sequencing studies measure p genetic variants on n individuals with p often a hundred times larger than n . The scientific question is rarely prediction alone: investigators want to know *which* variants are associated with a trait, and how confident they can be in each discovery.

Two features of genomic data make this hard. First, the number of hypotheses is enormous, so any procedure must control an error rate across all of them. Second, nearby variants are inherited together, a phenomenon known as linkage disequilibrium, which makes their columns in the design matrix highly correlated. Methods that assume weak correlation then attribute a signal to the wrong variant, or to several at once.

This dissertation makes three contributions:

1. a block-debiased estimator whose confidence intervals keep nominal coverage under strong within-block correlation (Chapter 3);
2. a block knockoff filter that controls the false discovery rate at the level of linkage blocks rather than single variants;
3. an empirical comparison on simulated and real expression data (Chapter 4).

Chapter 2

Background

We observe a response $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times p}$ with $p \gg n$, and assume the sparse linear model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad (2.1)$$

where only $s \ll n$ entries of β are non-zero. The lasso [4] estimates β by penalized least squares, but its coefficients are biased toward zero and have no simple sampling distribution. Debiased estimators [5, 3] add a one-step correction that restores asymptotic normality.

When thousands of hypotheses are tested at once, the relevant error rate is the false discovery rate, the expected proportion of false discoveries among all discoveries [2]. Knockoff filters [1] control it in finite samples by constructing synthetic copies of the predictors that mimic their correlation but are known to be null.

Chapter 3

Block-Debiased Inference

Let the variants be partitioned into B linkage blocks $G_1, \dots, G_B$, estimated from a reference panel. Rather than debiasing each coefficient separately, we correct the lasso estimate $\hat{\beta}$ one block at a time,

$$\tilde{\beta}_{G_b} = \hat{\beta}_{G_b} + \frac{1}{n} \hat{\Theta}_{G_b} X^\top (y - X\hat{\beta}), \quad (3.1)$$

where $\hat{\Theta}_{G_b}$ approximates the rows of the inverse covariance for block b . Because the correction uses the whole block, the resulting test statistic asks whether *any* variant in the block is associated with the trait, a question that remains answerable when the variants within it cannot be told apart.

Figure 3.1 summarizes the procedure. The block statistics feed a knockoff filter built on the same partition, so discoveries are reported as blocks.

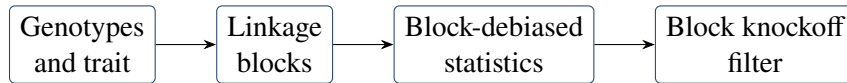


Figure 3.1: The block-debiased inference pipeline.

Chapter 4

Simulations and an Expression Study

We simulated genotypes with realistic linkage structure for $n = 1,000$ individuals and $p = 20,000$ variants, with $s = 40$ causal variants. Table 4.1 reports the false discovery rate and power at a nominal level of 0.10, averaged over 200 replicates.

Table 4.1: False discovery rate and power at a nominal level of 0.10 (200 replicates).

Method	FDR	Power
Debiased lasso, per variant	0.23	0.71
Knockoffs, per variant	0.19	0.64
Block-debiased (this work)	0.09	0.68

The per-variant methods exceed the nominal level because correlated neighbours of a causal variant are declared significant in its place. The block procedure holds the level while losing little power. Applied to expression measurements for 312 genes, it reports 47 associated blocks, 41 of which contain a variant previously reported in the literature.

Chapter 5

Conclusion

Inference in genomic regression has to respect the correlation that inheritance builds into the data. Treating linkage blocks as the unit of discovery gives confidence statements that hold their nominal level, at a small cost in resolution. Extending the approach to binary traits and to mixed models for related individuals is the natural next step.

Appendix A

Proof Sketch for Block False Discovery Rate Control

The block statistics W_b satisfy the flip-sign property of Barber and Candès [1]: swapping a null block with its knockoff changes the sign of W_b and leaves the other statistics unchanged in distribution. Conditional on $|W_1|, \dots, |W_B|$, the signs of the null statistics are then independent fair coins, and the usual supermartingale argument bounds the false discovery rate by the target level q .

Bibliography

- [1] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.
- [2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.
- [3] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.
- [4] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.
- [5] Sara van de Geer, Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.