

Ancestral Trees as Weighted Networks:

Scalable Screening for Genome Wide Association Studies



Ryan Christ

Supervisor: **Prof. Chris Holmes**

Prof. David Steinsaltz

Dr. Chris Spencer

Department of Statistics

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

St. Hilda's College

October 2017

I would like to dedicate this thesis to my parents, Bob and Linda, for their constant love and support.

Acknowledgements

I and this thesis are incredibly indebted to the help and support that has been given to me by my supervisors and friends at Oxford throughout my studies. Thank you Chris Holmes and Chris Spencer for your continued support from the beginning on this long journey. We set out to build a haplotype-based approach that could incorporate selection information. After many failed attempts, I believe we can finally say mission accomplished. The software presented in Chapter 2 is the result of a close collaboration between myself and Louis Aslett. He is architect and author of the implementation of the specialized, Li & Stephens Model motivated, Forward-Backward Algorithm. I need to give very special thanks to Louis and to David Steinsaltz, who by joining the project as collaborator and supervisor respectively, helped break down key barriers that made this thesis possible. You have both contributed unimaginably to this thesis and how I think about math, computation, haplotypes, and more. Our times working in front of the white board together will be some of my fondest memories of Oxford.

I also need to give a special thanks to George Busby, who helped mentor and advise me on the genetics front of this work. Also special thanks to my advisor and friend Susanna Mireau, who during my undergraduate and graduate work helped me to navigate my graduate studies and guided me through some of the toughest times. I'd also like to thank my undergraduate mentors Pardis Sabeti and Joe Blitzstein who ignited my interest in population genetics and statistics.

To my dearest Emily Mayne, thank you so much for your unwavering love and support in helping me through the struggles to complete this work. To my friends and housemates Holly Trochet, Nausad Velgy, Carolin Rindfleisch, and Amelie Gavard, thanks for all of the laughs and help that got me through those late nights.

Copyright & License

The copyright remains solely with the author, Ryan Christ. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Ryan Christ

October 2017

Abstract

Several haplotype-based methods have been developed to identify loci where multiple mutations of low to moderate frequency and effect size modulate disease susceptibility. Most such approaches either do not scale to hundreds of thousands of genomes or do not explicitly model recombination and the block-like structure of haplotypes. Using a novel checkpointing technique and a C-core, vectorized implementation of an Hidden Markov Model (HMM) based on the Li & Stephens Model, at each single nucleotide polymorphism (SNP) along the genome, we obtain a local genetic distance between all pairs of haplotypes in a phased dataset. To rapidly test this local distance matrix for association with a phenotype, we derive two finite sample central limit type theorems for quadratic forms which do not require any further assumptions on the matrix other than it is free of outliers, for which we have an easily calculable, formal condition. We combine these results with a novel concentration inequality for Gaussian quadratic forms to upper and lower bound p-values for quadratic forms while avoiding a full eigendecomposition of each matrix. Applying our HMM implementation and quadratic form screening method, we recover known loci associated with malaria susceptibility and uncover new potential associations in a pilot dataset of 6,136 haplotypes.

Table of contents

List of figures	xv
List of tables	xxiii
1 Haplotype Methods for GWAS	1
1.1 From Single Marker to Haplotype Tests	2
1.2 Random Effects Models	4
1.2.1 Derivation	5
1.2.2 Challenges Facing Standard Random Effects Models	8
1.3 Non-exchangeable Models	10
1.4 The Li & Stephens Model	15
1.4.1 PAC Likelihood	15
1.4.2 Emission Distribution	17
1.4.3 Transition Distribution	18
1.4.4 Pseudo-likelihood Approximation	20
1.4.5 LS Model Advantages	21
1.5 Correctly Testing Quadratic Forms	22
1.6 Overview of Thesis	24
1.7 Proofs	25
1.7.1 Proof of Lemma 1.2.1	25

1.7.2	Proof of Lemma 1.5.1	27
2	Building Local Distance Matrices	31
2.1	Local Distances from Posterior Decoding	32
2.1.1	LS Model Without Recombination	32
2.1.2	With Recombination	37
2.1.3	Interpretation of Cost Functions	42
2.1.4	Tuning Parameters for Local Resolution	45
2.1.5	Simulation Study to Show Local Resolution	47
2.2	Computation	50
2.2.1	Exact Inference Under LS Model	50
2.2.2	Checkpointing	54
2.3	Implementation	59
2.3.1	Rescaling	60
2.3.2	Parallelization & Memory Optimization	61
2.3.3	16 Row Blocks & AVX Instructions	62
2.3.4	Benchmarking	63
2.4	Discussion	65
2.5	Appendix	67
3	Scalable Screening Procedures for Quadratic Forms	71
3.1	Efficient Bounds for Gaussian Quadratic Forms	72
3.1.1	Overall Approach	72
3.1.2	Exponential Concentration for Gaussian Quadratic Forms	75
3.2	Non-Gaussian Random Variables	78
3.2.1	Hollow Matrices	81
3.2.2	Centered Random Variables	89

3.3	Discussion	95
3.4	Proofs	96
3.4.1	Proof of Lemma 3.1.3	96
3.4.2	Proof of Theorem 3.4.2	97
3.4.3	Proof of Theorem 3.2.1	99
3.4.4	Proof of Lemma 3.2.2	106
3.4.5	Proof of Theorem 3.2.6	108
4	Application to GWAS	113
4.1	Dediploiding	114
4.2	GWAS via Autologistic Models	123
4.2.1	Classic Autologistic Models	123
4.2.2	Centered Autologistic Models	124
4.3	Pilot on data from the MalariaGEN Consortium	126
4.3.1	Background on Malaria	126
4.3.2	Dataset & Methods	128
4.3.3	Recovering Known Results	129
4.3.4	Novel Results	136
4.4	Discussion	151
4.4.1	Comparing to SMTs	151
4.4.2	P-values vs. Score Statistics	152
4.4.3	P-value Inflation and Bound Effectiveness	153
4.5	Appendix	156
5	Conclusion & Future Directions	161
	References	165

List of figures

- 1.1 **Exchangeability Under SKAT Kernel:** Sample A depicts four pairs of haplotypes collected from four individuals (haplotype pairs are depicted by the colored brackets to the right of the haplotypes). In this region, there is a 8-SNP-long stretch where haplotypes may be either type blue or type orange. To the right of the four haplotype pairs are the corresponding four genotypes, to which we've applied the unweighted linear kernel to produce a local kernel matrix on the right. Note that the gold and red genotypes have the lowest pairwise similarity (11), reflecting the fact that the gold sample has two blue haplotypes and the red sample has two orange haplotypes. Sample B is identical to Sample A except that the columns have been permuted to break up the 8 SNP haplotype blocks. Notice that while this yields different genotypes, the implied kernel matrix, K , is identical, hence the exchangeability of loci under the SKAT kernel. An efficient compression of the 8 SNP region in sample A might use a binary indicator for either the blue or orange block.

- 2.1 **Marginal Tree Perspective of $d_{i,j}$ in Absence of Recombination:** Below we show the marginal tree describing the ancestry of 6 observed haplotypes. The colored markers on the tree represent mutations. Consider calculating the distance from recipient haplotype i to donor haplotype j . Three mutations separate those haplotypes on the tree. There is not one but two "nearest neighbours" to i , haplotypes q and m , who are only separated by two mutations. Since $-\log\left(\sum_i \exp(-z_i)\right)$ provides a lower bound on the distance to the nearest haplotype, it would evaluate to a number between zero and two. Thus, $d_{i,j}$ in this case would be a number in between 3 and 1. Calculating $d_{i,j}^l$ can be thought of as taking the local number of mutations that separate i and j and subtracting off the distance to a theoretical nearest neighbour of i , here drawn in as haplotype k 34
- 2.2 **Symmeterization for Outlier Robustness:** Blue circles represent observed haplotypes in an abstract haplotype space. Haplotype i is shown as an outlying haplotype (it has many mutations that separate it from all other haplotypes). In panel A, k is the hypothetical nearest neighbour of i (solid red circle). If we subtract the number of mutations separating i from k from the number of mutations separating i from j , we obtain $d_{i,j}$, which does not reflect the fact that i is an outlier. In panel B, the hypothetical nearest neighbour of j , k' , is much closer to j than k is to i . Thus, d_{ji} captures the fact that i is an outlier. Panels C and D show that $d_{i,j}$ becomes a more reliable measure when a new observation (shown as a green circle) is introduced near i . Thus, as long as no extreme outlier is alone, we can expect d_{ji} to be a meaningful measure of haplotype distance. 36

- 2.3 **Distance as Average over Copying Paths:** Blue lines represent different copying paths for i^{th} recipient through the remaining $N - 1$ donors across the L loci. Each of the paths shown have a different penalty associated with them: darker blue lines have lower penalties (fewer transitions and mutations) than lighter blue lines. Only paths that go through donor j and locus l , shown as a "gold gate," are considered. $-\log(\mathbb{P}(h_i, X_i^l))$ considers all possible paths passing through the gold gate and tightly lower bounds the penalty incurred by the best path. 40
- 2.4 **Tree to Distance Matrix Correspondence:** The trees from two simulations are shown at the top, where vertical height differences between nodes correspond to the simulated coalescence times. The corresponding matrices below show the inferred local structure where we generally see greater distances between rather than within clades. The histograms at the bottom plot the distribution of distances encoded in each matrix (the diagonal is excluded since it is set to zero by default). 49
- 2.5 **Optimized Checkpointing Performance:** The total distance required of the forward algorithm for locus by locus decoding using one or two checkpoints is plotted against the number of loci L . The total distance is plotted in units of L loci (using two checkpoints on a chromosome with 10 million loci would require the forward algorithm to cover approximately $300L$ loci). 57

- 2.6 **Benchmarking:** Panel 2.6a shows the time it takes the forward algorithm to propagate a full $N \times N - 1$ matrix of forward probabilities to across a segment of length L moving continuously from locus to locus using all 36 cores. We achieve the theoretical limit for the forward algorithm, which scales $\mathcal{O}(N^2)$. Rearranging the same data to show the times as a function of L , we see in panel 2.6b, that we also obtain the theoretical limit of $\mathcal{O}(L)$ scalability. Panel 2.6c, which presents timings for moving a forward table just a single locus, shows roughly the same $\mathcal{O}(N^2)$ behavior as panel 2.6a. Panel 2.6d shows the same incremental timings as 2.6c, just presented as a function of the number of cores used. Panel 2.6c and 2.6d suggest that for a problem of this size, the overhead of using 36 cores to propagate the forward table a single locus makes using just 9 cores almost as fast. This suggests that using an architecture with fewer cores but more memory per machine may be more optimal for incremental propagation of the forward table. . . . 64

- 4.1 **Information Loss During Convolution:** Here we present a sample of 4 individuals (each with a distinct ID number). Two of them have the blue phenotype and two have the red phenotype. Alongside their phenotypes are their local haplotypes in a causal region of the genome. At this locus, there are four distinct haplotypes. The purple haplotype is dominant and elicits the blue phenotype. The golden haplotype is dominant and elicits the red phenotype. The convolution of the haplotypes into genotypes is shown (paired haplotype rows are simply added). To the right of the genotypes is the kernel matrix K that would result followed by an application of the non-weighted linear kernel: $K_{ij} = \sum_k G_{ik} G_{jk}$. This is the same kernel used before in Chapter 1, Figure 1.1. As emphasized by the four red entries in the resulting kernel matrix, K , the genotypes predict that individuals 2 and 3 have similar phenotypes and that individuals 3 and 4 have different phenotypes. This reflects the fact that the genotypes obfuscate the genetic distinction between individuals 2 and 3 that drives their distinct phenotypes. This information cannot be recovered no matter which kernel function is used to process the genotypes. 115

- 4.2 **Classic Modes of Inheritance:** The expected phenotype corresponding to recessive, additive, and dominant inheritance is shown as a function of genotype. Strictly looking at this figure, it would appear that in order to maximize the power to detect a dominant mutation, ψ should evaluate heterozygotes (01) as close to (11) homozygotes but far from (00) homozygotes. However, if we flipped the labelling of the alleles, we would want to ψ to group heterozygotes (01) with (00) homozygotes. Since the coding of alleles is arbitrary, this means that testing for dominance or recessiveness is the same in the binary allele case. It amounts to assessing whether the expected phenotype of the heterozygote is closer to the expected phenotype of one of the homozygous states. 119
- 4.3 **Propagation of Information via Dediploidizing Functions:** Here we reconsider the configuration of haplotypes and phenotypes from Figure 4.1. Applying our HMM approach, we obtain a symmetric $N \times N = 8 \times 8$ distance matrix of ε and δ denoting small and large genetic distances respectively. After applying $\psi_{\min 1}$ or $\psi_{\min 2}$, we see that the resulting $N_s \times N_s$ matrix M that places a small distance between individuals with the same phenotype (highlighted with a green ε) and larger distances between individuals with different phenotypes. 121
- 4.4 **Chr 11 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark HBB. 131
- 4.5 **Chr 9 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark ABO. 132
- 4.6 **ABO Signal Using ψ_{dom} and ψ_{add} :** Bounds and score test shown for M obtained with ψ_{dom} compared to those obtained with ψ_{add} . Gray bars mark ABO. 133

- 4.7 **Hit-plots for *HBB* and *ABO*:** Gray vertical bars denote the limits of the gene of interest. Just as dark blue and dark red points denote upper and lower bounds respectively on the p-value for $(y - \tilde{\mu})^T M (y - \tilde{\mu})$, violet and orange points denote upper and lower bounds respectively on the p-value for $y^T M y$. 134
- 4.8 **Chr 4 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 2}$. Gray bars mark the outer borders of the regions of *GYPB* and *GYP A*. 135
- 4.9 **Chr 6 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *GCLC*. 139
- 4.10 **Hit-plots for *GCLC*:** All three plots are based on $\psi_{\min 1}$. Gray vertical bars denote the limits of the coding region of *GCLC*. Just as dark blue and dark red points denote upper and lower bounds respectively on the p-value for $(y - \tilde{\mu})^T M (y - \tilde{\mu})$, violet and orange points denote upper and lower bounds respectively on the p-value for $y^T M y$ 140
- 4.11 **Hit-plots for *GNPDA1*:** Gray vertical bars denote the limits of the *GNPDA1* coding region. $\psi_{\min 2}$ was used to generate both plots. 141
- 4.12 **Chr 19 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *LILRB1*. 142
- 4.13 **Hit-plots for *LILRB1*:** Gray vertical bars denote the limits of the *LILRB1* coding region. $\psi_{\min 1}$ was used to generate both plots. 143
- 4.14 **Chr 20 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *CDH4*. 144
- 4.15 **Hit-plots for *CDH4*:** Gray vertical bars denote the limits of the *CDH4* coding region. $\psi_{\min 1}$ was used to generate both plots. 145
- 4.16 **Chr 12 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *TMEM132B*. 146

- 4.17 **Hit-plots for *TMEM132B*:** Gray vertical bars denote the limits of the *TMEM132B* coding region. $\psi_{\min1}$ was used to generate both plots. 147
- 4.18 **Chr 10 SMT vs. Quad. Form Screening:** Bounds and score test shown for M obtained with ψ_{add} . Gray bars mark *ASAH2*. Gold bars mark *GRID1*. . . 148
- 4.19 **Hit-plots for *ASAH2*:** Gray vertical bars denote the limits of the *ASAH2* coding region. ψ_{add} was used to generate both plots. 149
- 4.20 **Hit-plots for *GRID1*:** Gold vertical bars denote the limits of the *GRID1* coding region. ψ_{add} was used to generate both plots. 150
- 4.21 **ECDF of p-value Bounds:** **a)** We compare the ECDF of the theoretical distribution of our $-\log_{10}$ -transformed p-values ($-\log_{10}(U) \sim \text{Expo}(\log(10))$ if $U \sim \text{Unif}$) against the empirical CDF of the $-\log_{10}$ -transformed p-value upper bounds (red) and lower bounds (blue) we obtained across chromosome 2. Inflation is evident in the discrepancy between the ECDF of the upper bounds and the theoretical CDF near the origin. **b)** The gap between the significance bounds at a given SNP is shown as a function of significance. Larger gaps between the bounds can occur at SNPs where the $-\log_{10}$ -transformed p-value lower bound is higher. 154

List of tables

4.1	The Five Table Classes: Let ε denote a relatively small haplotype distance and δ denote a relatively large haplotype distance. Each of the 2×2 tables indicate whether we would expect the distance between two haplotypes to be declared relatively low (ε) or relatively high (δ) using our HMM implementation under different haplotype configurations. Regardless of the dataset or locus of interest, ψ should always return a small distance for matching homozygous-homozygous tables and matching heterozygous-heterozygous tables. It should always return a large distance for non-matching tables where all four distances are δ . In contrast, partial matching heterozygous-homozygous tables and partial matching heterozygous-heterozygous tables should be treated differently by ψ depending on the suspected underlying genetic architecture.	117
4.2	Handling of Heterozygotes by ψ Candidates: Entries indicate the distance that a given ψ function would return for the three non-trivial haplotype table configurations.	120

Chapter 1

Haplotype Methods for GWAS

Genome-Wide Association Studies (GWASs) have guided medical and agricultural research by uncovering genes and molecular pathways that underlie various traits observed within species [1–4]. It is anticipated that the dropping cost of genomic sequencing and growing size of cohorts will dramatically boost the statistical power of GWASs to find genetic variants associated with a range of phenotypes [5]. Future association studies will require new statistical methods that can scale to hundreds of thousands of full genomes.

Despite recent advancements, most existing methods still struggle to identify genes in regions with allelic heterogeneity, where multiple mutations riding on distinct ancestral haplotypes modulate disease susceptibility [6, 7]. As we discuss in this chapter, this is partly because simple models that can scale to large datasets do not model recombination and haplotype structure while most methods that attempt to model haplotypes struggle to scale to large datasets. Existing haplotypes methods are also sensitive to how they divide the genome into functional regions, so called "windowing" [8]. This thesis aims to develop computational and theoretical tools which together open a class of haplotype methods that can scan the genome for regions with allelic heterogeneity, avoid windowing, and can feasibly be applied to hundreds of thousands of genomes. In our final chapter we implement one model from this class of screening methods and use it to identify putatively associated loci in a published

dataset. Although we do not pursue these putative associations further in this thesis, it is our intent that investigators may further interrogate the loci they identify using our approach with more precise, computationally intensive models.

There is an enormous range of diversity among existing GWAS methods ranging from frequentist linear mixed models to Bayesian regression to machine learning with random forests [9–11]. In this motivating chapter we will focus on a subset of examples to illustrate the need for our haplotype approach; however, at least some of the critiques that we make of existing methods also apply to methods not explicitly mentioned in the text.

1.1 From Single Marker to Haplotype Tests

Due to their simplicity, robustness and computational scalability, most GWASs use single marker tests (SMTs) to screen the genome for associations [8]. These SMTs regress the phenotype of interest onto each ascertained single nucleotide polymorphism (SNP) separately. In order to more formally introduce SMTs, let us introduce some notation. Let L denote the number of ascertained SNPs in a given dataset. We will also use N_s to denote the number of organisms or individuals sampled. Since it will be easier to work in terms of haplotypes rather than genotypes later in this thesis, we will use N to denote the number of haplotypes ascertained in a given dataset. In other words, for diploid organisms, $N = 2N_s$. We will also consider the inclusion of background covariates, $A \in \mathbb{R}^{N_s \times p}$, with corresponding parameters $\alpha \in \mathbb{R}^p$. For a given set of genotypes, $X \in \{0, 1, 2\}^{N_s \times L}$, at a given locus $l \in \{1, \dots, L\}$, and a binary phenotype, $y \in \{0, 1\}^{N_s}$, a standard additive effects SMT would fit the following logistic model.

$$y_i \sim \text{Bern}(\mu_i) \quad \log \left(\frac{\mu_i}{1 - \mu_i} \right) = A_{[i, \cdot]} \alpha + X_{[i, l]} \beta_l \quad (1.1)$$

Evidence for whether the SNP predicts the phenotype is typically assessed by calculating a p-value for the likelihood ratio test statistic or a Bayes Factor for the parameter of interest, β_l . In the case where the phenotype is caused by a single observed mutation, then the SMT is the most parsimonious and powerful model under the standard logistic regression model assumptions. However, testing every SNP independently ignores the complex dependence between nearby SNPs and limits the ability of SMTs to detect causal variation in regions with allelic heterogeneity. While SMTs may still detect associations driven by untyped causal variants with a correlated "tag" SNP, in regions with allelic heterogeneity, it is unlikely that a single SNP will tag all of the functionally distinct allelic subgroups since the causal mutations appear on different allelic backgrounds.

If we were able to visualize the ancestral tree underlying the variation of a locus with allelic heterogeneity, we would see clades under distinct causal mutations whose leaves exhibit disease risk that is different from the other individuals in the sample. However, there is a much more parsimonious representation of the local variation that would retain the same power to detect associations. For illustration, consider a sample of haploid genomes ($N_S = N$). Assume that we could define a functional unit of that genome, say for example the transcribed region of a gene, with some known left and right bounds and that we could know the change in phenotype expected for each unique sequence over that region. If we were then to cluster unique sequences with the same phenotypic effect into K functionally distinct haplotype groups, we could define a matrix of indicators, $Z \in \{0, 1\}^{N \times K}$, coding for group membership. For a continuous phenotype, we might fit a model such as

$$y = A\alpha + \theta Z\beta + \varepsilon \quad (1.2)$$

where the effect of each haplotype group, β_k , $k \in \{1, \dots, K\}$, is a known constant, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$, and $\theta \in \mathbb{R}$ is the parameter of interest ($\theta = 0$ under the null hypothesis). When investigators attempt to model interactions between genes, they are hoping to find combina-

tions of SNPs which together define one of these K , functionally distinct, haplotype groups. This example shows that we could hypothetically retain all of the relevant information in the underlying ancestral tree with just the assignments of the observed sequences into clusters of functionally distinct haplotypes. In other words, we do not need to reconstruct the entire tree. We will collectively refer to methods that target this haplotype clustering as "haplotype methods."

A GWAS method will have more power if it can compress the haplotype variation in each region into a representation that is likely to track with phenotypic variation, while remaining robust to phenotypically neutral variation. Of course this idealized model assumes that we know a) how to cluster distinct haplotypes into groups that have the same phenotypic effect, β_k , and b) how to draw left and right bounds on our region of interest. While these two pieces of "oracle" information can never be fully known in practice, methods hoping to target this haplotypic representation of local genetic variation must address them with assumptions. In the following section, we demonstrate how these two types of assumptions arise in a leading haplotype method, the Sequential Kernel Association Test (SKAT) [12]. We then proceed to show how these assumptions may be improved and refined with our proposed method.

1.2 Random Effects Models

In this section, we will use the popular Sequential Kernel Association Test (SKAT, [12]) as a prototypical example of a random effects model for testing local haplotype structure; however the analysis and critiques provided here generalize to other related random effects models.

1.2.1 Derivation

SKAT, and several variants such as SKAT-O, are designed to test regions of the genome for allelic heterogeneity and rare SNP effects [12, 13]. SKAT does not require phasing since it operates on genotype data. It takes a window of SNPs around a locus of interest l , say from position a to position b , and uses a positive definite kernel to construct a covariance matrix K that reflects the similarity between all pairs of genotypes over the $[a, b]$ interval. For continuous outcomes, SKAT implements a score test based on the following linear mixed model (in vector notation):

$$y = A\alpha + u + \varepsilon$$

where $u \sim N(0, \tau K)$ and $\varepsilon \sim N(0, \sigma_e^2 I_n)$. In order to test $H_0 : \tau = 0$, SKAT uses the score test statistic for τ . Below we derive the score test for a variance component in a slightly more general class of models since the result will be helpful in later discussions.

Let $y \sim N(\mu, \Sigma_\tau)$ where Σ_τ is some function of the parameter of interest, τ . Let $y \sim N(\mu, \Sigma_0)$ under the null hypothesis. Let $M = \left. \frac{\partial \Sigma_\tau}{\partial \tau} \right|_{\tau=0}$. Below we derive the score test statistic for τ , starting with the log likelihood, $l(y)$.

$$l(y) = -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_\tau| - \frac{1}{2} (y - \mu)^T \Sigma_\tau^{-1} (y - \mu)$$

$$\frac{\partial l(y)}{\partial \tau} = -\frac{1}{2} \text{tr} \left(\Sigma_\tau^{-1} \frac{\partial \Sigma_\tau}{\partial \tau} \right) + \frac{1}{2} (y - \mu)^T \Sigma_\tau^{-1} \frac{\partial \Sigma_\tau}{\partial \tau} \Sigma_\tau^{-1} (y - \mu)$$

$$\left. \frac{\partial l(y)}{\partial \tau} \right|_{\tau=0} = -\frac{1}{2} \text{tr} (\Sigma_0^{-1} M) + \frac{1}{2} (y - \mu)^T \Sigma_0^{-1} M \Sigma_0^{-1} (y - \mu)$$

Note that the tail probability of a statistic does not change with the addition or multiplication of constants that do not involve the data y , hence we arrive at

$$Q_{gauss} = (y - \mu)^T \Sigma_0^{-1} M \Sigma_0^{-1} (y - \mu). \quad (1.3)$$

Since this score test statistic applies for general Σ_0^{-1} , this test statistic can be used for generalized least squares models. In the case of SKAT where we assume homoskedasticity of the residuals, $\Sigma_\tau = \tau K + \sigma_e^2 I$, so $M = K$ and $\Sigma_0^{-1} = \sigma_e^{-2} I$. Factoring out σ_e^{-2} , we can write the SKAT score statistic for Gaussian phenotypes as

$$Q_{gaussSKAT} = (y - \hat{\mu})^T K (y - \hat{\mu}), \quad (1.4)$$

where $\hat{\mu} = A\hat{\alpha}_{null} = E(y)$ fit under the null model, as presented in the original SKAT paper by Wu et al. [12]. This test statistic will become large if $(y_i - \hat{\mu}_i)$ and $(y_j - \hat{\mu}_j)$ tend to be close when K_{ij} is large. In other words $Q_{gaussSKAT}$ will be significant when individuals who are more closely related according to K have relatively similar outcomes. Wu et al. also present a version of SKAT for binary outcomes motivated by a generalized linear mixed model for binary outcomes takes the form

$$y \sim \text{Bern}(\theta) \quad \text{logit}(\theta) = X\beta + u \quad (1.5)$$

where $u \sim N(0, \tau K)$.

Wu et al. do not provide the derivation of the score statistic for τ in model 1.5. They cite literature tracing back to Zhang and Lin (2003) [14], which provides a Laplace-approximation-based version of the score statistic for a more general class of models that includes variance components under the null model (note the absence of an error term ε or other null model variance components in Equation 1.5). We follow the same approach to derive the test statistic in the special case of Equation 1.5, which allows us to obtain the test statistic in a simpler closed form without the need for a Laplace approximation. This brings us to the following lemma (proof shown at the end of the present chapter).

Lemma 1.2.1. *Let $y \sim \text{Bern}(\mu)$, $\text{logit}(\mu) = X\beta + u$, and $u \sim N(0, \tau K)$. The score test statistic for $\tau = 0$ is*

$$Q_{\text{binarySKAT}} = (y - \hat{\mu})^T K (y - \hat{\mu}), \quad (1.6)$$

where $\hat{\mu} = \mathbb{E}(y)$ estimated under the null model when $\tau = 0$.

$Q_{\text{gaussSKAT}}$ and $Q_{\text{binarySKAT}}$ take the same form and both provide a convenient one-degree-of-freedom test for whether the structure in the covariance matrix K explains phenotypic variation. Although there are variants on the original form of SKAT that construct K slightly differently, to construct K for a given region they all require the user to define a fixed window of the genome. Then, for the p variants in that window, a kernel function is applied to each pair of genotypes based on some weighted distance metric. Let $G_{il} \in \{0, 1, 2\}$ be the genotype of sample i at locus l . For the covariance between genotypes i and j , a few examples given in [12] are as follows:

$$\begin{aligned} K_{ij} &= \sum_{l=1}^p w_l G_{il} G_{jl} && \text{(weighted linear kernel)} \\ K_{ij} &= \left(1 + \sum_{l=1}^p w_l G_{il} G_{jl} \right)^2 && \text{(quadratic kernel)} \\ K_{ij} &= \sum_{l=1}^p w_l (2 - |G_{il} - G_{jl}|) && \text{("additive" kernel)} \\ K_{ij} &= \sum_{l=1}^p w_l \mathbf{1}_{\text{IBD}} \{G_{il}, G_{jl}\} && \text{(IBD kernel)} \end{aligned}$$

Here, w_l refers to the weight given to locus l , which is usually taken to be a function of the allele frequency at that locus such as $w_l = (\text{MAF}_l (1 - \text{MAF}_l))^{-1}$ where MAF is the minor allele frequency. Assuming that there is some external method to assign identity-by-descent (IBD) to each pair of genotypes, $\mathbf{1}_{\text{IBD}} \{G_{il}, G_{jl}\}$ is the indicator of IBD. As we will see in the following chapter, our proposal for constructing a matrix K is most closely related to a probabilistic version of this IBD kernel.

1.2.2 Challenges Facing Standard Random Effects Models

Recalling our oracle model, Equation 1.2, we see that SKAT assumes that a region of interest in the genome can be defined by some fixed window of SNPs. In practice, these windows are usually centered on genes of interest, but this approach does not provide guidance in non-coding portions of the genome, where regulatory elements may still drive active associations. Some related methods attempt to circumvent this challenge by using a sliding window with overlapping segments along the genome. However, we are still left with the problem of choosing the width of such regions and how far to slide them with each increment. Drawing too narrow of a region could miss causal haplotype subgroups which are associated with different expected phenotypes. Drawing too wide of a region could lead us to consider spurious haplotypic variation that has no impact on the phenotype, obscuring true associations. Inevitably, all haplotype methods need to make some assumptions about how wide of a region to consider when comparing local haplotypes. Should the association signals driven by two nearby SNPs or haplotypes be combined into a single signal or resolved as distinct smaller signals? We will return to this question later after we present alternative approaches that haplotype methods use to define a local region of the genome.

If we compare SKAT to our Oracle Model, Equation 1.2, we see that SKAT assumes that the covariance between the phenotypes of two individuals will be some monotonic function of the percentage of their local genotypes that match. Compressing local genetic variation in this way treats the mutations in a region as exchangeable. In other words, the distance between the local genotypes of two samples does not depend on the order of their allelic mismatches (Figure 1.1). This exchangeability assumption throws away the information in underlying haplotype block structure.

Fig. 1.1 Exchangeability Under SKAT Kernel: Sample A depicts four pairs of haplotypes collected from four individuals (haplotype pairs are depicted by the colored brackets to the right of the haplotypes). In this region, there is a 8-SNP-long stretch where haplotypes may be either type blue or type orange. To the right of the four haplotype pairs are the corresponding four genotypes, to which we've applied the unweighted linear kernel to produce a local kernel matrix on the right. Note that the gold and red genotypes have the lowest pairwise similarity (11), reflecting the fact that the gold sample has two blue haplotypes and the red sample has two orange haplotypes. Sample B is identical to Sample A except that the columns have been permuted to break up the 8 SNP haplotype blocks. Notice that while this yields different genotypes, the implied kernel matrix, K , is identical, hence the exchangeability of loci under the SKAT kernel. An efficient compression of the 8 SNP region in sample A might use a binary indicator for either the blue or orange block.



Modelling loci as exchangeable and independent leads SKAT and related tests to lose power when many non-causal loci are included in a region of interest [15]. Rewriting the SKAT linear mixed model in terms of its Bayesian analogue, a connection pointed out by Listgarten (2012), explains this behavior [16]. Consider a Bayesian linear model where

$$y_i \sim \alpha + X_1 \beta_1 + \cdots + X_L \beta_L + \varepsilon. \quad (1.7)$$

Here, X_l is the genotype vector at locus l and we are considering testing a window with L SNPs. For continuous outcomes, we may write this regression model as $y \sim X\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$ and $\beta \sim N(0, \tau \text{Diag}(w))$ for some weights vector w . Under this Bayesian specification,

$$\text{Var}(y) = \mathbb{E}[\text{Var}(y|\beta)] + \text{Var}[\mathbb{E}(y|\beta)] = \sigma_\varepsilon^2 I + \text{Var}[X\beta] = \tau X \text{Diag}(w) X^T + \sigma_\varepsilon^2 I. \quad (1.8)$$

Recognizing that $X \text{Diag}(w) X^T$ is the SKAT weighted linear kernel shows that testing $\tau = 0$ in this Bayesian model is equivalent to testing $\tau = 0$ in the original SKAT linear mixed model framework. Adding many non-causal predictors to Model 1.7 would shrink the posterior of τ towards zero and reduce the power to reject $H_0 : \tau = 0$.

1.3 Non-exchangeable Models

If we believe that causal variation will tend to ride on haplotype blocks, then we should have greater power to detect associations if we have a means of compressing local variation that identifies sub-haplotypes and uses those rather than individual SNPs as predictors in a Bayesian regression model like 1.7. Essentially we'd have a model like 1.7 where a group of SNPs on a common haplotype would be collapsed into a single predictor coding for the "presence of Haplotype version A vs. Haplotype version B." This observation motivated several earlier haplotype methods, including the work of Browning & Browning (2007), which uses a variable-length Markov chain model to assign haplotypes to clusters which change along the genome [17]. Although Browning & Browning (2007) used a Fisher's Exact Test for testing whether the cluster assignments predicted case-control status, haplotypes assigned to the same local cluster could be declared to have IBD, allowing for these cluster assignments to be used by SKAT with the IBD kernel. Although this method does not require

windowing, it forces hard assignments to fixed clusters which does not fully account for uncertainty in the clusterings or the length over which haplotypes are compared.

Tree-based methods, such as *LATAG* [18] and *mARGarita* [19], attempt to account for the hierarchical, multi-resolution, tree-based structure and uncertainty in haplotype relatedness by integrating over marginal trees that could explain the local haplotypes. However, difficulty arises in testing these structures. For example, *mARGarita* finds the branch in each proposed tree which best separates cases from controls according to a χ^2 test and averages the trees at each locus to produce summary scores. A more precise tree sampling method, *TMARG* [20], samples uniformly from the set of time-ordered MinARGs (Ancestral Recombination Graphs which minimize the number of recombination events required to produce the observed sample) [21]. However, largely because trees lie within a discrete space, these methods do not readily scale to thousands of genomes.

While tree-based methods may "correctly" model the local dependence structure among SNPs, it is difficult to modify them to control for population structure. The problem arises when some non-genetic effect acts under a subtree of the marginal tree, causing any mutations on the branch above that subtree to show a spurious correlation with the phenotype [22]. Most regression-based GWAS approaches penalize variants that are liable to be confounded with environmental factors by incorporating fixed and random effects. Let X denote the N by L binary matrix indicating the SNP-status of every sampled individual at every locus. A leading approach is to include a random effect with a covariance structure given by XX^T or some customized SNP set [23–25]. A second, nearly equivalent approach is to use the first few principal components of XX^T as fixed effects [26, 27]. The underlying assumption of these two approaches of accounting for confounding environmental structure: SNPs which explain a large proportion of the variants will tend to lie on branches that roughly separate sub-populations or distinct environments. As we might expect, if a strong environmental

confounder operates on a small geographical scale, this assumption is violated and the principal components will not control for the confounding process [28].

One might attempt to account for fine-scale confounders by including rare variants in the SNP set that is used to generate the random effect covariance matrix or in the PCA calculation. However, this comes at the risk of obfuscating a true causal variant by including a SNP that is linked to it in the set of potential confounders. This "proximal contamination" is what has led to some methods to exclude SNPs from the current chromosome of interest in calculating the "background" relatedness matrix [9]. In practice, investigators apply this reasoning by adjusting for just enough genome-wide ancestry structure to eliminate the inflation of p-values along the genome. We will return to this balance of adjusting population structure in Chapter 4, but for now we simply emphasize that adjusting for population-structure-based principal components or random effects can be really challenging for methods that do not take a regression approach to testing, such as most tree-based approaches.

A groundbreaking compromise between the Markov-chain based latent factor model of Browning & Browning (2007) and tree-based methods that could easily use the PCA method to account for population structure came with the two-layer Hidden Markov Model (HMM) proposed by Xu & Guan (2014) [29]. Their approach models genotypes as being emitted from a fixed set of K ancestral haplotypes, which themselves are emitted from a set of S hypothetical populations. A recombination event between ancestral haplotypes from different hypothetical populations is modeled as less likely than a recombination event between ancestral haplotypes from the same hypothetical population. Essentially, the K lower level haplotypes of the HMM model tree structure near the leaves and the S upper level states model tree structure near the root. While this two-layer HMM models the structure of trees, critically, it also loads observed genotypes onto K inferred donor haplotypes. In other words, it provides a soft, probabilistic clustering of the observed genotypes into K clusters rather than a hard clustering. This enables a simple testing procedure that assesses whether

cases or controls preferentially associate with any of the K groups, which is relatively computationally efficient. Since the HMM integrates over recombination events (modeled as ancestral haplotype switches), it avoids windowing the genome, although it still must infer parameters relating to the recombination rate that determine the amount of ancestral information flow along a region of the genome. It also infers the ancestral haplotype loadings of each genotype at each SNP, so the number of tests is the same as in a conventional single-marker-based GWAS.

Since Xu & Guan's two-layer HMM approach models trees within a continuous space of copying probabilities rather than in their original discrete space, it can scale to tens-of-thousands of genomes. To achieve this scalability, their method assumes that all regions of the genome can essentially be represented by a fixed number of K underlying donor haplotypes. In this sense, the model may be considered too rigid. However, the model is already highly parameterized: not only does the model have to learn the latent haplotypes, each of which is very high dimensional, but there is also a matrix of transition probabilities for each layer in addition to mutation/emission parameters. As a probabilistic latent factor model with a very high-dimensional parameter space, it is very likely that EM will get caught in local modes. Although one might at first waive this criticism since this haplotype association method is really aimed at performing local ancestry inference at each region of the genome, the authors report that their inferred haplotype loadings varied drastically between EM iterations using the exact forward-backward algorithm. In other words, multimodality was a problem that produced substantially different local ancestry inferences. The authors partially circumvent this issue by averaging local ancestry inferences over multiple EM runs, which in simulation turns out to yield useful renderings of the local haplotype structure. However, the fact that the underlying model yields inconsistent local inferences suggests that the underlying model might not be a parsimonious representation of genetic variation.

Another key problem with the complexity of the two-layer HMM approach is that it required an approximate forward-backward algorithm that is linear rather than quadratic in the number of hidden haplotypes K to scale to tens-of-thousands of genomes. Xu & Guan call this method stochastic linear iterative marginalization (SLIM). Although the authors provide a figure showing that the local haplotype clusterings inferred via SLIM correlate with those produced by the exact forward-backward algorithm according to a single metric, this does not imply that this correspondence will hold on other datasets or that the inferred clusterings between the two methods can be used interchangeably. The two-layer HMM is an example of a relatively exact model that can only be approximately computed, rather than a simple, approximate model that can be exactly computed.

These limitations motivated us to look for a simpler HMM-based approach to inferring representations of local ancestries that would be amenable to exact inference while avoiding restrictive assumptions like the existence of strictly K hidden haplotypes. To do this, we returned to a highly-cited predecessor of Xu & Guan’s two-layer HMM, the Li & Stephens model [30, 31], which we introduce in the following section. Since the Li & Stephens model is based on haplotypes rather than genotypes, direct use of the model requires for the data to be phased *a priori*. While Xu & Guan’s two-layer HMM avoid phasing errors by working directly on genotypes, with the emergence of larger haplotype reference panels and more efficient phasing algorithms [32], it appears that requiring data to be phased might not introduce too many errors or be computationally prohibitive. Furthermore, recent work suggests that modeling phased haplotypes might explain a large proportion of the missing heritability in complex traits [33]. We will explicitly revisit this point in Chapter 4, where we explicitly show how information relevant to detecting association can be lost when haplotypes are reduced to genotypes.

1.4 The Li & Stephens Model

The Na Li & Matthew Stephens (LS) model has become a widely popular model for genetic diversity [31]. Although the model arises via an approximation to the coalescent with recombination, it has been proven sufficiently accurate in many applications ranging from recombination map construction to demographic inference. Its primary advantage is that the model lends itself to inference via a Hidden Markov Model, making exact inference possible for large genomics datasets. The use of derivation pathways, which can be seen as a special case of the LS model when the mutation rate is brought to zero, has also been shown to not only be scalable for applications in association testing, but also to compete favorably beside much more exact Ancestral Recombination Graph (ARG) inference methods [21, 20].

A key motivation for focusing on the LS model comes from the fact that the tree-based association software *TMARG* defaults to sampling ARGs based on a copying heuristic that is a special case of the LS model when mutation is not allowed. Under this LS-type model, *TMARG* performed approximately as well empirically as it did when sampling true minARGs [20]. Below we review the motivations and assumptions underlying the LS model [34] as an approximation to the ARG [21]. In Chapter 2, we will show how it may be adapted to yield a scalable approach for approximating local ancestries at each position in the genome.

1.4.1 PAC Likelihood

Consider a sample of N haploid genetic sequences of length L . Let μ be a vector of length L denoting the expected number of mutations to occur at site l in a single meiosis (one generation). Let r be a vector of length $L - 1$ such that $r(l)$ denotes the probability of a recombination event between the l^{th} and $(l + 1)^{th}$ locus in a single meiosis. Let $\mathcal{G}$ denote the true ARG underlying the sample of N individuals. Assume that all sampled sites are biallelic. Let the n^{th} haplotype, h_n , be a binary vector of length L denoting the sequence of alleles of the n^{th} sample. Let h_n^l denote the allelic state at a specific locus, l . Let H denote the

entire sample $\{h_1, \dots, h_N\}$. Define vectors $\rho = 4N_e r$ and $\theta = 4N_e \mu$, where N_e is the effective population size.

The LS model is motivated by a full-likelihood approach devised by Fearnhead & Donnelly, who sought to sample from $P(H|\rho, \theta)$ under the Coalescent with Recombination by integrating over all possible $\mathcal{G}$ [35]. In an attempt to overcome the computational barrier to this approach, Li & Stephens (2003) sought to approximate the Fearnhead & Donnelly likelihood with a “product of approximate conditionals” (PAC) likelihood:

$$\begin{aligned} P(H|\rho, \theta) &= P(h_1|\rho, \theta)P(h_2|h_1, \rho, \theta) \cdots P(h_N|h_1, \dots, h_{N-1}, \rho, \theta) \\ &\approx \pi(h_1|\rho, \theta) \pi(h_2|h_1, \rho, \theta) \cdots \pi(h_N|h_1, \dots, h_{N-1}, \rho, \theta) \end{aligned}$$

Each π approximates the probability of observing a specific haplotype having already observed n of N haplotypes in a sample. Li & Stephens used a Hidden Markov Model (HMM) to approximate each conditional likelihood by arguing that a haplotype could be approximated as a series of segments copied from similar, previously observed haplotypes, with segment boundaries determined by recombination events and miscopied sites corresponding to mutation events. Let us refer to the $(n+1)^{th}$ haplotype as the “recipient” haplotype and the haplotype that it copies from for a given segment its “donor” haplotype. The emission distribution (ψ) and the transition distribution (ϕ) are motivated by the coalescent where the donor haplotype is thought to be under the ancestral lineage that the recipient haplotype first coalesces with. In other words, the donor should be a “nearest-neighbor” or tied to be a “nearest-neighbor.” Let $X_{n+1}^l \in \{1, \dots, n\}$ be the identity of the donor haplotype by a new haplotype at locus l , the hidden state. By symmetry, without any information of the marginal coalescent tree to the left or right of a given position, every previously observed haplotype (each potential donor) has equal probability a priori of being the nearest neighbor. This motivated the choice of a discrete uniform prior ξ on the hidden states. In the standard form of a HMM likelihood, the LS model approximate conditional distribution can be written as

$$\begin{aligned}
\pi(h_{n+1}|h_1, \dots, h_n) &= \sum_{X_n^{1:L}} \pi(h_{n+1}, \{X_{n+1}^1, \dots, X_{n+1}^L\} | h_1, \dots, h_n) \\
&= \sum_{X_n^{1:L}} \left(\xi(X_{n+1}^1) \prod_{l=1}^{L-1} \phi(X_{n+1}^{l+1} | X_{n+1}^l) \prod_{l=1}^L \psi(h_{n+1}^l | X_{n+1}^l) \right).
\end{aligned}$$

1.4.2 Emission Distribution

The emission density, ψ , is rooted in research into optimal proposal distributions for Importance Sampling marginal trees under the Coalescent [36]. We need the probability that a new $(n+1)^{th}$ haplotype will have the same allelic state (0 or 1) as its specific donor haplotype in the previously observed set of n haplotypes. In other words, what is the probability that the recipient haplotype will copy exactly from its donor?

We can approximate this probability (Q_n) based on the coalescent mutation parameter θ by first calculating the probability that, in a tree of $n+1$ haplotypes sampled under the coalescent, there is no mutation on the path connecting the recipient haplotype to its donor haplotype. Since the donor haplotype must be under the ancestral lineage that the recipient haplotype first coalesces with, this is equal to one minus the probability that a mutation occurs on the ancestral lineage of the recipient or the donor before they coalesce. We can derive this probability by conditioning on the number of ancestors that remained in the coalescent when the $(n+1)^{th}$ sample joined the rest, yielding Q_n below.

$$Q_n = 1 - \binom{n}{2}^{-1} \sum_{k=2}^n (k-1) \prod_{j=k}^n \left(1 + \frac{\theta}{j(j-1)}\right)^{-1} \approx \frac{\theta}{n+\theta} \quad \text{for } n \text{ large.} \quad (1.9)$$

Numerical calculation confirms that Q_n is very well approximated by $\frac{\theta}{n+\theta}$ for even modest n and high mutation rates. Therefore, Li & Stephens approximate the probability that h_{n+1} directly copies from one of the $h_1, \dots, h_n$ with $\frac{n}{n+\theta}$.

If we assume parent-independent mutation (PIM), the allele of a mutant does not depend on the allele of its parent; therefore, when a mutation happens, we assume that the SNP flips to another allele or stays the same with probability $1/2$. These assumptions specify an emission kernel:

$$\psi(h_{n+1}^l | X_{n+1}^l) = \begin{cases} \frac{n}{n+\theta} + \frac{1}{2} \frac{\theta}{n+\theta} & : h_{n+1}^l = h_{X_{n+1}^l}^l \\ \frac{1}{2} \frac{\theta}{n+\theta} & : h_{n+1}^l \neq h_{X_{n+1}^l}^l \end{cases}.$$

Consistent with intuition, copying is perfect in the limit as the mutation rate approaches zero and $Pr(h_{n+1}^l = 0) = Pr(h_{n+1}^l = 1)$ in the limit as the mutation rate approaches infinity.

1.4.3 Transition Distribution

Although Li & Stephens do not clearly explain how they arrived at their transition density [34], here we provide a rough justification for their treatment of recombination. To begin, it is important to note that the LS model makes a Markov approximation of the Coalescent with Recombination that is closely related to a formalized stochastic process that was presented two years after the LS model: the sequentially Markov coalescent (SMC) [37]. Under the SMC, recombination events between loci $l-1$ and l are independent of the marginal trees above loci $\{1, \dots, l-2\}$ conditional on the marginal coalescent tree above locus $l-1$. Given the marginal coalescent tree above locus $l-1$, recombination events occur as a Poisson process on the tree with rate $\frac{\rho_l}{2}$ under the SMC, just like mutation events. If we assume that the recombination rate is low enough that events where more than one recombination event occur at the same locus may be ignored, the probability that a recombination event occurs on the recipient haplotype's lineage before its first coalescence, is $1 - \exp\left(\frac{-\rho_l}{2} (T^*)\right)$ where T^* is the time to recipient haplotype's first coalescence event. A classic calculation gives the expected value of T^* [38]. Let k here denote the number of lineages remaining when the recipient haplotype first coalesces.

$$\begin{aligned}
\mathbb{E}[T^*] &= \mathbb{E}[\mathbb{E}[T^*|k]] \\
&= \sum_{k=2}^n \mathbb{E}[T^*|k] \left(\frac{k-1}{\binom{k}{2}} \prod_{j=k+1}^n \left(1 - \frac{j-1}{\binom{j}{2}} \right) \right) \\
&= \sum_{k=2}^n \mathbb{E}[T^*|k] \left(\frac{2}{k} \prod_{j=k+1}^n \left(1 - \frac{2}{j} \right) \right) \\
&= \frac{2}{n(n-1)} \sum_{k=2}^n (k-1) \mathbb{E}[T^*|k] \\
&= \frac{2}{n(n-1)} \sum_{k=2}^n \left((k-1) \sum_{j=k}^n \frac{2}{j(j-1)} \right) \\
&= \frac{2}{n(n-1)} \sum_{k=2}^n \left((k-1) \left(\frac{2}{k-1} - \frac{2}{n} \right) \right) \\
&= \frac{2}{n}
\end{aligned}$$

If we now substitute T^* with its expectation, the probability of a recombination event comes to $1 - \exp\left(\frac{-\rho_l}{n}\right)$. Note that, under the SMC, this result is consistent with our treatment of mutation events. For small recombination rates and large n , $1 - \exp\left(\frac{-\rho_l}{n}\right) \approx \frac{\rho_l}{n+\rho_l}$, which is the probability of a recombination event before coalescence that we would have obtained from Equation 1.9 if we had treated recombination events like mutation events.

Given that there is a recombination event on the lineage of the recipient haplotype before it coalesces into the tree, the LS models assumes that the next donor that the recipient copies is chosen uniformly at random. This choice is consistent with making the copying density, ξ , discrete uniform on $\{1, \dots, n\}$; however, as is done in practice in ChromoPainter [39], these uniformity assumptions can be generalized to account for the fact that in a real dataset, not all individuals are equally related to each other within a population. Let δ_n^l be the indicator that there was at least one “recombination event” between the $(l-1)^{th}$ and l^{th} loci affecting the recent ancestry of the $(n+1)^{th}$ individual. Li & Stephens propose the following transition density.

$$\begin{aligned}
\phi(X_{n+1}^l | X_{n+1}^{l-1}) &= P(X_{n+1}^l = x' | X_{n+1}^{l-1} = x) \\
&= P(X_{n+1}^l = x' | X_{n+1}^{l-1} = x, \delta_{n+1}^l = 1)P(\delta_{n+1}^l = 1) + \\
&\quad P(X_{n+1}^l = x' | X_{n+1}^{l-1} = x, \delta_{n+1}^l = 0)P(\delta_{n+1}^l = 0) \\
&= \frac{1}{n} \left(1 - \exp \left[\frac{-\rho_l}{n} \right] \right) + \mathbf{1}(x' = x) \exp \left[\frac{-\rho_l}{n} \right]
\end{aligned}$$

When Li & Stephens use their model to actually estimate recombination rates, they found that they could improve the performance of the model by taking the recombination map to a power, say γ . So they defined, $\rho_l = 4N_e r_l^\gamma$, where as before, r_l is the distance between sites l and $l + 1$ in Morgans. A large $\gamma > 1$ makes the recombination map more heavily peaked whereas a small $\gamma < 1$ tempers the recombination map to make it more flat and smooth. When we return to the LS Model in Chapter 2, we will consider this more general parameterization of the LS model with a γ exponent.

1.4.4 Pseudo-likelihood Approximation

In order to make the N haplotypes observed in a sample exchangeable in $P(H|\rho, \theta)$ and avoid maximizing $P(H|\rho, \theta)$ over many permutations of the haplotypes, Li & Stephens also suggest using a pseudo-likelihood approximation [40] before invoking the approximate conditional distribution π :

$$P(H|\rho, \theta) \approx \prod_{n=1}^N P(h_n|\rho, \theta, H_{-n}) \approx \prod_{n=1}^N \pi(h_n|\rho, \theta, H_{-n})$$

where $H_{-n} = H \setminus h_n$.

This pseudo-likelihood approximation was made popular with ChromoPainter and fineSTRUCTURE [39], where it has proven very effective in resolving fine-scale structure in overall ancestry inference based on comparing whole genomes.

In all future mentions of the LS Model, we will implicitly be referring to the pseudo-likelihood extension in which each haplotype is painted as a mosaic of all of the other haplotypes in the sample. This approach corresponds to n independent HMMs, each with $n - 1$ hidden states. Similar to Xu & Guan two-level HMM approach [29], we will be interested in the posterior copying probabilities implied by these HMMs and use that as a basis for defining local relatedness.

1.4.5 LS Model Advantages

Although the ability of the posterior marginals of the LS model to capture local structure at a particular locus has not been thoroughly studied, it achieves better genome-wide ancestry resolution than more classic, genotype relatedness matrices, which is the genome-wide analog of the linear SKAT kernel, by accounting for linkage between sites [39]. ChromoPainter achieves this resolution by defining the relatedness between two individuals as the approximate number of haplotype blocks that they share rather than the number of SNPs they share. Similarly, we hope to obtain better local ancestry resolution (local to a region of the genome) by using the LS model to capture local haplotype sharing rather than local SNP sharing like SKAT kernels.

As we will show in Chapter 2, the HMM likelihood typically places a higher penalty if a donor haplotype mismatches the recipient haplotype on a rare SNP rather than on a common SNP. However, a more nuanced investigation of this effect reveals that the HMM places a higher penalty on a mismatch between donor and recipient if the recipient's local sequence is rare among the haplotypes in the sample. So while the LS model, like SKAT, must make assumptions about how to weight mutations of different frequencies in determining local genetic similarity, the LS model penalizes haplotype blocks and mutations of different frequencies with different penalties all based on a single, coalescent-motivated, mutation parameter. As with specifying a single mutation parameter, the LS model also summarizes the

flow of ancestral information along the genome with a single recombination parameter. This two-parameter specification makes optimizing the LS model for a given genetic dataset much easier and less arbitrary than specifying a series of weights. In Chapter 2, we will return to how we optimize the mutation and recombination parameters to capture local genealogies.

By using a HMM-based approach that accounts for recombination, like Xu & Guan's two-level HMM, we avoid hard windowing and are better able to capture the staggered structure of local haplotype blocks. If one set of samples share a very long local haplotype while other samples share a very short local haplotype, the LS model can incorporate the added information from the longer haplotype showing that those samples are very recently related without diluting the genetic similarity measured among the samples carrying the shorter haplotype with a wide, fixed-window.

This ability of the LS model to naturally incorporate haplotype length information means that it might be especially powerful for identifying haplotypes that have been under recent positive natural selection and are associated with the trait of interest. In Chapter 4, we will apply our new method to a Malaria GWAS dataset where the outcome, severe malaria susceptibility, has driven very strong recent positive selection at multiple loci. By capturing extended haplotype homozygosity, we expect our method to incorporate this selection information and thereby improve our power to find association signals.

1.5 Correctly Testing Quadratic Forms

In order to use a quadratic form like $Q_y = (y - \hat{\mu})^T K (y - \hat{\mu})$ as a test statistic in a Fisherian framework, it is necessary to obtain p-values under the null hypothesis. While it is simple to express the null as a linear combination of independent χ^2 random variables when y is Gaussian, when y is non-Gaussian, as in the case of binary case-control outcomes, the distribution of Q is difficult to obtain. SKAT and other related approaches usually circumvent this problem by asserting that, asymptotically, Q_y will converge to $Q_x = (x - \hat{\mu})^T K (x - \hat{\mu})$

in distribution, where each $x_i \sim N(\hat{\mu}_i, \text{Var}(y_i))$. This assumption then allows the authors to treat the quadratic forms for binary case-control outcomes as if they were quadratic forms in normal random variables and use the linear combination of χ^2 random variables representation to obtain a p-value. Below we explain why this central limit claim is not generally true and how examples that violate this claim could easily arise in applications where investigators might want to apply SKAT.

To get an immediate intuition for why Q_y will not converge to Q_x , one only has to compare the variance of the two test statistics.

Lemma 1.5.1. *Consider a real matrix $B \in \mathbb{R}^{N \times N}$ and a vector of independent random variables $z \in \mathbb{R}^N$. Let $b \in \mathbb{R}^N$ be a vector encoding the diagonal of B and $\tilde{B} = \frac{1}{2}(B + B^T - 2\text{diag}(B))$. Here, $\text{diag}(B)$ is the diagonal matrix with b down the diagonal, and for a vector $v \in \mathbb{R}^n$, we will use $\text{Diag}(v)$ to denote the diagonal matrix with v down the diagonal. Define the real vectors μ , σ^2 , γ and κ such that*

$$\begin{aligned}\mu_i &= \mathbb{E}(z_i), \\ \sigma_i^2 &= \mathbb{E}(z_i^2) - \mu_i^2, \\ \gamma_i &= b_i(\mathbb{E}(z_i^3) - \mu_i \mathbb{E}(z_i^2)), \\ \kappa_i &= \mathbb{E}(z_i^4) - (\mathbb{E}(z_i^2))^2.\end{aligned}$$

Then

$$\text{Var}(z^T B z) = \kappa^T (b * b) + 2\sigma^{2T} (\tilde{B} * \tilde{B}) \sigma^2 + 4\mu^{2T} (\tilde{B} * \tilde{B}) \sigma^2 + 4\gamma^T \tilde{B} \mu + 4\mu^T \tilde{B} \text{Diag}(\sigma^2) \tilde{B} \mu.$$

From Lemma 1.5.1, we can see that the variance is a function of the fourth moments of the random variables, even if $\mu = 0$. In general, the fourth order moments of a Gaussian will not match those of binary random variables. The influence of these fourth order moments will be especially important if the diagonal terms are non-negligible compared to the off

diagonal terms or the case-control prediction of certain individuals in the sample is close to 0 or 1 under the null model. This is a major obstacle for SKAT because, especially in a region of the genome with high diversity, the diagonal terms of a kernel matrix K can be very large relative to the off-diagonal elements of K . Note that since the matrix K used in SKAT is positive semi-definite, its diagonal terms must sum to a positive number, so this variance mismatch is a problem for all positive semi-definite K . Even if one were to rescale Q_x to match the variance of Q_y , pointing out the variance mismatch illustrates that there are higher-order interactions among the random variables in Q_y that cannot be matched by Q_x . This means that for all N , the moments will not match and therefore the distributions will not converge without additional assumptions.

In Chapter 3, we fill this gap in the literature with a set of theoretical and computational results which allow us to obtain p-values, or strict upper and lower bounds on p-values, for quadratic forms in non-Gaussian random variables.

1.6 Overview of Thesis

In this chapter we have motivated an HMM-based approach that will a) judiciously define regions of the genome over which to combine association signals based on local haplotype structure, b) generate a parsimonious representation of local haplotypic diversity by recognizing the non-exchangeability of loci, and c) has the potential to incorporate recent positive selection signals into association testing. Building on this, we present a series of results that allow for the direct use of the LS Model in haplotype association testing. Although we introduced the LS model from a coalescent perspective, in Chapter 2, we will examine it from a more algorithmic perspective with the aim of using the model to quickly calculate distances between all pairs of haplotypes at a given locus. We run N independent HMMs, allowing each individual to copy from everyone else in the sample. In this way, we avoid the K-haplotypes assumption by allowing each observation to be their own cluster. We show the

posterior marginal probabilities can be obtained exactly with reasonable CPU and memory requirements for simulated data with up to 20,000 individuals ($N = 40,000$). We also provide theoretical projections which show how our approach could scale to datasets with hundreds of thousands of individuals.

In Chapter 3, we develop a Krylov-subspace-based approach to efficiently upper and lower bounding the p-values of quadratic forms in Gaussian random variables. We then prove two invariance principles and present a related Rao-Blackwellization procedure which together allow us to apply our Krylov subspace approach for Gaussian quadratic forms to a large class of non-Gaussian quadratic forms.

In Chapter 4, we show how the machinery developed in Chapters 2 and 3 can be combined in a genetics context to quickly screen whether there is evidence of association at each locus along the genome. Although our haplotype-based approach requires phased haplotypes, we show in Chapter 4 how phasing provides access to relevant association information that is obscured by genotypes and allows us to test for different genetic architectures. We demonstrate the applicability of our approach to real genetics data in a pilot application to a malaria GWAS dataset.

Chapter 5 concludes with a few exciting new directions for this line of research, including detecting adaptive introgression between populations.

1.7 Proofs

1.7.1 Proof of Lemma 1.2.1

To derive the score statistic, first, rewrite model 1.5 as

$$y \sim \text{Bern}(\theta) \quad \log\left(\frac{\theta}{1-\theta}\right) = X\beta + \sqrt{\tau}K^{1/2}z \quad z \sim N(0, I) \quad (1.10)$$

The corresponding log-likelihood for a dataset with n observations $l(y, \tau)$, can be written up to a constant as

$$\log \left[\int_{\mathbf{R}^n} \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz \right], \quad (1.11)$$

where

$$h(y, \tau) = y^T \left(X\beta + \sqrt{\tau} K^{1/2} z \right) - \sum_i \log \left[1 + \exp \left(X_i^T \beta + \sqrt{\tau} K_i^{1/2} z \right) \right]. \quad (1.12)$$

This implies that

$$\frac{\partial}{\partial \tau} l(y, \tau) = \frac{\frac{\partial}{\partial \tau} \int_{\mathbf{R}^n} \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz}{\int_{\mathbf{R}^n} \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz} \quad (1.13)$$

$$= \frac{\int_{\mathbf{R}^n} \left[(y - \theta_\tau)^T K^{1/2} z \right] \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz}{2\sqrt{\tau} \int_{\mathbf{R}^n} \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz} \quad (1.14)$$

where $\theta_\tau = \mathbb{E}(y) = \frac{\exp(X\beta + \sqrt{\tau} K^{1/2} z)}{1 + \exp(X\beta + \sqrt{\tau} K^{1/2} z)}$. Note that as $\tau \rightarrow 0$, the denominator will tend to 0 because of the $\sqrt{\tau}$ term. The numerator will also tend to zero 0 as $\tau \rightarrow 0$ because for $\tau = 0$, the numerator is

$$\int_{\mathbf{R}^n} \left[(y - \theta_0)^T K^{1/2} z \right] \exp \left(h(y, 0) - \frac{1}{2} z^T z \right) dz \quad (1.15)$$

$$= \exp(h(y, 0)) (2\pi)^{n/2} \int_{\mathbf{R}^n} \left[(y - \theta_0)^T K^{1/2} z \right] \frac{1}{(2\pi)^{n/2}} \exp \left(-\frac{1}{2} z^T z \right) dz \quad (1.16)$$

$$= \exp(h(y, 0)) (2\pi)^{n/2} \mathbb{E} \left[(y - \theta_0)^T K^{1/2} z \right] = 0. \quad (1.17)$$

This gives us an indeterminate form, allowing us to apply L'Hopital's Rule. Applying a partial derivative with respect to τ to the numerator of Equation 1.14 yields

$$\frac{1}{2\sqrt{\tau}} \int_{\mathbf{R}^n} \left[\left((y - \theta_\tau)^T (K^{1/2} z) \right)^2 - [\theta_\tau (1 - \theta_\tau)]^T (K^{1/2} z)^2 \right] \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz$$

Applying a partial derivative with respect to τ to the denominator of Equation 1.14 yields

$$\left(\frac{1}{\sqrt{\tau}} \int_{\mathbf{R}^n} \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz \right) + \int_{\mathbf{R}^n} \left[(y - \theta_\tau)^T K^{1/2} z \right] \exp \left(h(y, \tau) - \frac{1}{2} z^T z \right) dz$$

Multiplying the new numerator and new denominator by $\sqrt{\tau}$ and noting that the second term in the denominator again goes to zero as $\tau \rightarrow 0$ as shown in Equation 1.17, we obtain

$$\begin{aligned} \lim_{\tau \rightarrow 0} \frac{\partial}{\partial \tau} l(y, \tau) &= \frac{\int_{\mathbf{R}^n} \left[\left((y - \theta_0)^T (K^{1/2} z) \right)^2 - [\theta_0 (1 - \theta_0)]^T (K^{1/2} z)^2 \right] \exp \left(h(y, 0) - \frac{1}{2} z^T z \right) dz}{2 \int_{\mathbf{R}^n} \exp \left(h(y, 0) - \frac{1}{2} z^T z \right) dz} \\ &= \frac{\int_{\mathbf{R}^n} \left[\left((y - \theta_0)^T (K^{1/2} z) \right)^2 - [\theta_0 (1 - \theta_0)]^T (K^{1/2} z)^2 \right] \exp \left(-\frac{1}{2} z^T z \right) dz}{2 \int_{\mathbf{R}^n} \exp \left(-\frac{1}{2} z^T z \right) dz} \\ &= \frac{1}{2} \mathbb{E} \left[\left((y - \theta_0)^T K^{1/2} z \right)^2 - [\theta_0 (1 - \theta_0)]^T (K^{1/2} z)^2 \right] \\ &= \frac{1}{2} \left((y - \theta_0)^T K (y - \theta_0) - [\theta_0 (1 - \theta_0)]^T \text{diag}(K) \right) \\ &= \frac{1}{2} \left((y - \theta_0)^T K (y - \theta_0) - \mathbb{E} \left[(y - \theta_0)^T K (y - \theta_0) \right] \right) \end{aligned}$$

Thus

$$Q_{\text{binarySKAT}} = (y - \hat{\theta}_0)^T K (y - \hat{\theta}_0)$$

1.7.2 Proof of Lemma 1.5.1

Lemma. Consider a real matrix $B \in \mathbb{R}^{N \times N}$ and a vector of independent random variables $z \in \mathbb{R}^N$. Let $b \in \mathbb{R}^N$ be a vector encoding the diagonal of B and $\tilde{B} = \frac{1}{2} (B + B^T - 2\text{diag}(B))$.

Here, $\text{diag}(B)$ is the diagonal matrix with b down the diagonal, and for a vector $v \in \mathbb{R}^n$, we will use $\text{Diag}(v)$ denote the diagonal matrix with v down the diagonal. Define the real vectors μ , σ^2 , γ and κ such that

$$\begin{aligned}\mu_i &= \mathbb{E}(z_i), \\ \sigma_i^2 &= \mathbb{E}(z_i^2) - \mu_i^2, \\ \gamma_i &= b_i (\mathbb{E}(z_i^3) - \mu_i \mathbb{E}(z_i^2)), \\ \kappa_i &= \mathbb{E}(z_i^4) - (\mathbb{E}(z_i^2))^2.\end{aligned}$$

Then

$$\text{Var}(z^T B z) = \kappa^T (b * b) + 2\sigma^{2T} (\tilde{B} * \tilde{B}) \sigma^2 + 4\mu^{2T} (\tilde{B} * \tilde{B}) \sigma^2 + 4\gamma^T \tilde{B} \mu + 4\mu^T \tilde{B} \text{Diag}(\sigma^2) \tilde{B} \mu.$$

Proof:

Let $B_S = \frac{1}{2} (B + B^T)$. Since $v^T B v = v^T B_S v \quad \forall v \in \mathbb{R}^N$, $\text{Var}(z^T B z) = \text{Var}(z^T B_S z)$. Thus, from here onwards, we assume without loss of generality that B is symmetric. Since the z_i are

independent, the only nonzero covariance terms are those involving a repeated term. Thus

$$\begin{aligned}
\text{Var}(z^T B z) &= \sum_i B_{ii}^2 \text{Var}(z_i^2) + 2 \sum_{j \neq k} B_{jk}^2 \text{Var}(z_j z_k) \\
&\quad + 4 \sum_i \sum_{j \neq i} B_{ii} B_{ij} \text{Cov}(z_i^2, z_j z_i) \\
&\quad + 4 \sum_{j, k, k' \text{ distinct}} B_{jk} B_{jk'} \text{Cov}(z_j z_k, z_j z_{k'}) \\
&= \sum_i B_{ii}^2 \kappa_i + 2 \sum_{j \neq k} B_{jk}^2 (\sigma_j^2 \sigma_k^2 + \mu_j^2 \sigma_k^2 + \mu_k^2 \sigma_j^2) \\
&\quad + 4 \sum_{i \neq j} B_{ij} \mu_j \gamma_i + 4 \sum_{j, k, k' \text{ distinct}} B_{jk} B_{jk'} \sigma_j^2 \mu_k \mu_{k'} \\
&= \kappa^T (b * b) + 2 \sigma^{2T} (\tilde{B} * \tilde{B}) \sigma^2 + 4 \mu^{2T} (\tilde{B} * \tilde{B}) \sigma^2 \\
&\quad + 4 \gamma^T \tilde{B} \mu + 4 \mu^T \tilde{B} \text{diag}(\sigma^2) \tilde{B} \mu.
\end{aligned}$$

Here we use $*$ to denote element-wise multiplication.

Chapter 2

Building Local Distance Matrices

Ascertaining the "genetic distance" between all pairs of phased haplotypes in a given region of the genome can be useful for detecting association signals. A simple algorithm for calculating haplotype distance would count the number of local mutations separating each pair of haplotypes and somehow account for the local recombination map. Below, we show how such an algorithm can be motivated by the Li & Stephens approximation to Kingman's Coalescent with Recombination. We show how a Hidden Markov Model derived from this model can be adapted to compute a matrix at each locus of the pairwise genetic distances between haplotypes.

This distance we will propose can be intuitively interpreted in the context of copying paths through a hidden Markov state space. For those familiar with applications of the Li & Stephens, it will be helpful to know up front that the distance we obtain between haplotypes i and j at locus l is the the added penalty of forcing i to copy j at l rather than allowing i to take its optimal path through the other haplotypes. We will return to this point in the text.

We then present an R Package which provides a convenient wrapper for a C core featuring low level vector CPU instructions, a special checkpointing scheme, and cache management to make these distance matrices tractable for modern-sized datasets.

2.1 Local Distances from Posterior Decoding

2.1.1 LS Model Without Recombination

Consider a sample of N haplotypes where we now paint the i^{th} haplotype as a mosaic of the $N - 1$ remaining haplotypes. Consider the posterior marginal distribution implied for a non-recombining chromosome. In that case, h_i must copy from the same haplotype for all loci in the sequence and the transition probability terms simply reduce to an indicator to enforce that constraint. This gives us the likelihood as

$$\pi(h_i, \{X_i^1, \dots, X_i^L = j\} | \rho = 0, \theta, H_{-i}) = \xi(X_i^1 = j) \prod_{l=1}^L \psi(h_i^l | X_i^l = j).$$

Since the posterior marginal distribution over the hidden state must be constant over the region in the absence of recombination, we can drop the locus superscripts on X .

$$\mathbb{P}(X_i = j | \rho = 0, \theta, H_{-i}, h_i) = \frac{\pi(h_i, X_i = j | \rho = 0, \theta, H_{-i})}{\sum_{j'} \pi(h_i, X_i = j' | \rho = 0, \theta, H_{-i})}$$

To consolidate notation, let us abbreviate the prior copying probability $\xi(X_i^1 = j)$ as ξ_j . Let $\zeta = \frac{1}{2} \frac{\theta}{N-1+\theta}$ and let $\delta_{i,j}^l = \mathbf{1} \{h_i^l \neq h_j^l\}$. Then

$$\begin{aligned} \mathbb{P}(X_i = j | \rho = 0, \theta, H) &= \frac{\xi_j \prod_l \zeta^{\delta_{i,j}^l} (1 - \zeta)^{1 - \delta_{i,j}^l}}{\sum_{j'} \left(\xi_{j'} \prod_l \zeta^{\delta_{i,j'}^l} (1 - \zeta)^{1 - \delta_{i,j'}^l} \right)} \\ &= \frac{\xi_j \left(\frac{\zeta}{1 - \zeta} \right)^{\sum_l \delta_{i,j}^l} (1 - \zeta)^L}{\sum_{j'} \left(\xi_{j'} \left(\frac{\zeta}{1 - \zeta} \right)^{\sum_l \delta_{i,j'}^l} (1 - \zeta)^L \right)} = \frac{\xi_j \left(\frac{\zeta}{1 - \zeta} \right)^{\sum_l \delta_{i,j}^l}}{\sum_{j'} \left(\xi_{j'} \left(\frac{\zeta}{1 - \zeta} \right)^{\sum_l \delta_{i,j'}^l} \right)} \end{aligned}$$

Let $C_\theta = -\log \left[\frac{\zeta}{1 - \zeta} \right]$. Since the prior only affects the model at the beginning on a chromosome, to make this illustration clear, let's also assume that ξ_j is the same for all j

so that the prior probabilities cancel. Then the posterior marginal probability can be simply rewritten as $\frac{\exp\left[-C_\theta \sum_l \delta_{i,j}^l\right]}{\sum_{j'} \exp\left[-C_\theta \sum_l \delta_{i,j'}^l\right]}$. Taking the negative logarithm yields

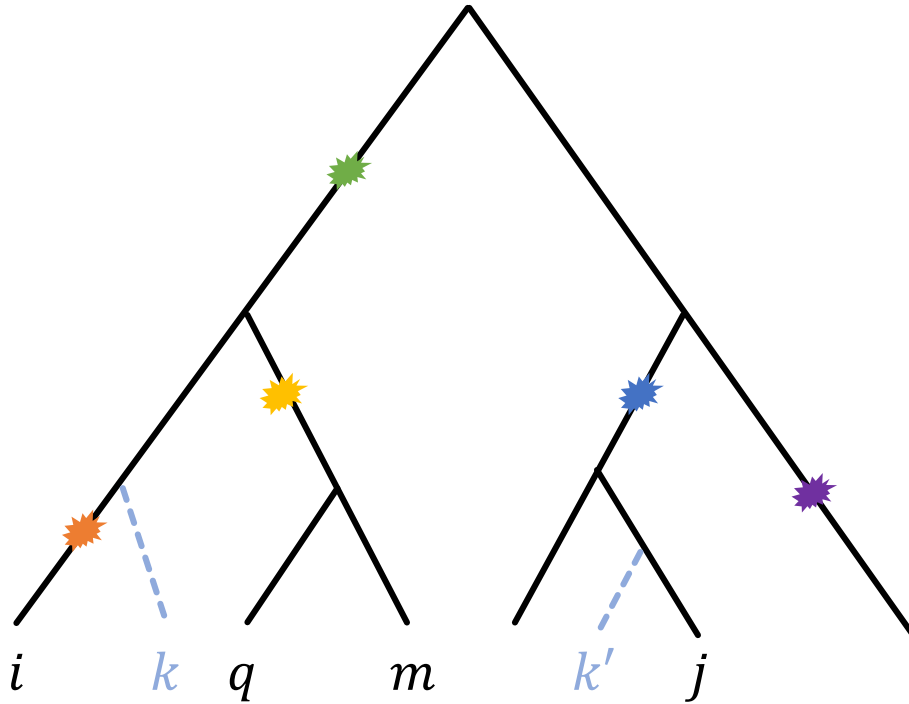
$$-\log [\mathbb{P}(X_i = j | \rho = 0, \theta, H)] = C_\theta \sum_l \delta_{i,j}^l + \log \left[\sum_{j'} \exp \left[-C_\theta \sum_l \delta_{i,j'}^l \right] \right]. \quad (2.1)$$

Note that the function $-\log \left(\sum_i \exp(-z_i) \right)$ is a lower bound on $\min_i z_i$ when all $z_i > 0$. This bound is very close to $\min$ when all of the arguments, the z_i , are sufficiently larger than zero and when there is enough separation between the smallest arguments for the minimum value to dominate. This observation yields the following approximation:

$$d_{i,j} = -\log [\mathbb{P}(X_i = j | \rho = 0, \theta, H)] \approx C_\theta \left(\sum_l \delta_{i,j}^l - \min_{j'} \sum_l \delta_{i,j'}^l \right). \quad (2.2)$$

Thus, without recombination, 2.1.1 is roughly proportional to the number of mismatched sites between the recipient and donor haplotype j in excess of the number of mismatched sites between the recipient and the closest donor. Since $-\log \left(\sum_i \exp(-z_i) \right)$ is a lower bound, it's more precise to say that 2.1.1 is proportional to the number of mismatched sites between the recipient and donor haplotype j minus the number of mismatched sites between the recipient and a theoretical donor that is slightly more similar to the recipient than the nearest donor haplotype. This is a natural consequence of the fact that $-\log(p) > 0 \forall p \in (0, 1)$. See Figure 2.1.

Fig. 2.1 Marginal Tree Perspective of $d_{i,j}$ in Absence of Recombination: Below we show the marginal tree describing the ancestry of 6 observed haplotypes. The colored markers on the tree represent mutations. Consider calculating the distance from recipient haplotype i to donor haplotype j . Three mutations separate those haplotypes on the tree. There is not one but two "nearest neighbours" to i , haplotypes q and m , who are only separated by two mutations. Since $-\log\left(\sum_i \exp(-z_i)\right)$ provides a lower bound on the distance to the nearest haplotype, it would evaluate to a number between zero and two. Thus, $d_{i,j}$ in this case would be a number in between 3 and 1. Calculating $d_{i,j}^l$ can be thought of as taking the local number of mutations that separate i and j and subtracting off the distance to a theoretical nearest neighbour of i , here drawn in as haplotype k .



While $d_{i,j}$ is a non-negative distance, it does not follow various properties of a metric. It will be non-zero even if a donor haplotype and a recipient haplotype are identical. While it's trivial to show that counting mutations between haplotypes on a marginal tree will respect the triangle inequality, since $d_{i,j}$ accounts for the theoretical nearest neighbour, we can only say that it should loosely respect the triangle inequality in moderate to large samples where the "noise" introduced by theoretical nearest neighbours should be relatively small. It might seem tempting to avoid normalizing $\mathbb{P}(X_i = j | \rho = 0, \theta, H)$, which would

remove the $-\log\left(\sum_i \exp(-z_i)\right)$ term from Equation and make leave us with a distance that's proportional to the number of mismatches, d_x , as we will show in the following section, the normalization allows us to obtain a local distance in the case of a chromosome with recombination.

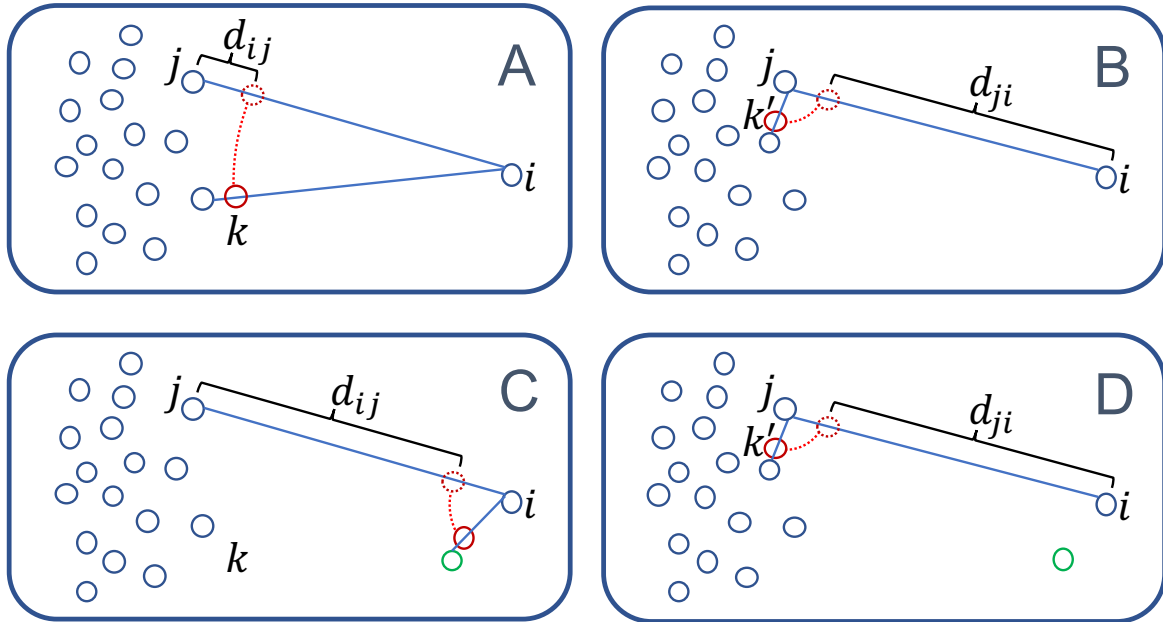
$\tilde{d}_{i,j}$ has a few key robustness properties to the inclusion of close relatives and biased population sampling. Representing a local relatedness structure with $\tilde{d}_{i,j}$ rather than using the original posterior copying probabilities has the advantage that it is much more robust to the addition of near or distant relatives. Whereas the posterior copying probabilities would shrink to numerical zeros for all individuals if an identical twin was introduced to the donor set, such an addition would only shift each of the original $\tilde{d}_{i,j}$ by a constant. As a consequence, $\tilde{d}_{i,j}$ retains the resolution to discriminate between genetically near and distant local haplotypes even in the presence of close relatives.

A second major benefit of working with negative log probabilities arises when working with a merged dataset of individuals from two diverged populations. The original posterior probabilities are sensitive to the number of individuals included in each population whereas $\tilde{d}_{i,j}$ will remain essentially unchanged if the ratio of samples from each population changes.

As shown in Figure 2.1, if we were to calculate, $d_{j,i}$ then we would subtract off the distance to a theoretical nearest neighbour of j , shown as k' . If we have a moderate to large sample size, it is reasonable to assume that the mutation distance separating i from k and j from k' to be roughly the same and small relative to the number of mutations separating i from j . Thus, it is reasonable to assume that $d_{i,j} \approx d_{j,i}$. Although $d_{i,j} \neq d_{j,i}$, symmeterizing the distance by taking the mean, $\tilde{d}_{i,j} = \frac{1}{2} (d_{i,j} + d_{j,i})$, not only gives us symmetry which may be helpful in downstream analysis, but also makes the distances more robust to the influence of the hypothetical nearest-neighbour. As shown in Figure 2.2, if i is a lone outlier in a dataset, then $d_{i,j}$ will under-report the distance between i and j since i 's nearest neighbour is in fact also far from i . However, $d_{j,i}$ will reflect the true haplotype distance between j and

i , so $\tilde{d}_{i,j}$ captures the extended distance. Thus, $\tilde{d}_{i,j}$ is a more reliable measure of haplotype distance.

Fig. 2.2 Symmeterization for Outlier Robustness: Blue circles represent observed haplotypes in an abstract haplotype space. Haplotype i is shown as an outlying haplotype (it has many mutations that separate it from all other haplotypes). In panel A, k is the hypothetical nearest neighbour of i (solid red circle). If we subtract the number of mutations separating i from k from the number of mutations separating i from j , we obtain $d_{i,j}$, which does not reflect the fact that i is an outlier. In panel B, the hypothetical nearest neighbour of j , k' , is much closer to j than k is to i . Thus, d_{ji} captures the fact that i is an outlier. Panels C and D show that $d_{i,j}$ becomes a more reliable measure when a new observation (shown as a green circle) is introduced near i . Thus, as long as no extreme outlier is alone, we can expect d_{ji} to be a meaningful measure of haplotype distance.



In the following section, we will consider the case of a non-zero recombination map, which will mean that posterior copying probabilities will no longer be constant along the entire length of the chromosome. Thus, in the next section we will begin indexing our distances $d_{i,j}$ with an l to denote the locus and write $d_{i,j}^l$, which we will formally define in the next section. It will be easier to provide an interpretation for $d_{i,j}^l$ first before introducing a symmeterized version, but we will return to the idea of using a symmeterized distance in

the more general case with recombination and use that symmeterized version of the distance in our later applications.

2.1.2 With Recombination

It is often natural to think of the negative log likelihood in terms of a cost function. Analogously, here $d_{i,j}^l$ which represents the "cost" of i copying j at a given locus l . As we already described, in the absence of recombination, $d_{i,j}^l$ is approximately a linear function of the number of typed mutations separating the donor from the recipient haplotype. Now we will consider the case with recombination and a recombination map. First, we rewrite the product of the transition densities to see how they contribute to the posterior marginal distribution from a cost or penalty perspective. Here, we use x^l to denote a realization of the hidden state copied at locus l .

$$-\log \left(\prod_{l=1}^{L-1} \phi \left(X_i^{l+1} = x^{l+1} | X_i^l = x^l \right) \right) \quad (2.3)$$

$$= -\log \left(\prod_{l:x^{l+1} \neq x^l} \frac{1}{n} \left(1 - \exp \left(\frac{-\rho_l}{n} \right) \right) \prod_{l:x^{l+1} = x^l} \exp \left(\frac{-\rho_l}{n} \right) \right) \quad (2.4)$$

$$= -\log \left(\prod_{l:x^{l+1} \neq x^l} \frac{1}{n} \left(1 - \exp \left(\frac{-\rho_l}{n} \right) \right) \exp \left(\frac{\rho_l}{n} \right) \prod_l \exp \left(\frac{-\rho_l}{n} \right) \right) \quad (2.5)$$

$$= \sum_l \frac{\rho_l}{n} - \sum_{l:x^{l+1} \neq x^l} \left[\frac{\rho_l}{n} + \log \left(\frac{1}{n} \left(1 - \exp \left(\frac{-\rho_l}{n} \right) \right) \right) \right] = C_M + \sum_{k=1}^{L-1} C_\rho^k \gamma_i(x^k, x^{k+1}) \quad (2.6)$$

where $C_M = \sum_l \frac{\rho_l}{n}$, $C_\rho^l = - \left(\frac{\rho_l}{n} + \log \left(\frac{1}{n} \left(1 - \exp \left(\frac{-\rho_l}{n} \right) \right) \right) \right)$, and

$$\gamma_i(x^k, x^{k+1}) = \begin{cases} 1 & x^k \neq x^{k+1} \\ 0 & x^k = x^{k+1} \end{cases}.$$

Now we can incorporate this contribution from the HMM state transitions into our evaluation of the posterior marginal likelihood. Let X_i^{-l} denote a vector for the entire path set of hidden states underlying haplotype i except for the hidden state at locus l . As before, define $d_{i,j}^l = -\log [\mathbb{P}(h_i, X_i^l = j) / \mathbb{P}(h_i)]$. The numerator gives us the following.

$$-\log (\mathbb{P} (h_i, X_i^l = j)) = -\log \left(\sum_{x^{-l}} \mathbb{P} (h_i, X_i^l = j, X_n^{-l} = x^{-l}) \right) \quad (2.7)$$

$$= -\log \left(\sum_{x^{-l}} \mathbb{P} (h_n | X_i^l = j, X_i^{-l} = x^{-l}) \mathbb{P} (X_i^l = j, X_i^{-l} = x^{-l}) \right) \quad (2.8)$$

$$= -\log \left(\sum_{x: x^l = j} \left[\xi_{x^l} \exp \left(- \left(C_\theta \sum_{k=1}^L \delta_{i, x^k}^k + C_M + \sum_{k=1}^{L-1} C_\rho^k \gamma_i (x^k, x^{k+1}) \right) \right) \right] \right) \quad (2.9)$$

Let us assume that the prior copying probability for each recipient haplotype is uniform:

$\xi_{x^l} = \frac{1}{n-1}$ for all paths x . Then $-\log (\xi_{x^l}) = \log(n-1)$. This yields,

$$= -\log \left(\sum_{x: x^l = j} \left[\exp \left(- \left(\log(n-1) + C_\theta \sum_{k=1}^L \delta_{i, x^k}^k + C_M + \sum_{k=1}^{L-1} C_\rho^k \gamma_i (x^k, x^{k+1}) \right) \right) \right] \right) \quad (2.10)$$

To evaluate $-\log (\mathbb{P} (h_i))$, we simply sum over all possible paths through the state space, not just those that copying through j at l .

$$-\log (\mathbb{P} (h_i)) = \quad (2.11)$$

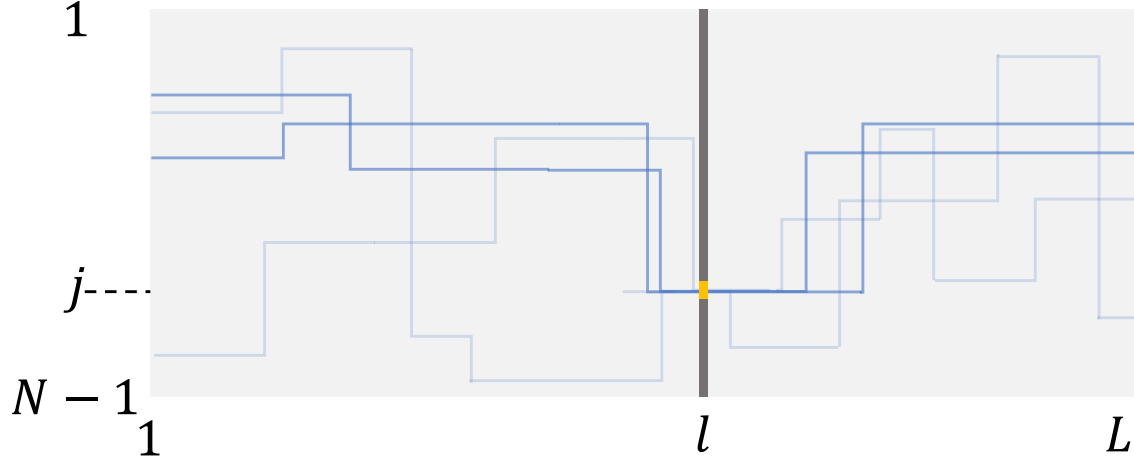
$$-\log \left(\sum_x \left[\exp \left(- \left(\log(n-1) + C_\theta \sum_{k=1}^L \delta_{i, x^k}^k + C_M + \sum_{k=1}^{L-1} C_\rho^k \gamma_i (x^k, x^{k+1}) \right) \right) \right] \right) \quad (2.12)$$

As discussed above, $-\log \left(\sum_i \exp(-z_i) \right)$ can be approximated by $\min$ when all of the arguments are sufficiently larger than zero and when there is enough separation between the smallest arguments for the minimum value to dominate. For even relatively small sample sizes (say $n = 1000$), $\log(n-1) \geq 6$, which means that even just the initial "penalty" coming from the prior ξ allows us to reasonably assume that we satisfy the first condition. Even if we inflate the mutation and recombination penalties by decreasing N_e , we still expect a fair number of paths to tie for the minimum in a large dataset. As discussed above LSE is robust to these repeats. Thus, with a little bit of simplification, we arrive at

$$\begin{aligned}
 d_{i,j}^l &= -\log \left(\mathbb{P} \left(h_i, X_i^l \right) \right) + \log \left(\mathbb{P} \left(h_i \right) \right) \\
 &\approx \min_{x: x^l = j} \left[C_\theta \sum_{k=1}^L \delta_{i,x^k}^k + \sum_{k=1}^{L-1} C_\rho^k \gamma_i \left(x^k, x^{k+1} \right) \right] \\
 &\quad - \min_x \left[C_\theta \sum_{k=1}^L \delta_{i,x^k}^k + \sum_{k=1}^{L-1} C_\rho^k \gamma_i \left(x^k, x^{k+1} \right) \right] \quad (2.13)
 \end{aligned}$$

This shows that $d_{i,j}^l$ can be intuitively interpreted as if we are trying to accurately represent recipient i 's haplotype as a mosaic of the donor haplotypes by minimizing the cost of the copying path. $d_{i,j}^l$ is the additional cost of deviating from the optimal copying path to force i to copy j at l . In other words, we obtain the "penalty" of a Viterbi-like path conditioned on passing through j at l . See Figure 2.3.

Fig. 2.3 Distance as Average over Copying Paths: Blue lines represent different copying paths for i^{th} recipient through the remaining $N - 1$ donors across the L loci. Each of the paths shown have a different penalty associated with them: darker blue lines have lower penalties (fewer transitions and mutations) than lighter blue lines. Only paths that go through donor j and locus l , shown as a "gold gate," are considered. $-\log(\mathbb{P}(h_i, X_i^l))$ considers all possible paths passing through the gold gate and tightly lower bounds the penalty incurred by the best path.



We can simplify $d_{i,j}^l$ slightly by recalling that for HMMs, the distribution over hidden states given a state k steps behind will tend to the unconditional distribution over states as k grows large, so called "Exponential Forgetting." Let $W = [w_l, w_r]$ be a window around l that is large enough such that

$$\mathbb{P}(X_i^{w_l} | X_i^l, h_i) \approx \mathbb{P}(X_i^{w_l} | h_i) \quad (2.14)$$

$$\mathbb{P}(X_i^{w_r} | X_i^l, h_i) \approx \mathbb{P}(X_i^{w_r} | h_i). \quad (2.15)$$

This implies that, outside of W , the optimal path(s) conditioned on going through j at l will essentially be the same as the optimal path(s) not conditioned on going through j at l . Since the conditioned path will only deviate from the unconditioned optimal path over W around locus l , we can cancel penalty terms incurred outside of W to arrive at

$$d_{i,j}^l \approx \min_{x: x^l=j} \left[C_\theta \sum_{k=w_l}^{w_r} \delta_{i,x^k}^k + \sum_{k=w_l}^{w_r-1} C_\rho^k \gamma_i(x^k, x^{k+1}) \right] - \min_x \left[C_\theta \sum_{k=w_l}^{w_r} \delta_{i,x^k}^k + \sum_{k=w_l}^{w_r-1} C_\rho^k \gamma_i(x^k, x^{k+1}) \right]. \quad (2.16)$$

Thus we can interpret $d_{i,j}^l$ as a local haplotype distance. If the genome that we are interested in analyzing with this HMM can be crudely represented by a series of non-recombining chunks separated by recombination hotspots, then it is reasonable to think of $d_{i,j}^l$ in terms of the marginal tree distance described in Figure 2.1. Crucially, $d_{i,j}^l$ does not pre-specify a specific window size for comparing haplotypes around l but considers all possible windowings by integrating over paths that begin copying j around l at all possible breakpoints. However, there is a tradeoff: we specify an effective population size and recombination rate map which controls the lengths of likely copying paths that pass through donor j at l .

Those familiar with Bayesian Non-parametrics might understand this as a standard tradeoff. Consider comparing all haplotypes using some window of loci around l , placing a prior over the window size, and integrating over the prior to obtain a pairwise haplotype distance. This approach corresponds to the classic Bayesian approach of explicitly specifying a distribution over the domain of the parameter space. Our approach does not place a prior over window sizes but as we lower the recombination rate, we force transitions around l to occur at positions where the recombination rate is locally high rather than allowing the transitions to happen just anywhere around l . Conversely, increasing the recombination rate would make the distribution over local recombination breakpoints much more diffuse or uniform. Therefore, just like the novelty parameter in the Chinese Restaurant Process controls the number of tables on which observations are concentrated, the recombination rate in our model governs how concentrated the breakpoints of likely paths will be around local

recombination spikes. This flexible specification of the prior makes it more adaptable to loci with different local recombination landscapes across the genome.

By the same arguments described above in the non-recombining case, we take

$$\tilde{d}_{i,j}^l = \frac{1}{2} (d_{i,j}^l + d_{j,i}^l) \quad (2.17)$$

as our distance measure for the remainder of the thesis.

2.1.3 Interpretation of Cost Functions

To understand how modulating n , N_e , μ , and r can affect $\tilde{d}_{i,j}^l$, let us consider the mutation and recombination penalties implied by typical values for a dataset of $N = 50,000$ haplotypes. Let $N_e = 40,000$ and $\mu = 1.25 \times 10^{-8}$. If we assume that the mutation rate is constant over the genome, then we have

$$C_\theta \approx \log \left(\frac{50000}{40000} \right) + \log \left((2 (1.25 \times 10^{-8}))^{-1} \right) \approx 17.7$$

Since the recombination rate varies dramatically along the genome, let us consider the penalty for a recombination in a hot region with rate $r_l = 2 \times 10^{-7}$

$$C_\rho^{hot} \approx \log \left(\frac{(50000)^2}{40000} \right) + \log \left((4 (2 \times 10^{-7}))^{-1} \right) \approx 25.1$$

and the penalty for a recombination in a cold region with rate $r_l = 2 \times 10^{-9}$

$$C_\rho^{cold} \approx \log \left(\frac{(50000)^2}{40000} \right) + \log \left((4 (2 \times 10^{-9}))^{-1} \right) \approx 29.7$$

As expected, the penalty for recombination events in regions with a high recombination rate will be lower than the penalty for recombination events in regions with a low recombination rate. Furthermore, under these standard parameters, we see that recombination events are

more penalized than mutation events. The fact that the recombination penalty is a function of n^2 while the mutation penalty is a function of n comes from uniform prior copying probability term in the LS Model likelihood. Recall from our derivation of the LS Model in Chapter 1 that the probability of a mutation event before a given lineage coalesces and the probability of a recombination before a given lineage coalesces shrink like $1/n$ since coalescent events happen at a higher rate at the bottom of the tree. However, given our uniform prior copying probability assumption, after a recombination event, the given lineage can coalesce with any other individual with probability on the order of $1/n$, making the recombination penalty a function of n^2 .

By rewriting C_θ and C_ρ^l , we can interpret how our parameters affect $\tilde{d}_{i,j}^l$. Let $\phi = \frac{4N_e}{n}$. For a given μ , C_θ can be written as

$$\begin{aligned} C_\theta &= -\log\left(\frac{\zeta}{1-\zeta}\right) = -\log\left(\frac{\frac{1}{2}\frac{\theta}{n-1+\theta}}{1-\frac{1}{2}\frac{\theta}{n-1+\theta}}\right) = \log\left(\frac{n-1}{2N_e\mu} + 1\right) \\ &\approx \log\left(\frac{2}{\phi\mu} + 1\right). \end{aligned} \quad (2.18)$$

To obtain a similar form for C_ρ^l , we can use the following series expansion at $x = 0$:

$$x + \log(1 - e^{-x}) = \log(x) + \frac{x}{2} + \frac{x^2}{24} - \frac{x^4}{2880} + \mathcal{O}(x^6). \quad (2.19)$$

Since $\frac{\rho_l}{n} = \frac{4N_e r_l^\gamma}{n} = \phi r_l^\gamma$ tends to be small even for modestly-sized datasets, we expand C_ρ^l as follows.

$$\begin{aligned}
C_\rho^l &= - \left(\frac{\rho_l}{n} + \log \left(\frac{1}{n} \left(1 - e^{-\frac{\rho_l}{n}} \right) \right) \right) \\
&= \log(n) - \left(\phi r_l^\gamma + \log \left(\left(1 - e^{-\phi r_l^\gamma} \right) \right) \right) \\
&= \log(n) - \log(\phi r_l^\gamma) - \frac{1}{2} \phi r_l^\gamma - \frac{1}{24} (\phi r_l^\gamma)^2 + \frac{1}{2880} (\phi r_l^\gamma)^4 + \mathcal{O}((\phi r_l^\gamma)^6) \\
&= \log(n) - \log(\phi) + \gamma(-\log(r_l)) + \mathcal{O}(\phi r_l^\gamma)
\end{aligned} \tag{2.20}$$

Note that θ only appears in Equation 2.18 as a coefficient of μ and C_ρ^l is independent of μ . Thus, the mutation rate μ can be set to modulate the mutation cost C_θ without affecting C_ρ^l . In other words, μ dictates a set cost that applies to every haplotype mismatch. We see that ϕ plays a similar role for C_ρ^l . Although ϕ is not a completely free parameter (it must satisfy $\phi < \frac{\log(n+1)}{\max_l [r_l]^\gamma}$ in order for the recombination penalty to be positive), it modulates the baseline penalty that is incurred for every donor haplotype switch. An additional penalty is dictated by the recombination rate where the switch occurs, $-\log(r_l)$, which is multiplied by γ . Thus γ governs how much less a recombination event at a recombination hotspot will be penalized than one in a region with low recombination. Although not considered in the LS model or present thesis, if variable mutation rates were known, one could consider adding an additional parameter to C_θ like γ that would dictate how much the penalty would increase per unit increase in $-\log(\mu)$.

Returning to Equation 2.16, we can predict the behavior of $\tilde{d}_{i,j}^l$ for extreme settings of μ , ϕ , and γ . Very high μ , low C_θ , would make the $\tilde{d}_{i,j}^l$ nearly equal for all donors j , completely washing out any local structure. Very low μ , high C_θ , would essentially override the recombination penalty. It would reduce $\tilde{d}_{i,j}^l$ to an indicator of whether or not i and j shared the same SNP at l since the copying path conditioned on going through donor j at l would immediately deviate and return to the optimal copying path for recipient i through the sequence. A similar phenomenon would also occur for very high ϕ , low C_ρ^l , which

would make all of the loci roughly independent and $\tilde{d}_{i,j}^l$ would again return to an indicator of haplotype matching at l . Very low ϕ , high C_ρ^l would make all of the loci linked and essentially $\tilde{d}_{i,j}^l$ would be nearly the same at every locus, reflecting the global amount of allele sharing between i and j . For moderate μ and ϕ , we obtain a local tradeoff between recombination and mutation penalties that captures local haplotype similarity. Our third parameter γ controls how much the local copying paths, the optimal one for recipient i and the optimal one for recipient i conditioned on passing through donor j at l , must respect the recombination map. High γ forces $\tilde{d}_{i,j}^l$ to reflect haplotype sharing for the full local region up to the nearest recombination rate spikes on the right and left of l .

2.1.4 Tuning Parameters for Local Resolution

Clearly some values of (μ, ϕ, γ) will be better than others for resolving local haplotype structure and performing association testing. The optimal values will vary between datasets based on several factors including the diversity of the sampled population, the recombination map, and the SNP density. The need to specify these parameters in a way makes our method adaptable to different datasets, as long as we can specify them with a reasonable criterion. It might be tempting to use standard EM to optimize these parameters, as is done by Xu & Guan [29]. However, in our case, since we are running N independent HMMs, one for each recipient, that would be optimizing up to $3N$ parameters. Not only would this approach be rather computationally intensive, optimizing the log-likelihood of the HMM along an entire chromosome does not reflect our primary objective: the local genealogy underlying each locus. First we will assume that (μ, ϕ, γ) are the same across all N HMMs. Although we cannot observe the tree at a given locus, the SNP at l captures a branch of the tree by encoding the leaves that must subtend that branch. If we could eliminate the contribution of SNP l from each local distance matrix $\left(\tilde{d}_{i,j}^l\right)^{N \times N}$ and optimize (μ, ϕ, γ) to capture the SNP at l , then our HMM would be using local haplotype information to predict cuts on

the underlying local tree, which should boost our power to detect clusters in the tree that modulate the phenotype. In our HMM framework, we can modify $\left(\tilde{d}_{i,j}^l\right)^{N \times N}$ to reflect local haplotype information without using SNP l by omitting the emission probability of h_i^l and forcing i to copy j . This yields a matrix $\left(\tilde{w}_{i,j}^l\right)^{N \times N}$ defined by

$$w_{i,j}^l = -\log\left(\mathbb{P}\left(h_i \setminus h_i^l, X_i^l = j\right)\right) + \log\left(\mathbb{P}\left(h_i \setminus h_i^l\right)\right), \quad (2.21)$$

$$\tilde{w}_{i,j}^l = \frac{1}{2} \left(w_{i,j}^l + w_{j,i}^l\right). \quad (2.22)$$

Given W , we now need a means of assessing how well it reflects the SNP at l . It is natural that this assessment should use the same test statistic that we will use to find association with phenotypes. If the SNP at l was the causal, fully-penetrant, Mendelian mutation for a given binary phenotype, but we hadn't genotyped it, then the SNP at locus l , encoded by $h^l \in \{0, 1\}^N$, would correspond exactly to the phenotype vector y . Thus, maximizing the power to detect an association via the p-value of $(y - \mu)^T W (y - \mu)$ would be equivalent to minimizing the p-value of $(h^l - \mu^l)^T W (h^l - \mu^l)$ where $\mu^l = \frac{1}{N} \sum h_i^l$. Thus, we can use the p-value of $(h^l - \mu)^T W (h^l - \mu)$ to assess how well HMM under the current parameter settings utilizes local haplotype information to infer the local tree structure, or in other words, impute the SNP at l . This motivates the following objective for optimizing (μ, ϕ, γ) . Let $a^l = (h^l - \mu^l)^T W (h^l - \mu^l)$ be the observed quadratic form at a given locus. For a given set of SNPs $\mathcal{S}$ that are sufficiently far apart in recombination distance as to be deemed independent, we seek to maximize

$$f(\mu, \phi, \gamma) = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} -\log\left(\mathbb{P}\left(\left(Y - \mu^l\right)^T W \left(Y - \mu^l\right) < a^l\right)\right), \quad (2.23)$$

where $Y \stackrel{iid}{\sim} \text{Bern}(\mu^l)$. In practice, one might want to look at the full distribution of the negative log p-values to make sure that a given parametrization does not perform very poorly on some small subset of SNPs. It might also be wise to stratify SNPs into a few frequency

bins and ensure that a given parametrization performs adequately in all of them. However, if one is mostly interested in contributions from rare mutations, one might only optimize $f(\mu, \phi, \gamma)$ based on rare SNPs sampled from the given dataset.

In tuning the recombination parameters ϕ and γ it is clear that there will be some optimal intermediate value. If ϕ is too low, then information from distant SNPs that are not informative about the local phylogeny will contaminate each $\tilde{w}_{i,j}^l$. If ϕ is too high, then not enough information from surrounding SNPs will be used to capture the local phylogeny. By optimizing $f(\mu, \phi, \gamma)$, we are training the model to throw away information from distant loci that are independent of the local phylogeny and to retain information from proximal SNPs that are predictive of the local phylogeny.

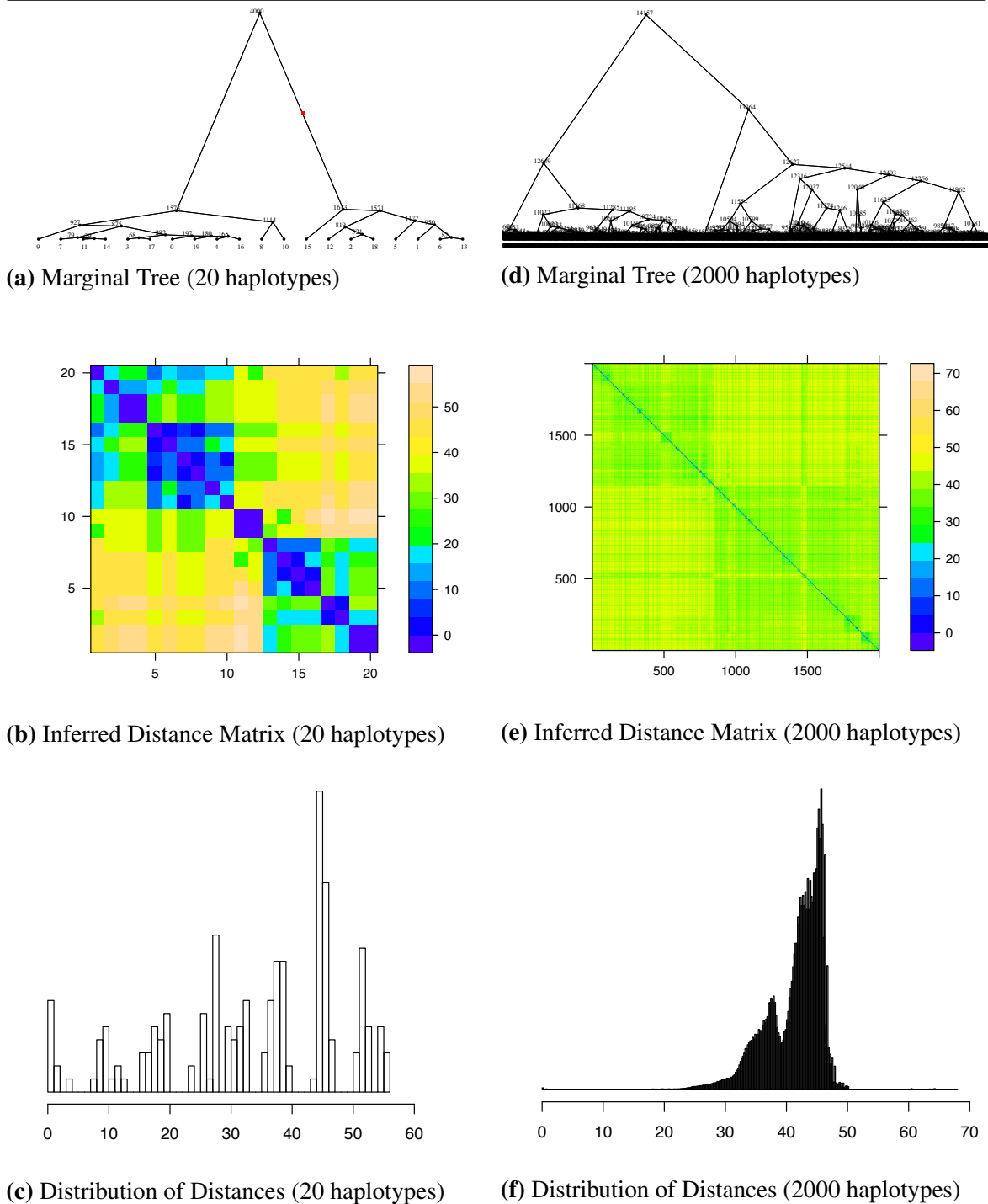
Thus, our approach to tuning the parameters (μ, ϕ, γ) has important implications for how we define a "genomic region." In Chapter 1, we emphasized that all haplotype methods must somehow decide when two nearby association signals should be grouped or considered disjoint. Tuning $f(\mu, \phi, \gamma)$ provides a principled response to this question by considering two association signals distinct if the haplotypes underlying those signals are inherited independently and to merge two signals into a single signal if they tend to ride on a common haplotype. If there is strong dependence between two loci, then it would be futile to try to distinguish the effect of one anyway (without functional experiments). This approach allows us to avoid breaking the genome up into windows.

2.1.5 Simulation Study to Show Local Resolution

In order to briefly demonstrate that the distance matrices we obtain reflect underlying tree structure, we provide a couple of examples below from simulated data. In each simulation, we used msprime [41] to sample a 1 Megabase region from the coalescent with uniform recombination rate 2×10^{-8} per locus per generation, mutation rate 1.25×10^{-8} per locus per generation, and effective population size 30,000. In Figure 2.4, we show the true tree

underlying the locus at the exact middle of the MB region in two simulations, one with 20 haplotypes and another with 2000 haplotypes. Using just the haplotypes from both simulations, we used our implementation of the LS model, to be described in the following section, to calculate our distance matrix $M = \left(\tilde{d}_{i,j}^l\right)^{N \times N}$ at the nearest SNP to the left of the middle of the simulated MB region. For comparison, those distance matrices are shown below their respective trees in Figure 2.4. We did not specifically optimize the HMM parameters used, but set $C_\theta = 2 \times 10^{-8}$, $\phi = 2$, and $\gamma = 1$. We assumed a uniform recombination map with rate 2×10^{-10} per locus. Note that since simulated mutations are different distances apart, this implies a proportionally different recombination distance between mutations. As we can see from Figure 2.4, we have a qualitative correspondance between our distance matrices and the underlying tree structure.

Fig. 2.4 Tree to Distance Matrix Correspondence: The trees from two simulations are shown at the top, where vertical height differences between nodes correspond to the simulated coalescence times. The corresponding matrices below show the inferred local structure where we generally see greater distances between rather than within clades. The histograms at the bottom plot the distribution of distances encoded in each matrix (the diagonal is excluded since it is set to zero by default).



2.2 Computation

The primary motivation for resorting to the LS model for capturing local haplotype structure is its relative computational simplicity. Storage is the primary obstacle to computing and testing local distance matrices based on the posterior marginals. Recall our overall approach: we will run N independent HMMs, which will model each haplotype as a mosaic of the other $N - 1$ haplotypes in the sample. If we were to attempt to save the posterior marginal copying matrices at each locus, we would have a $N \times N - 1 \times L$ array, which would overwhelm most storage systems even for modest sized datasets, not to mention the enormous amount of time that would be consumed by I/O.

We begin this section showing how the posterior marginal copying probabilities, and hence our local distance matrix, are amenable to exact inference via the Forward-Backward (FB) Algorithm [42]. The efficiency of the FB Algorithm is exploited in ChromoPainter to infer a genome-wide copying matrix [39]. After recapitulating these standard results, we proceed with a new checkpointing approach and vectorized implementation of the FB Algorithm for the LS model, which together enable us to efficiently calculate the distance matrices, one locus at a time. This completely avoids the storage obstacle and I/O bottleneck and, as we'll show in the following chapter, allows us to screen each distance matrix for association without storing it. Our implementation makes it quick to recover the distance matrices at loci of interest that show evidence of association. We demonstrate the scalability of our implementation on simulated haplotypes.

2.2.1 Exact Inference Under LS Model

As presented in the previous section, our local distance matrix of interest is a function of the posterior marginal copying probabilities:

$$\tilde{d}_{i,j} = \frac{1}{2} (d_{i,j} + d_{j,i}) = -\frac{1}{2} \left[\log \left(\mathbb{P} \left(X_i^l = j \mid h_i \right) \right) + \log \left(\mathbb{P} \left(X_j^l = i \mid h_j \right) \right) \right]. \quad (2.24)$$

As implied above, in order to calculate $\tilde{d}_{i,j}$ for all i and j to form our distance matrix, we must obtain the $N - 1$ marginal copying probabilities at locus l for each of the N HMMs, each corresponding to taking one of the haplotypes as the emitted recipient haplotype. A generic approach to this calculation would involve performing a single forward pass and a single backward pass for each of the N recipient haplotypes, storing the array of forward and backward copying probabilities, and combining them to yield the $N \times N - 1 \times L$ array of posterior marginals. Algorithm 1 presents the Forward and Backward Algorithms if we were to write the LS Model as a genetic HMM. Note that Algorithm 1 has $\mathcal{O}(N^2L)$ complexity.

Since in the LS Model, the transition probability matrix at any given locus only has two distinct entries, one "non-recombination" probability repeated down the diagonal, and a "recombination probability" in all of the off-diagonal elements, the generic Forward-Backward algorithm can be rewritten in a relatively simple form for the LS Model (see Algorithm 2).

Algorithm 1 Generic Forward-Backward Algorithm for Marginal Posterior Decoding

initialization

$\mathbf{P} \in \mathbb{R}_{N-1 \times L}$	▷ Output Array for Posterior Marginals
$\alpha \in \mathbb{R}_{N-1 \times L}$	▷ Storage Array for Forward Probabilities
$\theta \in \mathbb{R}_{N-1 \times L}$	▷ Emission Probabilities Matrix
$T \in \mathbb{R}_{N-1 \times N-1}$	▷ Transition Probabilities Matrix
$\vec{\pi} \in \mathbb{R}_{N-1}$	▷ Prior Copying Probabilities Vector
$\beta^{new}, \beta^{old} \in \mathbb{R}_{N-1}$	▷ Storage Vectors for Backward Probabilities
$g_L \in \mathbb{R}$	▷ Total Probability of the Observed Sequences

procedure FORWARD PASS($\alpha, T, \theta, \vec{\pi}$)

```

 $\alpha_{[:,1]} \leftarrow \vec{\pi} \cdot \theta_{[:,1]}$ 
 $l \leftarrow 1$ 
while  $l < L$  do
     $\alpha_{[:,l+1]} \leftarrow \theta_{[:,l]} \cdot (T^T \alpha_{[:,l]})$ 
     $l \leftarrow l + 1$ 
end while
 $g_L \leftarrow \sum_i \alpha_{[i,L]}$ 

```

end procedure**procedure** BACKWARD DECODING($\beta^{old}, \beta^{new}, T, \theta, \alpha$)

```

 $\beta^{old} \leftarrow \vec{1}$ 
 $l \leftarrow L - 1$ 
while  $l > 0$  do
     $\beta^{new} \leftarrow \mathbf{T}(\theta_{[:,l+1]} \cdot \beta^{old})$ 
     $\mathbf{P}_{[:,l]} \leftarrow (\alpha_{[:,l]} \cdot \beta^{new}) / g_L$ 
     $l \leftarrow l - 1$ 
end while

```

end procedure**return** $\mathbf{P}$

Algorithm 2 LS Model Vectorized Forward-Backward Algorithm for Posterior Decoding**initialization**

$\mathbf{P} \in \mathbb{R}_{N \times L}$	▷ Output Array for Posterior Marginals
$\alpha \in \mathbb{R}_{N \times L}$	▷ Storage Array for Forward Probabilities
$\theta \in \mathbb{R}_{N \times L}$	▷ Emission Probabilities Matrix
$\vec{\rho} \in \mathbb{R}_L$	▷ Expected Number of Recombination Events/Meiosis at each locus
$\vec{\pi} \in \mathbb{R}_N$	▷ Prior Copying Probabilities Vector
$\beta^{new}, \beta^{old} \in \mathbb{R}_N$	▷ Storage Vectors for Backward Probabilities
$g_L \in \mathbb{R}$	▷ Total Probability of the Observed Sequence

procedure FORWARD PASS($\alpha, \theta, \vec{\pi}, \vec{\rho}, g_L$)

$$\alpha_{[:,1]} \leftarrow \vec{\pi} \cdot \theta_{[:,1]}$$

$$l \leftarrow 1$$

while $l < L$ **do**

$$\alpha_{[:,l+1]} \leftarrow \theta_{[:,l]} \cdot \left((1 - \rho_l) \alpha_{[:,l]} + \rho_l \left(\sum_i \alpha_{[i,l]} \right) \vec{\pi} \right)$$

$$l \leftarrow l + 1$$

end while

$$g_L \leftarrow \sum_i \alpha_{[i,L]}$$

end procedure**procedure** BACKWARD DECODING($\beta^{old}, \beta^{new}, \alpha, \theta, \vec{\pi}, \vec{\rho}, g_L$)

$$\beta^{old} \leftarrow \vec{1}$$

$$l \leftarrow L - 1$$

while $l > 0$ **do**

$$\beta^{new} \leftarrow (1 - \rho_l) \theta_{[:,l+1]} \cdot \beta^{old} + \rho_l \left(\sum_i \theta_{[i,l+1]} \beta_i^{old} \right) \vec{\pi}$$

$$\mathbf{P}_{[:,l]} \leftarrow (\alpha_{[:,l]} \cdot \beta^{new}) / g_L$$

$$\beta^{new} \leftarrow \beta^{old}$$

$$l \leftarrow l - 1$$

end while**end procedure****return** $\mathbf{P}$

2.2.2 Checkpointing

As alluded to previously, the primary difficulty that faces our approach to local graph construction is storing the output. A naive implementation of posterior decoding would involve running the forward algorithm and storing the forward probabilities at each position, yielding a $N \times N \times L$ tensor, then running the backward algorithm to generate and test the posterior probability matrices locus by locus. Consider the UK Biobank dataset: 500,000 individuals typed at 820,967 markers. That's equivalent to 1 million haplotypes. If all of the markers were used and even if the forward probability tensor was stored in single precision, that would mean storing a tensor of over 3 million TB.

The standard response to this problem in the HMM literature is to use checkpointing [43]. These approaches store the forward probabilities at checkpoints along the genome (perhaps at every $10,000^{th}$ position), then, during backward decoding, recompute the forward probabilities at each position between the stored windows (fill in the gaps). Checkpointing allows us to obtain the posterior decoding at each position, test the implied distance matrix, and then remove it from memory.

There are a wide variety of checkpointing strategies; typically, the lower the memory usage, the more redundant computation required [43]. Let us measure the amount of recomputation required by a checkpointing strategy as of the total distance D in terms of SNPs that the Forward and Backward Algorithms would need to cover in order to calculate the graph at every position for a chromosome with L SNPs. If we didn't use checkpointing, which would have no computational redundancy but use an enormous amount of memory, we'd have to cover a distance of $D = 2L$. This is a lower bound. Assume that we have the memory to store K distinct $N \times N - 1$ matrices. We may first consider uniformly spacing the checkpoints along the sequence of L SNPs. That would require one $N \times N - 1$ matrix of forward probabilities and one $N \times N - 1$ matrix of backward probabilities per checkpoint, leaving space for $K/2$ checkpoints. We would calculate these by first running a single pass of

the forward algorithm and a single pass of the backward algorithm, then evaluating the loci in each interval by propagating from the adjacent checkpoints. This uniform checkpointing scheme would yield a total computational cost of

$$D_{\text{unif}} = 2L + (K/2) \frac{\frac{L}{K/2} \left(\frac{L}{K/2} + 1 \right)}{2} = 2.5L + \frac{L^2}{K}.$$

Since D_{unif} scales with $\mathcal{O}(L^2)$, it would quickly become infeasible for GWAS applications.

Toward Optimal Checkpointing

Let us now consider a checkpointing strategy that would seek to minimize D as a function of the number of checkpoints we can store, K . Although we have yet to formally prove that the following strategy is optimal and we only use a heuristic version of it in practice, it is helpful to present an idealized checkpointing strategy in order to motivate a heuristics with comparable efficiency.

As with other checkpointing approaches, we will only use checkpoints to store forward probability tables which we will reshuffle and "recycle" as they are passed during backward decoding. Since the backward decoding always requires a distance of L SNPs, let us write $D = L + f_K^*(L)$ where $f_K^*(L)$ denotes the total number of SNPs that must be covered by the forward algorithm to optimally set up and use K checkpoints. Note for the base case with no checkpoints $K = 0$, we have $f_0^*(L) = \frac{L(L+1)}{2}$ since the forward table must be propagated forward from start of the haplotypes to every other locus along the sequence. Let us now consider the case where we have a single checkpoint. It will be convenient for our derivation to make a continuous approximation to our problem so that we can place a checkpoint at any real number between 1 and L . Let $L\alpha$ be the position where we should place the first checkpoint for $\alpha \in (0, 1)$. After we have placed the first checkpoint, we will use that checkpoint to calculate the forward tables for all positions further than that checkpoint, costing $f_0^*(L(1 - \alpha))$. Then, after all of the positions to the right of our first checkpoint have

been evaluated, we are left with a sequence of length $L\alpha$ and we have a free checkpoint. We will again apply the same checkpoint strategy to this subsection of the sequence, placing the checkpoint now at $L\alpha^2$. Iterating in this way, we can express the total cost of forward propagation using one checkpoint as a function of α :

$$f_1(\alpha, L) = \sum_{i=0}^{\infty} [\alpha^{i+1}L + f_0^*(L(1-\alpha)\alpha^i)]. \quad (2.25)$$

If we can solve this equation for $\alpha_1^* = \underset{\alpha}{\operatorname{argmin}} f_1(\alpha, L)$, then, at least in this continuous approximation, we could efficiently schedule the placement of the checkpoint. Extending this approach to the case where $K > 1$ is straightforward. If we had two checkpoints, we would first have to place a checkpoint at some position $L\alpha$ along the sequence and then use our optimized 1-checkpoint approach on the rest of the sequence. Once the backward decoding has reached $L\alpha$, then again we are left with a sequence of length $L\alpha$ and two available checkpoints. Extending this recursive thinking to the case of K checkpoints we have

$$f_K(\alpha, L) = \sum_{i=0}^{\infty} \alpha^{i+1}L + f_{K-1}^*(L(1-\alpha)\alpha^i), \quad (2.26)$$

where as before, $f_0^*(L) = \frac{L(L+1)}{2}$. This recurrence relation implies a set of proportions $\{\alpha_1^*, \dots, \alpha_K^*\}$ such that $\alpha_K^* = \underset{\alpha}{\operatorname{argmin}} f_K(\alpha, L)$. Unfortunately, there is no explicit formula for α_K^* beyond α_1^* ; however, they can be obtained with numerical optimization.

In order to appreciate the massive impact of checkpointing on the overall computational burden, let us explicitly solve for α_1^* and $f_1^*(L)$. Substituting the definition of f_0^* into Equation 2.25 and evaluating the sum yields

$$f_1(\alpha, L) = L \left(\frac{1}{(1-\alpha)^2} - \frac{L}{(1+\alpha)^2} \right) \quad (2.27)$$

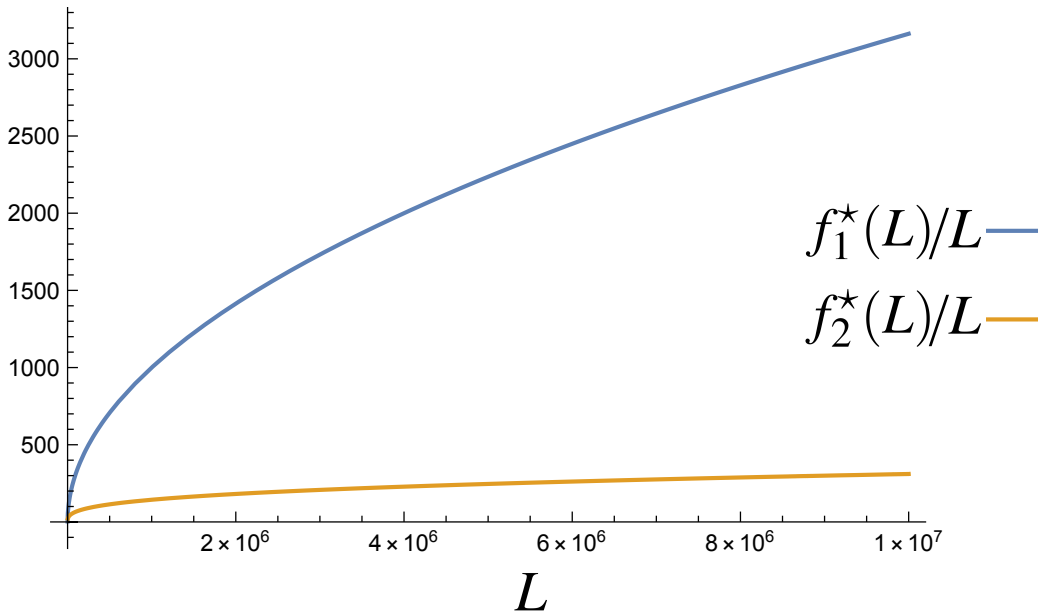
Maximizing via the derivative and taking the solution in the interval $(0, 1)$ yields $\alpha_1^* = \frac{\sqrt{L}-1}{\sqrt{L}+1}$. Note that this is a function of L . For arbitrary K , let $f_K^*(L) = \min_{\alpha} f_K(\alpha, L)$, the minimized

cost of using K checkpoints to decode a sequence of length L . For $K = 1$ we can substitute in α_1^* and arrive at

$$f_1^*(L) = L^{3/2}. \quad (2.28)$$

Of course there will be some slightly added cost when mapping this 1-checkpoint procedure back to a discrete sequence of loci since we would have to move the checkpoints to integers rather than any real number. Disregarding this negligible cost, with only a single checkpoint, we've been able to reduce the computational burden from $\mathcal{O}(L^2)$ to $\mathcal{O}(L^{3/2})$. In Figure 2.5 we plot $f_1^*(L)$ and $f_2^*(L)$, which we have solved numerically. Note the marked drop in the computational burden possible with even two checkpoints, which achieves nearly $\mathcal{O}(L)$ efficiency for large L .

Fig. 2.5 Optimized Checkpointing Performance: The total distance required of the forward algorithm for locus by locus decoding using one or two checkpoints is plotted against the number of loci L . The total distance is plotted in units of L loci (using two checkpoints on a chromosome with 10 million loci would require the forward algorithm to cover approximately $300L$ loci).



Geometric Checkpointing

Unfortunately the checkpointing scheme presented in the previous section requires nontrivial optimization and implementation when using more than two checkpoints. However, it suggests that an efficient checkpointing approach will recycle checkpoints and initialize checkpoints by placing one very close to the current locus of the backward table and space subsequent checkpoints further and further apart towards the first locus. Geometrically spacing checkpoints provides an intuitive and easily programmable approach to achieving a pattern. More specifically, consider initializing checkpoints K by placing them at $L - \frac{L}{2}, L - \frac{L}{4}, L - \frac{L}{8}, \dots, L - \frac{L}{2^K}$. Each of our checkpoints only store a $N \times N - 1$ matrix of forward probabilities, which we will refer to as a forward table. Since 2^K grows so quickly, this will place the right-most checkpoint very near the end of the sequence.

After precomputing these checkpoints, we perform a single sweep along the sequence using the Backward Algorithm, shifting the backward probability matrix (the backward table) one SNP at a time. During this backward march, as the backward table increments to a given SNP, we compute the forward table at that SNP by propagating the forward table stored at the nearest checkpoint forward, just as described in the "optimal" checkpointing approach from the previous section. Every time we pass a checkpoint on the backward march, it is "recycled" to re-establish a geometric progression of checkpoints within the interval currently being evaluated between the rightmost checkpoint and the position of the backward march. This approach essentially allows us to act as if we evenly divided the sequence at 2^K uniform checkpoints while only using K . This comes at the added cost of re-establishing the geometric series of checkpoints as required. As we show below, this tradeoff is worthwhile. We can upper bound the distance in loci required to be covered by the forward and backward algorithms in this approach as follows

$$\begin{aligned}
D_{\text{geom}} &= 2L + \frac{Lx(x+1)}{x} + x \sum_{i=1}^{K-1} (2^{K-i} - 1) 2^{i-1} \text{ where } x = \frac{L}{2^K} \\
&= 2L + \frac{1}{2}L(x+1) + x \left[\sum_{i=1}^{K-1} 2^{K-i} - \sum_{i=1}^{K-1} 2^{i-1} \right] \\
&= 2L + \frac{1}{2}L(x+1) + x [(K-1)2^{K-1} - (2^{K-1} - 1)] \\
&= 2L + \frac{1}{2}L(x+1) + x [(K-2)2^{K-1} + 1] \\
&= 2L + \frac{L^2}{2^{K+1}} + \frac{L}{2} + \frac{L(K-2)2^{K-1}}{2^K} + \frac{L}{2^K} \\
&= L \left(\frac{K-2}{2} + \frac{5}{2} + \frac{1}{2^K} \right) + \frac{L^2}{2^{K+1}} \tag{2.29}
\end{aligned}$$

Instead of scaling with $\mathcal{O}(L^2)$, this approach is $\mathcal{O}(L)$ as long as the number of stored checkpoints, K , is sufficiently large. Since 2^K grows so rapidly, somewhere between $K = 20$ and $K = 40$ checkpoints will work in the vast majority of applications. For a dataset of 20,000 individuals ($N = 40,000$), only 384Gb of memory would be required to store 30 checkpoints. We will use this geometric checkpointing approach in our pilot GWAS in Chapter 4. For our pilot, we found that setting $K = 14$ or $K = 15$ provided a convenient compromise between memory usage and computational efficiency.

2.3 Implementation

In the following section we outline features of our implementation of the LS model forward algorithm and LS model backward algorithm. The work in this section was done in close collaboration with Dr. Louis Aslett. Over many meetings, we worked together to structure the problem and eventually converge to a solution. He is responsible for the conception of the performance features discussed below, their assembly-level implementation, and the interfacing R package.

2.3.1 Rescaling

The previous sections have focused on making our local distance matrix construction tractable in terms of memory usage. Of course, this still leaves the problem of computational efficiency, which is especially critical if we need to perform even a modest amount of recomputation for checkpointing. A direct implementation of the LS model forward and backward algorithms detailed in Algorithm 2 without concern for numerical stability typically hit the boundary of machine precision after propagating for maybe 100 or 1000 loci because of the summation required to compute $\sum_i \alpha_{[i,l]}$. Since in most realistic datasets, some donor haplotypes are far better matches to the recipient haplotype than other donors, $\alpha_{[1,l]}$ can quickly become orders of magnitude apart from $\alpha_{[2,l]}$. Furthermore, as forward probabilities, these $\alpha_{[i,l]}$ shrink roughly exponentially in l , so we quickly underflow. To overcome this numerical barrier, it is common for HMMs to be implemented in log space [44]. However, this requires log and exp functions to be used at every step of the inner loop of both the forward and backward steps, costing an enormous number of clock cycles. The Forward-Backward Algorithm found in the ChromoPainter module of the software package fineSTRUCTURE used this technique as of April 2016.

In order to avoid this cost, we make use of the following rewriting of the inner loop update. Let $F_l = \sum_i \alpha_{[i,l]}$.

$$\begin{aligned}
 \alpha_{[\cdot, l+1]} &\leftarrow \theta_{[\cdot, l]} \cdot ((1 - \rho_l) \alpha_{[\cdot, l]} + \rho_l F_l \vec{\pi}) \\
 \frac{\alpha_{[\cdot, l+1]}}{\rho_l F_l} &\leftarrow \theta_{[\cdot, l]} \cdot \left((1 - \rho_l) \frac{1}{\rho_l F_l} \alpha_{[\cdot, l]} + \vec{\pi} \right) \\
 \frac{\alpha_{[\cdot, l+1]}}{\rho_l F_l} &\leftarrow \theta_{[\cdot, l]} \cdot \left((1 - \rho_l) \frac{\rho_{l-1} F_{l-1}}{\rho_l F_l} \frac{\alpha_{[\cdot, l]}}{\rho_{l-1} F_{l-1}} + \vec{\pi} \right) \\
 a_{[\cdot, l+1]} &\leftarrow \theta_{[\cdot, l]} \cdot (\eta_l a_{[\cdot, l]} + \vec{\pi})
 \end{aligned} \tag{2.30}$$

Here, $a_{[:,l]} = \frac{\alpha_{[:,l]}}{\rho_{l-1}F_{l-1}}$ and $\eta_l = (1 - \rho_l) \frac{\rho_{l-1}F_{l-1}}{\rho_l F_l}$. Updating η_l at the end of each iteration is simply given by an accumulation of a :

$$\eta_{l+1} = \left[\frac{\rho_{l+1}}{1 - \rho_{l+1}} \sum_i a_{[i,l+1]} \right]^{-1}. \quad (2.31)$$

Computing in this rescaled space allows our forward algorithm to maintain numerical stability along the genome. We use the analogous rescaling for the backward algorithm.

2.3.2 Parallelization & Memory Optimization

Recall that each row of this $N \times N - 1$ table corresponds to a given recipient haplotype and each of the $N - 1$ columns corresponds to a given donor haplotype. In order to propagate a $N \times N - 1$ table of forward probabilities to a given target locus, we use each available core to advance 16 rows at a time from the starting locus all the way to the target locus. This means that each core evaluates its own set of 16 independent HMMs, each with a distinct recipient haplotype, at a time. If a dataset of interest is too large to fit into memory on a given machine, chunking the table updates by rows makes current implementation readily scalable to parallelization across machines.

For a given core with its 16 rows, the algorithm updates the entries of the forward table by sweeping through all of the columns of the 16 haplotypes in parallel (sweeping through all of the donors), and then returning to the first column to propagate forward another locus. We apply the same approach to propagating the backward table. In order to prevent striding across cache lines during this propagation, our $N \times L$ haplotypes are stored locus by locus rather than haplotype by haplotype.

In order to further reduce memory bandwidth limitations, we used bit manipulation intrinsics in order to make it possible to send subsets of haplotypes from memory to the CPU as bits rather than bytes. Although this costs some additional computation on the CPU, the added speed from relieving the memory bottleneck far outweighs the computational

costs. For comparison, as of April 2016, ChromoPainter uses integers to store haplotypes rather than bits so we obtain a 32 times compression of the haplotypes which means that they consume less of the cache and we can import them faster from memory. Even for a sample of 40,000 haplotypes, stored as bits, each locus only consumes 1,250 bytes per locus. At 64 bytes per cache line, this means that only 20 memory fetches are required to load in a locus from memory.

2.3.3 16 Row Blocks & AVX Instructions

We propagate specifically 16 rows at a time in order to make optimal use of some of the vector unit capabilities introduced with Intel's Haswell generation processors. In order to explain this feature of the implementation, note that the computation required to propagate a single entry of the forward table as shown Equation (2.30) can be captured in only three essential operations. We must first do a fused multiply-add (FMA) in order to calculate $\theta_{[,l]}$ and then another FMA to calculate $(\eta_l a_{[,l]} + \tilde{\pi})$. From there we simply need to multiply these two results together. In order to then accumulate these products into η_{l+1} (see Equation (2.31)), we need to perform one addition per entry in each row. This yields a total of four essential operations that need to be iterated to sweep along the columns and advance to the next locus.

Haswell chips have eight execution ports [45]. Two of those ports, port 0 and port 1, can perform four floating point multiplications in parallel or four FMAs in parallel. Port 1 has the added capability of doing four floating point additions in parallel. In order to make use of these parallel operations, it makes sense to advance 4 rows at a time so that we can advance four rows with one instruction. However, as described in the previous paragraph, the four steps to advancing a single entry of the forward table involve two FMAs, which can be done in parallel, but then a multiplication and an addition must wait for the prior steps to complete. This may leave either port 0 or port 1 dormant in a given clock cycle.

To try to reduce this idle time, the Haswell chip has a reorder buffer, where instructions wait for execution before being sent to ports. The reorder buffer attempts to reshuffle instructions so that independent calculations have the opportunity to be executed on separate ports in parallel. This parallel execution is made possible by a Haswell feature referred to as its super scalar architecture: the processor can dispatch two instructions to ports rather than just one per clock cycle. Although the timing doesn't always make this possible, this means that instead of dispatching a four-unit FMA instruction to a port in a given clock cycle, if the processor knows that two sets of FMAs are independent, it will sometimes take advantage of the ability to dispatch two four-unit FMAs in a given clock cycle.

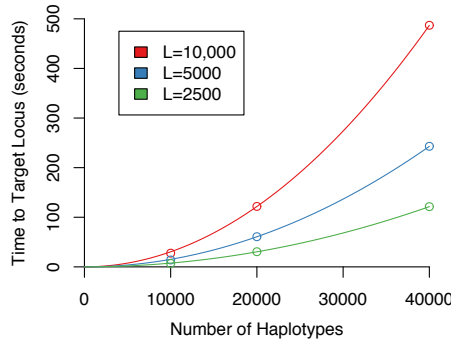
By streaming four sets of four HMMs (16 rows) to each core at a time, we increase the probability that the reorder buffer will recognize instances where one instruction for one set of four HMMs and another instruction for another set of four HMMs are independent and dispatch both instructions in one clock cycle. While this does not quite double our computational throughput, it still has a non-negligible impact on our performance.

2.3.4 Benchmarking

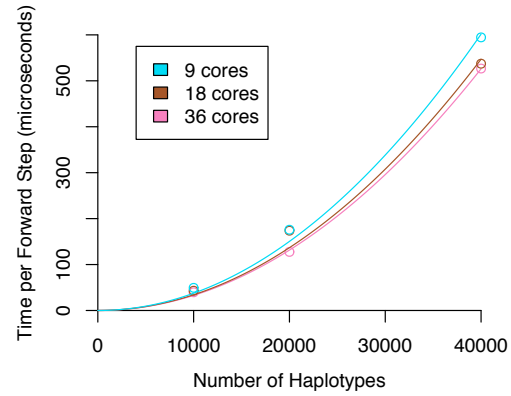
We benchmarked our implementation on the c4.8xlarge instances available on Amazon's EC2 service. Designed for computationally focused tasks, the c4.8xlarge features 36 virtual high frequency Intel Xeon E5-2666 v3 (Haswell) processors and 60 GiB of memory. Performing this benchmarking on Amazon Web Services allows other investigators to replicate our results with relative ease. As is shown in Figure 2.6, we obtain the theoretical scalability of the Forward algorithm when propagating the forward table long distances along the sequence. Note the scale on the y-axes. Projecting based on Figure 2.6b, using this architecture, our implementation can propagate a $40,000 \times 40,000 - 1$ table of forward probabilities across a chromosome of 100,000 SNPs in 83 minutes.

Using our convenient R wrapper, the code used to perform the simulations shown in Figure 2.6 is only about 100 lines long.

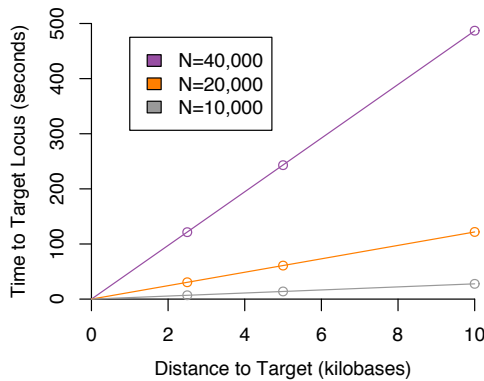
Fig. 2.6 Benchmarking: Panel 2.6a shows the time it takes the forward algorithm to propagate a full $N \times N - 1$ matrix of forward probabilities to across a segment of length L moving continuously from locus to locus using all 36 cores. We achieve the theoretical limit for the forward algorithm, which scales $\mathcal{O}(N^2)$. Rearranging the same data to show the times as a function of L , we see in panel 2.6b, that we also obtain the theoretical limit of $\mathcal{O}(L)$ scalability. Panel 2.6c, which presents timings for moving a forward table just a single locus, shows roughly the same $\mathcal{O}(N^2)$ behavior as panel 2.6a. Panel 2.6d shows the same incremental timings as 2.6c, just presented as a function of the number of cores used. Panel 2.6c and 2.6d suggest that for a problem of this size, the overhead of using 36 cores to propagate the forward table a single locus makes using just 9 cores almost as fast. This suggests that using an architecture with fewer cores but more memory per machine may be more optimal for incremental propagation of the forward table.



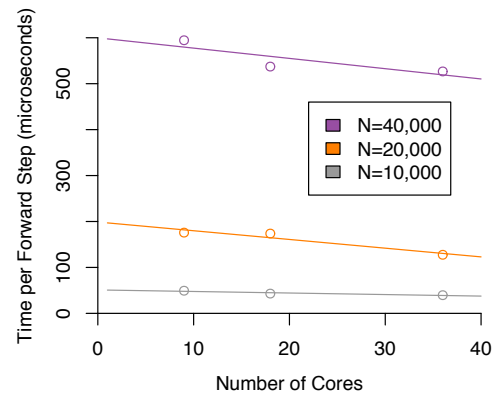
(a) Continuous Speed as $f(N)$



(c) Incremental Speed as $f(N)$



(b) Continuous Speed as $f(L)$



(d) Incremental Speed as $f(\# \text{ cores})$

2.4 Discussion

The primary bottleneck in computational speed faced by our current implementation is based on the size of the L3 cache. While the haplotypes and other components take up relatively little space (assuming uniform prior copying probabilities), storing the forward table itself is what takes up the vast majority of the cache during our computation since it is a $N \times N - 1$ array of doubles. If we consider the c4.8xlarge architecture that we used for our benchmarking, the L3 cache is 25 Mb large. Dividing that space by 8 bytes per double, then again by 36 cores, and again by 16 rows per core, we can estimate that each row of the forward table cannot be longer than 5,425 haplotypes before the processor has to begin fetching doubles from memory for forward table propagation. The smallest example shown in our benchmarking study featured 10,000 haplotypes so our implementation certainly had to resort to calls to memory. However, if we could use an architecture with a larger L3 cache to processor ratio, we could see a dramatic increase in the speed of our implementation. We hope to explore this possibility as we look to scale our approach to hundreds of thousands of genomes.

The negative logarithm of the marginal posterior probabilities implied by the LS Model summarizes the local (local to a region of the genome) relatedness between haplotypes. Using this distance measure, we obtain a local relatedness matrix that recognizes commonly inherited haplotype blocks and the non-exchangeability of loci. Furthermore, our approach is relatively robust to SNP density, much more so than standard window-based methods. To see this, simply consider adding several monomorphic sites in a region of the genome in order to dilute the local SNP density. Since our distances can be reinterpreted as the added penalty of copying through a given individual at locus l rather than taking the optimal path, the contributions of these added monomorphic sites cancel, leaving our distances unchanged and therefore robust to variations in SNP density. Our model ignores the absolute length of sequence between loci, only considering the recombination distance.

Alongside the benefits of using the LS model as the basis for our inference of local genetic distances, there are limitations that have been overcome with more modern methods such as SMC++ [46]. For example, we require phased haplotypes by using the LS model. SMC++, a recent demographic inference approach based on more recent work on the conditional site frequency spectrum directly models genotypes. However, phasing accuracy should continue to improve with the advent of larger reference panels and long read sequencing. Furthermore, as we'll see in Chapter 4, using the convolving haplotypes into genotypes can eliminate information relevant to association testing.

Instead of comparing each haplotype to all other $N - 1$ haplotypes as donors, SMC++ models each pair of genotypes separately, bringing in information from the remaining $N - 2$ haplotypes by incorporating the derived allele frequency at each locus. While this approach might be very attractive since it makes calculating each element of a local distance matrix completely independent, we expect that the added computational cost of using such a relatively complex model would render this approach intractable for tens of thousands of samples, especially since checkpointing would be required.

Computational concerns aside, while carefully modeling allele frequencies may improve ancient demographic inference, it's not clear that it will improve local haplotype resolution useful for association testing (see Chapter 1). Furthermore, SMC++ heavily relies on the assumptions of the coalescent when incorporating the derived allele frequencies. With our implementation of the LS model, we use to compare each haplotype to all of the others in the sample (an "one vs. all" implementation). In order to compare this "one vs. all" approach to pairwise models like SMC++, we can rewrite the LS model as a pairwise model (see Chapter 2 Appendix). This shows that by comparing all sampled haplotypes simultaneously, the Li & Stephens model implicitly adjusts the mutation penalty depending on how common each local haplotype sequence is in the population. In other words, if two haplotypes have two distinct, relatively rare local sequences, the distance calculated between them will be much

larger than if their two distinct sequences were relatively common. As a consequence, our distances reflect the allele frequency at each position: haplotypes that share a rare SNP will have a significantly smaller distance calculated between them than haplotypes that share a common SNP. By simply weighting mutation penalties by "local matchability," our model should be relatively robust to violations of the coalescent.

Most importantly, we have obtained a scalable approach based on the LS Model that yields meaningful representations of local haplotype diversity. In Chapter 4, we will endeavor to show that this representation is useful for association studies.

2.5 Appendix

Algorithm 3, rewrites the forward and backward algorithms of the LS model assuming that we know the normalizing constant for the forward probabilities $F_{[i,l]}$ at each locus l for each recipient haplotype i and the normalizing constant for the backward probabilities $G_{[i,l]}$ at each locus l for each recipient haplotype i . In practice these would have to be precomputed from a prior run of the forward and backward algorithms. However, conditional on knowing $F_{[i,l]}$ and $G_{[i,l]}$ for all l and i , the forward and backward probabilities under the LS model can be calculated for each pair of haplotypes without any additional information about the other haplotypes in the sample, making it a "pairwise algorithm."

Working in this conditional space is mainly interesting because we may be able to approximate $\left(\frac{F_{[i,l]}}{F_{[i,l+1]}}\right)$ and $\left(\frac{G_{[i,l]}}{G_{[i,l-1]}}\right)$ at each locus to achieve another order of parallelism. If $\left(\frac{F_{[i,l]}}{F_{[i,l+1]}}\right)$ and $\left(\frac{G_{[i,l]}}{G_{[i,l-1]}}\right)$ were known $\forall i$, there would be no need to run the forward-backward algorithm at any point in our local graph construction. We could simply run the Conditional Propagation algorithms on each pair of haplotypes independently.

To give some intuition behind the interpretation of $\left(\frac{F_{[i,l]}}{F_{[i,l+1]}}\right)$, note that it is the probability of the observed emissions up to position l divided by the probability of the observed emissions up to position $l + 1$.

Algorithm 3 Conditional Forward & Backward Propagation Functions

```

function CFP( $i, t, F_{[i,\cdot]} | H, \Pi, \vec{\mu}, \vec{\rho}$ )
   $l \leftarrow 1$ 
   $\theta \leftarrow \mathbf{1}\{H_{[i,l]} = H_{[-i,l]}\} \cdot (1 - \mu_l) + \mathbf{1}\{H_{[i,l]} \neq H_{[-i,l]}\} \cdot \mu_l$ 
   $a^{old} \leftarrow \Pi_{[i,-i]} \cdot \theta / F_{[i,1]}$ 
  while  $l < t$  do
     $\theta \leftarrow \mathbf{1}\{H_{[i,l]} = H_{[-i,l]}\} \cdot (1 - \mu_l) + \mathbf{1}\{H_{[i,l]} \neq H_{[-i,l]}\} \cdot \mu_l$ 
     $a^{new} \leftarrow \left( \frac{F_{[i,l]}}{F_{[i,l+1]}} \right) \theta \cdot ((1 - \rho_l) a^{old} + \rho_l \Pi_{[i,-i]})$ 
     $a^{old} \leftarrow a^{new}$ 
     $l \leftarrow l + 1$ 
  end while
  return  $F_{[i,t]} \left( a_{[1:i-1]}^{old}, 0, a_{[i+1:N]}^{old} \right) \triangleright N$  vector, zero reflects that a haplotype cannot copy itself.
end function

function CBP( $i, t, G_{[i,\cdot]} | H, \Pi, \mu, \rho$ )
   $l \leftarrow L$ 
   $\theta \leftarrow \mathbf{1}\{H_{[i,l]} = H_{[-i,l]}\} \cdot (1 - \mu_l) + \mathbf{1}\{H_{[i,l]} \neq H_{[-i,l]}\} \cdot \mu_l$ 
   $b^{old} \leftarrow \theta / G_{[i,L]}$ 
  while  $l > t$  do
     $b^{new} \leftarrow \left( \frac{G_{[i,l]}}{G_{[i,l-1]}} \right) ((1 - \rho_{l-1}) \theta \cdot b^{old} + \rho_{l-1} \Pi_{[i,-i]})$ 
     $\theta \leftarrow \mathbf{1}\{H_{[i,l-1]} = H_{[-i,l-1]}\} \cdot (1 - \mu_{l-1}) + \mathbf{1}\{H_{[i,l-1]} \neq H_{[-i,l-1]}\} \cdot \mu_{l-1}$ 
     $b^{old} \leftarrow b^{new}$ 
     $l \leftarrow l - 1$ 
  end while
  return  $G_{[i,t]} \left( (b_{[1:i-1]}^{old}, 0, b_{[i+1:N]}^{old} \right) \triangleright N$  vector, zero reflects that a haplotype cannot copy itself.
end function

```

$$\left(\frac{F_{[i,l]}}{F_{[i,l+1]}} \right) = \frac{P(H_{[i,1:l]})}{P(H_{[i,1:l+1]})} \geq 1$$

This ratio essentially measures how “matchable” the recipient haplotype is to the donor haplotypes at the $l + 1$ position; how “important” it is for the recipient to match a donor at a given position. If the ratio is much greater than 1, that implies that matching a donor at the $l + 1$ position will be strongly preferred in among paths sampled from the HMM.

$$\begin{aligned} \left(\frac{F_{[i,l]}}{F_{[i,l+1]}} \right) &= \frac{\mathbf{1}^T \alpha^l}{(\theta^{l+1})^T ((1 - \rho_l) \alpha^l + \rho_l (\mathbf{1}^T \alpha^l) \Pi_{[i,-i]})} \\ &= \left((\theta^{l+1})^T \left((1 - \rho_l) \frac{\alpha^l}{(\mathbf{1}^T \alpha^l)} + \rho_l \Pi_{[i,-i]} \right) \right)^{-1} \end{aligned}$$

Note that if we can approximate the normalized forward probabilities $\frac{\alpha^l}{(\mathbf{1}^T \alpha^l)}$ based on a local update rule, then we might be able to completely eliminate communication outside of pairwise comparisons between haplotypes.

A totally naive approximation would just be to assume that $\frac{\alpha^l}{(\mathbf{1}^T \alpha^l)} \approx \Pi_{[i,-i]}$, the prior. In that case.

$$\left(\frac{F_{[i,l]}}{F_{[i,l+1]}} \right) \approx \left((\theta^{l+1})^T \Pi_{[i,-i]} \right)^{-1}$$

Note that the elements of θ^{l+1} are μ for each donor that mismatches the i^{th} haplotype at position $l + 1$ and $1 - \mu$ for all donors that match. Let x be the proportion of donors that match the i^{th} haplotype at position $l + 1$. If $\Pi_{[i,-i]}$ is uniform, $\frac{1}{N-1} \vec{1}$, then

$$\left((\theta^{l+1})^T \Pi_{[i,-i]} \right)^{-1} = ((1 - \mu)x + \mu(1 - x))^{-1} = ((1 - 2\mu)x + \mu)^{-1}$$

Note that $\lim_{x \rightarrow 1} ((1 - 2\mu)x + \mu)^{-1} = \frac{1}{1-\mu} \approx 1$. In other words, if donor's allele at the position is common in the sample, then it is not very important for the donor to choose a haplotype that matches at the $l + 1$ position. However $\lim_{x \rightarrow 0} ((1 - 2\mu)x + \mu)^{-1} = \frac{1}{\mu} \gg 1$; therefore, if haplotype i has a rare allele at position $l + 1$, then that position will contribute very strongly to the conditional forward propagation.

Chapter 3

Scalable Screening Procedures for Quadratic Forms

Chapter 2 introduced a procedure for computing a distance matrix at each SNP along a chromosome that encodes pair-wise genetic distances between all individuals in a sample. Since it is impossible to simultaneously store each $N \times N$ local distance matrix obtained at each SNP, we need to be able to quickly screen these distance matrices for an association signal one at a time as they are computed. In this chapter, we develop a set of methods for obtaining lower and upper bounds on the CDF of a quadratic form in independent random variables under a widely applicable set of assumptions. Due to its utility outside of genetics, we present this work as a general method for screening for dependence between a weighted graph or covariance structures and a vector of outcomes. In Chapter 4, we will return to the genomics context and show how this result allows for the efficient screening of loci for association signals under a range of plausible genetics models.

Let $y_i \in \mathbb{R}$, $i = 1, \dots, n$, be independent random variables and $M \in \mathbb{R}^{n \times n}$ a given matrix. In the special case where each $y_i \sim N(\mu_i, \sigma_i^2)$, we provide an efficient screening method for obtaining upper and lower bounds on the CDF of the quadratic form, $y^T M y$ that avoids a full eigendecomposition. The result does not require any additional constraints on M , such as

positive semi-definiteness. We then provide two invariance principles which together allow us to extend our framework to many applications involving non-Gaussian random variables under explicit assumptions.

3.1 Efficient Bounds for Gaussian Quadratic Forms

3.1.1 Overall Approach

Let $y_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$. Several methods that have been proposed to precisely calculate $\mathbb{P}(y^T M y \geq a)$ require the full eigendecomposition of M (see [47] for an review). In order to avoid this $\mathcal{O}(n^3)$ operation, we perform a top k eigenvector decomposition that allows us to partition $y^T M y$ into a truncated part, T_k , from the top k eigenvalues and a remainder part, R_k . Since $v^T M v = v^T (\frac{1}{2}(M + M^T)) v$ for any $v \in \mathbf{R}^n$, we can assume that M is symmetric without loss of generality when estimating the CDF. Henceforth, we assume that every M is symmetric. Using the Implicitly Restarted Lanczos Method, a Krylov subspace method optimized for symmetric matrices, the top k eigenvalues and their corresponding eigenvectors of a symmetric matrix can be achieved in a fraction of the time [48]. By the "top k " eigenvalues, we mean the k eigenvalues of M with the largest magnitude. If most of the variance of M is captured in the top k eigenvalues for $k \ll N$, then intuitively, $\mathbb{P}(T_k \geq a)$ should be close to $\mathbb{P}(y^T M y \geq a)$. Let $\mathbb{E}(y) = \mu \in \mathbf{R}^n$. We make use of an exponential concentration identity related to the Hanson-Wright inequality to bound the potential contribution R_k , achieving bounds, $\{\mathcal{L}, \mathcal{U}\}$.

$$\mathcal{L}(M, y) \leq \mathbb{P}(y^T M y \geq a) \leq \mathcal{U}(M, y). \quad (3.1)$$

In practice, when given an asymmetric matrix W , we perform a pre-processing step to set $M = \frac{1}{2}(W + W^T)$ and use the fact that $\mathbb{P}(y^T W y \geq a) = \mathbb{P}(y^T M y \geq a)$. Note that this symmetrization step is only a $\mathcal{O}(n^2)$ operation, not a $\mathcal{O}(n^3)$ operation. Furthermore, we can also assume that each y_i has unit variance since we can pre-standardize the rows and columns of M , another $\mathcal{O}(n^2)$ operation. For now, let us assume that $y_i \sim N(\mu_i, 1)$. Let $V\Lambda V^T$ be the orthogonal diagonalization of M and $Z_j \stackrel{iid}{\sim} N(0, 1)$, $j = 1, \dots, n$. Exploiting the spherical symmetry of the centered multivariate gaussian with identity covariance, we have a standard representation of the quadratic form:

$$y^T M y \sim (Z + \mu)^T M (Z + \mu) = \sum_{j=1}^n \lambda_j (v_j^T Z + v_j^T \mu)^2 \sim \sum_{j=1}^n \lambda_j (Z_j + v_j^T \mu)^2. \quad (3.2)$$

Thus we can write $y^T M y = T_k + R_k$ where

$$T_k = \sum_{j=1}^k \lambda_j (Z_j + v_j^T \mu)^2, \quad R_k = \sum_{j=k+1}^n \lambda_j (Z_j + v_j^T \mu)^2.$$

With a top k eigendecomposition, the Davies method makes the CDF of T_k accessible by numerically inverting the characteristic function of T_k , which is simply a linear combination of independent non-central χ_1^2 random variables [49]. If we can obtain a function $H(r) : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\mathbb{P}(R_k > r) \leq H(r) \quad \text{and} \quad \mathbb{P}(-R_k > r) \leq H(r), \quad (3.3)$$

then we can control the contribution from R_k to $\mathbb{P}(y^T M y < a)$ with either of the following two lemmas.

Lemma 3.1.1. *For any two real random variables, X, Y , if there exists a function H such that $\mathbb{P}(X > \varepsilon) \leq H(\varepsilon)$ and $\mathbb{P}(-X > \varepsilon) \leq H(\varepsilon)$ for all $\varepsilon \geq 0$, then*

$$\mathbb{P}(Y < a - \varepsilon) - H(\varepsilon) \leq \mathbb{P}(X + Y < a) \leq \mathbb{P}(Y < a + \varepsilon) + H(\varepsilon)$$

for all $\varepsilon \geq 0$.

Proof:

$$\begin{aligned} \mathbb{P}(Y < a - \varepsilon) - H(\varepsilon) &\leq \\ \mathbb{P}(Y < a - \varepsilon) - \mathbb{P}(X > \varepsilon) &\leq \mathbb{P}(X + Y < a) \leq \mathbb{P}(Y < a + \varepsilon) + \mathbb{P}(X < -\varepsilon) \leq \\ &\mathbb{P}(Y < a + \varepsilon) + H(\varepsilon) \quad \square \end{aligned}$$

By optimizing these bounds in ε , this lemma can be used to obtain relatively tight lower and upper bounds if the $\text{Var}(X) \ll \text{Var}(Y)$. Since we also know that R_k and T_k are independent, we can achieve a tighter bound that also avoids optimizing over a free parameter and, as shown Lemma 3.1.2, can be obtained if we can integrate over the density of T_k using the Davies method.

Lemma 3.1.2. *For any two independent, real random variables, X, Y , if there exists a function H such that $\mathbb{P}(X > \varepsilon) \leq H(\varepsilon)$ and $\mathbb{P}(-X > \varepsilon) \leq H(\varepsilon)$ for all $\varepsilon \geq 0$, then*

$$\mathbb{P}(Y < a) - \int_{-\infty}^a H(a - y) f_Y(y) dy \leq \mathbb{P}(Y + X < a) \leq \mathbb{P}(Y < a) + \int_a^{\infty} H(y - a) f_Y(y) dy.$$

Proof: Using the independence of X and Y to factorize their density, we have the following bounds.

$$\begin{aligned} \mathbb{P}(Y < a) - \int_{-\infty}^a H(a-y)f_Y(y) &\leq \\ \mathbb{P}(Y < a) - \int_{-\infty}^a \mathbb{P}(X > a-y)f_Y(y) &\leq \mathbb{P}(X+Y < a) \leq \mathbb{P}(Y < a) + \int_a^{\infty} \mathbb{P}(X < a-y)f_Y(y) \leq \\ &\mathbb{P}(Y < a) + \int_a^{\infty} H(y-a)f_Y(y) \quad \square \end{aligned}$$

This implies that if we can find a way to upper bound the tails of R_k , then we'll be able to upper and lower bound the CDF of $y^T M y$. We obtain such bounds in the following section.

3.1.2 Exponential Concentration for Gaussian Quadratic Forms

Consider a real, symmetric matrix M . In order to construct an upper bound on the tail of $X^T M X$ where $x \sim N(\mu, I)$, we need the following definitions. Let $\|M\|_{HS}$ denote the Hilbert-Schmidt norm of M . Let $M = V \Lambda V^T$ be the diagonalization of M such that λ_i is the eigenvalue corresponding to eigenvector v_i . Let $b > 0$ be an upper bound on the operator norm of M : $\|M\| = \max_i |\lambda_i| \leq b$. With these definitions, we are ready to state our primary concentration inequality. The proof, as with many inequalities, is based on bounding the MGF of the quadratic form, and will require the following lemma.

Lemma 3.1.3. *Let $Z \sim N(0, 1)$. Then, for all $t \in (0, \frac{1}{2\sigma^2})$,*

$$E \left(e^{t(\sigma Z + \mu)^2} \right) = (1 - 2\sigma^2 t)^{-1/2} \exp \left\{ \frac{\mu^2 t}{1 - 2\sigma^2 t} \right\}. \quad (3.4)$$

Using this MGF, we use an approach inspired by notes by David Pollard to obtain the following concentration inequality.

Theorem 3.1.4. *Let $X \sim N(\mu, I)$ and M be a real, symmetric matrix. Let $v = 2 \left(4 \|M\mu\|_2^2 + 2 \|M\|_{HS}^2 \right)$. Then, for all $q > \mathbb{E} [X^T M X]$,*

$$\mathbb{P}(X^T M X > q) \leq \begin{cases} \exp\left(-\frac{1}{2} \frac{(q - \mathbb{E}[X^T M X])^2}{v}\right) & \mathbb{E}[X^T M X] < q \leq \frac{v}{4b} + \mathbb{E}[X^T M X] \\ \exp\left(\frac{1}{2} \frac{v}{(4b)^2}\right) \exp\left(-\frac{q - \mathbb{E}[X^T M X]}{4b}\right) & q > \frac{v}{4b} + \mathbb{E}[X^T M X]. \end{cases}$$

Furthermore, for all $q < \mathbb{E}[X^T M X]$,

$$\mathbb{P}(X^T M X < q) \leq \begin{cases} \exp\left(-\frac{1}{2} \frac{(\mathbb{E}[X^T M X] - q)^2}{v}\right) & \mathbb{E}[X^T M X] - \frac{v}{4b} \leq q < \mathbb{E}[X^T M X] \\ \exp\left(\frac{1}{2} \frac{v}{(4b)^2}\right) \exp\left(-\frac{\mathbb{E}[X^T M X] - q}{4b}\right) & q < \mathbb{E}[X^T M X] - \frac{v}{4b}. \end{cases}$$

Importantly, this concentration inequality allows us to achieve exponential bounds on the tail of $X^T M X$ in terms of quantities that can be readily computed for even very large matrices: the largest eigenvalue (in magnitude), the Hilbert-Schmidt norm, and the norm of a matrix applied to a vector. Now that we have this machinery for bounding quadratic forms in Gaussian random variables, in the following section, we show how this result can be useful for bounding quadratic forms in non-Gaussian random variables. Below we pause to briefly present how our approach to deriving Theorem 3.4.2 may be generalized to functions of matrices.

Extending to $X^T f(M) X$

A powerful generalization of this concentration inequality is possible for functions of M that still only rely on the trace and norms of M , all of which are easily calculable. In this context, when we say a function of M , $f(M)$, let that denote the matrix where f is applied elementwise to the eigenvalues of M . $f(M) = V \text{diag}(f(\lambda_j)) V^T$. Typically, calculating functions of M such as the matrix exponential are computationally prohibitive since, in general, they require a full eigendecomposition. Perhaps surprisingly, it is possible to avoid the eigendecomposition and still obtain a bound on the tail probabilities of $X^T f(M) X$ without

ever computing $f(M)$. Below we recapitulate part of the proof of Theorem 3.4.2 in the more general case where we are interested in bounding the CDF of $X^T f(M)X$. Unfortunately, this extension does not yield a bound in closed form that can be written explicitly for all f , but it does imply a numerical procedure for achieving such bounds for a specific choice of f . We have not fully formalized this approach, but we give a brief intuition for the extension below.

Theorem 3.4.2 relies on the ability to bound the MGF of $X^T MX$ in terms of quadratics in λ_j . Since quadratics in λ_j can be expressed in terms of basic matrix functions, as done in the proof of Theorem 3.4.2, bounding the MGF in this way yields an easily calculable concentration inequality for $X^T MX$. Following the same steps as in the proof of Theorem 3.4.2, we can obtain the MGF of $X^T f(M)X$:

$$\mathbb{E} \left[e^{tX^T f(M)X} \right] = \exp \left(\sum_{j=1}^n (v_j^T \mu)^2 \frac{tf(\lambda_j)}{1-2tf(\lambda_j)} - \frac{1}{2} \log(1-2tf(\lambda_j)) \right).$$

Let us assume without loss of generality that we've obtained the largest magnitude eigenvalue of M and have normalized the matrix so that $|\lambda_j| \leq 1 \forall j$. Following the same approach as the proof of Theorem 3.4.2, bounding $\frac{tf(\lambda_j)}{1-2tf(\lambda_j)}$ and $-\frac{1}{2} \log(1-2tf(\lambda_j))$ with quadratics in λ_j is sufficient to obtain a bound in terms of easily calculable matrix norms. This can be easily done using Taylor's Theorem as demonstrated below

$$\frac{tf(\lambda_j)}{1-2tf(\lambda_j)} \leq \frac{tf(0)}{1-2tf(0)} + \frac{\lambda_j}{(1-2tf(0))^2} + \frac{1}{2} \max_{\lambda_j \in [-1,1]} \left[\frac{4tf'(\lambda_j)}{(1-2tf(\lambda_j))^3} \right] \lambda_j^2.$$

Note that the coefficients of this quadratic are simply functions of t , which can be numerically optimized to obtain bounds on the CDF of $X^T f(M)X$ without ever forming $f(M)$.

3.2 Non-Gaussian Random Variables

In order to exploit our CDF bounding approach for Gaussian random variables in the non-Gaussian case, we will provide two non-asymptotic bounds which guarantee when the CDF of a quadratic form of interest will be close to the CDF of a quadratic form in Gaussians. A large body of work has already been done on the asymptotics of quadratic forms and can be summarized in terms of invariance classes of quadratic forms. Consider two random vectors $y, x \in R^n$. Let each component y_i be independent and come from some distribution F_i such that $\mathbb{E}(y_i) = \mu_i$ and $\text{Var}(y_i) = \sigma_i^2$. Let each component x_i be independent and come from some distribution G_i such that $\mathbb{E}(x_i) = \mu_i$ and $\text{Var}(x_i) = \sigma_i^2$. Let M be a real, $n \times n$ matrix, $v_y^2 = \text{Var}(y^T M y)$, and $v_x^2 = \text{Var}(x^T M x)$.

$$Q_y^{(n)} = \frac{1}{v_y} y^T M y, \quad Q_x^{(n)} = \frac{1}{v_x} x^T M x$$

,

$$d_M(x, y) = \max_a \left| \mathbb{P} \left(Q_y^{(n)} \leq a \right) - \mathbb{P} \left(Q_x^{(n)} \leq a \right) \right|.$$

If the Kolmogorov Distance, $d_M(x, y)$, between the distributions of the $Q_y^{(n)}$ and $Q_x^{(n)}$ tends to zero for n large, then x and y are said to belong to the invariance class of M [50]. Of course such a result will only be guaranteed and an invariance class will only exist for a given set of sufficient conditions on M_n , the set of distributions $\{F_i\}$, and the set of distributions $\{G_i\}$.

Wet and Venter (1973) defined an invariance class of M , assuming difficult-to-verify conditions on the spectrum of M , that the y_i are all i.i.d. $y_i \sim F$, and that the x_i are all i.i.d. $x_i \sim G$ [51]. Rotar (1973) derives a similar result in the i.i.d. case but provides an explicit, non-asymptotic bound on $d_M(x, y)$. This non-asymptotic bound provides a set of sufficient conditions on M that are much easier to verify in practice. Gamkrelidze and Rotar (1978) and Rotar and Shervashidze (1985), extend this result to the more general, non-i.i.d. case but

with the added assumptions that M is hollow ($M_{ii} = 0 \forall i$) and that $\mathbb{E}(y_i) = \mathbb{E}(x_i) = 0 \forall i$ [52, 53].

As explained in Rotar (1973) and later in Rotar's work extending these results to higher dimensional tensors (polylinear forms), the zero mean condition can be easily relaxed [50, 54]. To provide the intuition for this result, any quadratic form in a random vector with a non-zero mean can be rewritten as the sum of a centered quadratic form and a linear form in a centered random vector as follows.

$$y^T M y = (y - \mu)^T M (y - \mu) + 2\mu^T M (y - \mu) + \mu^T M \mu \quad (3.5)$$

However, none of these existing works extend on this observation to provide an explicit, non-asymptotic bound on $d_M(x, y)$ in the non-i.i.d. case in terms of $\mu = \mathbb{E}(y)$. So while Rotar shows that such an asymptotic result should hold, it is important in practice to be able to flag instances where the convergence may not be valid without a bound that includes μ . Although our bound will require the existence of the fourth moment rather just the third moment for each y_i as in Rotar's work, this is a valid assumption for the vast majority of applications. In addition, none of these prior works provide a non-asymptotic bound for the non-i.i.d. case when M has a non-zero diagonal, which of course would be the case for all covariance matrices.

We answer these open questions with two sets of results. The first set provides bounds on $d_M(x, y)$ when M is hollow. The second set of results provides bounds on $d_M(x, y)$ when $\mu = 0$ but M is not hollow. As we will see, this second set of results also requires that the second, third, and fourth moments of x and y match. Although this may appear to imply that the distribution of the quadratic form $y^T M y$ cannot, in general, be approximated by a quadratic form in normal random variables, we provide a Rao-Blackwellization procedure which allows us to meet the fourth order moment matching condition of this second theorem while still applying our Gaussian bounding approach to obtain fast tail bounds on $\mathbb{P}\left(Q_y^{(n)} \leq a\right)$.

Our non-asymptotic bounds are crucial in practice because they capture all of the terms that could make approximating $Q_y^{(n)}$ with $Q_x^{(n)}$ hazardous for finite n . As we will see below, Theorem 3.2.1 and Theorem 3.2.6 both give bounds that imply sufficient conditions for the asymptotic convergence of $d_M(x, y)$ to zero. However, when we extend these theorems to derive explicit, non-asymptotic bounds on $d_M(x, y)$ where we assume x is Gaussian, we find that the spectrum of M actually impacts the convergence of $Q_y^{(n)}$ to $Q_x^{(n)}$ (see Lemma 3.2.3). Explicitly, we find that if M has only two eigenvalues and that they have opposite signs, it is not safe to invoke the convergence of $Q_y^{(n)}$ to $Q_x^{(n)}$. This property is completely washed over in the purely asymptotic results. Thus, our non-asymptotic bounds on $d_M(x, y)$ provide practitioners with an explicit set of statistics that can be used to diagnose when it may not be valid to approximate the distribution of $Q_y^{(n)}$ with the distribution of $Q_x^{(n)}$.

3.2.1 Hollow Matrices

Theorem 3.2.1. *Let M be a hollow, real matrix and ϕ be a real function with a bounded third derivative: $|\phi'''| \leq K_3$. For two random vectors with independent entries, y, x , let the first four moments of each y_i and each x_i exist. If the first two moments of x_i match the first two moments of y_i for all i , then*

$$\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right| \leq \frac{1}{\sqrt{n}} \frac{\psi K_3}{6} \frac{27^{1/4} \max(\tau^3, 1)}{\sigma_\star^3 \min \left(\min_j \left(\left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right)^{3/2} \right), 1 \right)} \rho_1^3. \quad (3.6)$$

where

$$\begin{aligned} \sigma_\star^2 &= \min_j \sigma_j^2, \\ \tau &\geq \max \left(\max_j (||x_j||_4), \max_j (||y_j||_4) \right), \\ \psi &\geq \max_j \left[\mathbb{E} \left[|x_j|^3 \right] + \mathbb{E} \left[|y_j|^3 \right] \right], \\ \rho_1 &= \frac{\max_j \left[||M_j||_2 + |M_j^T \mu| \right]}{\min_j \left[\sqrt{||M_j||_2^2 + (M_j^T \mu)^2} \right]}. \end{aligned}$$

Since Theorem 3.2.1 applies to any ϕ with bounded third moments, it applies to all sin and cos functions. As long as $\frac{3^{3/4} \sqrt{8} \max(\tau^3, 1)}{\sigma_\star^3 \min(\sigma_\star^3, \sqrt{8})} \rho_1^3$ does not grow faster than $\sqrt{n}$, this implies asymptotic convergence in cumulant generating functions and hence convergence in distribution. This result does not assume that either x or y come from any specific distribution, hence this is not a central limit theorem in a traditional sense but rather an invariance principle.

Since n will of course be finite in any application, we will now proceed to use Theorem 3.2.1 to derive non-asymptotic bounds on $d_M(x, y)$. Our overall approach will be to choose a continuous function ϕ that closely approximates a step function in order to bound the difference in CDFs. This idea is presented in Lemma 3.2.2 below. For some $\varepsilon > 0$ let

$$\phi_{a,\varepsilon}(x) = \begin{cases} 1 & x \leq a \\ \frac{1}{2} \left(\cos \left(\frac{\pi}{\varepsilon}(x-a) \right) + 1 \right) & a < x \leq a + \varepsilon \\ 0 & x > a + \varepsilon \end{cases}.$$

The first, second, and third, partial derivatives of $\phi_{a,\varepsilon}$ with respect to x exist and are bounded.

Define $K_3 = \sup_{x \in \mathbb{R}} |\phi_{a,\varepsilon}'''(x)|$. Explicitly, the third partial derivative is:

$$\phi_{a,\varepsilon}'''(x) = \begin{cases} 0 & x < a \\ \frac{1}{\varepsilon^3} \frac{\pi^3}{2} \sin \left(\frac{\pi}{\varepsilon}(x-a) \right) & a \leq x \leq a + \varepsilon \\ 0 & x > a \end{cases}$$

Thus, $K_3 = \frac{1}{2} \left(\frac{\pi}{\varepsilon} \right)^3$. For this specific definition of $\phi_{a,\varepsilon}$, we can bound the Kolmogorov distance between two distributions using the following lemma.

Lemma 3.2.2. *Consider two random variables X and Y . Let Y be continuous and its density be upper bounded by a constant U . If there exists a constant $C \in \mathbb{R}$ such that $|\mathbb{E}(\phi_{a,\varepsilon}(X)) - \mathbb{E}(\phi_{a,\varepsilon}(Y))| \leq K_3 C \forall a \in \mathbb{R}$ and $\forall \varepsilon > 0$, then*

$$|\mathbb{P}(X \leq a) - \mathbb{P}(Y \leq a)| \leq \frac{4}{3} U \left(\frac{3\pi^3 C}{2} \right)^{1/4}$$

for all $a \in \mathbb{R}$.

It is critical to note that Lemma 3.2.2 requires one of the random variables to be continuous with a density function bounded above by some U . While such a bound is hard to obtain for a quadratic form in arbitrary independent random variables, such a bound is possible for

most instances of Gaussian quadratic forms. We emphasize "most cases," because there are a few easy counter-examples which show that not all Gaussian quadratic forms have bounded densities. For example, if M is a rank one matrix and $\mu = 0$, then $(Z + \mu)^T M (Z + \mu) = Z^T M Z \sim \lambda_1 \chi_1^2$, which has a density that is unbounded at zero. A less trivial counterexample that might be disturbing to those hoping to use our results for genetics is the case where

$$M = \begin{bmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 1 & \cdots & 1 \\ 0 & \ddots & 0 & 1 & \ddots & 1 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 1 & \ddots & 1 & 0 & \ddots & 0 \\ 1 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix}. \quad (3.7)$$

Such an M might arise in genetics where there are two equally sized sub-populations of clones or where M represents the distance between individuals based on a single SNP of frequency 0.5 in a haploid population. It's easy to show that the eigenvalues of M are $N/2$ and $-N/2$. Again assuming that $\mu = 0$, we have, $Z^T M Z \sim \frac{N}{2} (X - Y)$, which we show has an unbounded density at zero with the following lemma.

Lemma 3.2.3. *Let $X \sim \alpha_1 (Z_1 + \eta_1)^2$, $Y \sim \alpha_2 (Z_2 + \eta_2)^2$, where $\eta, \alpha \in \mathbb{R}^2$, $\alpha_1, \alpha_2 > 0$, and $Z_1, Z_2 \sim N(0, 1)$. If $W = X - Y$, then the density of W is unbounded at 0.*

Proof:

First we need the density of X .

$$\begin{aligned}
\mathbb{P}(X \leq x) &= \mathbb{P}\left(\alpha_1 (Z_1 + \eta_1)^2 \leq x\right) \\
&= \mathbb{P}\left(-\left(\sqrt{\frac{x}{\alpha_1}} + \eta_1\right) \leq Z_1 \leq \sqrt{\frac{x}{\alpha_1}} - \eta_1\right) \\
&= \Phi\left(\sqrt{\frac{x}{\alpha_1}} - \eta_1\right) - \Phi\left(-\left(\sqrt{\frac{x}{\alpha_1}} + \eta_1\right)\right)
\end{aligned}$$

$$\begin{aligned}
f_X(x) &= \frac{\partial}{\partial x} \mathbb{P}(X \leq x) \\
&= \frac{1}{2\sqrt{\alpha_1 x}} \left[\phi\left(\sqrt{\frac{x}{\alpha_1}} - \eta_1\right) + \phi\left(-\left(\sqrt{\frac{x}{\alpha_1}} + \eta_1\right)\right) \right] \\
&= \frac{1}{2\sqrt{\alpha_1 x}} \left[\phi\left(\sqrt{\frac{x}{\alpha_1}} - \eta_1\right) + \phi\left(\sqrt{\frac{x}{\alpha_1}} + \eta_1\right) \right]
\end{aligned}$$

Since $X, Y \geq 0$, we can write the density of W as

$$\begin{aligned}
f_W(w) &= \int_w^\infty f_X(\tau) f_Y(\tau - w) d\tau \\
&= \frac{1}{4\sqrt{\alpha_1 \alpha_2}} \int_w^\infty \frac{1}{\sqrt{\tau(\tau - w)}} \left[\phi\left(\sqrt{\frac{\tau}{\alpha_1}} - \eta_1\right) + \phi\left(\sqrt{\frac{\tau}{\alpha_1}} + \eta_1\right) \right] \\
&\quad \left[\phi\left(\sqrt{\frac{\tau - w}{\alpha_2}} - \eta_2\right) + \phi\left(\sqrt{\frac{\tau - w}{\alpha_2}} + \eta_2\right) \right] d\tau
\end{aligned}$$

If $C = \min \left[\phi\left(\sqrt{\frac{\tau}{\alpha_1}} - \eta_1\right), \phi\left(\sqrt{\frac{\tau}{\alpha_1}} + \eta_1\right), \phi\left(\sqrt{\frac{\tau}{\alpha_2}} - \eta_2\right), \phi\left(\sqrt{\frac{\tau}{\alpha_2}} + \eta_2\right) \right]$, then

$$f_W(0) \geq \frac{C^2}{4\sqrt{\alpha_1 \alpha_2}} \int_0^\infty \frac{1}{\tau} d\tau = \infty. \quad \square$$

However, as shown in Lemma 3.2.4, if we can assume that M has three or more eigenvalues, or two of the same sign, then we can be sure that the density is bounded. We state Lemma

3.2.4 while allowing for μ to be an arbitrary random vector because it will be important for the justification of our Rao-Blackwellization routine at the end of the chapter.

Lemma 3.2.4. *Let $Z \sim N(0, I)$ and μ be a random vector in $\mathbf{R}^n$. Let M have two or more eigenvalues with the same sign and let $\zeta = \max_{i,j} \lambda_i \lambda_j$. Then the density of $Q = (Z + \mu)^T M (Z + \mu)$ is bounded above by $\frac{1}{2\sqrt{\zeta}}$.*

Proof:

Let λ_1, λ_2 be the two eigenvalues of the same sign with corresponding eigenvectors v_1, v_2 . Without loss of generality, assume $\lambda_1, \lambda_2 > 0$. Let $X \sim \lambda_1 (Z_1 + v_1^T \mu)^2$, $Y \sim \lambda_2 (Z_2 + v_2^T \mu)^2$, where $Z_1, Z_2 \sim N(0, 1)$. From the proof of Lemma 3.2.3, we have

$$\begin{aligned} f_X(x) &= \frac{1}{2\sqrt{\lambda_1 x}} \left[\phi\left(\sqrt{\frac{x}{\lambda_1}} - v_1^T \mu\right) + \phi\left(\sqrt{\frac{x}{\lambda_1}} + v_1^T \mu\right) \right] \\ &\leq \frac{1}{\sqrt{2\pi\lambda_1 x}}. \end{aligned}$$

Let $W = X + Y$. Since $X, Y \geq 0$,

$$\begin{aligned} f_W(w) &= \int_0^w f_X(\tau) f_Y(w - \tau) d\tau \\ &\leq \int_0^w \frac{1}{\sqrt{2\pi\lambda_1 \tau}} \frac{1}{\sqrt{2\pi\lambda_2 (w - \tau)}} d\tau \\ &= \frac{1}{2\pi\sqrt{\lambda_1 \lambda_2}} \int_0^w \frac{1}{\sqrt{\tau(w - \tau)}} d\tau \\ &= \frac{1}{2\sqrt{\lambda_1 \lambda_2}} \end{aligned}$$

This proves our base case. Now let us define $W = X + Y$ where $X = \sum_{j=1}^k \lambda_j (Z_j + v_j^T \mu)^2$ for some $k \in \{2, \dots, n-1\}$ and $Y = \lambda_{k+1} (Z_{k+1} + v_{k+1}^T \mu)^2$. Using our assumption that the density of X is bounded by some constant C , then

$$f_W(w) = \int_{-\infty}^{\infty} f_X(\tau) f_Y(w - \tau) d\tau \leq C \int_{-\infty}^{\infty} f_Y(w - \tau) d\tau = C. \quad (3.8)$$

Since, $Q = (Z + \mu)^T M (Z + \mu) \sim \sum_{j=1}^n \lambda_j \left(Z_j + v_j^T \mu \right)^2$, by induction, we have that the density of Q given μ is bounded:

$$f_{Q|\mu}(q) \leq \frac{1}{2\sqrt{\lambda_1 \lambda_2}}.$$

Mixing over μ , we still have

$$f_Q(q) = \int f_{Q|\mu}(q) f_{\mu} d\mu \leq \frac{1}{2\sqrt{\lambda_1 \lambda_2}} \int f_{\mu} d\mu = \frac{1}{2\sqrt{\lambda_1 \lambda_2}}. \quad \square$$

Now we have an explicit bound which can be combined with Theorem 3.2.1 and Lemma 3.2.2 to obtain the following bound on $d_M(x, y)$.

Theorem 3.2.5. *Let M be hollow and have at least two eigenvalues of the same sign. Let the first four moments of each y_i exist and $x_i \sim N(\mu_i, \sigma_i^2)$. If the first two moments of x_i match the first two moments of y_i for all i , then*

$$d_M(x, y) \leq \frac{1}{n^{1/8}} \frac{\rho_1^{3/4}}{\sqrt{\zeta}} g_1(\tau, \sigma_\star^2, \sigma, \mu, \psi),$$

where

$$\begin{aligned} \sigma_\star^2 &= \min_j \sigma_j^2, \\ \tau &\geq \max \left(\max_j (||x_j||_4), \max_j (||y_j||_4) \right), \\ \psi &\geq \max_j \left[\mathbb{E} [|x_j|^3] + \mathbb{E} [|y_j|^3] \right], \\ \rho_1 &= \frac{\max_j [||M_j||_2 + |M_j^T \mu|]}{\min_j \left[\sqrt{||M_j||_2^2 + (M_j^T \mu)^2} \right]}, \\ \zeta &= \max_{i,j} [\lambda_i \lambda_j], \\ g_1(\tau, \sigma_\star^2, \sigma, \mu, \psi) &= \frac{2}{3} \left(\frac{27^{1/4} \psi \pi^3 \max(\tau^3, 1)}{4 \sigma_\star^3 \min \left(\min_j \left(\left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right)^{3/2} \right), 1 \right)} \right)^{1/4}. \end{aligned}$$

Proof:

Since each x_i is Gaussian and we have that M has at least two eigenvalues of the same sign, we can be sure that the density of $\frac{1}{\sqrt{x}} x^T M x$ is bounded by Lemma 3.2.4. Since the Gaussianity of x also implies $\frac{1}{\sqrt{x}} x^T M x$ is a continuous random variable, we can combine this result with Lemma 3.2.2 and Theorem 3.2.1 to obtain

$$\begin{aligned}
d_M(x, y) &\leq \frac{4}{3} \frac{1}{2\sqrt{\zeta}} \left(\frac{3\pi^3}{2} \frac{1}{\sqrt{n}} \frac{\psi}{6} \frac{27^{1/4} \max(\tau^3, 1)}{\sigma_*^3 \min\left(\min_j \left(\left(\frac{1}{2}\sigma_j^2 + \mu_j^2\right)^{3/2}\right), 1\right)} \rho_1^3 \right)^{1/4} \\
&= \frac{1}{n^{1/8}} \frac{\rho_1^{3/4}}{\sqrt{\zeta}} \frac{2}{3} \left(\frac{27^{1/4} \psi \pi^3 \max(\tau^3, 1)}{4\sigma_*^3 \min\left(\min_j \left(\left(\frac{1}{2}\sigma_j^2 + \mu_j^2\right)^{3/2}\right), 1\right)} \right)^{1/4} \quad \square
\end{aligned}$$

Using Theorem 3.2.5, we can see when it should be safe to assume in practice that $y^T M y$ is close in distribution to $x^T M x$, where $x_i \sim N(\mu_i, \sigma_i^2)$. Informally, we would expect the distribution of $y^T M y$ to be close to the corresponding quadratic form in Gaussian random variables when

1. none of the y_i are heavy tailed,
2. no y_i is very close to a pointmass (each y_i must have some minimal variance),
3. there are no outliers in the graph structure implied by M (ρ_1 grows slower than $n^{1/6}$).

Intuitively, these conditions ensure that no single component of y can have too much influence on the quadratic form. Although one might be concerned by the fact that ζ is a function of the spectrum of M , since the first few largest in magnitude eigenvalues of M will be non-decreasing in n in most applications, we can essentially take it to be at worst a constant. In genetics applications, we can reduce these conditions to three major criteria:

1. avoid models where the fourth moment of y_i doesn't exist (eg: t-distribution with 4 degrees of freedom),
2. if the y_i are modeled as outcomes in some generalized linear model, then make sure none of them are perfectly predicted by the base model (each y_i must have some variance),

3. remove all samples that are lone outliers genetically or phenotypically from the dataset.

Importantly, the bound from Theorem 3.2.5 can be quickly and easily calculated for any M , using the Implicitly Restarted Lanczos Method to obtain the leading three eigenvalues and calculate ζ . Thus, in screening applications where M may be arriving online, we can monitor ζ and ρ_1 in order to detect matrices for which the true CDF may diverge significantly from the CDF given by Gaussian random variables.

3.2.2 Centered Random Variables

Instead of assuming that M is hollow, we can derive a similar invariance principle by assuming that $\mathbb{E}(x) = \mathbb{E}(y) = 0$. Although we will now consider a general real M , it will be helpful to separate the diagonal from the non-diagonal elements, so let $\tilde{M}$ denote a copy of M where the diagonal has been set to zero. Applying Lemma 1.5.1 we again have that $v_y^2 = v_x^2$ if y and x also have the same first four moments. Using the same variable swapping approach that we use for Theorem 3.2.1, we have the following result.

Theorem 3.2.6. *Let M be a real matrix and ϕ be a real function with a bounded third derivative: $|\phi'''| \leq K_3$. For two random vectors with independent entries, y, x , let the first eight moments of each y_i and each x_i exist. If $\mathbb{E}(y_i) = \mathbb{E}(x_i) = 0 \forall i$ and the first four moments of x_i match the first four moments of y_i for all i , then*

$$\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right| \leq \frac{1}{\sqrt{n}} \frac{K_3 \psi \max \left(8 \left(3^{3/4} \right) \tau^3, 1 \right)}{3 \left(\min \left(\kappa_*, 2\sigma_*^4 \right) \right)^{3/2}} \rho_2^3,$$

where

$$\begin{aligned} \sigma_*^2 &= \min_j \sigma_j^2 \\ \kappa_j &= \text{Var}(y_j^2) = \text{Var}(x_j^2) \\ \kappa_* &= \min_j \kappa_j \\ \tau &\geq \max_j \left(\max \left(\|x_j\|_4, \|y_j\|_4 \right) \right) \\ \psi &\geq \max_j \left(\max \left(\sqrt{\mathbb{E}[x_j^8]}, \mathbb{E}[x_j^8], \sqrt{\mathbb{E}[y_j^8]}, \mathbb{E}[y_j^8] \right) \right) \\ \rho_2 &= \frac{\max_j \left[\|\tilde{M}_j\|_2 + |m_{jj}| \right]}{\min_j \left[\sqrt{\|\tilde{M}_j\|_2^2 + m_{jj}^2} \right]} \end{aligned}$$

Note how ρ_2 weights the diagonal elements slightly more than the off diagonal elements. Comparing ρ_2 to ρ_1 from Theorem 3.2.1, we see that the diagonal elements have taken the place of $|\tilde{M}_j^T \mu|$. Just as we derived Theorem 3.2.5, we can use Theorem 3.2.6 to arrive at Theorem 3.2.7 below.

Theorem 3.2.7. *Let M have at least two eigenvalues of the same sign. Let the first eight moments of each y_i exist and $\mathbb{E}(y_i) = 0 \forall i$. For each y_i , let $x_i \sim Z_i + \mu_i$ such that $Z_i \sim N(0, 1)$ is independent of $\mu_i \sim F_i$. If marginally the first four moments of x_i match the first four moments of y_i , then*

$$d_M(x, y) \leq \frac{1}{n^{1/8}} \frac{\rho_2^{3/4}}{\sqrt{\zeta}} g_2(\psi, \tau, \sigma_\star^2, \kappa_\star),$$

where

$$\sigma_\star^2 = \min_j \sigma_j^2$$

$$\kappa_j = \text{Var}(y_j^2) = \text{Var}(x_j^2)$$

$$\kappa_\star = \min_j \kappa_j$$

$$\tau \geq \max_j (\max(\|x_j\|_4, \|y_j\|_4))$$

$$\psi \geq \max_j \left(\max \left(\sqrt{\mathbb{E}[x_j^8]}, \mathbb{E}[x_j^8], \sqrt{\mathbb{E}[y_j^8]}, \mathbb{E}[y_j^8] \right) \right)$$

$$\rho_2 = \frac{\max_j [\|\tilde{M}_j\|_2 + |m_{jj}|]}{\min_j [\sqrt{\|\tilde{M}_j\|_2^2 + m_{jj}^2}]}$$

$$\zeta = \max_{i,j} [\lambda_i \lambda_j]$$

$$g_2(\psi, \tau, \sigma_\star^2, \kappa_\star) = \frac{2}{3} \left(\frac{\pi^3}{2} \frac{\psi \max(2\tau, 1)}{(\min(\kappa_\star, 2\sigma_\star^4))^{3/2}} \right)^{1/4}.$$

Proof:

Since each Z_i is Gaussian and we have that M has at least two eigenvalues of the same sign, we can be sure that the density of $\frac{1}{v_x} x^T M x$ is bounded by Lemma 3.2.4. Since the Gaussianity of each Z_i also implies $\frac{1}{v_x} x^T M x$ is a continuous random variable, we can combine this result with Lemma 3.2.2 and Theorem 3.2.6 to obtain

$$\begin{aligned}
d_M(x, y) &\leq \frac{4}{3} \frac{1}{2\sqrt{\zeta}} \left(\frac{3\pi^3}{2} \frac{1}{\sqrt{n}} \frac{\psi \max(2\tau, 1)}{3 (\min(\kappa_*, 2\sigma_*^4))^{3/2}} \rho_2^3 \right)^{1/4}, \\
&= \frac{1}{n^{1/8}} \frac{\rho_2^{3/4}}{\sqrt{\zeta}} \frac{2}{3} \left(\frac{\pi^3}{2} \frac{\psi \max(2\tau, 1)}{(\min(\kappa_*, 2\sigma_*^4))^{3/2}} \right)^{1/4}. \quad \square
\end{aligned}$$

Theorem 3.2.6 implies sufficient conditions for convergence that are very similar to those implied by Theorem 3.2.1, except that now μ must be introduced in order to match the 3rd and 4th moments of y . In the following section, we present a Rao-Blackwellization procedure that shows how to explicitly construct μ and apply Theorem 3.2.6 to achieve fast upper and lower bounds on the CDF for quadratic forms in centered random variables. We recommend this approach for investigators hoping to use models like SKAT for testing random effects in GLMMs [12].

Rao-Blackwellization

Below we show how Theorem 3.2.6 can be used in combination with our Gaussian concentration inequality, Theorem 3.4.2, to bound the tails of quadratic forms in centered random variables. As required by Theorem 3.2.6, we assume that the first eight moments of each y_i exist.

Let $x_i = \sigma \left(\sqrt{1 - \omega_i^2} z_i + \omega_i b_i \right)$ where $z_i \sim N(0, 1)$ and $b_i \sim \frac{g_i - \mathbb{E}(g_i)}{\sqrt{\text{Var}(g_i)}}$, $g_i \sim \text{Beta}(\alpha_i, \beta_i)$. Using our invariance principle, we can approximate $\mathbb{P}(y^T M y < a)$ with $\mathbb{P}(x^T M x < a) = \mathbb{E}[\mathbf{1}\{x^T M x < a\}]$ if both of the vectors are centered and the remaining first four moments match. It's easy to show with our current construction that $\mathbb{E}(y_i) = \mathbb{E}(x_i) = 0$ and $\text{Var}(y_i) = \text{Var}(x_i) = \sigma_i^2$. The difficulty arises in how to choose α_i and β_i to match the third and fourth moments. Before addressing this point, first we will show how this construction is useful for obtaining bounds on $\mathbb{E}[\mathbf{1}\{x^T M x < a\}]$.

$$\mathbb{E} [\mathbf{1} \{x^T M x < a\}] = \mathbb{E} \left[\mathbf{1} \left\{ \left(\sigma \left(\sqrt{1-\omega^2} z + \omega b \right) \right)^T M \left(\sigma \left(\sqrt{1-\omega^2} z + \omega b \right) \right) < a \right\} \right]. \quad (3.9)$$

If we let $M^\dagger = M * \left(\sigma \sqrt{1-\omega^2} \right) \left(\sigma \sqrt{1-\omega^2} \right)^T$, then we can rewrite this as

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1} \left\{ \left(z + \sqrt{\frac{\omega^2}{1-\omega^2}} b \right)^T M^\dagger \left(z + \sqrt{\frac{\omega^2}{1-\omega^2}} b \right) < a \right\} \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\mathbf{1} \left\{ \left(z + \sqrt{\frac{\omega^2}{1-\omega^2}} b \right)^T M^\dagger \left(z + \sqrt{\frac{\omega^2}{1-\omega^2}} b \right) < a \right\} \middle| x \right] \right] \\ &= \mathbb{E} \left[\mathbb{P} \left(\left(z + \sqrt{\frac{\omega^2}{1-\omega^2}} b \right)^T M^\dagger \left(z + \sqrt{\frac{\omega^2}{1-\omega^2}} b \right) < a \middle| b \right) \right], \end{aligned}$$

where the outer expectation is over b . Let $\mathcal{L}(M, \mu, a)$ be a function of matrix M , a fixed vector μ , and a scalar a such that $\mathcal{L}(M, \mu, a) \leq \mathbb{P} \left((Z + \mu)^T M (Z + \mu) < a \right) \forall a \in \mathbb{R}$ where $Z \sim N(0, I)$. Let $\mathcal{U}(M, \mu, a)$ be an analogous function that upper bounds the same CDF. Since we have such functions from our lower and upper bounding approach for Gaussian quadratic forms, we apply them here inside the expectation over b to obtain

$$\mathbb{E} \left[\mathcal{L} \left(M, \sqrt{\frac{\omega^2}{1-\omega^2}} b, a \right) \right] \leq \mathbb{P} (y^T M y < a) \leq \mathbb{E} \left[\mathcal{U} \left(M, \sqrt{\frac{\omega^2}{1-\omega^2}} b, a \right) \right]. \quad (3.10)$$

This result demonstrates that we can obtain upper and lower bounds on $\mathbb{P} (y^T M y < a)$ when M has a non-zero diagonal by simulating b , and for each draw, apply our bounding approach for Gaussian quadratic forms. Note that since the eigenvalues and eigenvectors of M do not change between draws of b , they only need to be calculated once. Since we are simulating b , we will not be able to completely remove the Monte-Carlo error. In estimating

the bounds, we are left with two options. Either we take the mean of our sampled upper bounds (or the sampled lower bounds) to be our estimate and appeal to the central limit theorem in order to put a confidence interval on the estimate. However, since we have been careful enough to use exact bounds up to this point, it is possible for a user to use Hoeffding's inequality to provide exact probabilistic bounds on the Monte-Carlo error.

We use normalized Beta random variables to produce the third and fourth order moment correction partly because they have bounded support, limiting the sampling variability that they can contribute to our bound estimates. This Rao-Blackwellization approach can only produce useful bounds as long as most of the variability in x is driven by the normal component z , which we integrate over analytically, rather than by the beta component b . This motivates the objective of minimizing ω_i . Since x_i and y_i match on the first two moments, it is sufficient to match them on their skewness and excess kurtosis in order to match on the first four moments. This yields the following objective for each y_i . Find $\omega_i, \alpha_i, \beta_i$ that minimize ω_i while satisfying

$$\begin{aligned} \text{Skew}(y_i) &= \text{Skew}(x_i) = \omega_i^3 \text{Skew}(g_i) = \omega_i^3 \frac{2(\beta_i - \alpha_i) \sqrt{\alpha_i + \beta_i + 1}}{(\alpha_i + \beta_i + 2) \sqrt{\alpha_i \beta_i}}, \\ \text{ExKurt}(y_i) &= \text{ExKurt}(x_i) = \mathbb{E} \left(\frac{x_i^4}{\sigma_i^4} \right) - 3 \\ &= (1 - \omega_i^2)^2 \mathbb{E}(z_i^4) + 6(1 - \omega_i^2) \omega_i^2 \mathbb{E}(z_i^2) \mathbb{E}(b_i^2) + \omega_i^4 \mathbb{E}(b_i^4) - 3 \\ &= 3(1 - \omega_i^2)^2 + 6(1 - \omega_i^2) \omega_i^2 + \omega_i^4 \text{Kurt}(g_i) - 3 \\ &= 3 - 6\omega_i^2 + 3\omega_i^4 + 6\omega_i^2 - 6\omega_i^4 + \omega_i^4 \text{Kurt}(g_i) - 3 \\ &= \omega_i^4 \text{ExKurt}(g_i) = \omega_i^4 \left(\frac{6(\beta_i - \alpha_i)^2 (\alpha_i + \beta_i + 1) - \alpha_i \beta_i (\alpha_i + \beta_i + 2)}{\alpha_i \beta_i (\alpha_i + \beta_i + 2) (\alpha_i + \beta_i + 3)} \right). \end{aligned}$$

For brevity, let us temporarily drop the i subscripts and reparameterize such that $\kappa = \alpha + \beta$ and $\gamma = \alpha - \beta$. We use the identity that $\frac{1}{2} \sqrt{\kappa^2 - \gamma^2} = \sqrt{\alpha \beta}$. We must minimize ω such that

$$\text{Skew}(y_i) = \frac{4\omega^3\gamma\sqrt{\kappa+1}}{(\kappa+2)\sqrt{\kappa^2-\gamma^2}} \quad (3.11)$$

$$\text{ExKurt}(y_i) = \frac{\omega^4}{\kappa+3} \left[\frac{24\gamma^2(\kappa+1)}{(\kappa+2)(\kappa^2-\gamma^2)} - 1 \right] \quad (3.12)$$

This optimization problem can be solved numerically. Note that even if our optimization routine did not find the triple (α, β, ω) that minimizes ω , as long as the third and fourth moment conditions are satisfied, this Rao-Blackwellization approach is valid, it just may take more samples to achieve the same precision on $\mathbb{P}(y^T My < a)$.

3.3 Discussion

We have presented a computationally efficient method for obtaining exact bounds on the CDF of quadratic forms in Gaussian random variables. We have also shown how this result can be extended to apply to quadratic forms in non-Gaussian random variables when the matrix of interest is hollow or when the random variables are centered. It is critical to note that the invariance principles that we've provided are robust in practice: the convergence can be expected to be tight even if the inlier-outlier ratio, ρ_1 or ρ_2 , grows logarithmically in n .

Together, these results allow for the efficient screening of quadratic forms in a wide variety of settings. Outside of genetics, these include screening for predictive random projections in high-dimensional classification and even trend detection on large social networks. Other screening applications include selecting good initialization points for the kernel parameters in Gaussian Process Regression. We will present these applications and related results in a forthcoming publication.

3.4 Proofs

3.4.1 Proof of Lemma 3.1.3

Lemma. *Let $Z \sim N(0, 1)$. Then, for all $t \in (0, \frac{1}{2\sigma^2})$,*

$$E\left(e^{t(\sigma Z + \mu)^2}\right) = (1 - 2\sigma^2 t)^{-1/2} \exp\left\{\frac{\mu^2 t}{1 - 2\sigma^2 t}\right\}$$

Proof:

$$E\left(e^{t(\sigma Z + \mu)^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{t(y+\mu)^2} e^{-\frac{y^2}{2\sigma^2}} dy = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\left(\frac{1}{2\sigma^2} - t\right)y^2 + 2\mu ty + \mu^2 t} dy$$

We can complete the square using $-ax^2 + 2bx = -a\left(x - \frac{b}{a}\right)^2 + \frac{b^2}{a}$. Also noting that $\left(\frac{1}{2\sigma^2} - t\right) = \frac{1}{2\sigma^2(1-2\sigma^2 t)^{-1}}$ yields

$$\exp\left\{\frac{\mu^2 t}{1 - 2\sigma^2 t}\right\} \sqrt{(1 - 2\sigma^2 t)^{-1}} \frac{1}{\sqrt{2\pi\sigma^2(1 - 2\sigma^2 t)^{-1}}} \int_{-\infty}^{\infty} \exp\left\{-\frac{\left(y - \frac{2\sigma^2 \mu t}{1 - 2\sigma^2 t}\right)^2}{2\sigma^2(1 - 2\sigma^2 t)^{-1}}\right\} dy.$$

Thus, for $t < \frac{1}{2\sigma^2}$, we have

$$E\left(e^{t(\sigma Z + \mu)^2}\right) = (1 - 2\sigma^2 t)^{-1/2} \exp\left\{\frac{\mu^2 t}{1 - 2\sigma^2 t}\right\}.$$

□

3.4.2 Proof of Theorem 3.4.2

Theorem. Let $X \sim N(\mu, I)$ and M be a real, symmetric matrix. Let $v = 2 \left(4 \|M\mu\|_2^2 + 2 \|M\|_{HS}^2 \right)$. Then, for all $q > \mathbb{E}[X^T M X]$,

$$\mathbb{P}(X^T M X > q) \leq \begin{cases} \exp\left(-\frac{1}{2} \frac{(q - \mathbb{E}[X^T M X])^2}{v}\right) & \mathbb{E}[X^T M X] < q \leq \frac{v}{4b} + \mathbb{E}[X^T M X] \\ \exp\left(\frac{1}{2} \frac{v}{(4b)^2}\right) \exp\left(-\frac{q - \mathbb{E}[X^T M X]}{4b}\right) & q > \frac{v}{4b} + \mathbb{E}[X^T M X]. \end{cases}$$

Furthermore, for all $q < \mathbb{E}[X^T M X]$,

$$\mathbb{P}(X^T M X < q) \leq \begin{cases} \exp\left(-\frac{1}{2} \frac{(\mathbb{E}[X^T M X] - q)^2}{v}\right) & \mathbb{E}[X^T M X] - \frac{v}{4b} \leq q < \mathbb{E}[X^T M X] \\ \exp\left(\frac{1}{2} \frac{v}{(4b)^2}\right) \exp\left(-\frac{\mathbb{E}[X^T M X] - q}{4b}\right) & q < \mathbb{E}[X^T M X] - \frac{v}{4b}. \end{cases}$$

Proof:

We credit David Pollard for the proof approach used below (reference: <http://www.stat.yale.edu/pollard/Books/Mini/Basic.pdf>).

$$X^T M X = (Z + \mu)^T M (Z + \mu) = \sum_{j=1}^n \lambda_j (v_j^T Z + v_j^T \mu)^2 \sim \sum_{j=1}^n \lambda_j (Z_j + v_j^T \mu)^2$$

Where $Z_j \sim N(0, 1)$. The last step relies on the spherical symmetry of the standard multivariate normal distribution. Using Lemma 3.1.3, for $t < \frac{1}{2 \max_j |\lambda_j|}$,

$$\begin{aligned} \mathbb{E}[e^{t X^T M X}] &= \prod_{j=1}^n (1 - 2t \lambda_j)^{-1/2} \exp\left(\frac{(v_j^T \mu)^2 t \lambda_j}{1 - 2t \lambda_j}\right) \\ &= \exp\left(\sum_{j=1}^n (v_j^T \mu)^2 \frac{t \lambda_j}{1 - 2t \lambda_j} - \frac{1}{2} \log(1 - 2t \lambda_j)\right). \end{aligned}$$

Let us further restrict t such that $0 < t \leq \frac{1}{4b}$. Since $\frac{1}{4b} \leq \frac{1}{4\max_j |\lambda_j|}$, we have that $-\frac{1}{4} \leq \lambda_j t \leq \frac{1}{4} \forall j$. For $x \in [-\frac{1}{4}, \frac{1}{4}]$,

$$-\frac{1}{2} \log(1 - 2x) \leq 2x^2 + x,$$

$$\frac{x}{1 - 2x} \leq 4x^2 + x.$$

Therefore,

$$\begin{aligned} \mathbb{E} \left[e^{tX^T M X} \right] &\leq \exp \left(\sum_{j=1}^n (v_j^T \mu)^2 \left(4(\lambda_j t)^2 + \lambda_j t \right) + 2(\lambda_j t)^2 + \lambda_j t \right) \\ &\leq \exp \left(4t^2 \sum_{j=1}^n \lambda_j^2 (v_j^T \mu)^2 + t \sum_{j=1}^n \lambda_j (v_j^T \mu)^2 + 2t^2 \sum_{j=1}^n \lambda_j^2 + t \sum_{j=1}^n \lambda_j \right) \\ &\leq \exp \left(4t^2 \|M\mu\|_2^2 + t \mu^T M \mu + 2t^2 \|M\|_{HS}^2 + t \operatorname{tr}(M) \right) \end{aligned}$$

To simplify our notation, let $\nu = 2 \left(4\|M\mu\|_2^2 + 2t^2 \|M\|_{HS}^2 \right)$ and $\psi = \mathbb{E}(X^T M X) = \mu^T M \mu + \operatorname{tr}(M)$. With these definitions we have

$$\mathbb{E} \left[e^{t(X^T M X - \psi)} \right] \leq \exp \left(\frac{\nu}{2} t^2 \right).$$

By the Markov Inequality, for any $q \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(X^T M X > q) &= \mathbb{P}(X^T M X - \psi > q - \psi) = \mathbb{P}(e^{X^T M X - \psi} > e^{q - \psi}) \\ &\leq \exp \left(\frac{\nu}{2} t^2 - (q - \psi)t \right) \quad \forall t \in \left[0, \frac{1}{4b} \right] \end{aligned}$$

For $q \leq \psi$ we have the trivial bound $\mathbb{P}(X^T M X > q) \leq 1$. We can optimize our bound in t to find

$$t_\star = \min \left[\frac{q - \psi}{\nu}, \frac{1}{4b} \right].$$

Thus, for $\psi < q \leq \frac{\nu}{4b} + \psi$, we have

$$\exp\left(\frac{\nu}{2}t_{\star}^2 - (q - \psi)t_{\star}\right) = \exp\left(\frac{(q - \psi)^2}{2\nu} - \frac{(q - \psi)^2}{\nu}\right) = \exp\left(-\frac{1}{2}\frac{(q - \psi)^2}{\nu}\right),$$

a Gaussian bound. For $q > \frac{\nu}{4b} + \psi$, we have

$$\exp\left(\frac{\nu}{2}t_{\star}^2 - (q - \psi)t_{\star}\right) = \exp\left(\frac{1}{2}\frac{\nu}{(4b)^2}\right) \exp\left(-\frac{(q - \psi)}{(4b)}\right),$$

an exponential bound. Combining these results yields our upper bound on $\mathbb{P}(X^T M X > q)$ for $q \geq \psi = \mathbb{E}(X^T M X)$. Since $\mathbb{P}(X^T M X < q) = \mathbb{P}(X^T (-M) X > -q)$, repeating the proof for $-M$ and $-q$ yields the upper bound on the lower tail, $\mathbb{P}(X^T M X < q)$, for $q \leq \psi = \mathbb{E}(X^T M X)$. $\square$

3.4.3 Proof of Theorem 3.2.1

Theorem. Let M be a hollow, real matrix and ϕ be a real function with a bounded third derivative: $|\phi'''| \leq K_3$. For two random vectors with independent entries, y, x , let the first four moments of each y_i and each x_i exist. If the first two moments of x_i match the first two moments of y_i for all i , then

$$\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right| \leq \frac{1}{\sqrt{n}} \frac{\psi K_3}{6} \frac{27^{1/4} \max(\tau^3, 1)}{\sigma_{\star}^3 \min \left(\min_j \left(\left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right)^{3/2} \right), 1 \right)} \rho_1^3. \quad (3.13)$$

where

$$\begin{aligned}\sigma_*^2 &= \min_j \sigma_j^2, \\ \tau &\geq \max \left(\max_j (||x_j||_4), \max_j (||y_j||_4) \right), \\ \psi &\geq \max_j \left[\mathbb{E} [|x_j|^3] + \mathbb{E} [|y_j|^3] \right], \\ \rho_1 &= \frac{\max_j [||M_j||_2 + |M_j^T \mu|]}{\min_j \left[\sqrt{||M_j||_2^2 + (M_j^T \mu)^2} \right]}.\end{aligned}$$

Proof:

By Lemma 1.5.1, $v_x = v_y$, so we can drop the subscript and use v to denote the variance of either quadratic form. Since $Q_y^{(n)} = \frac{1}{v} y^T M y = \frac{1}{2v} y^T (M + M^T) y$, we can assume that M is symmetric without loss of generality. . We will proceed using a standard variable swapping approach reminiscent of Gibbs-sampling.

$$\text{Let } \Delta_j^{(n)} = 2 \left[\sum_{i=1}^{j-1} m_{ij} y_i + \sum_{i=j+1}^n m_{ij} x_i \right].$$

$$S_j^{(n)} = \frac{1}{v} \begin{pmatrix} y_1 \\ \vdots \\ y_{j-1} \\ 0 \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix}^T M \begin{pmatrix} y_1 \\ \vdots \\ y_{j-1} \\ 0 \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix}$$

Under these definitions we not only have that $Q_x^{(n)} = S_1^{(n)} + v^{-1} x_1 \Delta_1^{(n)}$ and $Q_y^{(n)} = S_n^{(n)} + v^{-1} y_n \Delta_n^{(n)}$, but also that

$$S_j^{(n)} + \mathbf{v}^{-1} y_j \Delta_j^{(n)} = S_{j+1}^{(n)} + \mathbf{v}^{-1} x_{j+1} \Delta_{j+1}^{(n)} \quad \forall j \in 1, \dots, n$$

This allows us to rewrite $\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right|$ as a telescoping sum:

$$\left| \mathbb{E} \left[\sum_{j=1}^n \left[\phi \left(S_j^{(n)} + \mathbf{v}^{-1} x_j \Delta_j^{(n)} \right) - \phi \left(S_j^{(n)} + \mathbf{v}^{-1} y_j \Delta_j^{(n)} \right) \right] \right] \right|$$

Taking the expectation through the sum and Taylor expanding ϕ around $S_j^{(n)}$, we can rewrite $\mathbb{E} \left[\phi \left(S_j^{(n)} + \mathbf{v}^{-1} x_j \Delta_j^{(n)} \right) \right]$ as

$$\begin{aligned} \mathbb{E} \left[\phi \left(S_j^{(n)} \right) \right] &+ \frac{1}{\mathbf{v}} \mathbb{E} \left[\phi' \left(S_j^{(n)} \right) x_j \Delta_j^{(n)} \right] \\ &+ \frac{1}{2\mathbf{v}^2} \mathbb{E} \left[\phi'' \left(S_j^{(n)} \right) \left(x_j \Delta_j^{(n)} \right)^2 \right] + \frac{1}{6\mathbf{v}^3} \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)} \right) \left(x_j \Delta_j^{(n)} \right)^3 \right] \end{aligned}$$

for some function of x_j , $\hat{S}_j(x_j) \in [S_j^{(n)}, S_j^{(n)} + \mathbf{v}^{-1} x_j \Delta_j^{(n)}]$. Using independence, we have

$$\begin{aligned} \mathbb{E} \left[\phi \left(S_j^{(n)} + \mathbf{v}^{-1} x_j \Delta_j^{(n)} \right) \right] &= \\ \mathbb{E} \left[\phi \left(S_j^{(n)} \right) \right] &+ \frac{1}{\mathbf{v}} \mathbb{E} [x_j] \mathbb{E} \left[\phi' \left(S_j^{(n)} \right) \Delta_j^{(n)} \right] + \frac{1}{2\mathbf{v}^2} \mathbb{E} [x_j^2] \mathbb{E} \left[\phi'' \left(S_j^{(n)} \right) \left(\Delta_j^{(n)} \right)^2 \right] \\ &+ \frac{1}{6\mathbf{v}^3} \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)}(x_j) \right) \left(x_j \Delta_j^{(n)} \right)^3 \right] \end{aligned}$$

Note that we can similarly expand $\mathbb{E} \left[\phi \left(S_j^{(n)} + \mathbf{v}^{-1} y_j \Delta_j^{(n)} \right) \right]$ with $\hat{S}_j(y_j) \in [S_j^{(n)}, S_j^{(n)} + \mathbf{v}^{-1} y_j \Delta_j^{(n)}]$ in the third order term. Since the first and second moments of x_i match those of y_i , we can cancel all first and second order terms of each expansion. Using Jensen's Inequality and

independence, we have,

$$\begin{aligned}
& \left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right| = \\
& \frac{1}{6v^3} \left| \sum_{j=1}^n \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)}(x_j) \right) \left(x_j \Delta_j^{(n)} \right)^3 \right] - \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)}(y_j) \right) \left(y_j \Delta_j^{(n)} \right)^3 \right] \right| \\
& \leq \frac{1}{6v^3} \sum_{j=1}^n \mathbb{E} \left[\left| \phi''' \left(\hat{S}_j^{(n)}(x_j) \right) \right| |x_j|^3 \left| \Delta_j^{(n)} \right|^3 \right] + \mathbb{E} \left[\left| \phi''' \left(\hat{S}_j^{(n)}(y_j) \right) \right| |y_j|^3 \left| \Delta_j^{(n)} \right|^3 \right] \\
& \leq \frac{K_3}{6v^3} \sum_{j=1}^n \mathbb{E} \left[|x_j|^3 \left| \Delta_j^{(n)} \right|^3 \right] + \mathbb{E} \left[|y_j|^3 \left| \Delta_j^{(n)} \right|^3 \right] \\
& \leq \frac{K_3}{6v^3} \sum_{j=1}^n \mathbb{E} \left[|x_j|^3 \right] \mathbb{E} \left[\left| \Delta_j^{(n)} \right|^3 \right] + \mathbb{E} \left[|y_j|^3 \right] \mathbb{E} \left[\left| \Delta_j^{(n)} \right|^3 \right] \\
& \leq \frac{\psi K_3}{6v^3} \sum_{j=1}^n \left\| \Delta_j^{(n)} \right\|_4^3. \tag{3.14}
\end{aligned}$$

In order to evaluate $\left\| \Delta_j^{(n)} \right\|_4$, let M_i denote the i^{th} column of M and M_i^2 denote the i^{th} column of M squared elementwise. Let $\tilde{y} = y - \mu$ and $\tilde{x} = x - \mu$. Let Z and $\tilde{Z}$ both be $n \times n$ matrices such that the j^{th} column,

$$Z_j = \begin{pmatrix} y_1 \\ \vdots \\ y_{j-1} \\ 0 \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad \tilde{Z}_j = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_{j-1} \\ 0 \\ \tilde{x}_{j+1} \\ \vdots \\ \tilde{x}_n \end{pmatrix}.$$

$$\left\| \Delta_j^{(n)} \right\|_4^4 = 16 \mathbb{E} \left[\left(M_j^T Z_j \right)^4 \right] = 16 \mathbb{E} \left[\left(M_j^T \tilde{Z}_j + M_j^T \mu \right)^4 \right]$$

Expanding yields

$$\begin{aligned}\mathbb{E} \left[(M_j^T \tilde{Z}_j + M_j^T \mu)^4 \right] &= \mathbb{E} \left[(M_j^T \tilde{Z}_j)^4 \right] + 4\mathbb{E} \left[(M_j^T \tilde{Z}_j)^3 \right] (M_j^T \mu) \\ &\quad + 6\mathbb{E} \left[(M_j^T \tilde{Z}_j)^2 \right] (M_j^T \mu)^2 + \mathbb{E} \left[(M_j^T \tilde{Z}_j) \right] (M_j^T \mu)^3 + (M_j^T \mu)^4.\end{aligned}$$

Since $\mathbb{E} [\tilde{z}_{ij}] = 0 \forall i, j$, all terms including a monomial in the above expectations are zero, this allows us to simplify this to

$$\mathbb{E} \left[(M_j^T \tilde{Z}_j)^4 \right] + 4\mathbb{E} \left[(M_j^3)^T \tilde{Z}_j^3 \right] (M_j^T \mu) + 6\mathbb{E} \left[(M_j^2)^T \tilde{Z}_j^2 \right] (M_j^T \mu)^2 + (M_j^T \mu)^4$$

Here when we write M_j^k we mean taking the elements of M_j to the k th power elementwise. Note again that here we have already removed all monomial terms since they evaluate to zero.

To evaluate the first term, we again use the fact that all terms with monomials will be zero.

$$\begin{aligned}\mathbb{E} \left[\left(\sum_{i=1}^{j-1} m_{ij} \tilde{y}_i + \sum_{i=j+1}^n m_{ij} \tilde{x}_i \right)^4 \right] &= \sum_{i=1}^{j-1} m_{ij}^4 \mathbb{E} (\tilde{y}_i^4) + \sum_{i=j+1}^n m_{ij}^4 \mathbb{E} (\tilde{x}_i^4) \\ &\quad + 6 \sum_{i=1}^{j-1} \sum_{\substack{1 \leq k \leq j-1 \\ k \neq i}} m_{ij}^2 m_{kj}^2 \mathbb{E} (\tilde{y}_i^2) \mathbb{E} (\tilde{y}_k^2) + 6 \sum_{i=1}^{j-1} \sum_{k=j+1}^n m_{ij}^2 m_{kj}^2 \mathbb{E} (\tilde{y}_i^2) \mathbb{E} (\tilde{x}_k^2) \\ &\quad + 6 \sum_{i=j+1}^n \sum_{k=1}^{j-1} m_{ij}^2 m_{kj}^2 \mathbb{E} (\tilde{x}_i^2) \mathbb{E} (\tilde{y}_k^2) + 6 \sum_{i=j+1}^n \sum_{\substack{j+1 \leq k \leq n \\ k \neq i}} m_{ij}^2 m_{kj}^2 \mathbb{E} (\tilde{x}_i^2) \mathbb{E} (\tilde{x}_k^2)\end{aligned}$$

Recalling that $\tau \geq \max \left(\max_j (\|x_j\|_4), \max_j (\|y_j\|_4) \right)$, which also makes $\tau^2 \geq \sigma_i^2 \forall i$, we have

$$\mathbb{E} \left[\left(\sum_{i=1}^{j-1} m_{ij} \tilde{y}_i + \sum_{i=j+1}^n m_{ij} \tilde{x}_i \right)^4 \right] \leq 3\tau^4 \left(\sum_{i=1}^n m_{ij}^2 \right)^2 = 3\tau^4 \|M_j\|_2^4$$

To evaluate the second term, we recall that $\left| \mathbb{E}(\tilde{y}_j^3) \right| \leq \mathbb{E}(|\tilde{y}_j|^3) \leq \tau^3$.

$$\begin{aligned} 4\mathbb{E}[(M_j^3)^T \tilde{Z}_j^3] (M_j^T \mu) &\leq 4(|M_j|^3)^T \begin{pmatrix} |\mathbb{E}(\tilde{y}_1^3)| \\ \vdots \\ |\mathbb{E}(\tilde{y}_{j-1}^3)| \\ 0 \\ |\mathbb{E}(\tilde{x}_{j+1}^3)| \\ \vdots \\ |\mathbb{E}(\tilde{x}_n^3)| \end{pmatrix} |M_j^T \mu| \\ &\leq 4\tau^3 \|M_j\|_3^3 |M_j^T \mu| \\ &\leq 4\tau^3 \|M_j\|_2^3 |M_j^T \mu|. \end{aligned}$$

We have a similar bound for the third term using $\mathbb{E}(\tilde{y}_j^2) \leq \tau^2$.

$$\begin{aligned} 6\mathbb{E}[(M_j^2)^T \tilde{Z}_j^2] (M_j^T \mu)^2 &\leq 6(M_j^2)^T \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_{j-1}^2 \\ 0 \\ \sigma_{j+1}^2 \\ \vdots \\ \sigma_n^2 \end{pmatrix} (M_j^T \mu)^2 \\ &\leq 6\tau^2 \|M_j\|_2^2 (M_j^T \mu)^2 \end{aligned}$$

Putting the terms back together yields

$$\begin{aligned}\mathbb{E} \left[(M_j^T \tilde{Z}_j + M_j^T \mu)^4 \right] &\leq 3\tau^4 \|M_j\|_2^4 + 4\tau^3 \|M_j\|_2^3 |M_j^T \mu| + 6\tau^2 \|M_j\|_2^2 (M_j^T \mu)^2 + (M_j^T \mu)^4 \\ &\leq 3 \max(\tau^4, 1) (\|M_j\|_2 + |M_j^T \mu|)^4.\end{aligned}$$

This in turn implies that

$$\left\| \Delta_j^{(n)} \right\|_4^4 \leq 48 \max(\tau^4, 1) (\|M_j\|_2 + |M_j^T \mu|)^4.$$

Using our Lemma 1.5.1 and that $M_{ii} = 0 \forall i$,

$$\begin{aligned}v^2 &\geq 2 \sigma^{2T} (M * M) \sigma^2 + 4 (\mu^2)^T (M * M) \sigma^2 + 4 (M \mu)^T \text{Diag}(\sigma^2) (M \mu) \\ &= \sum_{i=1}^n \sigma_i^2 \left(4 (M_i^2)^T \left(\frac{1}{2} \sigma^2 + \mu^2 \right) + 4 (M_i^T \mu)^2 \right) \\ &\geq 4 \sigma_*^2 \sum_{i=1}^n (M_i^2)^T \left(\frac{1}{2} \sigma^2 + \mu^2 \right) + (M_i^T \mu)^2 \\ &\geq 4 \sigma_*^2 \sum_{i=1}^n \min_j \left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right) \|M_i\|_2^2 + (M_i^T \mu)^2 \\ &\geq 4 \sigma_*^2 \min \left(\min_j \left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right), 1 \right) \sum_{i=1}^n \|M_i\|_2^2 + (M_i^T \mu)^2 \\ &\geq 4n \sigma_*^2 \min \left(\min_j \left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right), 1 \right) \min_i (\|M_i\|_2^2 + (M_i^T \mu)^2)\end{aligned}$$

Applying these bounds we have,

$$\begin{aligned}
v^{-3} \sum_{j=1}^n \left\| \Delta_j^{(n)} \right\|_4^3 &\leq v^{-3} \sum_{j=1}^n \left(48 \max(\tau^4, 1) (\|M_j\|_2 + |M_j^T \mu|)^4 \right)^{3/4} \\
&\leq 48^{3/4} \max(\tau^3, 1) v^{-3} \sum_{j=1}^n (\|M_j\|_2 + |M_j^T \mu|)^3 \\
&\leq 3^{3/4} 8 \max(\tau^3, 1) v^{-3} n \max_j \left((\|M_j\|_2 + |M_j^T \mu|)^3 \right) \\
&\leq \frac{3^{3/4} 8 n \max(\tau^3, 1) \max_j \left((\|M_j\|_2 + |M_j^T \mu|)^3 \right)}{\left(4n \sigma_*^2 \min \left(\min_j \left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right), 1 \right) \min_i (\|M_i\|_2^2 + (M_i^T \mu)^2) \right)^{3/2}} \\
&\leq \frac{1}{\sqrt{n}} \frac{27^{1/4} \max(\tau^3, 1)}{\sigma_*^3 \min \left(\min_j \left(\left(\frac{1}{2} \sigma_j^2 + \mu_j^2 \right)^{3/2} \right), 1 \right)} \left(\frac{\max_j (\|M_j\|_2 + |M_j^T \mu|)}{\min_i (\sqrt{\|M_i\|_2^2 + (M_i^T \mu)^2})} \right)^3.
\end{aligned}$$

Substituting this into our bound on $\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right|$ (Equation 3.14) completes the proof.

3.4.4 Proof of Lemma 3.2.2

Lemma. Consider two random variables X and Y . Let Y be continuous and its density be upper bounded by a constant U . If there exists a constant $C \in \mathbb{R}$ such that $|\mathbb{E}(\phi_{a,\varepsilon}(X)) - \mathbb{E}(\phi_{a,\varepsilon}(Y))| \leq K_3 C \forall a \in \mathbf{R}$ and $\forall \varepsilon > 0$, then

$$|\mathbb{P}(X \leq a) - \mathbb{P}(Y \leq a)| \leq \frac{4}{3} U \left(\frac{3\pi^3 C}{2} \right)^{1/4}$$

Proof:

Since $\frac{1}{2} \left(\cos \left(\frac{\pi}{\varepsilon} (x - a) \right) + 1 \right) \in [0, 1]$ for $x \in (a, a + \varepsilon]$, we have

$$\mathbb{P}(X \leq a) \leq \mathbb{E}(\phi_{a,\varepsilon}(X)) \leq \mathbb{E}(\phi_{a,\varepsilon}(Y)) + K_3 C \leq \mathbb{P}(Y \leq a) + U\varepsilon + K_3 C.$$

Going the other direction,

$$\mathbb{P}(X \leq a) \geq \mathbb{E}(\phi_{a-\varepsilon, \varepsilon}(X)) \geq \mathbb{E}(\phi_{a-\varepsilon, \varepsilon}(Y)) - K_3 C \geq \mathbb{P}(Y \leq a) - U\varepsilon - K_3 C.$$

Therefore,

$$|\mathbb{P}(X \leq a) - \mathbb{P}(Y \leq a)| \leq U\varepsilon + K_3 C = U\varepsilon + \frac{\pi^3 C}{2} \varepsilon^{-3} \quad (3.15)$$

Optimizing our bound in ε ,

$$\frac{d}{d\varepsilon} \left(U\varepsilon + \frac{\pi^3 C}{2} \varepsilon^{-3} \right) = U - \frac{3\pi^3 C}{2} \varepsilon^{-4} = 0,$$

we find that we should set ε to $\left(\frac{3\pi^3 C}{2U} \right)^{1/4}$. Thus,

$$\begin{aligned} |\mathbb{P}(X \leq a) - \mathbb{P}(Y \leq a)| &\leq U \left(\frac{3\pi^3 C}{2U} \right)^{1/4} + \frac{\pi^3 C}{2} \left(\frac{3\pi^3 C}{2U} \right)^{-3/4} \\ &\leq \left(\frac{3\pi^3 C}{2U} \right)^{1/4} \left(U + \left(\frac{3}{U} \right)^{-1} \right) \\ &\leq \frac{4}{3} U \left(\frac{3\pi^3 C}{2} \right)^{1/4}. \end{aligned}$$

□

3.4.5 Proof of Theorem 3.2.6

Theorem. Let M be a real matrix and ϕ be a real function with a bounded third derivative: $|\phi'''| \leq K_3$. For two random vectors with independent entries, y, x , let the first eight moments of each y_i and each x_i exist. If $\mathbb{E}(y_i) = \mathbb{E}(x_i) = 0 \forall i$ and the first four moments of x_i match the first four moments of y_i for all i , then

$$\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right| \leq \frac{1}{\sqrt{n}} \frac{K_3 \psi \max \left(8 \left(3^{3/4} \right) \tau^3, 1 \right)}{3 \left(\min(\kappa_*, 2\sigma_*^4) \right)^{3/2}} \rho_2^3,$$

where

$$\begin{aligned} \sigma_*^2 &= \min_j \sigma_j^2 \\ \kappa_j &= \text{Var}(y_j^2) = \text{Var}(x_j^2) \\ \kappa_* &= \min_j \kappa_j \\ \tau &\geq \max_j \left(\max(|x_j|_4, |y_j|_4) \right) \\ \psi &\geq \max_j \left(\max \left(\sqrt{\mathbb{E}[x_j^8]}, \mathbb{E}[x_j^8], \sqrt{\mathbb{E}[y_j^8]}, \mathbb{E}[y_j^8] \right) \right) \\ \rho_2 &= \frac{\max_j [||\tilde{M}_j||_2 + |m_{jj}|]}{\min_j \left[\sqrt{||\tilde{M}_j||_2^2 + m_{jj}^2} \right]} \end{aligned}$$

Proof: By Lemma 1.5.1, $v_x = v_y$, so we can drop the subscript and use v to denote the variance of either quadratic form. Since $Q_y^{(n)} = \frac{1}{v} y^T M y = \frac{1}{2v} y^T (M + M^T) y$, we can assume that M is symmetric without loss of generality. Let $\tilde{M} = (M - \text{diag}(M))$ so that $\tilde{m}_{jj} = 0 \forall j$. Let us temporarily consider a more general three-times differentiable function ϕ with bounded derivatives up to fourth order. Let $K_3 = \sup_{x \in \mathbb{R}} |\phi'''(x)|$. We will proceed using the same variable

swapping approach as in the previous proof. Define $\Delta_j^{(n)} = 2 \left[\sum_{i=1}^{j-1} m_{ij} y_i + \sum_{i=j+1}^n m_{ij} x_i \right]$,

$$S_j^{(n)} = \frac{1}{\mathbf{v}} \begin{pmatrix} y_1 \\ \vdots \\ y_{j-1} \\ 0 \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix}^T M \begin{pmatrix} y_1 \\ \vdots \\ y_{j-1} \\ 0 \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix}.$$

Under these definitions we have

$$\begin{aligned} S_j^{(n)} + \mathbf{v}^{-1} \left(y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right) &= S_{j+1}^{(n)} + \mathbf{v}^{-1} \left(x_{j+1} \Delta_{j+1}^{(n)} + x_{j+1}^2 m_{j+1,j+1} \right), \\ Q_x^{(n)} &= S_1^{(n)} + \mathbf{v}^{-1} \left(x_1 \Delta_1^{(n)} + x_1^2 m_{11} \right), \\ Q_y^{(n)} &= S_n^{(n)} + \mathbf{v}^{-1} \left(y_n \Delta_n^{(n)} + y_n^2 m_{nn} \right). \end{aligned}$$

This allows us to rewrite $\left| \mathbb{E} \left[\phi \left(Q_x^{(n)} \right) - \phi \left(Q_y^{(n)} \right) \right] \right|$ as a telescoping sum:

$$\left| \mathbb{E} \left[\sum_{j=1}^n \left[\phi \left(S_j^{(n)} + \mathbf{v}^{-1} \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right) \right) - \phi \left(S_j^{(n)} + \mathbf{v}^{-1} \left(y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right) \right) \right] \right] \right|.$$

Taking the expectation through the sum and Taylor expanding ϕ around $S_j^{(n)}$, we can rewrite $\mathbb{E} \left[\phi \left(S_j^{(n)} + \mathbf{v}^{-1} \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right) \right) \right]$ as

$$\begin{aligned} \mathbb{E} \left[\phi \left(S_j^{(n)} \right) \right] + \mathbb{E} \left[\phi' \left(S_j^{(n)} \right) v^{-1} \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right) \right] \\ + \frac{1}{2} \mathbb{E} \left[\phi'' \left(S_j^{(n)} \right) v^{-2} \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right)^2 \right] \\ + \frac{1}{6} \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)} \right) v^{-3} \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right)^3 \right] \end{aligned}$$

for some function of x_j , $\hat{S}_j(x_j) \in \left[S_j^{(n)}, S_j^{(n)} + v^{-1} \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right) \right]$. Using independence, we have

$$\begin{aligned} \mathbb{E} \left[\phi \left(S_j^{(n)} \right) \right] + v^{-1} \mathbb{E} [x_j] \mathbb{E} \left[\phi' \left(S_j^{(n)} \right) \Delta_j^{(n)} \right] + v^{-1} \mathbb{E} [x_j^2] m_{jj} \mathbb{E} \left[\phi' \left(S_j^{(n)} \right) \right] \\ + \frac{1}{2} v^{-2} \mathbb{E} [x_j^2] \mathbb{E} \left[\phi'' \left(S_j^{(n)} \right) \left(\Delta_j^{(n)} \right)^2 \right] + v^{-2} \mathbb{E} [x_j^3] m_{jj} \mathbb{E} \left[\phi'' \left(S_j^{(n)} \right) \Delta_j^{(n)} \right] \\ + \frac{1}{2} v^{-2} \mathbb{E} [x_j^4] m_{jj}^2 \mathbb{E} \left[\phi'' \left(S_j^{(n)} \right) \right] + \frac{1}{6} v^{-3} \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)}(x_j) \right) \left(x_j \Delta_j^{(n)} \right)^3 \right]. \end{aligned}$$

Note that we can similarly expand $\mathbb{E} \left[\phi \left(S_j^{(n)} + v^{-1} \left(y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right) \right) \right]$ with $\hat{S}_j(y_j) \in \left[S_j^{(n)}, S_j^{(n)} + v^{-1} \left(y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right) \right]$ in the third order term. Since the first, second, third and fourth moments of x_i match those of y_i , we can cancel all first and second order terms of each expansion and apply Jensen's Inequality to yield

$$\begin{aligned} \frac{1}{6v^3} \left| \sum_{j=1}^n \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)}(x_j) \right) \left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right)^3 \right] - \mathbb{E} \left[\phi''' \left(\hat{S}_j^{(n)}(y_j) \right) \left(y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right)^3 \right] \right| \\ \leq \frac{1}{6v^3} \sum_{j=1}^n \mathbb{E} \left[\left| \phi''' \left(\hat{S}_j^{(n)}(x_j) \right) \right| \left| x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right|^3 \right] + \mathbb{E} \left[\left| \phi''' \left(\hat{S}_j^{(n)}(y_j) \right) \right| \left| y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right|^3 \right] \\ \leq \frac{K_3}{6v^3} \sum_{j=1}^n \left(\mathbb{E} \left[\left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right)^4 \right] \right)^{3/4} + \left(\mathbb{E} \left[\left(y_j \Delta_j^{(n)} + y_j^2 m_{jj} \right)^4 \right] \right)^{3/4}. \end{aligned}$$

Expanding the first term, we can use independence and the fact that $\mathbb{E} [\Delta_j^{(n)}] = 0$ to yield

$$\begin{aligned}
\mathbb{E} \left[\left(x_j \Delta_j^{(n)} + x_j^2 m_{jj} \right)^4 \right] &= \mathbb{E} [x_j^4] \mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right] \\
&\quad + 4m_{jj} \mathbb{E} [x_j^5] \mathbb{E} \left[\left(\Delta_j^{(n)} \right)^3 \right] + 6m_{jj}^2 \mathbb{E} [x_j^6] \mathbb{E} \left[\left(\Delta_j^{(n)} \right)^2 \right] \\
&\quad + 4m_{jj}^3 \mathbb{E} [x_j^7] \mathbb{E} [\Delta_j^{(n)}] + m_{jj}^4 \mathbb{E} [x_j^8] \\
&\leq \sqrt{\mathbb{E} [x_j^8]} \mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right] + 4|m_{jj}| \mathbb{E} [x_j^5] \mathbb{E} \left[\left| \Delta_j^{(n)} \right|^3 \right] \\
&\quad + 6m_{jj}^2 (\mathbb{E} [x_j^8])^{3/4} \sqrt{\mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right]} + m_{jj}^4 \mathbb{E} [x_j^8] \\
&\leq \sqrt{\mathbb{E} [x_j^8]} \mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right] + 4|m_{jj}| (\mathbb{E} [x_j^8])^{5/8} \mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right]^{3/4} \\
&\quad + 6m_{jj}^2 (\mathbb{E} [x_j^8])^{3/4} \sqrt{\mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right]} + m_{jj}^4 \mathbb{E} [x_j^8] \\
&\leq \max \left(\sqrt{\mathbb{E} [x_j^8]}, \mathbb{E} [x_j^8] \right) \left(\mathbb{E} \left[\left(\Delta_j^{(n)} \right)^4 \right]^{1/4} + |m_{jj}| \right)^4 \\
&= \max \left(\sqrt{\mathbb{E} [x_j^8]}, \mathbb{E} [x_j^8] \right) \left(\left\| \Delta_j^{(n)} \right\|_4 + |m_{jj}| \right)^4.
\end{aligned}$$

This gives us

$$\left| \mathbb{E} \left[\phi \left(\mathcal{Q}_x^{(n)} \right) - \phi \left(\mathcal{Q}_y^{(n)} \right) \right] \right| \leq \frac{K_3 \psi}{3v^3} \sum_{j=1}^n \left(\left\| \Delta_j^{(n)} \right\|_4 + |m_{jj}| \right)^3.$$

As shown in the previous proof, we have

$$\left\| \Delta_j^{(n)} \right\|_4^4 = 16 \mathbb{E} \left[\left(\sum_{i=1}^{j-1} m_{ij} y_i + \sum_{i=j+1}^n m_{ij} x_i \right)^4 \right] \leq 48 \tau^4 \|\tilde{M}_j\|_2^4.$$

Thus, $\left\| \Delta_j^{(n)} \right\|_4 \leq 2 \left(3^{1/4} \right) \tau \|\tilde{M}_j\|_2$, implying that

$$\begin{aligned} \left| \mathbb{E} \left[\phi \left(\mathcal{Q}_x^{(n)} \right) - \phi \left(\mathcal{Q}_y^{(n)} \right) \right] \right| &\leq \frac{K_3 \psi}{3\nu^3} \sum_{j=1}^n \left(2 \left(3^{1/4} \right) \tau \|\tilde{M}_j\|_2 + |m_{jj}| \right)^3 \\ &\leq n \frac{K_3 \psi \max \left(8 \left(3^{3/4} \right) \tau^3, 1 \right)}{3\nu^3} \max_j \left(\|\tilde{M}_j\|_2 + |m_{jj}| \right)^3. \end{aligned}$$

Using Lemma 1.5.1,

$$\begin{aligned} \nu^2 &\geq \sum_{i=1}^n \left(\kappa_i m_{ii}^2 + 2\sigma_i^2 \left(\sum_{j=1}^n \tilde{m}_{ij}^2 \sigma_j^2 \right) \right) \\ &\geq \sum_{i=1}^n \left(\kappa_\star m_{ii}^2 + 2\sigma_\star^4 \|\tilde{M}_i\|_2^2 \right) \\ &\geq \min(\kappa_\star, 2\sigma_\star^4) \sum_{i=1}^n \left(m_{ii}^2 + \|\tilde{M}_i\|_2^2 \right) \\ &\geq n \min(\kappa_\star, 2\sigma_\star^4) \min_i \left(\|\tilde{M}_i\|_2^2 + m_{ii}^2 \right). \end{aligned}$$

Substituting in this lower bound yields

$$\left| \mathbb{E} \left[\phi \left(\mathcal{Q}_x^{(n)} \right) - \phi \left(\mathcal{Q}_y^{(n)} \right) \right] \right| \leq \frac{1}{\sqrt{n}} \frac{K_3 \psi \max \left(8 \left(3^{3/4} \right) \tau^3, 1 \right)}{3 \left(\min(\kappa_\star, 2\sigma_\star^4) \right)^{3/2}} \left(\frac{\max_j \left(\|\tilde{M}_j\|_2 + |m_{jj}| \right)}{\min_i \left(\|\tilde{M}_i\|_2^2 + m_{ii}^2 \right)} \right)^3.$$

□

Chapter 4

Application to GWAS

In Chapter 2, we presented an approach for calculating a genetic distance between all pairs of haplotypes in a sample at each locus along the genome. In Chapter 3, we developed a screening approach for quadratic forms that can test whether a set of outcomes respects the weighted graph structure implied by a given matrix. In this chapter, we make the final steps to show how these two methods provide the foundation for a class of haplotype-based screening models for GWASs. We pilot one model from this class on a dataset of Kenyan haplotypes from the Malaria Genomic Epidemiology Network [55]. We will focus on binary phenotypes, specifically the diagnosis of severe malaria.

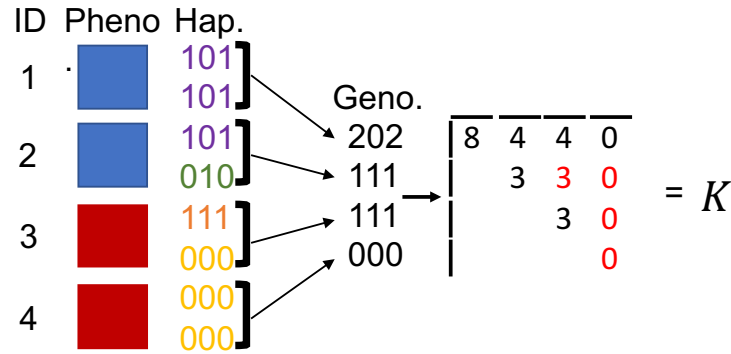
The HMM parameters we use, how we adjust for population structure, and other features of the model have not been optimized, and our present analysis is by no means a full or robust analysis of this dataset. We simply aim to show that our framework is applicable to real data and recovers a few known associated loci. While our implementation of the screening approach derived from Chapter 3 has not been computationally optimized, our pilot shows that coupling our HMM decoding with our quadratic form screening approach is tractable on a real GWAS dataset. We also highlight a few interesting novel signals and discuss future steps that we will take to improve robustness and performance.

4.1 Dediploiding

Recalling our notation established at the beginning of Chapter 1, let N_s refer to the number of organisms sampled in a study, and N be the number of haplotypes derived from that sample. In diploid organisms, $N = 2N_s$. Note that the distance matrices we calculate between haplotypes in Chapter 2, $\left(\tilde{d}_{i,j}^l\right)^{N \times N}$, are $N \times N$. In other words, there are four distances between any two individuals. Since we only have one outcome per individual, testing quadratic forms like those in Chapter 3 requires us to construct a meaningful genetic distance between each pair of individuals based on the four distances between their pairs of haplotypes. We will refer to the process of generating a $N_s \times N_s$ distance matrix from $\left(\tilde{d}_{i,j}^l\right)^{N \times N}$ as "dediploiding."

At first, the need to perform dediploiding may appear to be an unnecessary complication arising from our choice to model haplotypes. Both SKAT and Xu & Guan's approach for haplotype testing seem to circumvent this step by inferring local relatedness structures directly from genotypes rather than from phased haplotypes [12, 29]. While modeling genotypes avoids the computational burden and switch errors that accompany phasing, we deliberately model haplotypes rather than genotypes because convolving haplotypes into genotypes often obscures the grouping of the underlying haplotypes. This eliminates information relevant to association testing (see Figure 4.1). As we will show below, we can choose transformations $\psi : \mathbf{R}^{2 \times 2} \rightarrow \mathbf{R}$ that will map the four distances between two individuals to a single distance that remembers whether two individuals share a common haplotype.

Fig. 4.1 Information Loss During Convolution: Here we present a sample of 4 individuals (each with a distinct ID number). Two of them have the blue phenotype and two have the red phenotype. Alongside their phenotypes are their local haplotypes in a causal region of the genome. At this locus, there are four distinct haplotypes. The purple haplotype is dominant and elicits the blue phenotype. The golden haplotype is dominant and elicits the red phenotype. The convolution of the haplotypes into genotypes is shown (paired haplotype rows are simply added). To the right of the genotypes is the kernel matrix K that would result followed by an application of the non-weighted linear kernel: $K_{ij} = \sum_k G_{ik} G_{jk}$. This is the same kernel used before in Chapter 1, Figure 1.1. As emphasized by the four red entries in the resulting kernel matrix, K , the genotypes predict that individuals 2 and 3 have similar phenotypes and that individuals 3 and 4 have different phenotypes. This reflects the fact that the genotypes obfuscate the genetic distinction between individuals 2 and 3 that drives their distinct phenotypes. This information cannot be recovered no matter which kernel function is used to process the genotypes.



Among the set of potential dediploiding functions, ψ , we want to choose those that will maximize our power to detect haplotypic associations. Consider a population with k distinct haplotypes that elicit different expected phenotypes. We will label these different alleles with capital letters: $A, B, C, \dots$. For a population with k distinct haplotypes, there are $\frac{k(k+1)}{2}$ possible pairs of alleles that an individual might carry, k of them being homozygous and $\binom{k}{2}$ of them being heterozygous. On average, our HMM implementation will add a mutation penalty and declare a relatively high distance between mismatched haplotypes and a relatively low distance between matching haplotypes. The resulting four local distances calculated by our HMM between any two individuals in our sample may be summarized in a 2×2 table. We will denote a generic 2×2 table as having the following form where $a, b, c, d \in \mathbb{R}$.

a	b
c	d

Using this notation, we will consider functions $\psi(a, b, c, d)$.

We only need to consider how ψ acts on five different configurations of tables because of three key symmetries. These are the same symmetries that arise in 2×2 tables in classical hypothesis testing. First, the order of rows and columns is arbitrary: the haplotypes within an individual may be denoted in either order. Similarly, which pair of alleles is assigned to be written across the rows versus across the columns is arbitrary: if we transpose each 2×2 table, the genetic relationship between two individuals is invariant. These two symmetries reflect the fact that in any given sample of haplotypes, the order of individuals within that haplotype file and the order of the two haplotypes within each individual can be changed without changing the underlying phylogeny. Finally, the labelling of the k groups is arbitrary. Even in a population with k distinct haplotypes, accounting for these symmetries leaves us with only the five classes of 2×2 tables in Table 4.1. These symmetries constrain our family of candidate dediploiding functions ψ : they must be invariant to changes in row/column order, transposition, and allele labelling. We will say that such functions are 2×2 table symmetric. We note that a slightly different approach toward compressing haplotypes should be taken for regions of the genome that are hemizygous, such as the X chromosome in males and imprinted regions, but we will not consider those special cases here.

Matching Hom-Hom <table><tr><td></td><td>A</td><td>A</td></tr><tr><td>A</td><td>ϵ</td><td>ϵ</td></tr><tr><td>A</td><td>ϵ</td><td>ϵ</td></tr></table>		A	A	A	ϵ	ϵ	A	ϵ	ϵ	Non-matching <table><tr><td></td><td>A</td><td>A</td></tr><tr><td>B</td><td>δ</td><td>δ</td></tr><tr><td>B</td><td>δ</td><td>δ</td></tr></table><table><tr><td></td><td>A</td><td>A</td></tr><tr><td>B</td><td>δ</td><td>δ</td></tr><tr><td>C</td><td>δ</td><td>δ</td></tr></table><table><tr><td></td><td>A</td><td>B</td></tr><tr><td>C</td><td>δ</td><td>δ</td></tr><tr><td>D</td><td>δ</td><td>δ</td></tr></table>				A	A	B	δ	δ	B	δ	δ		A	A	B	δ	δ	C	δ	δ		A	B	C	δ	δ	D	δ	δ
	A	A																																					
A	ϵ	ϵ																																					
A	ϵ	ϵ																																					
	A	A																																					
B	δ	δ																																					
B	δ	δ																																					
	A	A																																					
B	δ	δ																																					
C	δ	δ																																					
	A	B																																					
C	δ	δ																																					
D	δ	δ																																					
Matching Het-Het <table><tr><td></td><td>A</td><td>B</td></tr><tr><td>A</td><td>ϵ</td><td>δ</td></tr><tr><td>B</td><td>δ</td><td>ϵ</td></tr></table>		A	B	A	ϵ	δ	B	δ	ϵ	P. Match. Het-Hom <table><tr><td></td><td>A</td><td>A</td></tr><tr><td>A</td><td>ϵ</td><td>ϵ</td></tr><tr><td>B</td><td>δ</td><td>δ</td></tr></table>		A	A	A	ϵ	ϵ	B	δ	δ	P. Match. Het-Het <table><tr><td></td><td>A</td><td>C</td></tr><tr><td>A</td><td>ϵ</td><td>δ</td></tr><tr><td>B</td><td>δ</td><td>δ</td></tr></table>		A	C	A	ϵ	δ	B	δ	δ										
	A	B																																					
A	ϵ	δ																																					
B	δ	ϵ																																					
	A	A																																					
A	ϵ	ϵ																																					
B	δ	δ																																					
	A	C																																					
A	ϵ	δ																																					
B	δ	δ																																					

Table 4.1 The Five Table Classes: Let ϵ denote a relatively small haplotype distance and δ denote a relatively large haplotype distance. Each of the 2×2 tables indicate whether we would expect the distance between two haplotypes to be declared relatively low (ϵ) or relatively high (δ) using our HMM implementation under different haplotype configurations. Regardless of the dataset or locus of interest, ψ should always return a small distance for matching homozygous-homozygous tables and matching heterozygous-heterozygous tables. It should always return a large distance for non-matching tables where all four distances are δ . In contrast, partial matching heterozygous-homozygous tables and partial matching heterozygous-heterozygous tables should be treated differently by ψ depending on the suspected underlying genetic architecture.

To maximize power, we want to select a set of ψ functions that will call individuals with the same expected phenotype genetically similar. In the matching homozygous-homozygous (matching hom-hom) case, it is obvious that the two individuals should be called genetically similar by ψ since all of their alleles are identical. In contrast, two individuals should be called genetically distant in the non-matching case where they do not share any alleles. Many ψ would map these two configurations appropriately. For example, taking the mean of the four distances would call matching hom-hom tables close and non-matching tables far. Importantly, $\psi_{\text{mean}} = \frac{1}{4}(a + b + c + d)$, is also 2×2 table symmetric. However, ψ_{mean} does not perform optimally in the matching heterozygous-heterozygous (matching het-het) case where both individuals would have the same alleles and the same expected phenotype but ψ_{mean} would declare them to be some medium genetic distance apart.

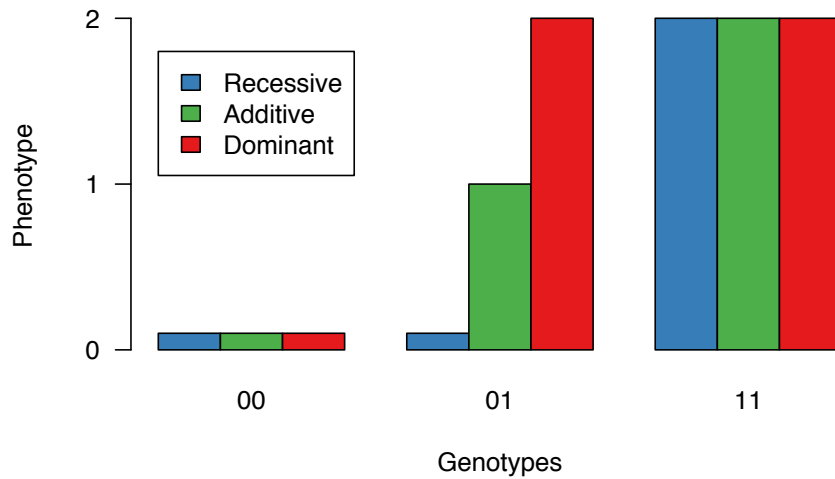
As we will see below, there are several 2×2 table symmetric functions that can return a small distance when given a matching het.-het. table as well as the appropriate distances for matching hom-hom and non-matching tables. However, the desirable behavior of ψ for the remaining two cases, partial matching heterozygous-homozygous (partial het-hom) and partial matching heterozygous-heterozygous (partial het-het), varies depending on the genetic architecture.

In essence, ψ should handle these two types of tables depending on how phenotypically similar two individuals should be based on sharing one allele. More specifically, for partial het-hom tables, we must decide whether heterozygotes will be phenotypically similar to homozygotes with which they share an allele, their "neighboring" homozygotes. Similarly, for the partial het-het case, we must decide whether to declare heterozygotes that only share one allele as close or far. Geneticists routinely make judgements about whether heterozygotes should be modeled as phenotypically equivalent to homozygotes when fitting dominant allele and recessive allele models (see Figure 4.2).

To connect dominant and recessive allele models to our tables, consider the partial het-hom case as depicted in Figure 4.1. If B was a dominant allele, then it would be optimal for ψ to return a large distance. However, note that a function ψ that calls AB far from AA, would also call AB far from BB due to allele labeling invariance, which if B were dominant, should share the same phenotype. The same problem arises if B was recessive in the partial het-hom table. In that case, it would be optimal for ψ to return a large distance. However, a function ψ that calls AB close to AA will also call it close to BB, which would be false if B is recessive. This reflects the fact that 2×2 tables are invariant to allele labeling while definitions of dominance and recessiveness are not invariant to allele labeling. In other words, with only a matrix of pairwise distances, the directionality encoding which allele is increasing in copy number is lost. This means that we cannot model recessiveness and dominance as distinct phenomena, but only whether heterozygotes are expected to have a phenotype that is

a compromise between the phenotypes of their neighboring homozygotes or a phenotype that is completely distinct from the phenotypes of their neighboring homozygotes.

Fig. 4.2 Classic Modes of Inheritance: The expected phenotype corresponding to recessive, additive, and dominant inheritance is shown as a function of genotype. Strictly looking at this figure, it would appear that in order to maximize the power to detect a dominant mutation, ψ should evaluate heterozygotes (01) as close to (11) homozygotes but far from (00) homozygotes. However, if we flipped the labelling of the alleles, we would want to ψ to group heterozygotes (01) with (00) homozygotes. Since the coding of alleles is arbitrary, this means that testing for dominance or recessiveness is the same in the binary allele case. It amounts to assessing whether the expected phenotype of the heterozygote is closer to the expected phenotype of one of the homozygous states.



We do not see this as a true limitation of the method. Dominant and recessive models capture when a heterozygote should be modeled as phenotypically more similar to one neighboring homozygote than the other. Once we move beyond the biallelic case to where we have k potential haplotypes and an entire $N_s \times N_s$ matrix of comparisons, testing for dominance and recessiveness would consume many degrees of freedom and require the identification of homozygous individuals (AA, BB, etc.) to establish directionality of allele copy number with a hard classification algorithm. Like our approach, SKAT cannot model recessive or dominant effects because SKAT is working with a matrix of differences between

individuals, not counting the number of copies of specific alleles. Assumptions must be made to achieve a one-degree of freedom test. However, since we model haplotypes rather than genotypes, our approach retains some control over how we model heterozygotes through our dediploiding functions.

Below we present four functions, which, as delineated in Table 4.2, provide us with control over how partial het-hom and partial het-het tables are evaluated. Note that all of these functions are 2×2 table symmetric and return a large distance in the trivial case of a non-matching hom-hom table and a small distance in the case of a matching hom-hom table. Unlike ψ_{mean} , they also return a low distance in the matching het-het case.

$$\psi_{\text{min1}} = \text{sort}(a, b, c, d)[1]$$

$$\psi_{\text{min2}} = \text{sort}(a, b, c, d)[2]$$

$$\psi_{\text{add}} = \min\left(\frac{a+d}{2}, \frac{b+c}{2}\right)$$

$$\psi_{\text{dom}} = \min(\max(a, d), \max(b, c))$$

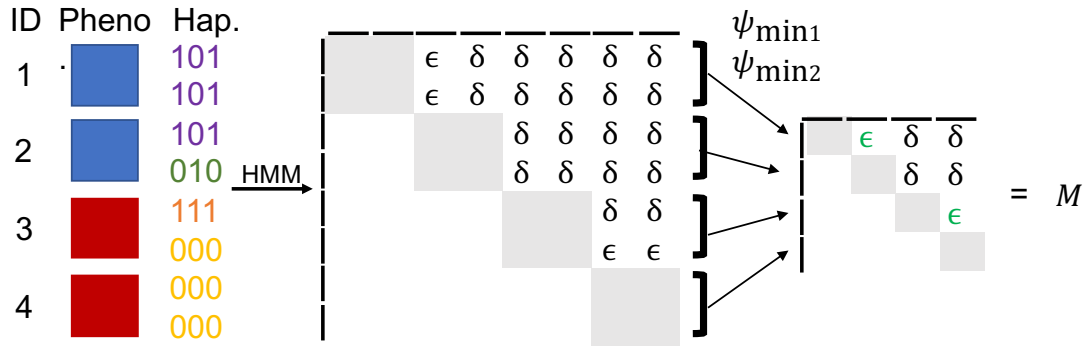
To clarify our use of the sort function, ψ_{min1} returns the smallest distance in the 2×2 table where as ψ_{min2} returns the second smallest distance in the 2×2 table.

ψ	Matching	Het-Het		P. Match.	Het-Hom		P. Match.	Het-Het	
	A	B		A	A		A	C	
	A	ϵ	δ	A	ϵ	ϵ	A	ϵ	δ
	B	δ	ϵ	B	δ	δ	B	δ	δ
ψ_{mean}	$(\epsilon + \delta)/2$			$(\epsilon + \delta)/2$			$(\epsilon + 3\delta)/4$		
ψ_{min1}	ϵ			ϵ			ϵ		
ψ_{min2}	ϵ			ϵ			δ		
ψ_{add}	ϵ			$(\epsilon + \delta)/2$			$(\epsilon + \delta)/2$		
ψ_{dom}	ϵ			δ			δ		

Table 4.2 **Handling of Heterozygotes by ψ Candidates:** Entries indicate the distance that a given ψ function would return for the three non-trivial haplotype table configurations.

In order to observe what is gained by using our haplotype-based rather than genotype-based HMM and then using dediploiding functions, we revisit the hypothetical example from Figure 4.1, where we saw the loss of information in reducing haplotypes to genotypes at the start of the chapter, in Figure 4.3.

Fig. 4.3 Propagation of Information via Dediploiding Functions: Here we reconsider the configuration of haplotypes and phenotypes from Figure 4.1. Applying our HMM approach, we obtain a symmetric $N \times N = 8 \times 8$ distance matrix of ϵ and δ denoting small and large genetic distances respectively. After applying $\psi_{\min 1}$ or $\psi_{\min 2}$, we see that the resulting $N_s \times N_s$ matrix M that places a small distance between individuals with the same phenotype (highlighted with a green ϵ) and larger distances between individuals with different phenotypes.



It may be advisable to try each of our proposed ψ functions at a given locus since they are tuned to distinguish distinct genetic signals. However, if an investigator wants to only use a single ψ function in order to minimize the multiple testing burden, the ψ_{add} function provides a sort of compromise by assigning a medium distance to partial matching tables. As can be seen in Table 4.2, we have a function $\psi_{\min 2}$, which assigns a low distance to partial het-hom tables and a high distance to partial het-het tables, so we do not need to treat both partial matching cases the same. We do not have a ψ that assigns a high distance to partial het-hom tables and a low distance to partial het-het tables. Although it may be difficult to countenance a genetics context in which two individuals who ostensibly have less allele sharing at a position should be declared to have a more similar phenotype, we provide

potential example in Example 4.1. Example 4.1 also illustrates a few of the characteristics of a genetic architecture to which our ψ functions are sensitive.

Example 4.1 Consider a locus where partial or full knock-outs of a particular gene are protective. Consider a genetic architecture where the functional, ancestral allele A is common with frequency $1 - p$ and there are $k - 1$ potential knockout alleles B, C, D, etc, each with equal frequency $\left(\frac{p}{k-1}\right)$. In a population with k alleles, there are $k(k - 1)$ potential partial het-hom pairings and $k \binom{k-1}{2}$ potential partial het-het pairings. In this hypothetical example, let us assume that these alleles are all in Hardy-Weinberg equilibrium. Since we assumed that the mutant alleles are dominant over the A allele, an ideal ψ would assign a large distance to partial het-hom pairings where A is the primary haplotype (AA-AX, for some other allele X). As a proportion of the 2×2 tables at this locus, (AA-AX) comparisons make up $4(1 - p)^3 p$ of tables. In contrast XX-AX and XX-YX tables (for some alleles $X \neq A$ and $Y \neq X$), only make up $4(k - 1) \left(\frac{p}{k-1}\right)^3 \left(1 - \frac{p}{k-1}\right)$ of tables. Recall that we assumed that p is small. Although it would be sub-optimal to declare the XX-AX and XX-YX tables as far apart, since AA - AX terms are much more common, it may be superior to declare all partial het-hom pairings as far. No calculation is necessary to consider the partial het-het tables because in this scenario, all heterozygotes are protected, and should be declared similar. In summary, this is a hypothetical case where calling partial het-hom tables far and partial het-het tables similar may be advantageous.

Given our four dediploiding functions, we proceed in the following section to show how the resulting $N_s \times N_s$ distance matrix M may be used in modeling phenotypes.

4.2 GWAS via Autologistic Models

In this section, we consider a couple of models that allow us to incorporate our distance matrix M into a logistic regression framework.

4.2.1 Classic Autologistic Models

Consider modeling a binary phenotype $Y \in \{0, 1\}$ with a logistic regression model that includes a covariate, indexed by a parameter of interest τ , which takes a weighted sum over the other phenotypes in the dataset. While previously in the thesis it has been sufficient to use y to denote a random vector of phenotypes, since we will be dealing with phenotypes both as a random vector but also as a realization of that random vector in a dataset, here let Y denote the random vector of phenotypes and y denote an observed realization of Y . Let $A \in \mathbb{R}^{n \times p}$ be a matrix of background covariates with corresponding coefficients $\alpha \in \mathbb{R}^p$. The model

$$Y_i | Y_{-i} \sim \text{Bern}(\eta_i) \quad \text{where} \quad \text{logit}(\eta_i) = \sum_{k=1}^p A_{ik} \alpha_k + \tau \sum_{j \neq i} M_{ij} y_j, \quad (4.1)$$

has been used in countless applied problems, but is sometimes used without proper normalization [56]. As pointed out by [57] in the context of probabilistic nearest neighbour models, this can lead to unreliable inference. The full conditional distribution implied by 4.1 corresponds to a joint likelihood of the form

$$P(Y = y | A, \alpha, M, \tau) = \frac{\exp(y^T A \alpha + \tau y^T M y)}{\sum_{z \in \{0,1\}^N} \exp(z^T A \alpha + \tau z^T M z)}, \quad (4.2)$$

where M is a hollow matrix (the diagonal entries are all zero). Equation 4.2 makes it clear that Model 4.1 automatically ignores any asymmetry in M since $v^T M v = \frac{1}{2} v^T (M + M^T) v \quad \forall v \in \mathbb{R}^n$. Therefore, we can assume that M is symmetric. If we are given an asymmetric distance or weight matrix W , we can always just set $M = \frac{1}{2} (W + W^T)$. This justifies the symmeterization

step that we perform in calculating our haplotype distance matrices: the skew component of M is ancillary in our model.

The normalizing constant is famously intractable, hence why proper normalization of this likelihood is often ignored in practice [58, 59]. However, the likelihood is normalizable on the slice of the parameter space where $\tau = 0$. Approximating the log-likelihood surface with a paraboloid based on the gradient and Hessian at the MLE under the null ($\tau = 0$), we obtain following first order approximation to the likelihood ratio test statistic, commonly known as the score statistic, for τ (see Appendix at end of chapter for derivation).

Proposition 4.2.1. *The score statistic for τ under the logistic automodel likelihood in Equation (4.2) is*

$$S_y = \frac{1}{1 - r_{\tilde{\alpha}, M}^2} \left(\frac{y^T M y - E_{\tilde{\alpha}}(Y^T M Y)}{\sqrt{\text{Var}_{\tilde{\alpha}}(Y^T M Y)}} \right)^2. \quad (4.3)$$

Here $\tilde{\alpha}$ are the maximum-likelihood parameters estimated under the null model. Note that $E_{\tilde{\alpha}}(Y^T M Y)$, $\text{Var}_{\tilde{\alpha}}(Y^T M Y)$, and $r_{\tilde{\alpha}, M}^2$ (see Appendix for definition) are all functions of $\tilde{\alpha}$ and M , not functions of the observed y . Therefore, given $\tilde{\alpha}$, S_y is a deterministic function of $y^T M y$ so the p-value of the score statistic is simply the tail probability of the quadratic form $\mathbb{P}(Y^T M Y \leq y^T M y)$ where each Y_i is independent and drawn from the fitted null model.

4.2.2 Centered Autologistic Models

The classic autologistic model presented above has been criticized for the lack of interpretability of the interaction parameter τ [60]. From a genetics perspective, it's helpful to decompose $y^T M y$ into its centered and mean components. Since we can assume that M is symmetric, we have

$$y^T M y = (y - \mu)^T M (y - \mu) + 2y^T M \mu + \mu^T M \mu \quad (4.4)$$

Therefore, $y^T M y$ will reach extreme values when either $(y - \mu)^T M (y - \mu)$ or $y^T M \mu$ are large. The centered quadratic form captures what we typically think of as heritability. It tests whether an individual is more likely to have a disease beyond the risk predicted by his or her background covariates if their close relatives also have the disease beyond the risk predicted by their background covariates. This would imply that there is some latent genetic process underpinning disease risk in the population. $y^T M \mu$ captures something more subtle: it tests whether an individual is more likely to have a disease if their genetic relatives are at a high environmental risk for the disease. While this type of gene-environment interaction is certainly of interest, especially in the context of infectious disease, for now we wish to focus on purely the latent genetic process. To this end, we resort to a modification of the classic autologistic model, the centered autologistic model [60, 58]. Its likelihood takes the form

$$Y_i | Y_{-i} \sim \text{Bern}(\eta_i) \quad \text{where} \quad \text{logit}(\eta_i) = \sum_{k=1}^p A_{ik} \alpha_k + \tau \sum_{j \neq i} M_{ij} (y_j - \mu_j). \quad (4.5)$$

Here we define μ_j as the expected value of Y_j if the effect of τ was absent: $\mu_j = \frac{\exp\left(\sum_{k=1}^p A_{jk} \alpha_k\right)}{1 + \exp\left(\sum_{k=1}^p A_{jk} \alpha_k\right)}$. As in the previous model, we are interested in cases where $\tau < 0$, since for a given distance matrix M , implies that individuals who are more genetically similar tend to similar outcomes. As before, note that here M is also hollow. Critically, the ability to include background covariates A , allows us to adjust for population structure [27]. Just as for Model 4.2, we obtain a quadratic form for the score statistic for τ under Model 4.5 which takes the form

$$(y - \tilde{\mu})^T M (y - \tilde{\mu}). \quad (4.6)$$

Since M has a zero diagonal, after removing genetically outlying individuals from a dataset, we can directly appeal to Theorem 3.2.5 and apply our Gaussian quadratic form screening approach to obtain bounds on the p-value of $(y - \tilde{\mu})^T M (y - \tilde{\mu})$.

4.3 Pilot on data from the MalariaGEN Consortium

4.3.1 Background on Malaria

Malaria is a life-threatening disease caused by Plasmodium parasites, which are spread between humans by infected female *Anopheles* mosquitoes. Although five species of Plasmodium are known to cause malaria in humans, the overwhelming majority of the disease burden worldwide is driven by two species, *P. falciparum* and *P. vivax* [61]. Although individuals living in areas where malaria is endemic often develop resistance during early childhood, for non-immune individuals, symptoms usually appear between one to two weeks after the initial mosquito bite and include fever, chills, and vomiting. For individuals infected by *P. falciparum*, these symptoms can escalate to multisystem organ failure and death [61]. Although the incidence rate and mortality rate have dropped globally between 2000 and 2015, the WHO estimates that there were over 214 million new cases of malaria and 438,000 malaria deaths worldwide in 2015 [62].

The phylogenetic ancestors of malaria parasites are thought to have began infecting early forms of mosquitoes and blood-sucking flies over 100 million years ago and evolved a two-stage life cycle in accordance with the blood feeding of the host insects, leading them to parasitize a wide range of land vertebrates, including reptiles, birds, and mammals [63]. Although malaria parasites are thought to have sporadically infected even the earliest apes and humans, it is suspected that the advent of agriculture and human settlements around 10,000 years ago provided malaria parasites with an unprecedented opportunity for expansion. By placing relatively large numbers of human hosts in close proximity to bodies of water, the site of mosquito replication, agriculture facilitated a dramatic increase in malaria transmission between mosquitoes and humans. Consistent with this hypothesis, many of the resistance mutations that have risen to prevalence in modern humans are only thought to have emerged in the past 10,000 to 5,000 years [63–65].

Among the genetic variants that are known to be protective from malarial infection are some of the most classic examples of natural selection: the process by which a certain trait or causal genetic variant becomes more prevalent in a population over generations due to a reproductive advantage it confers to carriers. *HBB* is the gene that codes for the two β -globin chains in hemoglobin, the oxygen-transporting protein complex in red blood cells. Three of its alleles, hemoglobin *S*, hemoglobin *E*, and hemoglobin *C*, are all known to confer resistance to malaria [64]. Perhaps the most famous allele, the *S* allele, otherwise known as the sickle-cell allele, corresponds to a glutamic acid to valine substitution that disrupts the natural structure of hemoglobin, causing red blood cells to take on a sickled confirmation in homozygotes. While being homozygous for the *S* allele yields sickle cell anemia, being heterozygous for the allele confers strong protection against malaria infection without significant anemia. This has made the *S* allele a textbook example of balancing selection: natural selection that acts on variant that is beneficial in a heterozygous state but deleterious in a homozygous state [64]. Since we anticipate that our HMM-based approach will be sensitive to long haplotypes that exhibit extended homozygosity due to recent positive selection [66], severe malaria susceptibility is a phenotype where our approach might be especially sensitive compared to current approaches.

Not only is the *S* allele interesting from a selection perspective, but the valine substitution of the *S* allele is thought to have arisen from five independent mutation events on distinct haplotype backgrounds in five distinct geographical regions from West Africa to India [64]. In other words, if we were to genotype individuals from across these regions and did not know to ascertain the valine substitution as a marker on our SNP chip, it is very unlikely that any nearby SNP would effectively tag the causal substitution. This type of extreme allelic heterogeneity is also seen at other known malaria risk loci, including *G6PD*, *ABO*, and *HLA*, which has made it difficult for standard single-SNP-based approaches to clearly identify or quantify the effects of variants at these risk loci [7, 67, 64, 68, 55]. Although more

than twenty regions of the genome have been reported to be associated with malaria risk, several associations have not been consistently replicated [69]. Only associations at six genes have been successfully replicated and validated: *ABO*, *HBB*, *G6PD*, *APT2B4*, *CD40LG*, and *GYP A/GYP B* [69, 70]. We hope to show that our method provides statistical power and consistency in these regions that exhibit allelic heterogeneity compared to single marker tests. This makes severe malaria susceptibility a prime test case for piloting our approach.

4.3.2 Dataset & Methods

We reanalyzed the phased and filtered 6,136 Kenyan haplotypes (3068 samples), used by Band et al. in their 2013 GWAS for loci that modulate severe malaria susceptibility [55]. Cases recruited into this dataset presented with cerebral malaria or severe malarial anemia. The control samples were population controls (obtained from umbilical cord blood). SNPs were also subsampled for this pilot study to reduce the overall computational time, though it is certainly tractable to apply our method to a larger dataset. For example, Chr2, with $L=27,242$ SNPs in our subsampled data, took 8 hours to run using 96 cores distributed across four machines. The majority of computational resources used went towards association testing since our implementation of the p-value bounding procedure has not been optimized to interface efficiently with our R package responsible for propagating the HMM. For each $N \times N$ distance matrix, we applied the five dediploiding functions mentioned earlier in the chapter $\{\psi_{\text{mean}}, \psi_{\text{min } 1}, \psi_{\text{min } 2}, \psi_{\text{add}}, \psi_{\text{dom}}\}$. For each of these dediploid transformations, we obtained upper and lower bounds on both $\mathbb{P}(Y^T M Y < y^T M y)$ and $\mathbb{P}\left((Y - \mu)^T M (Y - \mu) < (y - \mu)^T M (y - \mu)\right)$. We also calculated the score statistic from Equation 4.3 based on $y^T M y$.

In this pilot, we did not tune the mutation or recombination parameters used in the HMM, but simply set the prior copying probability to uniform for all donors across all recipients.

We set $\mu = 8 \times 10^{-7}$ and the exponent on the recombination map $\gamma = 1$. Of course these parameters can and should be optimized as outlined in Chapter 2.

For every locus and each of our dediploiding transformations, we bounded the p-values of the resulting quadratic form using the top 50 eigenvalues and eigenvectors for each distance matrix M . The bound on the potential contribution from remaining eigenvalues and eigenvectors was then optimized using Lemma 3.1.1. We added the five principal components calculated from the genome-wide genotype sharing matrix and added them as background covariats to Model (4.5) to account for population structure [27, 22].

Alongside our method, we present Manhattan plots of single marker tests applied to the same SNPs that were provided to our haplotype method. One important criticism of our pilot is that we do not give the single marker test the added benefit of performing imputation. However, it is still informative of our method's ability to see more complex haplotypic structure to compare the single marker tests to our approach on the same SNP set. At each locus we applied four different single marker tests (SMTs) corresponding to different modes of inheritance: dominant, recessive, additive and general (the two-degree of freedom model for arbitrary heterozygote and homozygote SNP effects).

4.3.3 Recovering Known Results

We begin our analysis with the most classic malaria susceptibility loci: *HBB* and *ABO*. As shown in Figure 4.4, the association signal is so strong at the *HBB* locus that it is detected by all SMTs and by our haplotype test regardless of the dediploiding function used. However, some dediploiding functions found *HBB* more of an outlier than others. In Figure 4.4 we report one of the stronger haplotype signals, found using $\psi_{\min 1}$.

Moving on to *ABO*, in Figure 4.5, we find that our haplotype approach clearly identifies the association at the locus while the association at *ABO* was missed by all four of the SMTs. Although this is not a strictly fair comparison since we have not imputed the SNPs for the

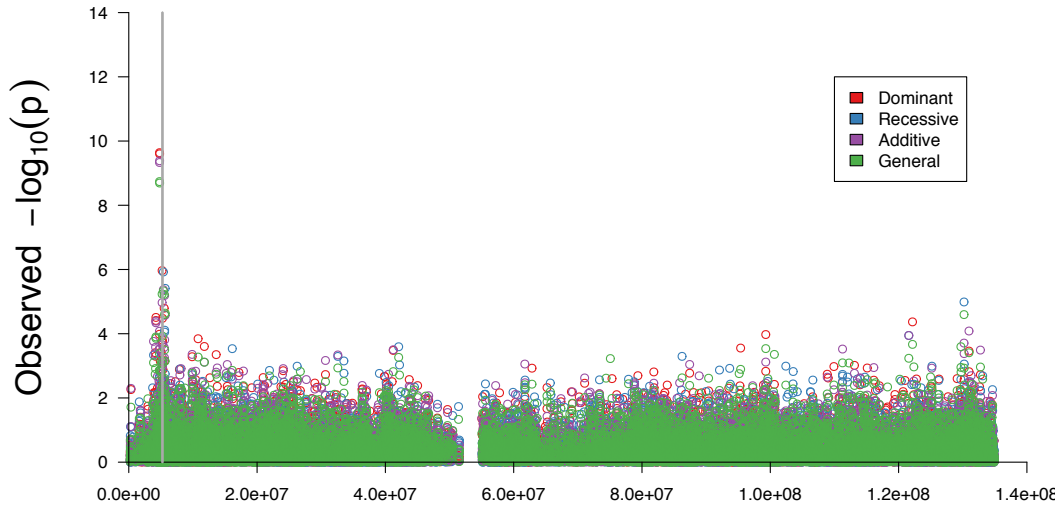
SMTs, that would increase the multiple testing burden on the SMTs and make it somewhat more difficult to compare p-values between results. At least this signal at *ABO* shows that when given the same set of SNPs, our approach can uncover a classic signal that SMTs do not. Whereas the signal at *HBB* was obvious no matter which ψ was used, the signal at *ABO* stood out considerably more using $\psi_{\min 1}$ or $\psi_{\min 2}$ than ψ_{dom} or ψ_{add} (see Figure 4.6).

Hit plots for *HBB* and *ABO* are shown in Figure 4.7. This figure not only includes a local Manhattan plot of the p-value bounds on the centered quadratic form, $(y - \tilde{\mu})^T M (y - \tilde{\mu})$, but also a local Manhattan plot of the p-value bounds on non-centered quadratic form, $y^T M y$ (shown in violet and orange in the bottom panels). As we can see qualitatively, the p-value bounds for the centered and non-centered quadratic forms tend to agree on both the magnitude and localization of signals. However, the score statistic for $y^T M y$ sharply disagrees with the p-value bounds for several SNPs in both regions. The score statistic also appears to more closely localize the signals to *HBB* and *ABO*.

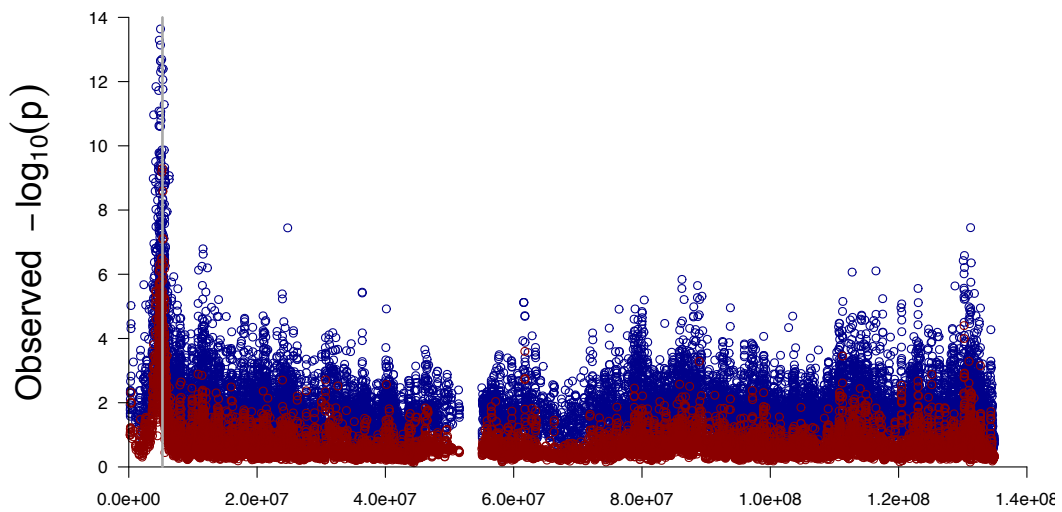
Of course if we were performing a full GWAS on these data rather than a pilot to characterize our method, we would use our HMM to quickly recover the distance matrices at *HBB* and *ABO* loci. We would then exactly evaluate the p-values at the two loci by performing a full eigendecomposition of each promising distance matrix and characterize the haplotypes driving the association signal. We leave that work to future investigations.

While our approach proved to be relatively sensitive to the signal around *ABO*, it did not strongly isolate the association on chromosome 4 at glycophorin A and glycophorin B like the SMTs. The p-value bounds on the centered quadratic form using $\psi_{\min 2}$ still appear to have captured some of the signal (see Figure 4.8). To make our plots easier to view, for the chromosome 4 SMT Manhattan plot and all subsequent Manhattan plots, we remove the legend coding for the different SMT models shown in the Manhattan plots for chromosomes 11 and 9. We maintain the same color codings for the different SMT models throughout.

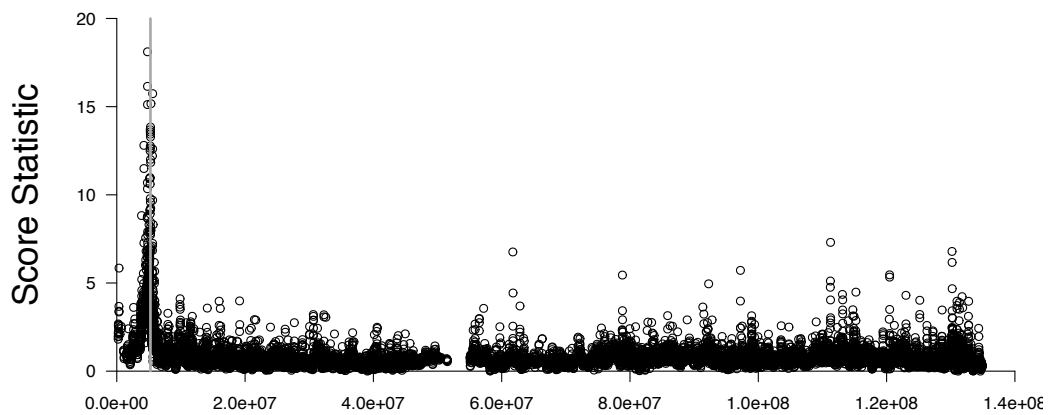
Fig. 4.4 Chr 11 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark HBB.



(a) SMT Manhattan Plot

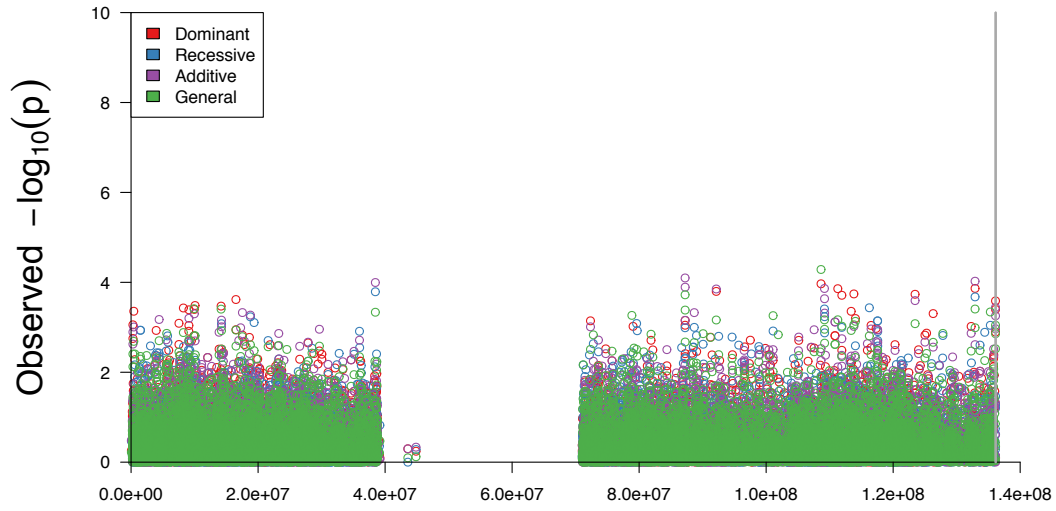


(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$

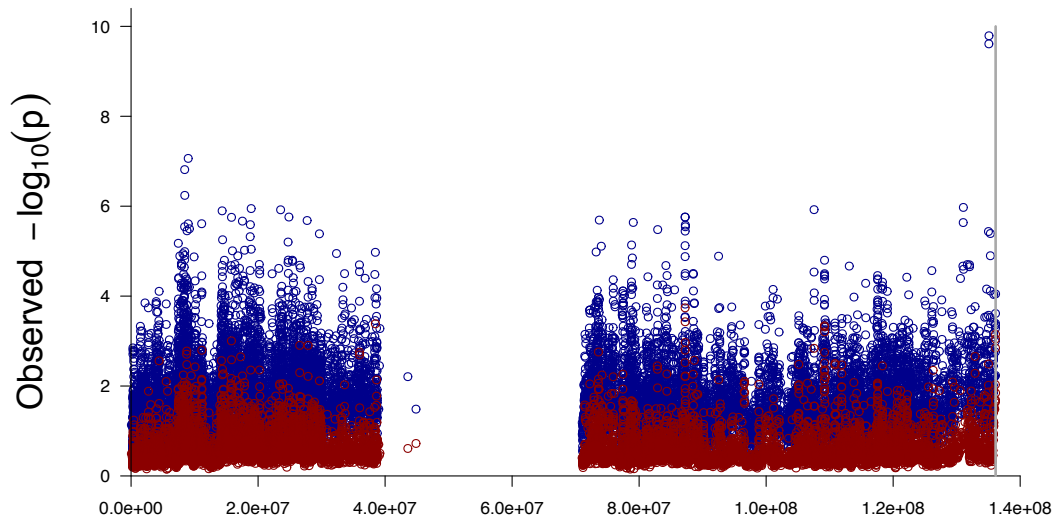


(c) Score Statistic for $y^T M y$

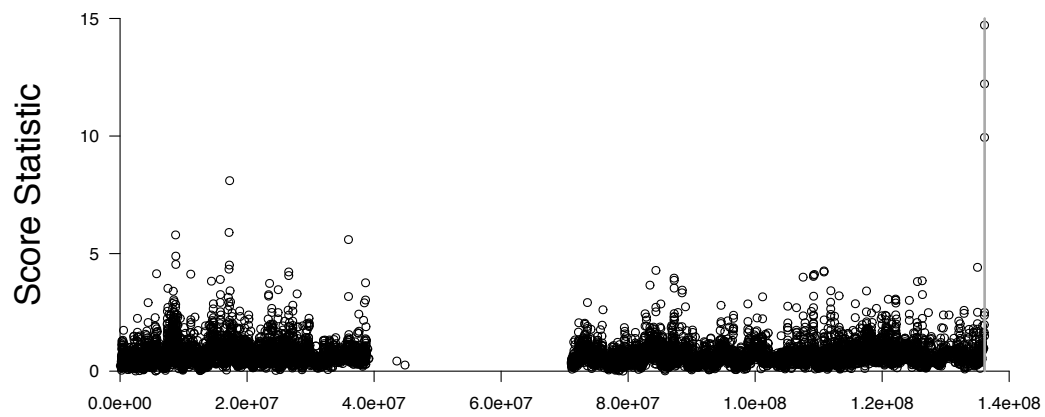
Fig. 4.5 Chr 9 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark ABO.



(a) SMT Manhattan Plot

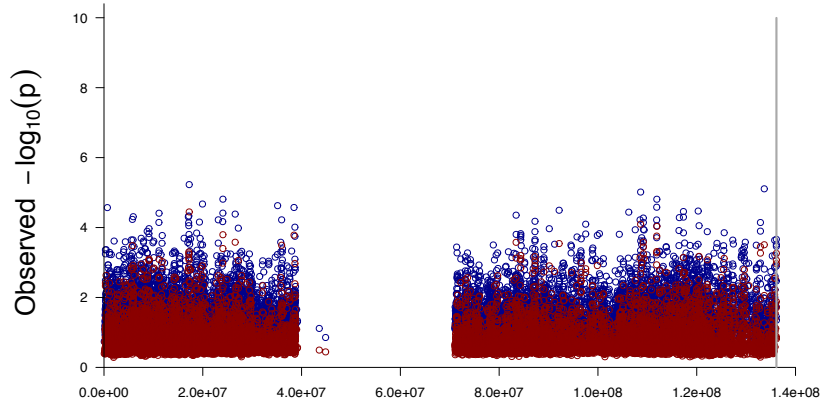


(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$

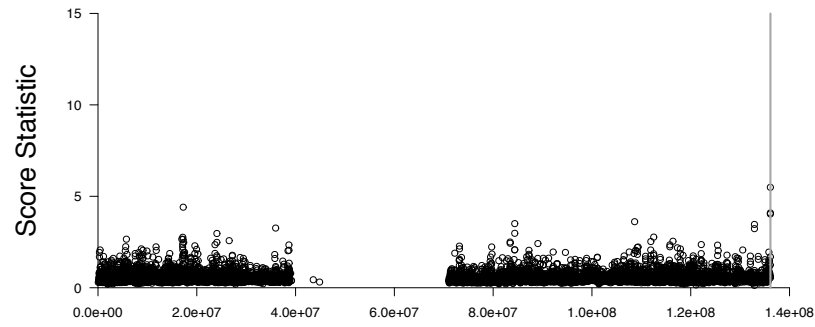


(c) Score Statistic for $y^T M y$

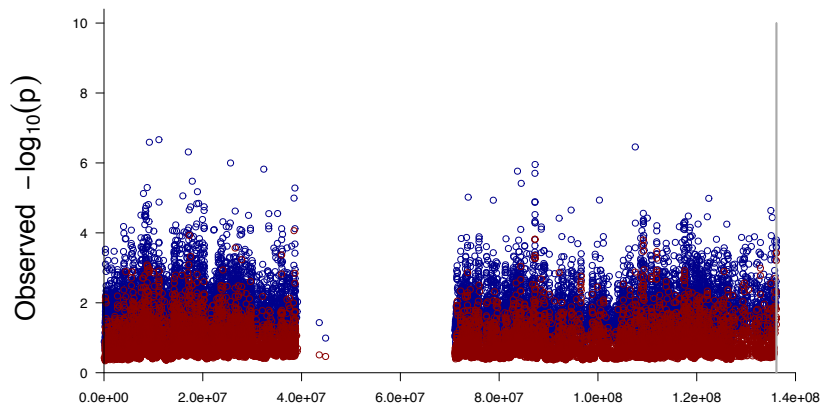
Fig. 4.6 ABO Signal Using ψ_{dom} and ψ_{add} : Bounds and score test shown for M obtained with ψ_{dom} compared to those obtained with ψ_{add} . Gray bars mark ABO.



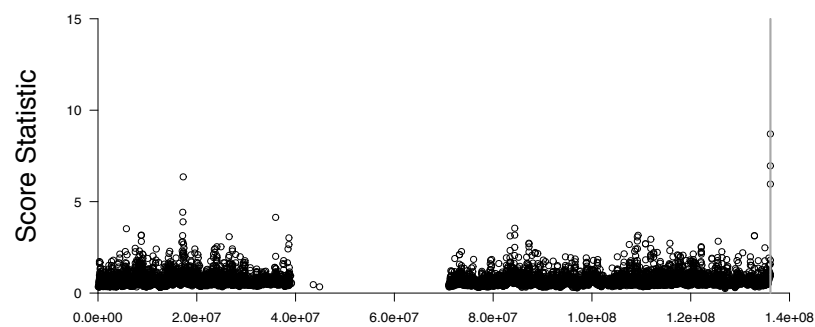
(a) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$ calculated using ψ_{dom}



(b) Score Statistic for $y^T M y$ calculated using ψ_{dom}



(c) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$ calculated using ψ_{add}



(d) Score Statistic for $y^T M y$ calculated using ψ_{add}

Fig. 4.7 Hit-plots for *HBB* and *ABO*: Gray vertical bars denote the limits of the gene of interest. Just as dark blue and dark red points denote upper and lower bounds respectively on the p-value for $(y - \tilde{\mu})^T M (y - \tilde{\mu})$, violet and orange points denote upper and lower bounds respectively on the p-value for $y^T M y$.

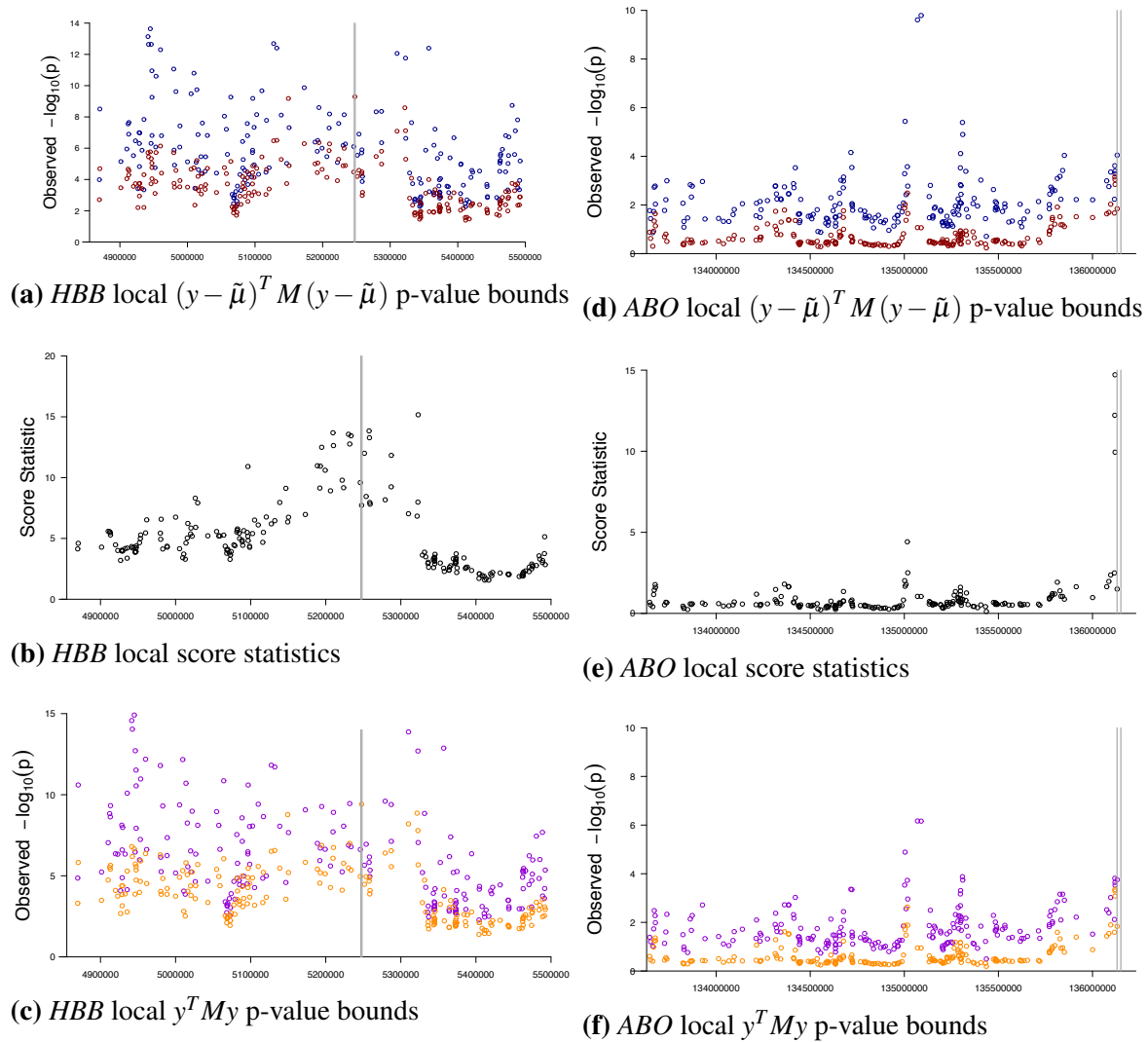
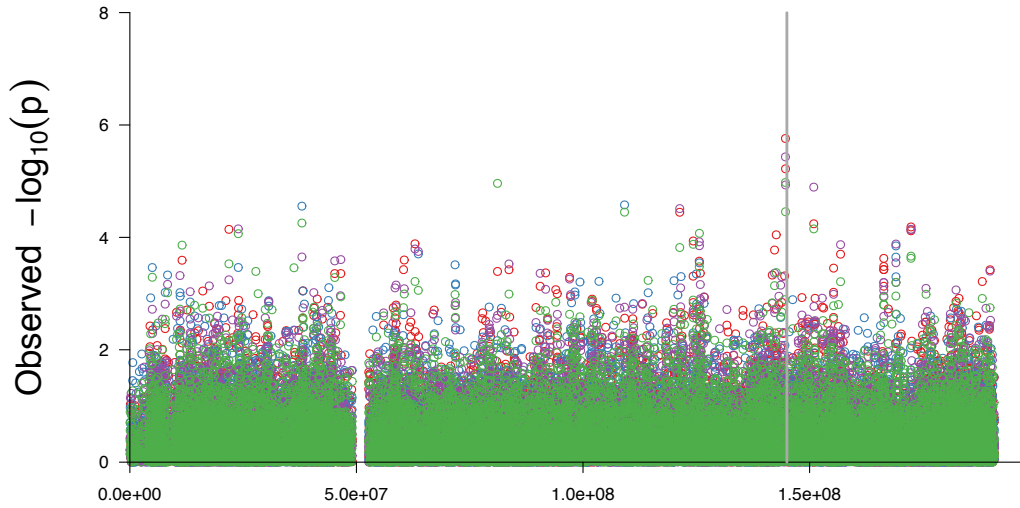
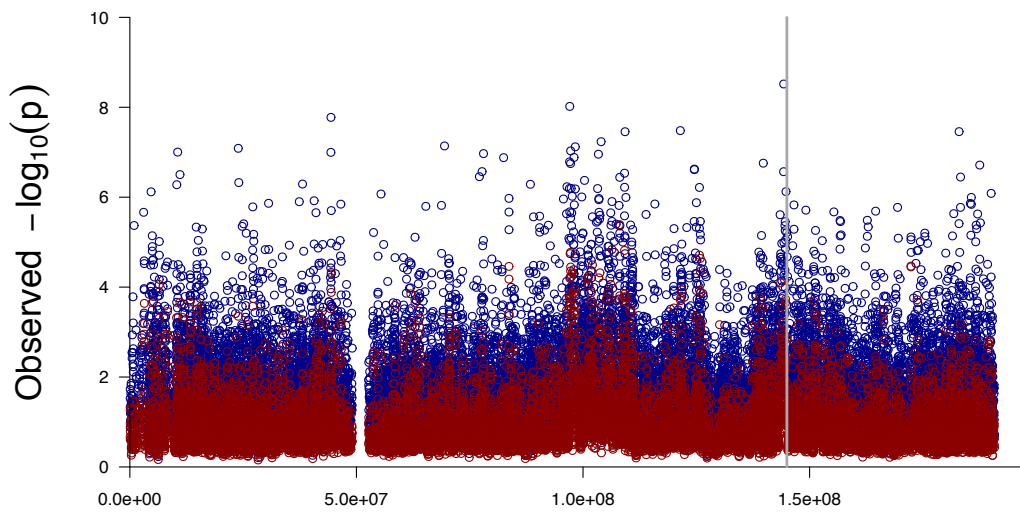


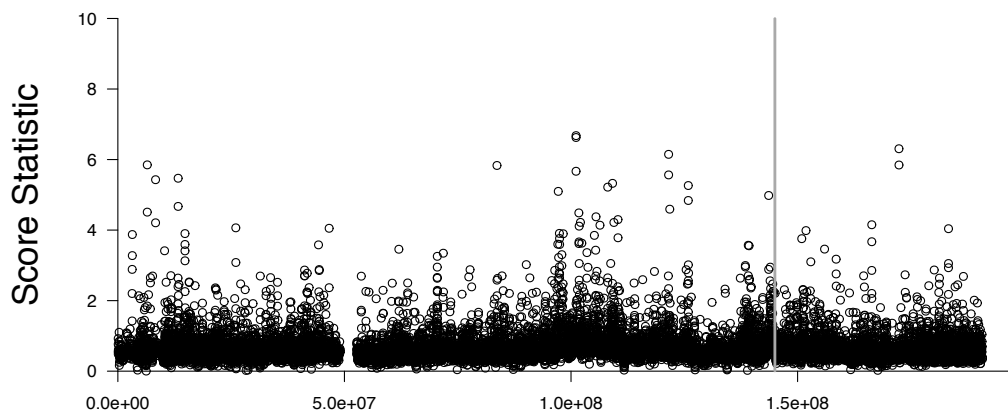
Fig. 4.8 Chr 4 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 2}$. Gray bars mark the outer borders of the regions of *GYPB* and *GYP A*.



(a) SMT Manhattan Plot



(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$



(c) Score Statistic for $y^T M y$

4.3.4 Novel Results

While we found several novel putative associations signals during our screen, we focus on seven here due to their relatively compelling biological significance. We group these genes into three categories: glutathione production, immune response, and cerebral malaria.

Glutathione Production

Here we present a gene that falls in the same glutathione synthesis pathway as Glucose-6-phosphate dehydrogenase, *G6PD*. Glutathione is crucial for neutralizing oxidative species in red blood cells [71]. *G6PD* deficiency is protective against malaria since it reduces glutathione production, which promotes oxidative stress in red blood cells, leading to premature cell lysis [71]. On chromosome 6, using $\psi_{\min 1}$, we identified an association signal at glutamate-cysteine ligase, *GCLC*, which catalyzes the rate limiting step in glutathione production [72]. See Figure 4.9. While variation in *GCLC* has been repeatedly associated with hemolytic anemia, the gene has not yet been associated with malaria susceptibility. While *GCLC* does not appear to stand out on plot of p-value bounds (Figure 4.9b), if one just considers the red lower bounds, then it is the second highest peak on the chromosome. Focusing on the *GCLC* locus, we see that the p-value bounds for both the centered and non-centered quadratic forms, as well as the score statistic, report a signal at *GCLC* (see Figure 4.10). In further support of this locus, the association signal was replicated when we applied any of our other three dediploiding transformations, albeit, the signals were weaker than that shown for $\psi_{\min 1}$.

Casting further spotlight on the pathways to glutathione production, the locus on Chromosome 5 with the largest $\psi_{\min 2}$ score statistic was less than 4kb downstream of glucosamine-6-phosphate deaminase 1 (*GNPDA1*, see Figure 4.11). This enzyme converts glucosamine-6-phosphate to fructose-6-phosphate [73]. While fructose-6-phosphate is a key intermediate in

several pathways, most notably glycolysis, it is also a key metabolite in the pentose phosphate pathway that leads to the production of glucathione [74].

Immune Response

Using $\psi_{\min 1}$, we identified signals within the boundaries of two important immunological genes, *LILRB1* and *CDH4* (see Figure 4.14). Both putative associations were weakly reported by the SMTs. Significantly increased levels of LILRB1+ CD19+ B cells have been found in patients with severe malaria versus those with uncomplicated malaria or healthy controls [75]. These B cells have been implicated in the excessive release of cytokines and impairment of immune memory [75]. See Figure 4.13 and Figure 4.15. Genetic variation in *CDH4* has been found to modulate blood levels of the classic cytokine, monocyte chemoattractant protein-1 (MCP-1), which is critical for recruiting macrophages to sites of infection [76, 77].

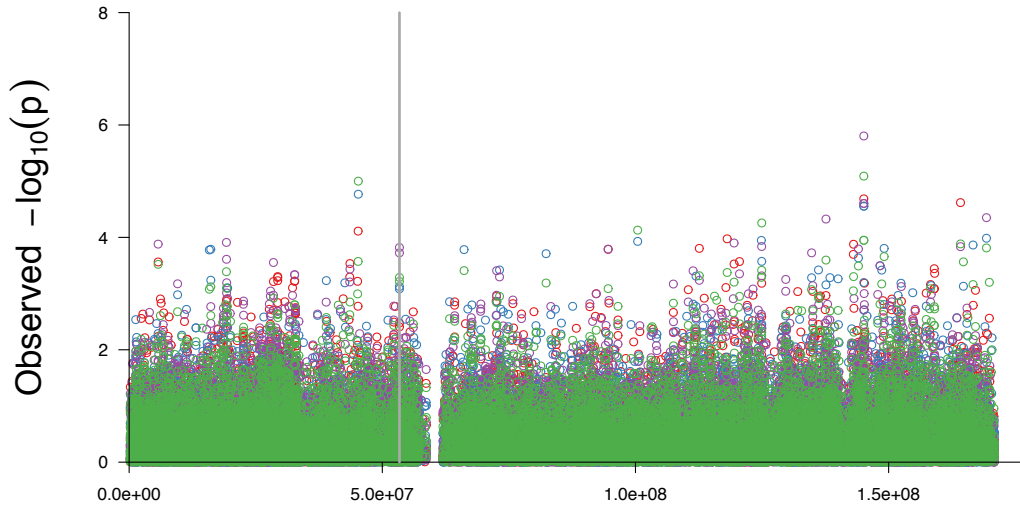
Cerebral Malaria

Using our dediploiding function $\psi_{\min 1}$, on chromosome 12 we identified an association at *TMEM132B* (see Figure 4.16). The SMTs also reported an association at this locus. This gene may play an important role in vascular integrity in cerebral malaria given that it was found to be differentially expressed in patients who had experienced an intracranial aneurysm [78]. See Figure 4.17.

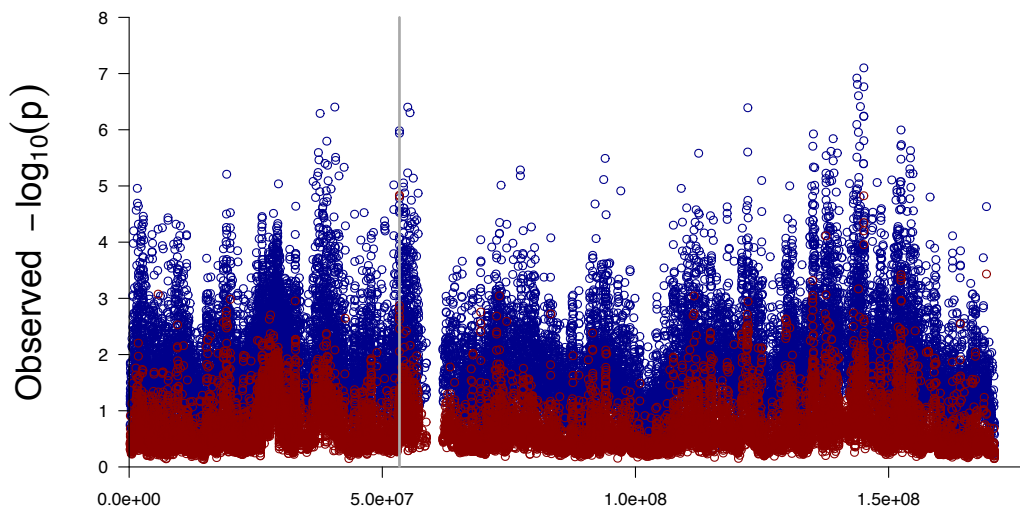
Using our dediploiding function ψ_{add} , on chromosome 10, we identified a signal at Homo sapiens N-acylsphingosine amidohydrolase 2 (*ASAH2*), which is responsible for the cell surface production of shingosine, the precursor for shingosine-1-phosphate (S1P) [79, 80]. Ugandan children with cerebral malaria were found to have decreased S1P levels compared to children with uncomplicated malaria [81]. In a mouse model of cerebral malaria, mice that had induced higher levels of S1P showed improved survival [81]. See Figure 4.18 and Figure 4.19.

We found a second, even stronger signal on chromosome 10 with ψ_{add} at homo sapiens glutamate receptor, ionotropic, delta 1 (*GRID1*). This signal was not at all detected by SMTs. A recent scan for soft selective sweeps found that several glutamate receptors, including *GRID1*, showed evidence for recent selection in humans [82]. Of course, further analysis will be required to confirm whether such a selection signal is present at this locus in our data and if those are the same haplotypes that are driving the association signal we see at *GRID1*. If so, it would support our current speculation that our haplotype approach is particularly powerful for detecting associated haplotypes under recent, positive selection. See Figure 4.18 and Figure 4.20.

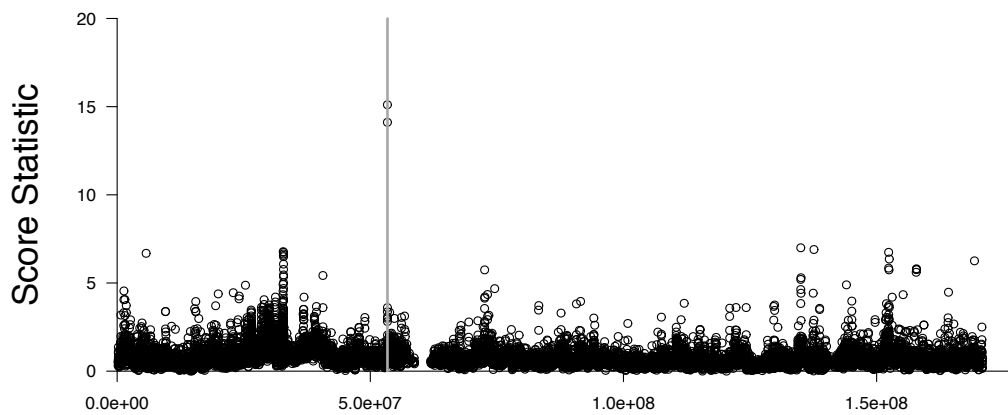
Fig. 4.9 Chr 6 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark GCLC.



(a) SMT Manhattan Plot

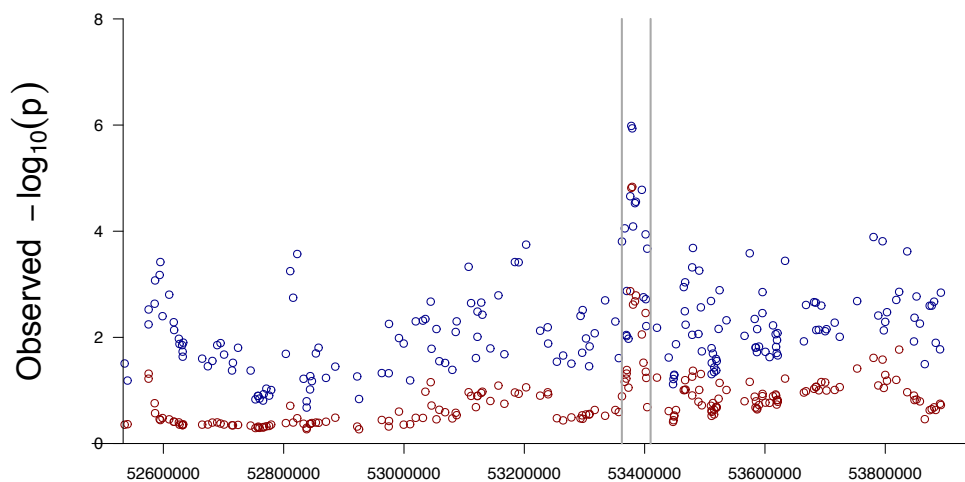


(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$

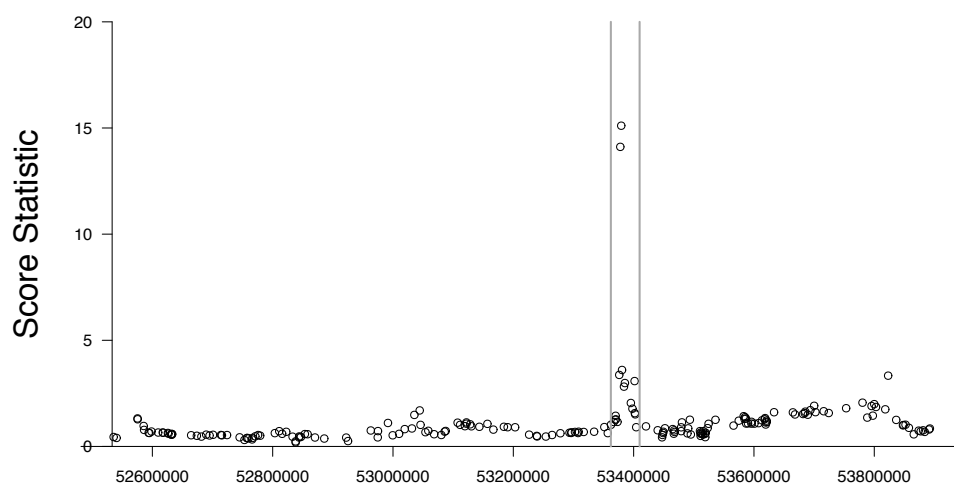


(c) Score Statistic for $y^T M y$

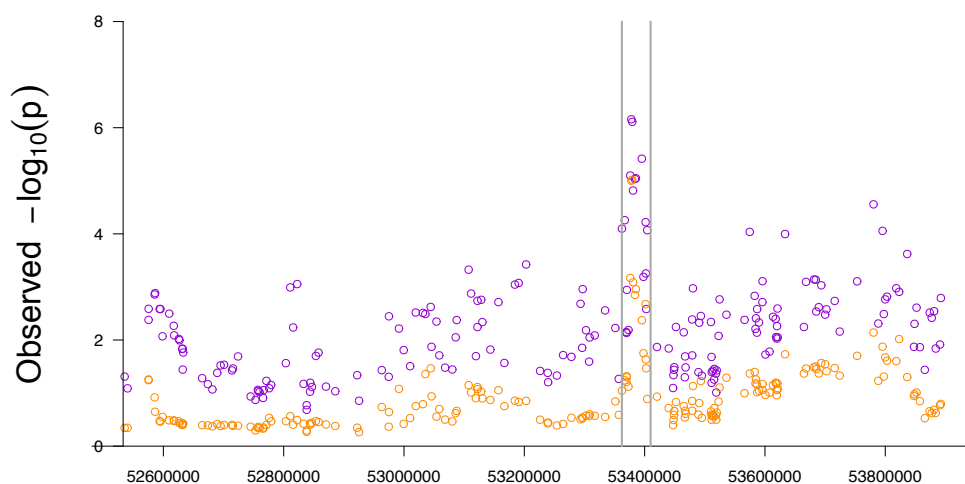
Fig. 4.10 Hit-plots for *GCLC*: All three plots are based on $\psi_{\min 1}$. Gray vertical bars denote the limits of the coding region of *GCLC*. Just as dark blue and dark red points denote upper and lower bounds respectively on the p-value for $(y - \tilde{\mu})^T M (y - \tilde{\mu})$, violet and orange points denote upper and lower bounds respectively on the p-value for $y^T M y$.



(a) *GCLC* local $(y - \tilde{\mu})^T M (y - \tilde{\mu})$ p-value bounds

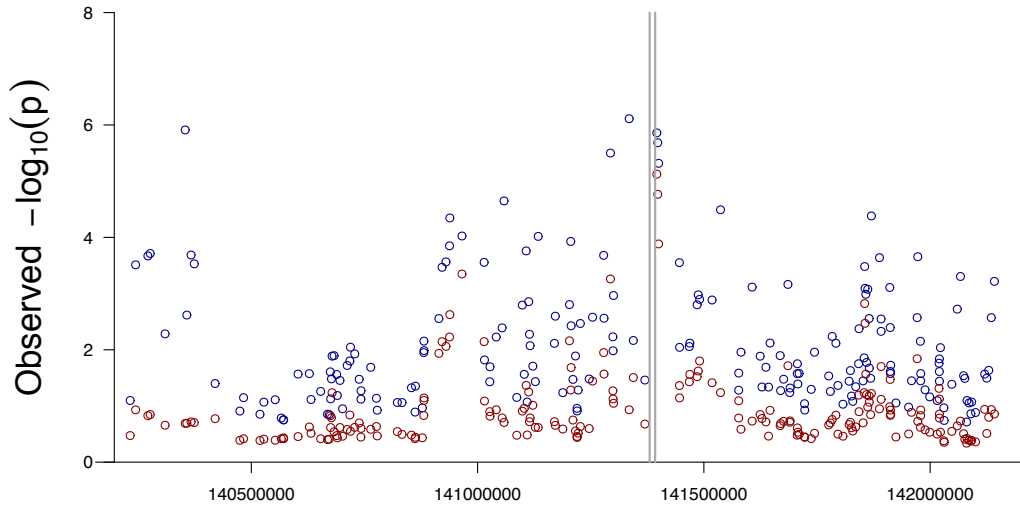


(b) *GCLC* local score statistics

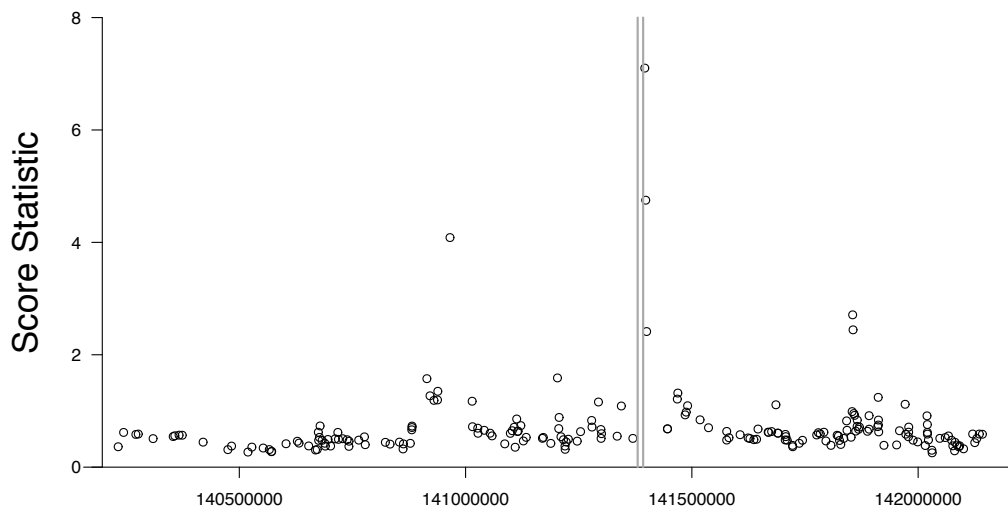


(c) *GCLC* local $y^T M y$ p-value bounds

Fig. 4.11 Hit-plots for *GNPDA1*: Gray vertical bars denote the limits of the *GNPDA1* coding region. $\psi_{\min 2}$ was used to generate both plots.

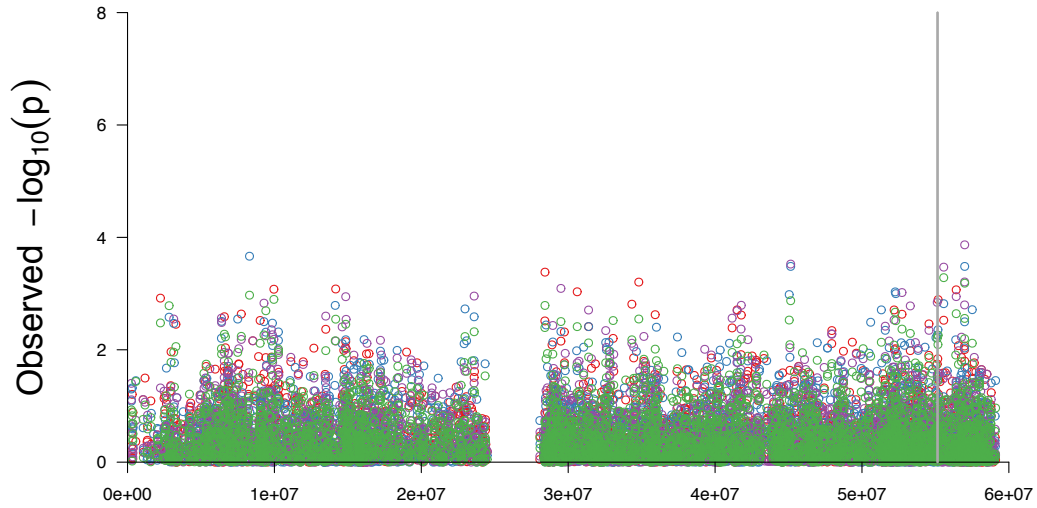


(a) *GNPDA1* local $(y - \tilde{\mu})^T M (y - \tilde{\mu})$ p-value bounds

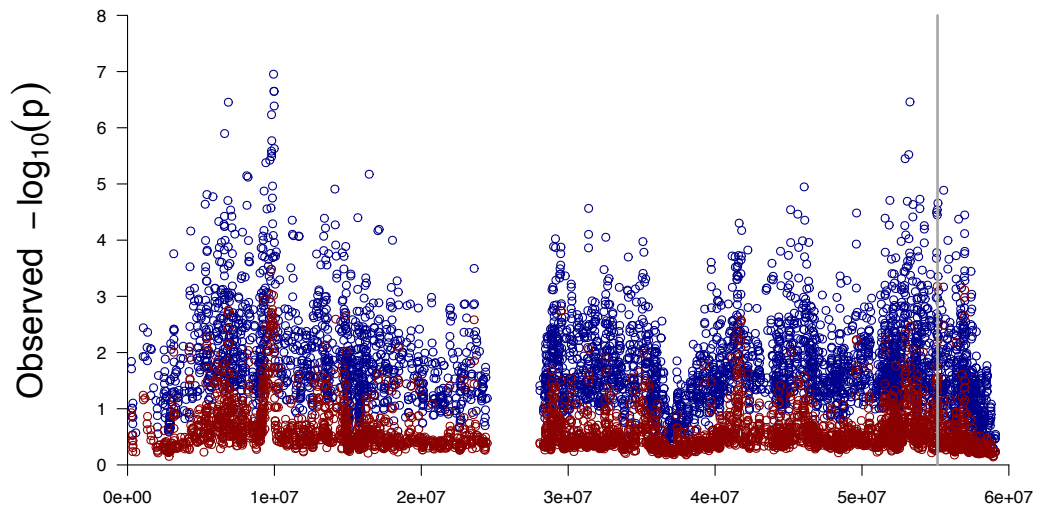


(b) *GNPDA1* local score statistics

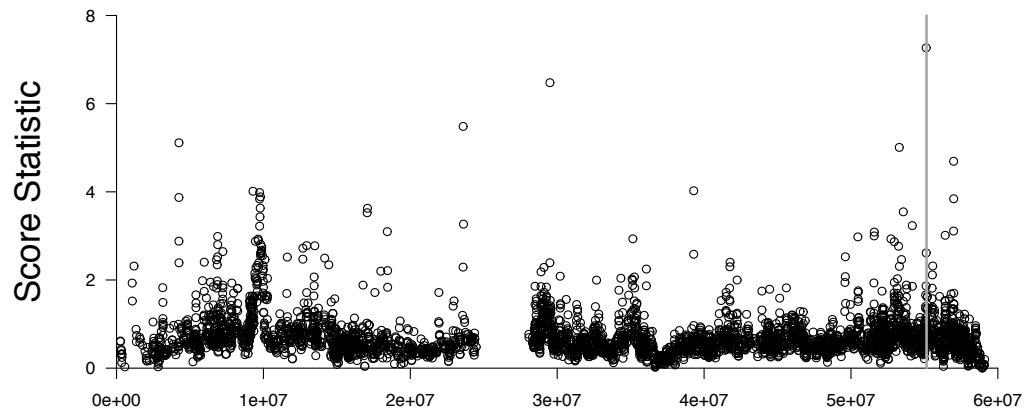
Fig. 4.12 Chr 19 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *LILRB1*.



(a) SMT Manhattan Plot

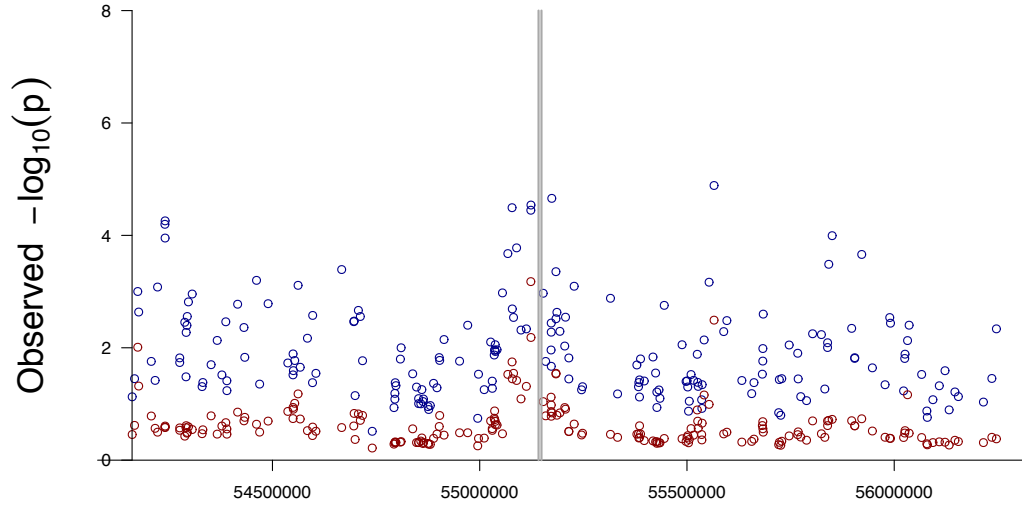


(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$

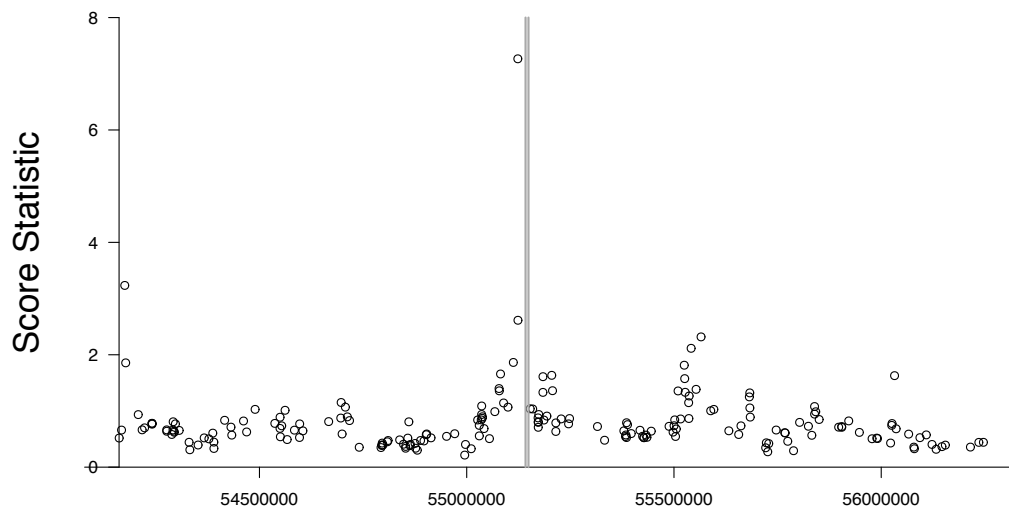


(c) Score Statistic for $y^T M y$

Fig. 4.13 Hit-plots for *LILRB1*: Gray vertical bars denote the limits of the *LILRB1* coding region. $\psi_{\min 1}$ was used to generate both plots.

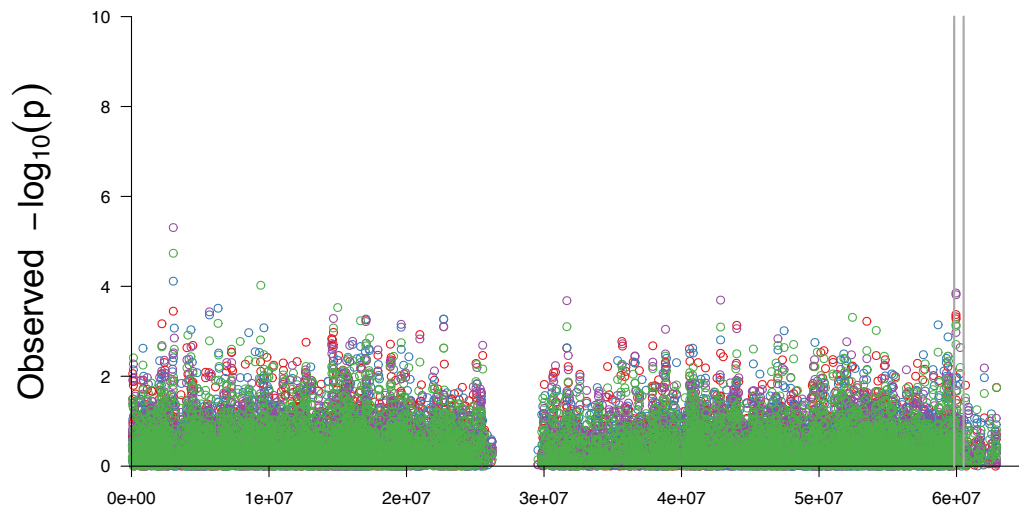


(a) *LILRB1* local $(y - \hat{\mu})^T M (y - \hat{\mu})$ p-value bounds

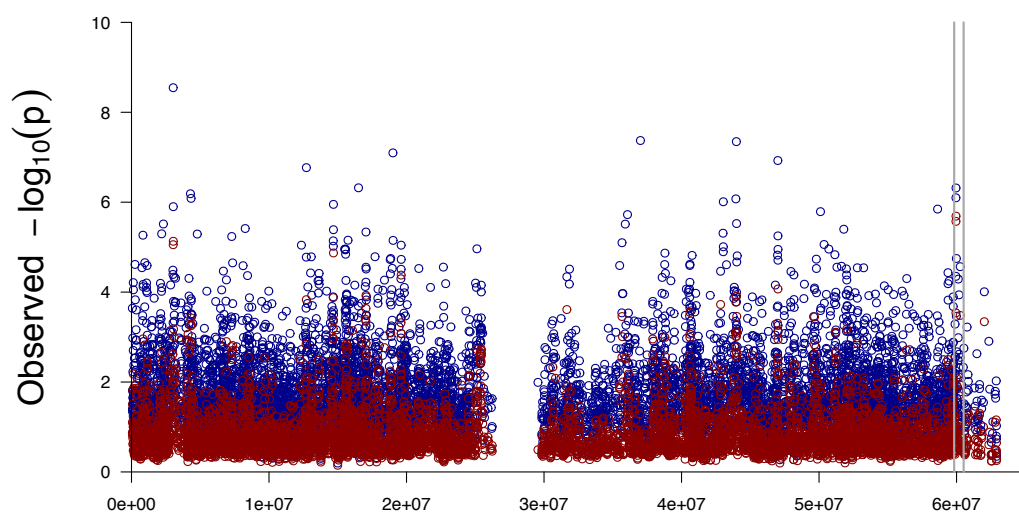


(b) *LILRB1* local score statistics

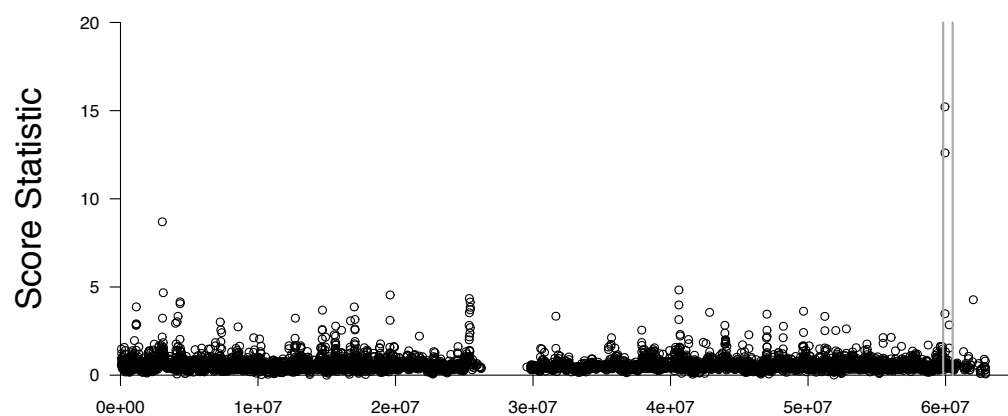
Fig. 4.14 Chr 20 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *CDH4*.



(a) SMT Manhattan Plot



(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$



(c) Score Statistic for $y^T M y$

Fig. 4.15 Hit-plots for *CDH4*: Gray vertical bars denote the limits of the *CDH4* coding region. $\psi_{\min 1}$ was used to generate both plots.

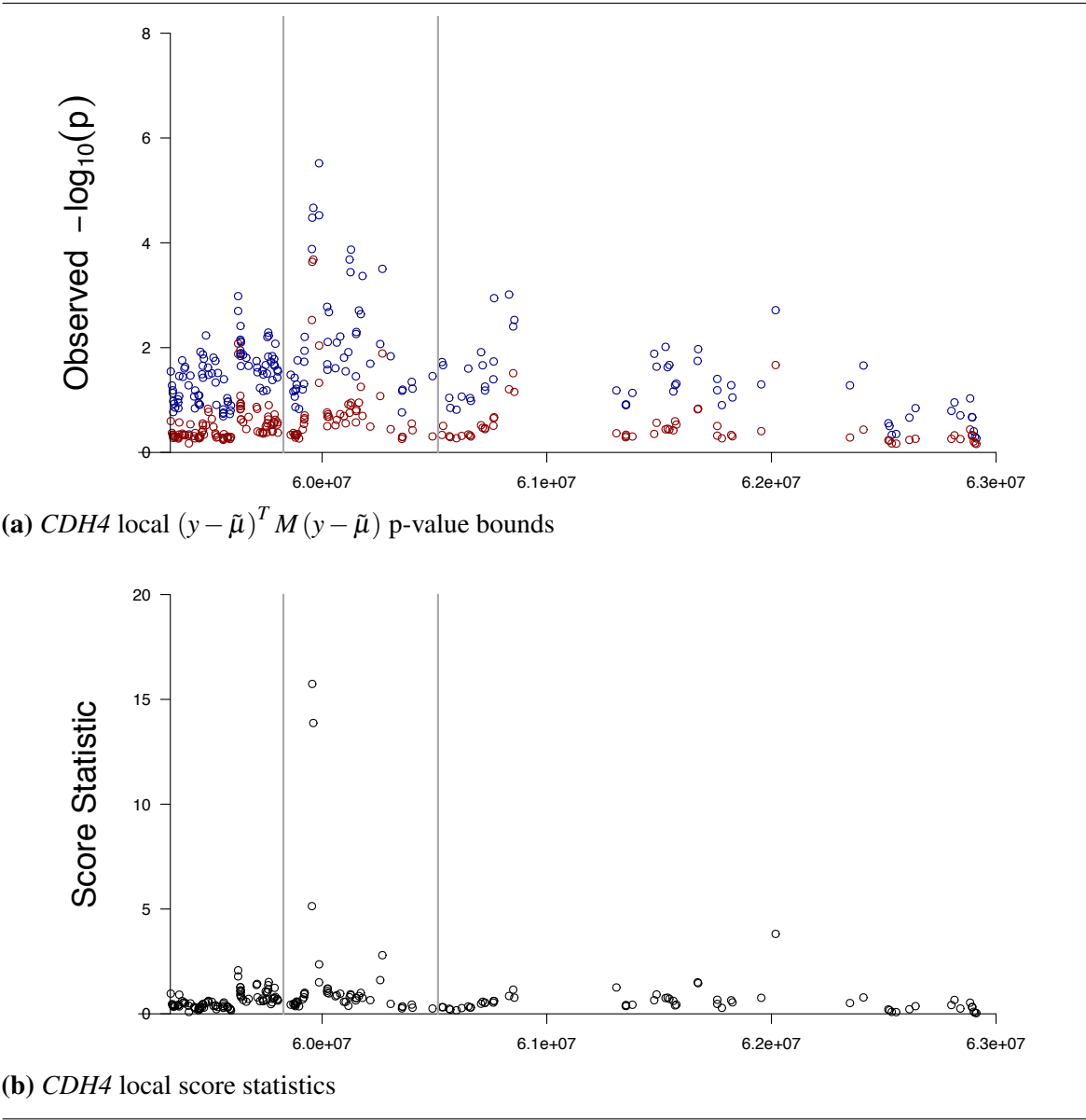
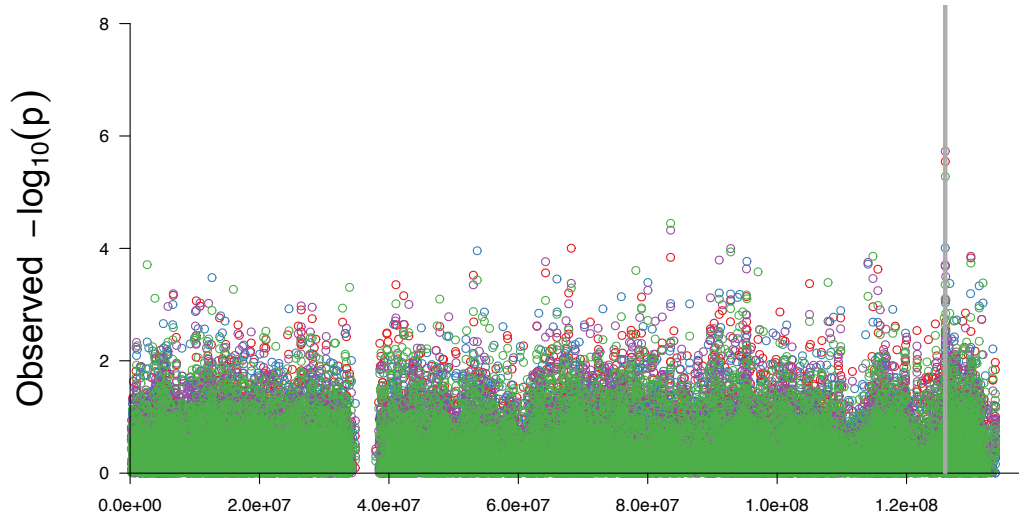
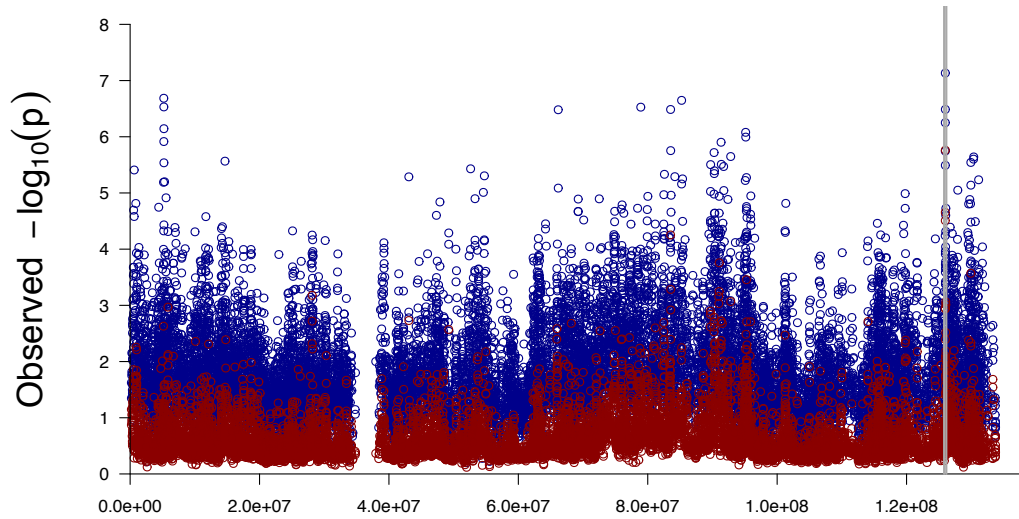


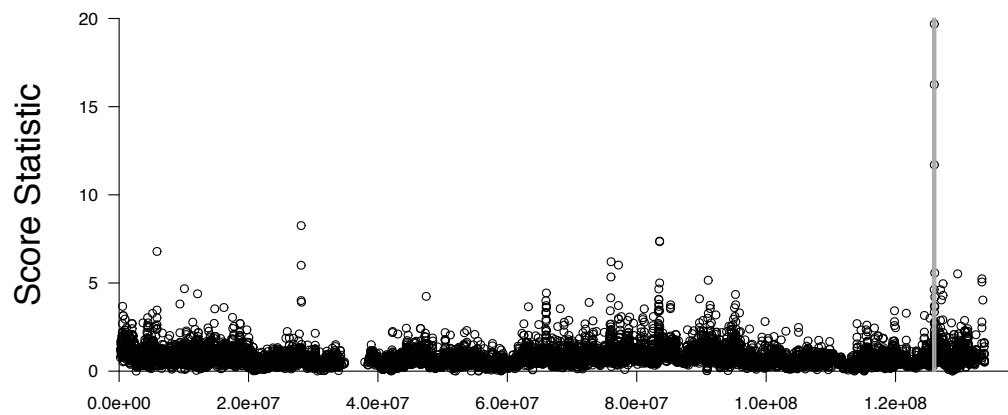
Fig. 4.16 Chr 12 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with $\psi_{\min 1}$. Gray bars mark *TMEM132B*.



(a) SMT Manhattan Plot



(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$



(c) Score Statistic for $y^T M y$

Fig. 4.17 Hit-plots for *TMEM132B*: Gray vertical bars denote the limits of the *TMEM132B* coding region. $\psi_{\min 1}$ was used to generate both plots.

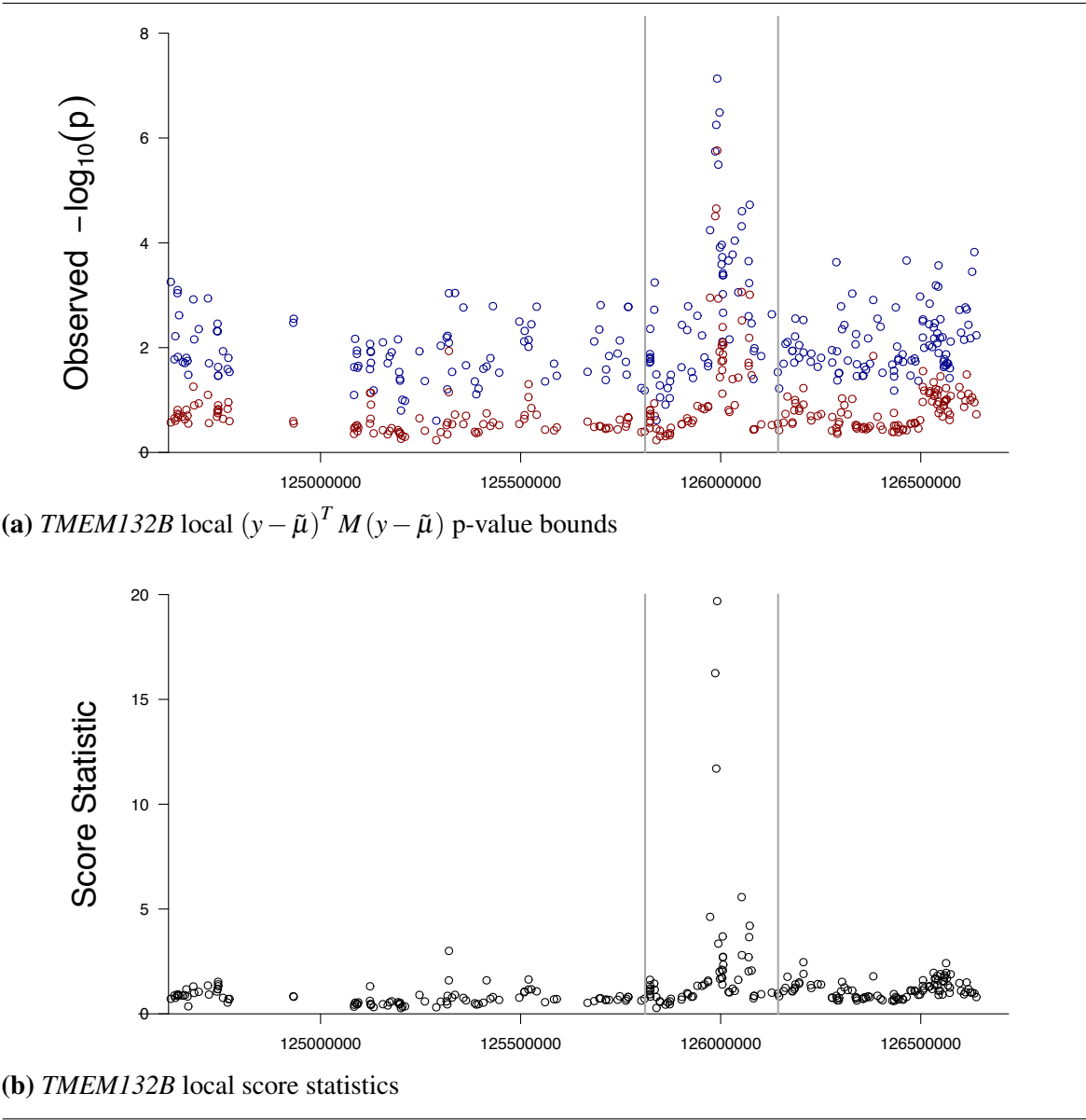
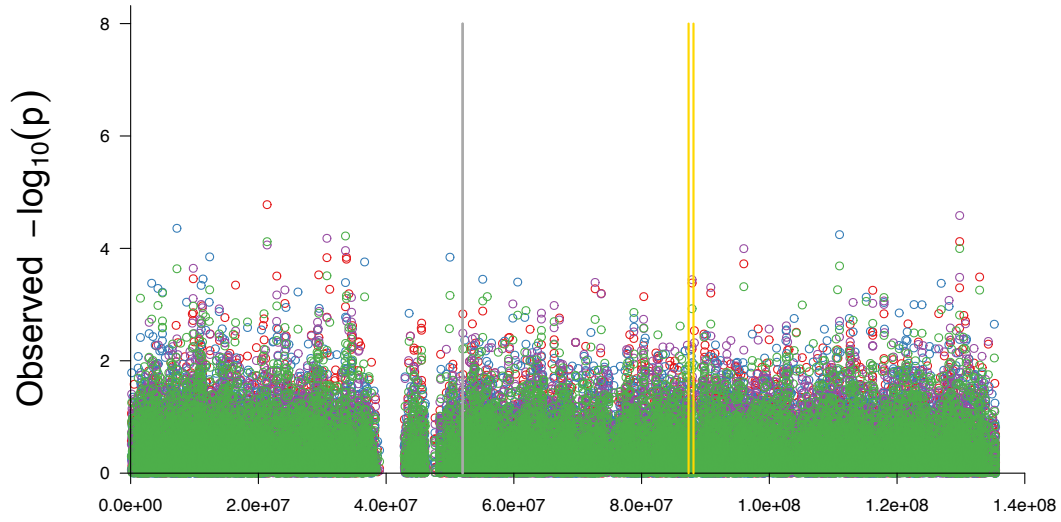
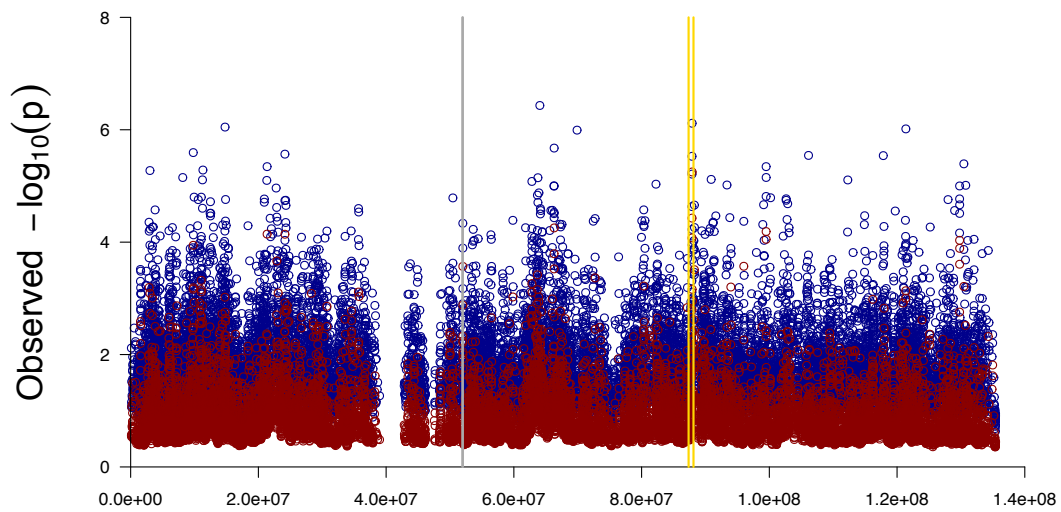


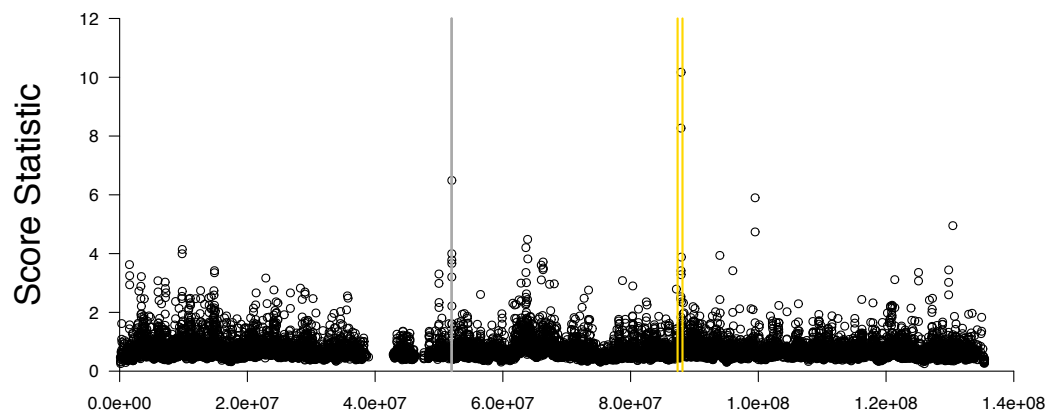
Fig. 4.18 Chr 10 SMT vs. Quad. Form Screening: Bounds and score test shown for M obtained with ψ_{add} . Gray bars mark *ASAH2*. Gold bars mark *GRID1*.



(a) SMT Manhattan Plot

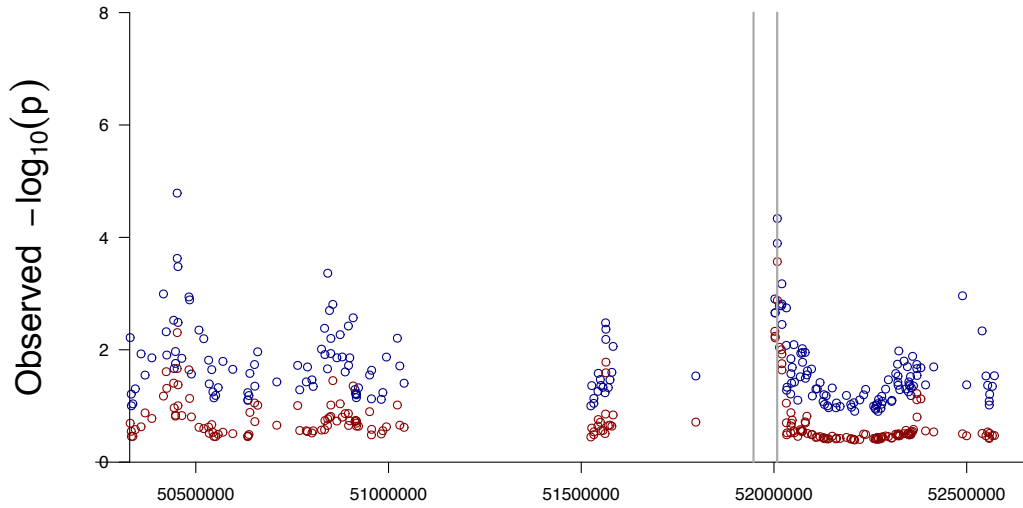


(b) Upper (red) and Lower (blue) P-value Bounds on $(y - \tilde{\mu})^T M (y - \tilde{\mu})$

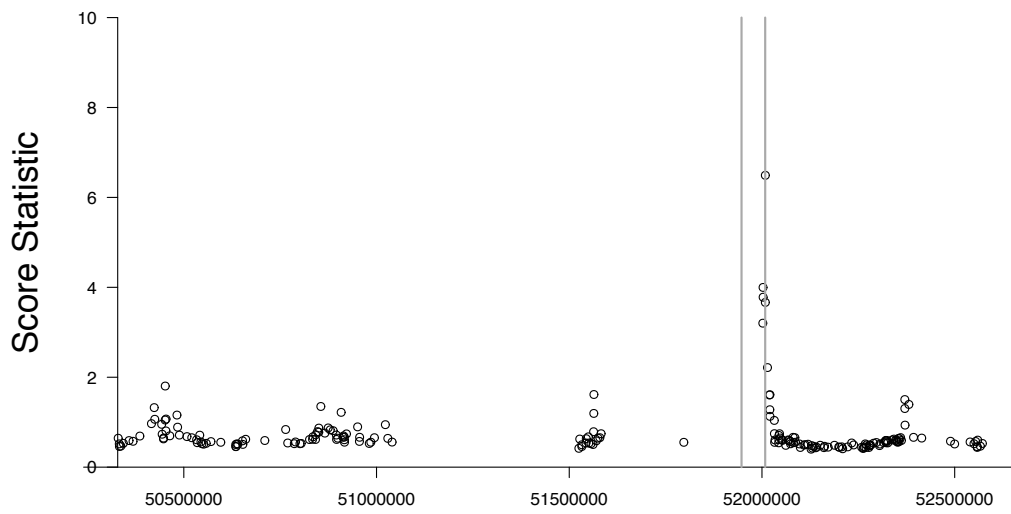


(c) Score Statistic for $y^T M y$

Fig. 4.19 Hit-plots for *ASAH2*: Gray vertical bars denote the limits of the *ASAH2* coding region. ψ_{add} was used to generate both plots.

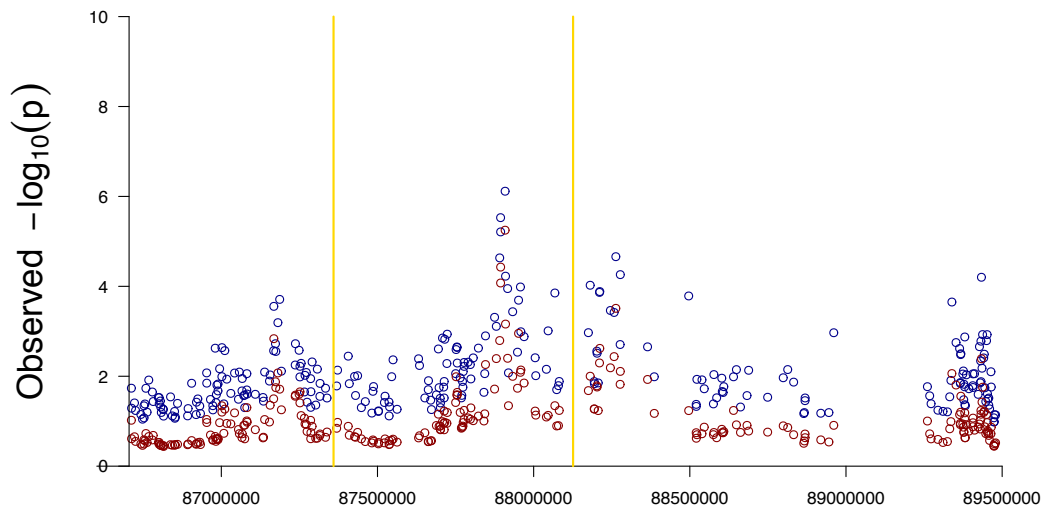


(a) *ASA H2* local $(y - \tilde{\mu})^T M (y - \tilde{\mu})$ p-value bounds

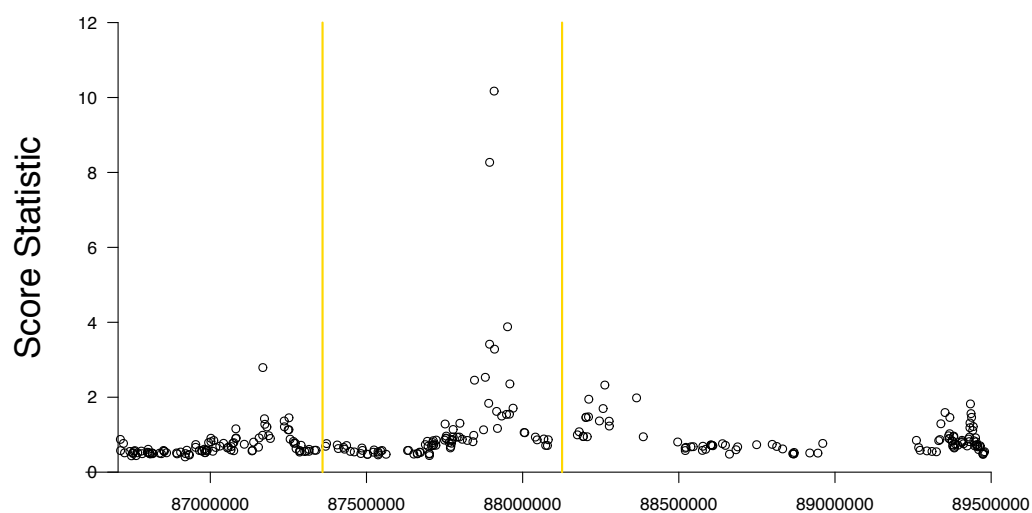


(b) *ASA H2* local score statistics

Fig. 4.20 Hit-plots for *GRID1*: Gold vertical bars denote the limits of the *GRID1* coding region. ψ_{add} was used to generate both plots.



(a) *GRID1* local $(y - \tilde{\mu})^T M (y - \tilde{\mu})$ p-value bounds



(b) *GRID1* local score statistics

4.4 Discussion

4.4.1 Comparing to SMTs

We have recovered classic association signals for malaria susceptibility and identified a few exciting new candidates. Much further indepth analysis of the distance matrices underlying our putative hits will need to be conducted in order to identify and charaterize the haplotypes driving these signals.

More importantly, from this pilot, we have verified that our overall approach can identify signals in a real dataset. Most of the association signals featured above were strongest when $\psi_{\min1}$ or $\psi_{\min2}$ were used, with the one exception of ψ_{add} . The fact that the dediploidizing functions used had such a strong effect on the results validates our attention given to them. It also warrants extensive simulations to investigate the performance of these various dediploidizing functions and verify the assumptions underlying their derivation. It would be helpful for practitioners to provide one or two default dediploidizing functions that work in the vast majority of cases since using too many functions on the same dataset increases the multiple testing burden.

At least in its current form, our haplotype approach is not strictly better than SMTs: our approach missed the signal at the glycophorin region which the SMTs detected. While this result may be improved if we optimize the HMM parameters, it is important to note that our approach need not be in opposition to SMTs. Since our model for testing local matrices M is simply an extension of logistic regression, we can fit a SMT model at a given position and see if there is a SMT association present. Then, taking the fitted SMT model as our new null model, we can test whether the local haplotype structure captured in distance matrix M shows evidence of a more sophisticated signal of association. Recall from Chapter 2, that we can construct a version of our distance matrix without including information from the SNP at

that precise locus (Equation (2.21)). This matrix could then be added into a model with the core SNP as a separate predictor while avoiding redundancy and excess collinearity.

4.4.2 P-values vs. Score Statistics

Perhaps the most unexpected statistical feature of the Manhattan plots presented is the parsimony of the score statistics. Complicating the situation, there is inconsistency in the loci reported by the p-value bounds and the score statistics across each chromosome. Even over a small region, score statistics can localize association signals to different SNPs than p-value bounds. As emphasized in Figure 4.7, the p-value bounds and the score statistic for a given distance matrix M can report conflicting results.

For each chromosome we analyzed, the score statistics tended to only highlight a few loci on each chromosome and remain virtually flat for the remainder of the chromosome. This putative "signal-to-noise" ratio was qualitatively better than that of the SMT p-values and our p-value bounds. Of course, beyond the known associated genes, we cannot be sure whether the loci that the score statistic isolated are true signals, but this observation still warrants further investigation.

Theoretically, it is not as principled from a Fisherian testing perspective to directly interpret a test statistic rather than its tail probability under the null. Measuring signals in the common currency of p-values might be especially important in our applications since each local distance matrix we form has a slightly different spectrum. Thus, each quadratic form has a different null distribution, which may make raw score statistics incomparable. However, since the score statistic is a first order approximation to the likelihood ratio (the ratio between the likelihood of the best model in the alternative model space and the likelihood of the best model in the null model space). One may argue that it is a measure of the evidence of association that may be taken on its own merit.

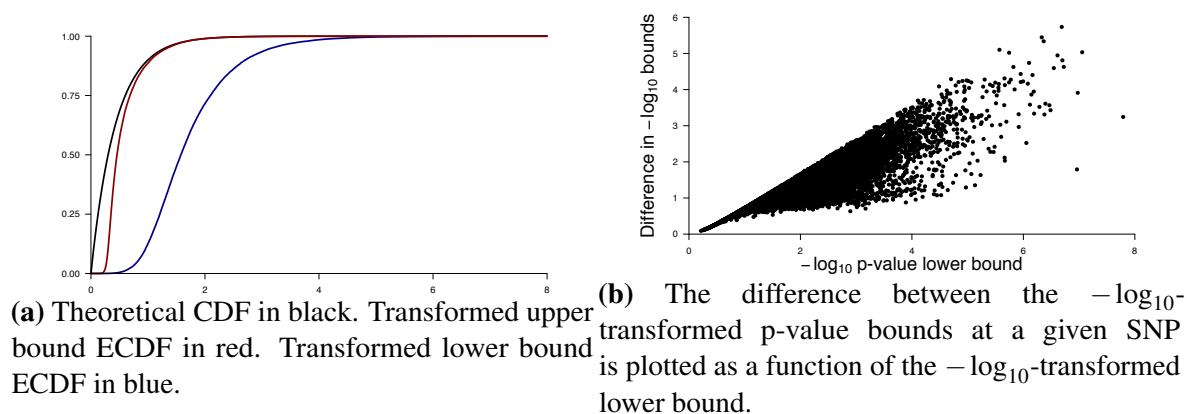
To help explain the discrepancy between p-value bounds and score statistics, note that the score statistic directly incorporates how similar the local distance matrix is to the global population structure via the term $\frac{1}{1-r_{\tilde{\alpha},M}^2}$. This information is partly propagated to our p-value bounds via the mean vector under the null $\tilde{\mu}$. The derivation and definition of $r_{\tilde{\alpha},M}^2$ are in the Appendix. This factor essentially rescales the variance so that local graphs whose leading eigenvectors closely align with the global principal components will not be penalized as strongly as local graphs whose leading eigenvectors are orthogonal to the global principal components. Ostensibly, this might lead one to presume that the score statistic would yield many false positives by boosting signals for local graphs whose structure resembles the genome-wide relatedness structure. Surprisingly, the score statistic remained relatively flat across most of the genome. As a reminder, we did not artificially introduce the $r_{\tilde{\alpha},M}^2$ term, but rather, it arose directly from taking derivatives of the likelihood under the null when we derived the score statistic. Quantifying the reliability of the score statistic is a subject for future work.

4.4.3 P-value Inflation and Bound Effectiveness

As the reader may have noticed by looking at the upper p-value bounds plotted in red on the Manhattan plots presented, our p-values are inflated by at least two orders of magnitude. Note that in our Manhattan plots, lower bounds on our p-values, after being $-\log_{10}$ -transformed, are plotted in blue as upper bounds on the evidence of association. Of course the exact amount of inflation cannot be determined without fully eigendecomposing each matrix and calculating each p-value exactly, but the CDF of the lower bounds clearly shows that there's inflation off of the theoretical expectation for transformed uniform p-values. See Figure 4.21a. This reflects the fact that our current implementation does not fully adjust for population structure.

Moving forward, there are several means by which we can hope to better correct for population structure. Since our local distance matrices are so much more information rich than single SNPs, it is not obvious that the top 10 or so eigenvectors of a single global matrix will be able to fully account for the confounding that could arise as a by-product of the fine resolution of our approach. A relatively straightforward idea to help account for the finer resolution of our distance matrices is to use top eigenvectors from a variety of local matrices on other chromosomes. Since local matrices are likely to capture a wider variability of "normal" relatedness than a single global matrix, their top eigenvectors may help to better control for local confounding. If we wish to remove non-linear confounding effects, it may also be important to include $M_{local}\tilde{\mu}$ as an additional background predictor. If we modified our HMM implementation to include an EM procedure that could estimate the prior copying probabilities, we could even calculate a global distance matrix M_{global} and include $M_{global}\tilde{\mu}$ as a predictor to help account for confounding.

Fig. 4.21 ECDF of p-value Bounds: **a)** We compare the ECDF of the theoretical distribution of our $-\log_{10}$ -transformed p-values ($-\log_{10}(U) \sim \text{Expo}(\log(10))$ if $U \sim \text{Unif}$) against the empirical CDF of the $-\log_{10}$ -transformed p-value upper bounds (red) and lower bounds (blue) we obtained across chromosome 2. Inflation is evident in the discrepancy between the ECDF of the upper bounds and the theoretical CDF near the origin. **b)** The gap between the significance bounds at a given SNP is shown as a function of significance. Larger gaps between the bounds can occur at SNPs where the $-\log_{10}$ -transformed p-value lower bound is higher.



Another important limitation is that our p-value bounds can be rather far apart, limiting our ability to estimate the underlying p-values (see Figure 4.21b). One direct way to attempt to remedy this would be to increase the number of eigenvalues and eigenvectors evaluated at each position; however this would greatly add to the computational burden. One optimization that we can make without computing additional eigenvalues involves using Lemma 3.1.2, which yields a tighter bound, rather than Lemma 3.1.1. Since Lemma 3.1.2 involves integration where as Lemma 3.1.1 only requires one-dimensional optimization, it was easier to implement the approach presented in Lemma 3.1.1 in time for this pilot study.

A more effective strategy to overcome the inflation and the lack of precision in our current p-value bounds is to do a few successive screens. While the p-value bounds that we obtained with the top 50 eigenvalues may not have proven very tight, they succeeded in screening out roughly 1 in 250 loci (fewer than 1 in 250 loci had a lower bound on their p-value below 10^{-5} on chromosome 2). This means that if we only use 50 eigenvalues to eliminate loci in our first pass, we can then run a second iteration just targeting the remaining candidate SNPs and evaluate 500 eigenvalues for each of them. Since weakly associated SNPs tend to be clustered together, we will get a marked improvement in the HMM speed by moving many loci at a time during the second screen.

While several aspects of this pilot model require improvement before mainstream application, this trial shows that our overall approach of building HMM-based distance matrices and applying our screening methods to them has the promise to identify real associations.

4.5 Appendix

Proposition. *The score statistic for τ under the logistic automodel likelihood in Equation (4.2) is*

$$S_y = \frac{1}{1 - r_{\tilde{\alpha}, M}^2} \left(\frac{y^T M y - \mathbb{E}_{\tilde{\alpha}}(Y^T M Y)}{\sqrt{\text{Var}_{\tilde{\alpha}}(Y^T M Y)}} \right)^2. \quad (4.7)$$

Proof:

Starting with our log likelihood:

$$\mathcal{L}_\theta = \log[P(y|\alpha, \tau)] = y^T (A\alpha + \tau M y) - \log \left[\sum_z \exp(z^T (A\alpha + \tau M z)) \right]$$

Here we use θ to refer to the total set of parameters $\theta = \{\alpha, \tau\}$.

$$\nabla \mathcal{L}_\theta = \begin{pmatrix} A^T y \\ y^T M y \end{pmatrix} - \mathbb{E}_\theta \begin{pmatrix} A^T z \\ z^T M z \end{pmatrix}$$

Since $-\nabla \nabla \mathcal{L}_\theta$ turns out not to be a function of y , $\mathcal{I}_\theta = \mathbb{E}_y(-\nabla \nabla \mathcal{L}_\theta) = -\nabla \nabla \mathcal{L}_\theta$.

$$\mathcal{I}_\theta = \mathbb{E}_\theta \left(\begin{bmatrix} A^T z \\ z^T M z \end{bmatrix} \begin{bmatrix} A^T z \\ z^T M z \end{bmatrix}^T \right) - \left[\mathbb{E}_\theta \begin{pmatrix} A^T z \\ z^T M z \end{pmatrix} \right] \left[\mathbb{E}_\theta \begin{pmatrix} A^T z \\ z^T M z \end{pmatrix} \right]^T$$

$$\text{If we let } U_\theta = \mathbb{E}_\theta \left(\begin{bmatrix} A^T z \\ z^T M z \end{bmatrix} \begin{bmatrix} A^T z \\ z^T M z \end{bmatrix}^T \right) \quad \text{and} \quad V_\theta = \mathbb{E}_\theta \begin{pmatrix} A^T z \\ z^T M z \end{pmatrix},$$

we can simply write $\mathcal{I}_\theta = U_\theta - V_\theta V_\theta^T$.

By Jensen's Inequality, each term of the Fisher Information is non-negative, meaning that the Hessian of the log likelihood has non-positive entries at all points of the parameter space. This confirms that the log-likelihood function is convex with one global maximum.

Since our log likelihood takes the form of $\theta^T \eta - \log Z(\theta)$, where θ is a vector of parameters and η is a vector of statistics, our likelihood is an Exponential Family member. For all Exponential Family Members, the likelihood ratio test statistic Λ can be written as

$$\Lambda_{\hat{\theta}} = 2 (\hat{\theta}^T \eta - \log Z(\hat{\theta}) - (\tilde{\theta}^T \eta - \log Z(\tilde{\theta})))$$

where $\hat{\theta}$ is the MLE under the alternative hypothesis and $\tilde{\theta}$ is the constrained MLE found in the subspace dictated by the null hypothesis. Consider Taylor expanding $\Lambda_{\hat{\theta}}$ from $\tilde{\theta}$, treating it as a function of $\hat{\theta}$:

$$\begin{aligned} \Lambda_{\hat{\theta}} &\approx 2 \left((\hat{\theta} - \tilde{\theta})^T \nabla \Lambda_{\tilde{\theta}} + \frac{1}{2} (\hat{\theta} - \tilde{\theta})^T \nabla \nabla \Lambda_{\tilde{\theta}} (\hat{\theta} - \tilde{\theta}) \right) \\ &\approx 2 \left((\hat{\theta} - \tilde{\theta})^T (\eta - \mathbb{E}_{\tilde{\theta}}(\eta)) - \frac{1}{2} (\hat{\theta} - \tilde{\theta})^T \mathcal{J}_{\tilde{\theta}} (\hat{\theta} - \tilde{\theta}) \right) \end{aligned}$$

If we assume that the paraboloid surface given by this approximation closely resembles the log-likelihood surface, then we can estimate $\Lambda_{\hat{\theta}}$ by finding the height of the apex of this paraboloid and subtracting the log-likelihood at $\tilde{\theta}$. Note that in the asymptotic limit where we would have many observations of y rather than just one, then the log-likelihood surface would converge to a quadratic around $\hat{\theta}$ (under the usual regularity assumptions) and our quadratic approximation to the log-likelihood surface would be accurate. With this motivation, we set the derivative of our approximation to zero and find

$$\hat{\theta}_{max} = \mathcal{J}_{\tilde{\theta}}^{-1} (\eta - \mathbb{E}_{\tilde{\theta}}(\eta)) + \tilde{\theta}$$

Substituting this yields:

$$2 \left((\mathcal{J}_{\tilde{\theta}}^{-1} (\eta - \mathbb{E}_{\tilde{\theta}}(\eta)))^T (\eta - \mathbb{E}_{\tilde{\theta}}(\eta)) - \frac{1}{2} (\mathcal{J}_{\tilde{\theta}}^{-1} (\eta - \mathbb{E}_{\tilde{\theta}}(\eta)))^T \mathcal{J}_{\tilde{\theta}} (\mathcal{J}_{\tilde{\theta}}^{-1} (\eta - \mathbb{E}_{\tilde{\theta}}(\eta))) \right)$$

$$= (\eta - \mathbb{E}_{\tilde{\theta}}(\eta))^T \mathcal{J}_{\tilde{\theta}}^{-1} (\eta - \mathbb{E}_{\tilde{\theta}}(\eta))$$

This form matches that of a classic Score Statistic, otherwise known as the Lagrange-Multiplier Statistic, which is typically written as

$$\mathcal{S}_y = \nabla \mathcal{L}_{\theta}^T \mathcal{J}_{\theta}^{-1} \nabla \mathcal{L}_{\theta} \big|_{\theta=\tilde{\theta}}$$

where $\tilde{\theta}$ is still the MLE under the null constraint. The subscript y is used to emphasize the fact that $\mathcal{S}_y$ is a function of the observed data y (a statistic).

Since the Score Statistic evaluates gradient and Fisher Information at the constrained MLE, we have that

$$\nabla \mathcal{L}_{\tilde{\theta}} = \begin{pmatrix} 0 \\ y^T M y - \mathbb{E}_{\tilde{\theta}}(z^T M z) \end{pmatrix} \quad \text{which implies} \quad V_{\tilde{\theta}} = \begin{pmatrix} A^T y \\ \mathbb{E}_{\tilde{\theta}}(z^T M z) \end{pmatrix}$$

Thus,

$$\mathcal{S}_{\tilde{\theta}} = \nabla \mathcal{L}_{\tilde{\theta}}^T \mathcal{J}_{\tilde{\theta}}^{-1} \nabla \mathcal{L}_{\tilde{\theta}} = (y^T M y - \mathbb{E}_{\tilde{\theta}}(z^T M z))^2 [\mathcal{J}_{\tilde{\theta}}^{-1}]_{22}$$

where $[\mathcal{J}_{\tilde{\theta}}^{-1}]_{22}$ denotes the lower-left corner entry of the inverse of the Fisher Information Matrix. Plugging in $V_{\tilde{\theta}}$ yields,

$$\begin{aligned} \mathcal{S}_{\tilde{\theta}} &= \mathbb{E}_{\theta} \left(\begin{bmatrix} A^T z \\ z^T M z \end{bmatrix} \begin{bmatrix} A^T z \\ z^T M z \end{bmatrix}^T \right) - \begin{bmatrix} A^T y \\ \mathbb{E}_{\tilde{\theta}}(z^T M z) \end{bmatrix} \begin{bmatrix} A^T y \\ \mathbb{E}_{\tilde{\theta}}(z^T M z) \end{bmatrix}^T = \\ &= \begin{bmatrix} \langle A^T z z^T A \rangle - A^T y y^T A & \langle A^T z z^T M z \rangle - A^T y \langle z^T M z \rangle \\ \langle z^T M z z^T A \rangle - \langle z^T M z \rangle y^T A & \langle (z^T M z)^2 \rangle - \langle z^T M z \rangle^2 \end{bmatrix} \end{aligned}$$

where we use $\langle \rangle$ to abbreviate the expectation with respect to $\pi(y|\tilde{\theta})$. By definition, $\langle (z^T Mz)^2 \rangle - \langle z^T Mz \rangle^2 = \text{Var}_{\tilde{\theta}}(z^T Mz)$. Using block matrix inversion, we rewrite $[\mathcal{J}_{\tilde{\theta}}^{-1}]_{22}$ as

$$\left(\text{Var}_{\tilde{\theta}}(z^T Mz) - [\langle A^T z z^T Mz \rangle - A^T y \langle z^T Mz \rangle]^T [\mathcal{J}_{\tilde{\theta}11}]^{-1} [\langle A^T z z^T Mz \rangle - A^T y \langle z^T Mz \rangle] \right)^{-1}$$

Therefore, our score statistic can simply be expressed as

$$\mathcal{S}_{\tilde{\theta}} = \frac{(y^T M y - \langle z^T M z \rangle)^2}{\text{Var}_{\tilde{\theta}}(z^T Mz) - [\langle A^T z z^T Mz \rangle - A^T y \langle z^T Mz \rangle]^T [\mathcal{J}_{\tilde{\theta}11}]^{-1} [\langle A^T z z^T Mz \rangle - A^T y \langle z^T Mz \rangle]}$$

where

$$[\mathcal{J}_{\tilde{\theta}11}]^{-1} = [\langle A^T z z^T A \rangle - A^T y y^T A]^{-1}$$

. This can be rewritten as

$$\mathcal{S}_y = \frac{1}{1 - r_{\tilde{\theta}}^2} \left(\frac{y^T M y - \mathbb{E}_{\tilde{\theta}}(z^T M z)}{\sqrt{\text{Var}_{\tilde{\theta}}(z^T M z)}} \right)^2$$

where

$$r_{\tilde{\theta}}^2 = \frac{[\langle A^T z z^T Mz \rangle - A^T y \langle z^T Mz \rangle]^T [\langle A^T z z^T A \rangle - A^T y y^T A]^{-1} [\langle A^T z z^T Mz \rangle - A^T y \langle z^T Mz \rangle]}{\langle (z^T Mz)^2 \rangle - \langle z^T Mz \rangle^2}.$$

This quantity $r_{\tilde{\alpha},M}^2$ is the same as the coefficient of determination from classic linear regression where $y^T M y$ is the outcome being predicted. Hence, $r_{\tilde{\alpha},M}^2$ is a measure of how well the background covariates can predict $y^T M y$. This function often appears in linear regression where it can be expressed as the proportion of the variation in $y^T M y$ that can be

“explained” by the background covariates. It is helpful for interpretation to note that we could also estimate it by a simple Monte Carlo scheme:

1. Draw S samples from $\pi(y|\tilde{\theta})$ and index them by $s \in \{1, \dots, S\}$.
2. For each y_s , calculate the scalar $y_s = y_s^T M y_s$ and the vector $\vec{x}_s = A^T y_s$.
3. Regress y_s onto $\vec{x}_s$ using standard linear regression and store the calculated coefficient of determination r_s^2 .
4. Estimate $r_{\alpha, M}^2 \approx \frac{1}{S} \sum_s r_s^2$

Chapter 5

Conclusion & Future Directions

We have developed several theoretical and computational components which together enable a scalable approach to genome-wide association studies that directly models recombination and captures local haplotype structure.

In Chapter 2, we recasted the likelihood of an HMM based on the Li & Stephens model in terms of penalty functions that reflect local haplotype similarity. We derived new checkpointing algorithms, which are applicable to any HMM, to avoid having to store a $N \times N \times L$ array. As part of a collaboration with Louis Aslett, we developed a highly optimized implementation of our HMM that can be easily accessed via R. Future directions for this work include benchmarking and tuning our implementation for larger GWAS datasets as well as working to improve our haplotype distances by assessing their performance in simulation. Several extensions to potentially improve the robustness are possible: we may want to say that haplotype i copies j at l only if i continues to copy j for at least one or two SNPs in either direction. Such extensions will be a focus of future work.

In Chapter 3, we presented an approach to screening Gaussian quadratic forms based on Krylov subspace methods and a novel concentration inequality which we extended to a much more general class of quadratic forms via two invariance principles and a Rao-Blackwellization routine. We hope to demonstrate the utility of this approach in applications

in network analysis and classification. Early computer scientists coined the adage, "never compute something more precisely than you have to." However, this advice has seems to only partially have worked its way into the math and models of statistical inference. In several applications beyond quadratic forms, it may be sufficient for an investigator to know if a p-value is within a certain tight window while requiring a fraction of the computational cost of full precision. We believe that our approach of using a decomposition to capture the major components of a statistic (in our case, the leading eigenvectors and eigenvalues) and then bounding the remaining components with a concentration inequality may prove fruitful in several other domains.

In Chapter 4, we developed a conceptual framework for working with distances between haplotypes that allowed us to bring our haplotype distance matrices from Chapter 2 into the GWAS setting with our quadratic form screening procedure. To our knowledge, our dediploiding framework is distinct from common paradigms for thinking about haplotypic relationships. This work motivated four functions, two of which were essential to identifying the signals we found in our pilot study. As mentioned in Chapter 4, there are several domains in which our pilot method must be improved, but it is capable of identifying known associated genes and shows promise in uncovering new association signals.

One particularly exciting line of research for this thesis includes detecting adaptive introgression between populations. Adaptive introgression events occur when an advantageous haplotype is introduced to population B from population A via migration which then arises to a high frequency in population B due to selection [83]. This phenomenon creates a unique signal in the genome at the position of the selected haplotype. While individuals from populations A and B may look genetically dissimilar at most regions of the genome, at this selected locus, individuals will appear closely related across populations. In order to find such signals, we don't need dediploiding functions but can model the haplotypes from both populations directly (assuming we can phase them). Then we can treat population status

as a "phenotype," except we would be looking for signals in the opposite direction from what we look for in a GWAS. To detect introgression, we don't want to identify regions where close genetic relatedness implies the same population status: individuals from both populations with tend to be more related to those from their own population along the entire genome. For introgression detection, we want to identify regions where genetic relatedness is less predictive of population status than usual. This can be done by simply calculating a chromosome-wide relatedness matrix between haplotypes and including the top principal components in a logistic regression model predicting population status. Taking this as our null model, we then screen along the genome for local haplotype distance matrices that have significantly large quadratic forms.

Within the realm of association testing and in other applications, we hope that approximating local phylogenies with HMM-derived distance matrices will yield a fruitful approach for modeling future genomic datasets.

References

- [1] PM Visscher, MA Brown, MI McCarthy, and J Yang. Five years of GWAS discovery. *American journal of human genetics*, 90(1):7–24, 2012.
- [2] SV Nuzhdin, ML Friesen, and LM McIntyre. Genotype–phenotype mapping in a post-GWAS world. *Trends in Genetics*, 28(9):421–426, 2012.
- [3] Y Ma, F Qin, and LP Tran. Contribution of genomics to gene discovery in plant abiotic stress responses. *Molecular plant*, 5(6):1176–8, 2012.
- [4] D Altshuler, MJ Daly, and ES Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, 2008.
- [5] KA Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2017.
- [6] JK Pritchard and NJ Cox. The allelic architecture of human disease genes: common disease–common variant or not? *Human molecular genetics*, 11(20):2417–2423, 2002.
- [7] TG Clark, AE Fry, S Auburn, S Campino, M Diakite, A Green, A Richardson, YY Teo, K Small, J Wilson, M Jallow, F Sisay-Joof, M Pinder, P Sabeti, DP Kwiatkowski, and KA Rockett. Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility. *European journal of human genetics : EJHG*, 17(8):1080–5, 2009.
- [8] S Konigorski, YE Yilmaz, and T Pischon. Comparison of single-marker and multi-marker tests in rare variant association studies of quantitative traits. *PLOS ONE*, 12(5):e0178504, 2017.
- [9] P Loh, G Tucker, BK Bulik-Sullivan, BJ Vilhjálmsón, HK Finucane, RM Salem, DI Chasman, PM Ridker, BM Neale, B Berger, N Patterson, and AL Price. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015.
- [10] M Stephens and DJ Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.
- [11] V Botta, G Louppe, P Geurts, and L Wehenkel. Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies. *PLoS ONE*, 9(4):e93379, 2014.
- [12] MC Wu, S Lee, T Cai, Y Li, M Boehnke, and X Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

- [13] S Lee, MJ Emond, MJ Bamshad, KC Barnes, MJ Rieder, DA Nickerson, DC Christiani, MM Wurfel, and X Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.
- [14] D Zhang and X Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.
- [15] K Liu, S Fast, M Zawistowski, and NL Tintle. A geometric framework for evaluating rare variant tests of association. *Genetic epidemiology*, 37(4):345–57, 2013.
- [16] J Listgarten, C Lippert, CM Kadie, RI Davidson, E Eskin, and D Heckerman. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–6, 2012.
- [17] BL Browning and SR Browning. Efficient Multilocus Association Testing for Whole Genome Association Studies Using Localized Haplotype Clustering. 375:365–375, 2007.
- [18] S Zöllner and JK Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 2005.
- [19] MJ Minichiello and R Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, 2006.
- [20] Y Wu. Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. *Journal of computational biology : a journal of computational molecular cell biology*, 15(7):667–84, 2008.
- [21] RC Griffiths and P Marjoram. An ancestral recombination graph. *Progress in Population Genetics and Human Evolution*, eds Donnelly P., Tavaré S., editors., 1997.
- [22] AL Price, NA Zaitlen, D Reich, and N Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 2010.
- [23] HM Kang, NA Zaitlen, CM Wade, A Kirby, D Heckerman, MJ Daly, and E Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–1723, 2008.
- [24] HM Kang, JH Sul, SK Service, NA Zaitlen, S Kong, NB Freimer, C Sabatti, and E Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–54, 2010.
- [25] J Yu, G Pressoir, WH Briggs, I Vroh Bi, M Yamasaki, JF Doebley, MD McMullen, BS Gaut, DM Nielsen, JB Holland, S Kresovich, and ES Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–8, 2006.
- [26] G McVean. A genealogical interpretation of principal components analysis. *PLoS genetics*, 5(10):e1000686, 2009.

- [27] AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, and D Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–9, 2006.
- [28] I Mathieson and G Mcvean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–246, 2012.
- [29] H Xu and Y Guan. Detecting local haplotype sharing and haplotype association. *Genetics*, 197(3):823–38, 2014.
- [30] N Li and M Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165:2213–2233, 2003.
- [31] YS Song. Na Li and Matthew Stephens on Modeling Linkage Disequilibrium. *Genetics*, 203(3):1005–1006, 2016.
- [32] P Loh, P Danecek, PF Palamara, C Fuchsberger, YA Reshef, HK Finucane, S Schoenherr, L Forer, S McCarthy, GR Abecasis, R Durbin, and AL Price. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.
- [33] G Bhatia, A Gusev, P Loh, BJ Vilhjálmsson, S Ripke, E Stahl, M Daly, TR de Candia, KS Kendler, MC O’Donovan, S Hong Lee, NR Wray, BM Neale, MC Keller, NA Zaitlen, B Pasaniuc, J Yang, and AL Price. Haplotypes of common SNPs can explain missing heritability of complex diseases. *Preprint available on BioRxiv.*, 2015.
- [34] N Li and M Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 2003.
- [35] P Fearnhead and P Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–318, 2001.
- [36] M Stephens and P Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635, 2000.
- [37] G McVean and NJ Cardin. Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459):1387–1393, 2005.
- [38] S Myers. Stochastic Models in Mathematical Genetics Lecture Notes, University of Oxford, Department of Statistics, 2014.
- [39] DJ Lawson, G Hellenthal, S Myers, and D Falush. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453, 2012.
- [40] J Besagt. Statistical Analysis of Non-lattice Data. *Journal of the Royal Statistical Society*, 24(3):179–195, 1975.
- [41] J Kelleher, AM Etheridge, and G McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):e1004842, 2016.

- [42] L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [43] R Wheeler and R Hughey. Optimizing reduced-space sequence analysis. *Bioinformatics*, 16(12):1082–90, 2000.
- [44] R Durbin, S Eddy, A Krogh, and G Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK, 1998.
- [45] P Hammarlund, AJ Martinez, AA Bajwa, DL Hill, E Hallnor, H Jiang, M Dixon, M Derr, M Hunsaker, R Kumar, RB Osborne, R Rajwar, R Singhal, R D’Sa, R Chappell, S Kaushik, S Chennupaty, S Jourdan, S Gunther, T Piazza, and T Burton. Haswell: The Fourth-Generation Intel Core Processor. *IEEE Micro*, 34(2):6–20, 2014.
- [46] J Terhorst, JA Kamm, and YS Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.
- [47] P Duchesne and P Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54(4):858–862, 2010.
- [48] D Calvetti, L Reichel, and DC Sorensen. An Implicitly Restarted Lanczos Method for Large Symmetric Eigenvalue Problems. *Etna*, 2:1–21, 1994.
- [49] RB Davies. Algorithm AS 155: The Distribution of a Linear Combination of χ^2 Random Variables. *Source Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323–333, 1980.
- [50] VI Rotar. Some Limit Theorems for Polynomials of Second Degree. *Theory of Probability & Its Applications*, 18(3):499–507, 1973.
- [51] T De Wet and JH Venter. Asymptotic distributions for quadratic forms with applications to tests of fit. *The Annals of Statistics*, 1(2):380–387, 1973.
- [52] N. G. Gamkrelidze and V.I. Rotar. On the Rate of Convergence in the Limit Theorem for Quadratic Forms. *Theory Probab. Appl.*, 22(2):394–397, 1978.
- [53] VI Rotar and TL Shervashidze. Some Estimates of Distributions of Quadratic Forms. *Theory of Probability & Its Applications*, 30:585–591, 1985.
- [54] VI Rotar. Limit theorems for polylinear forms. *Journal of Multivariate Analysis*, 9(4):511–530, 1979.
- [55] G Band, QS Le, L Jostins, M Pirinen, K Kivinen, M Jallow, F Sisay-Joof, K Bo-jiang, M Pinder, G Sirugo, D Conway, V Nyirongo, D Kachala, M Molyneux, T Taylor, C Ndila, N Peshu, K Marsh, TN Williams, D Alcock, R Andrews, S Edkins, E Gray, C Hubbard, A Jeffreys, K Rowlands, K Schuldt, TG Clark, KS Small, YY Teo, DP Kwiatkowski, KA Rockett, JC Barrett, and CCA Spencer. Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genetics*, 9(5):e1003509, 2013.

- [56] C C Holmes and N M Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Royal Statistical Society Series B (Statistical Methodology)*, 64(2):295–306, 2002.
- [57] L Cucala, JM Marin, CP Robert, and DM Titterton. A Bayesian Reassessment of Nearest-Neighbor Classification. *Journal of the American Statistical Association*, 104(485):263–273, 2009.
- [58] J Hughes, M Haran, and PC Caragea. Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871, 2011.
- [59] CC Holmes and NM Adams. Likelihood inference in nearest-neighbour classification models. *Biometrika*, 90(1):99–112, 2003.
- [60] PC Caragea and MS Kaiser. Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(3):281–300, 2009.
- [61] WHO. Malaria Fact Sheet, 2016.
- [62] WHO. World Malaria Report 2015. Technical report, 2015.
- [63] R Carter and KN Mendis. Evolutionary and Historical Aspects of the Burden of Malaria. *Clinical Microbiology Reviews*, 15(4):564–594, 2002.
- [64] PW Hedrick. Population genetics of malaria resistance in humans. *Heredity*, 107(4):283–304, 2011.
- [65] EK Karlsson, DP Kwiatkowski, and PC Sabeti. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, 15(6):379–393, 2014.
- [66] PC Sabeti, DE Reich, JM Higgins, HZP Levine, DJ Richter, and SF Schaffner. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419, 2002.
- [67] SS Shah, KA Rockett, M Jallow, F Sisay-Joof, KA Bojang, M Pinder, A Jeffreys, R Craik, C Hubbart, TE Wellems, DP Kwiatkowski, MD Cappellini, G Fiorelli, E Beutler, TJ Vulliamy, E Beutler, ET Nkhoma, C Poole, V Vannappagari, SA Hall, E Beutler, MJ Mackinnon, TW Mwangi, RW Snow, K Marsh, TN Williams, TI Sørensen, GG Nielsen, PK Andersen, TW Teasdale, DP Kwiatkowski, SA Tishkoff, R Varkonyi, N Cahinhinan, S Abbes, G Argyropoulos, G Destro-Bisol, PC Sabeti, DE Reich, JM Higgins, HZP Levine, DJ Richter, SF Schaffner, C Louicharoen, E Patin, R Paul, I Nuchprayoon, B Witoonpanich, C Peerapittayamongkol, U Bienzle, O Ayeni, A Lucas, L Luzzatto, C Ruwende, SC Khoo, RW Snow, SN Yates, D Kwiatkowski, S Gupta, A Guindo, RM Fairhurst, OK Doumbo, TE Wellems, DA Diallo, T Clark, A Fry, S Auburn, S Campino, M Diakite, A Green, MK Johnson, TD Clark, D Njama-Meya, PJ Rosenthal, S Parikh, A Manjurano, N Sepulveda, B Nadjm, G Mtove, H Wangai, C Maxwell, G Sirugo, IM Predazzi, J Bartlett, A Tacconelli, M Walther, SM Williams, SK Martin, LH Miller, D Alling, VC Okoye, GJ Esan, BO Osunkoya, AC Allison, DF Clyde, M Kruatrachue, P Charoenlarp, T Chongsuphajaisiddhi, C Harinasuta, HM Gilles, KA Fletcher, RG Hendrickse, R Lindner, S Reddy, N Allan, K Marsh,

- D Forster, C Waruiru, I Mwangi, M Winstanley, V Marsh, E Beutler, M Yeh, V Fairbanks, JC Gall, GJ Brewer, RJ Dern, J May, CG Meyer, L Grossterlinden, OG Ademowo, FP Mockenhaupt, PE Olumese, SS Shah, SAS Diakite, K Traore, M Diakite, DP Kwiatkowski, KA Rockett, SS Shah, A Macharia, J Makale, S Uyoga, K Kivinen, R Craik, Malaria Genomic Epidemiology Network, M Jallow, C Casals-Pascual, H Ackerman, B Walther, M Walther, M Pinder, CD Araujo, F Migot-Nabias, J Guitard, S Pelleau, T Vulliamy, R Ducrocq, S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, D Clayton, E Beutler, W Kuhl, GF Sáenz, W Rodríguez, E Beutler, W Kuhl, JL Vives-Corrons, JT Prchal, PJ Mason, JM Bautista, F Gilsanz, SK Mishra, CRJC Newton, M Friedman, G Pasvol, DJ Weatherall, RJ Wilson, A Allison, TN Williams, TW Mwangi, S Wambua, ND Alexander, M Kortok, RW Snow, J May, JA Evans, C Timmann, C Ehmen, W Busch, T Thye, E Beutler, R Dern, C Flanagan, TN Williams, TW Mwangi, DJ Roberts, ND Alexander, DJ Weatherall, S Wambua, J Carlson, GB Nash, V Gabutti, F al Yaman, M Wahlgren, R Cholera, NJ Brittain, MR Gillrie, TM Lopera-Mesa, SAS Diakité, T Arie, L Luzzatto, FA Usanga, S Reddy, EF Roth, C Raventos-Suarez, A Rinaldi, RL Nagel, MJ Friedman, M Cappadoro, G Giribaldi, E O'Brien, F Turrini, F Mannu, D Ulliers, JA Leopold, J Walker, AW Scribner, B Voetsch, YY Zhang, AJ Loscalzo, JA Leopold, A Dam, BA Maron, AW Scribner, R Liao, DE Handy, P Agudelo-Flórez, BT Costa-Carvalho, JA López, J Redher, PE Newburger, and ST Olalla-Saad. Heterogeneous alleles comprising G6PD deficiency trait in West Africa exert contrasting effects on two major clinical presentations of severe malaria. *Malaria Journal*, 15(1):13, 2016.
- [68] AE Fry, MJ Griffiths, S Auburn, M Diakite, JT Forton, A Green, A Richardson, J Wilson, M Jallow, F Sisay-Joof, M Pinder, N Peshu, TN Williams, K Marsh, ME Molyneux, TE Taylor, KA Rockett, and DP Kwiatkowski. Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Human Molecular Genetics*, 17(4):567–576, 2007.
- [69] KA Rockett, GM Clarke, K Fitzpatrick, C Hubbart, AE Jeffreys, K Rowlands, R Craik, M Jallow, DJ Conway, KA Bojang, M Pinder, S Usen, F Sisay-Joof, G Sirugo, O Toure, MA Thera, S Konate, S Sissoko, A Niangaly, B Poudiougou, VD Mangano, EC Bougouma, SB Sirima, D Modiano, LN Amenga-Etego, A Ghansah, K Koram, MD Wilson, A Enimil, J Evans, O Amodu, S Olaniyan, T Apinjoh, R Mugri, A Ndi, CM Ndila, S Uyoga, A Macharia, N Peshu, TN Williams, A Manjurano, E Riley, C Drakeley, H Reyburn, V Nyirongo, D Kachala, M Molyneux, SJ Dunstan, NH Phu, NTN Quyen, CQ Thai, TT Hien, L Manning, M Laman, P Siba, H Karunajeewa, S Allen, A Allen, T Davis, P Michon, I Mueller, A Green, S Molloy, KJ Johnson, A Kerasidou, V Cornelius, L Hart, A Vanderwal, M SanJoaquin, G Band, SQ Le, M Pirinen, N Sepúlveda, C Spencer, TG Clark, T Agbenyega, E Achidi, O Doumbo, J Farrar, K Marsh, T Taylor, and DP Kwiatkowski. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature Genetics*, 46(11):1197–1204, 2014.
- [70] EM Leffler, G Band, GBJ Busby, K Kivinen, QS Le, GM Clarke, KA Bojang, DJ Conway, M Jallow, F Sisay-Joof, EC Bougouma, VD Mangano, D Modiano, SB Sirima, E Achidi, TO Apinjoh, K Marsh, CM Ndila, N Peshu, TN Williams, C Drakeley, A Manjurano, H Reyburn, E Riley, D Kachala, M Molyneux, V Nyirongo, T Taylor, N Thornton, L Tilley, S Grimsley, E Drury, J Stalker, V Cornelius, C Hubbart, AE Jeffreys, K Rowlands, KA Rockett, CCA Spencer, DP Kwiatkowski, and Malaria

- Genomic Epidemiology Network. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, 356(6343):eaam6393, 2017.
- [71] J Bulp, M Jen, and K Matuszewski. Caring for Glucose-6-Phosphate Dehydrogenase (G6PD)-Deficient Patients: Implications for Pharmacy. *P & T : a peer-reviewed journal for formulary management*, 40(9):572–574, 2015.
- [72] SC Lu. Glutathione synthesis. *Biochimica et biophysica acta*, 1830(5):3143–3153, 2013.
- [73] R Arreola, B Valderrama, ML Morante, and E Horjales. Two mammalian glucosamine-6-phosphate deaminases: a structural and genetic study. *FEBS letters*, 551(1-3):63–70, 2003.
- [74] JM Berg, JL Tymoczko, GJ Gatto, and L Stryer. *Biochemistry*. W.H. Freeman & Company, 7th edition, 2015.
- [75] Y Kalmbach, ABW Boldt, R Fendel, B Mordmüller, PG Kremsner, and JFJ Kun. Increase in annexin V-positive B cells expressing LILRB1/ILT2/CD85j in malaria. *European cytokine network*, 17(3):175–180, 2006.
- [76] S Naitza, E Porcu, M Steri, DD Taub, A Mulas, X Xiao, J Strait, M Dei, S Lai, F Busonero, A Maschio, G Usala, M Zoledziewska, C Sidore, I Zara, M Pitzalis, A Loi, F Viridis, R Piras, F Deidda, MB Whalen, L Crisponi, A Concas, C Podda, S Uzzau, P Scheet, DL Longo, E Lakatta, GR Abecasis, A Cao, D Schlessinger, M Uda, S Sanna, and F Cucca. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS genetics*, 8(1):e1002480, 2012.
- [77] SL Deshmane, S Kremlev, S Amini, and BE Sawaya. Monocyte chemoattractant protein-1 (MCP-1): an overview. *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research*, 29(6):313–326, 2009.
- [78] JL Farlow, H Lin, L Sauerbeck, D Lai, DL Koller, E Pugh, K Hetrick, H Ling, R Kleinloog, P van der Vlies, P Deelen, MA Swertz, BH Verweij, L Regli, GJE Rinkel, YM Ruigrok, K Doheny, Y Liu, J Broderick, T Foroud, and FIA Study Investigators. Lessons Learned from Whole Exome Sequencing in Multiplex Families Affected by a Complex Genetic Disorder, Intracranial Aneurysm. *PLOS ONE*, 10(3):e0121104, 2015.
- [79] M Tani, T Sano, M Ito, and Y Igarashi. Mechanisms of sphingosine and sphingosine 1-phosphate generation in human platelets. *Journal of Lipid Research*, 46(11):2458–2467, 2005.
- [80] Y Osawa, H Uchinami, J Bielawski, RF Schwabe, YA Hannun, and DA Brenner. Roles for C 16 -ceramide and Sphingosine 1-Phosphate in Regulating Hepatocyte Apoptosis in Response to Tumor Necrosis Factor- α . *Journal of Biological Chemistry*, 280(30):27879–27887, 2005.
- [81] C Finney, C Hawkes, DC Kain, A Dhabangi, C Musoke, C Cserti-Gazdewich, T Oravec, WC Liles, and KC Kain. S1P is associated with protection in human and experimental cerebral malaria. *Molecular Medicine*, 17(7-8):717–725, 2011.

-
- [82] DR Schrider and AD Kern. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular biology and evolution*, 34(8):1863–1877, 2017.
- [83] F Racimo, D Marnetto, and E Huerta-Sánchez. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Molecular biology and evolution*, 34(2):296–317, 2017.