OPEN ACCESS Freely available online

PLoS one

Conditional Random Fields for Fast, Large-Scale Genome-Wide Association Studies

Jim C. Huang, Christopher Meek, Carl Kadie, David Heckerman*

Microsoft Research, Redmond, Washington, United States of America

Abstract

Understanding the role of genetic variation in human diseases remains an important problem to be solved in genomics. An important component of such variation consist of variations at single sites in DNA, or single nucleotide polymorphisms (SNPs). Typically, the problem of associating particular SNPs to phenotypes has been confounded by hidden factors such as the presence of population structure, family structure or cryptic relatedness in the sample of individuals being analyzed. Such confounding factors lead to a large number of spurious associations and missed associations. Various statistical methods have been proposed to account for such confounding factors such as linear mixed-effect models (LMMs) or methods that adjust data based on a principal components analysis (PCA), but these methods either suffer from low power or cease to be tractable for larger numbers of individuals in the sample. Here we present a statistical model for conducting genome-wide association studies (GWAS) that accounts for such confounding factors. Our method scales in runtime quadratic in the number of individuals being studied with only a modest loss in statistical power as compared to LMM-based and PCA-based methods when testing on synthetic data that was generated from a generalized LMM. Applying our method to both real and synthetic human genotype/phenotype data, we demonstrate the ability of our model to correct for confounding factors while requiring significantly less runtime relative to LMMs. We have implemented methods for fitting these models, which are available at http://www.microsoft.com/science.

Citation: Huang JC, Meek C, Kadie C, Heckerman D (2011) Conditional Random Fields for Fast, Large-Scale Genome-Wide Association Studies. PLoS ONE 6(7): e21591. doi:10.1371/journal.pone.0021591

Editor: Momiao Xiong, University of Texas, United States of America

Received February 15, 2011; **Accepted** June 3, 2011; **Published** July 12, 2011

Funding: Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors are affiliated with Microsoft Corporation. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: heckerma@microsoft.com

Introduction

Population structure, family structure and/or cryptic relatedness are well-known confounding factors that cause spurious associations to be found in GWAS [1–6]. Standard statistical hypothesis testing of association between markers and phenotypes can produce a large number of false positive associations, as SNP markers may be correlated with phenotype purely as a result of confounding factor effects. As the cost of genotyping drops and the sizes of such studies continue to grow above tens of thousands of individuals [7–9], the influence of such confounding effects on GWAS will become more acute, requiring statistical analysis methods that will both scale for large numbers of individuals while accounting for the confounders.

The standard techniques for dealing with confounding factors fall into several classes. An effective class of methods includes approaches formulated as LMMs [10], which model confounding factors using pairwise similarity measures between every pair of individuals. As the effects of confounders are all encoded in the set of SNPs carried by all individuals, the set of similarities can then be used in a regression model to distinguish between spurious and true SNP-phenotype associations. Other methods have been proposed that use a principal components analysis of individuals' SNPs [4], perform a post-hoc correction of test statistics such as Genomic Control [2], or cluster individuals before performing an aggregate association between clusters and phenotypes [11]. These methods, while accounting for confounding factors under different assumptions, have been shown to either suffer from insufficient statistical power when the confounding effects are strong [4,5] or are unable to fully capture their effects altogether, such that many false positives are produced [3,5,12]. In several recent studies [3,5,12,13], methods based on LMMs were found to produce fewer false positives and had higher statistical power as compared to other methods for modeling confounding factors, making LMMs a popular class of GWAS methods that have high statistical power and low false positive rates.

Although LMMs have been shown to effectively model and correct for confounding factors in GWAS, an important problem that remains to be solved is how to minimize the computational costs of such methods. Methods based on LMMs typically incur high computational costs, particularly for studies with larger numbers of individuals, as the matrix operations required for parameter estimation scale cubically with the number of individuals. In the regime where the number of individuals grows large and where confounding factors exert strong effects, this may hinder the applicability of LMMs. One possible approach to the above problem is to turn to alternative classes of models that allow us to model similarities between individuals in order to account for confounding factors in the data (as do LMMs) while eschewing the need for costly matrix operations during parameter estimation. In particular, probabilistic graphical models are a natural class of statistical models that allow both for modeling similarities between

individuals and fast parameter estimation. In this paper we propose a probabilistic graphical model and parameter estimation method for associating SNPs to phenotype that both accounts for confounding factors and runs significantly faster than current LMM-based methods for larger numbers of individuals, allowing the method to scale to larger study sizes. Unlike LMM-based methods (which present local optima in parameter estimation [3]) or PCA-based methods [4], our method for parameter estimation is not prone to local optima and is also guaranteed to yield unique, globally optimal parameter estimates. We will apply our model to real and synthetic human genotype datasets, where we show significantly lower runtimes for our method as compared to LMM-based methods for larger study sizes, with only a modest loss in statistical power relative to LMM-based methods when testing on synthetic data that was generated from a generalized LMM. Finally, we have implemented methods for fitting these models, which are available at http://www.microsoft.com/science.

Results

We present a model for relating individuals' phenotypic labels as a function of a given SNP marker and other covariates. The output of our model will be some statistic for the SNP marker, so that we can perform a GWAS by applying our model to each SNP marker in a large set of interest. Given a set of individuals, we assume that phenotypes consist of binary labels corresponding to the absence/presence of a phenotype in an individual, although the model can easily be generalized to polytomous discrete or continuous phenotypes. For a given locus, our model specifies a joint probability over individuals' observed phenotypes, conditioned on each individual's SNP and covariates. The joint probability will be a function of all pairs of individuals' phenotypes and each individual's SNP and covariates. Under our model, the contribution of each pair of individual phenotypes will increase or decrease as a function of the genetic similarity between the pair of individuals. Analogously, the contribution of each individual's SNP and covariates will vary as a function of how strongly the SNP and covariates influence that individual's phenotype, taking into account genetic similarity between individuals. The dependencies between individuals due to genetic similarity, in addition to the influence of genetic variation and covariates in generating phenotypes, can be modelled using a graph in which nodes correspond to observed phenotypes and covariates. Edges in the graph denote dependencies between phenotypes and covariates (Figure 1).

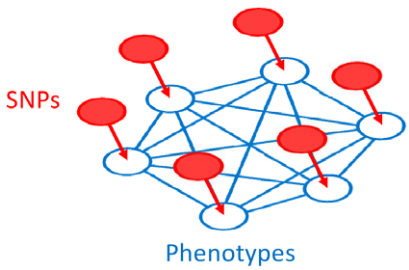

Figure 1. The graphical model for relating genetic variation to phenotype. Nodes correspond to variables in the model and edges correspond to dependencies between variables under the model. Shaded nodes correspond to observed variables under the model. Conditioned on each individual's SNP and covariates, phenotypic labels are modeled using a fully connected undirected graphical model.
doi:10.1371/journal.pone.0021591.g001

The goal of associating SNPs to phenotypes then corresponds to parameter estimation under our model in which genetic similarity between individuals is accounted for (see Methods for more details). For a given SNP, the model parameters can be assigned a p-value which we will use as a test statistic of significance of association between the given SNP and individuals' phenotypes under the null hypothesis that no associations hold between genetic variation and phenotype (see Methods). To test the utility of the proposed model for association studies, we describe in the next section a series of experiments that measure the degree to which the above model accounts for confounding factors and its computational cost for larger studies.

Experiments

Given our probabilistic model for estimating associations between SNPs and phenotype, we would like to test two aspects of the model. The first is that of *calibration*, or whether the distribution of p-values is uniform under the null hypothesis for each SNP. On synthetic data, it is straightforward to guarantee this condition. On real data, we use our prior belief that very few SNPs are associated with the phenotype to obtain this condition. As is standard practice in GWAS, we summarize the departure of an observed p-value distribution from the theoretical null distribution by use of the λ statistic, or genomic inflation factor [2], which measures how much smaller the observed median p-value is compared to that expected in the theoretical null distribution. Therefore, on data containing no (or very few) associations, $\lambda > 1$ suggests that the p-value distribution is inflated (too many small, significant p-values), which can happen when confounding factors are inadequately modeled. Conversely, $\lambda < 1$ implies deflated p-values (too few small p-values). In general, small variations from $\lambda = 1$ are expected to occur even in synthetically generated datasets with no associations due to sampling error for a finite number of SNPs.

The second aspect we wish to test is that of *discrimination*, or whether the model can distinguish spurious associations from real ones. To do this, we must apply our method to data where the ground truth as to the strength of associations to be found is known at the outset. Ideally, we would sample individuals' phenotypes under the undirected graphical model, conditioned on their SNPs and covariates. However, obtaining samples from the correct joint probability is in general intractable (see Methods). An alternative is to instead generate synthetic data from a generalized linear mixed model (GLMM), which is tractable, and then assess the statistical power of our method in distinguishing between spurious and true associations for this dataset (see Methods for details on how synthetic data were generated). One caveat is that sampling from the GLMM would mean that our model is misspecified and would suffer some loss of power relative to a LMM when both are applied to the sampled data. However, provided that we are able to generate data similar to real genotype/phenotype data, the analysis on synthetic data will inform us about whether the method will have significant statistical power on real data. Furthermore, if the outputs of our model are similar between synthetic and real data, then this would suggest that the model has adequately captured the statistics of the data in the sense of modeling confounding factors.

To test the above two aspects, we used both real data and synthetic phenotypes generated from a GLMM using real human genotype and phenotype data. The GAW14 [14] dataset consists of 7,579 SNP markers for 1,261 individuals from four distinct subpopulations (white non-Hispanic, black non-Hispanic, Hispanic, and other), where an individual's phenotype corresponds to whether he/she smokes or not. We also used the GOLDN dataset

[15] which consisted of 647 SNP markers for 1,114 individuals from two National Heart, Lung and Blood Institute (NHLBI) Family Heart Study (FHS) field centers, where an individual's phenotype corresponds to whether he/she is above or below the population median height. In both datasets, due to a large amount of population structure and family structure, it is expected that the effects of confounding factors will be strong.

We applied our model to the above real datasets and to the synthetically-generated data, where for all datasets, individual age covariates were binned into five ranges $0-21$, $21-30$, $30-45$, $45-65$, $65+$ and each individual's age group, encoded as a binary 5-vector, was used in the regression. All covariates and SNP values were standardized to have zero mean and unit standard deviation across individuals. We see that the distributions of p-values in both real and synthetic datasets are not significantly different from the uniform distribution of p-values that is expected under the null hypothesis, as measured by both one-sample Kolmogorov-Smirnov tests ($p = 0.16, 0.13$ for the synthetic and real GAW14 data, $p = 0.74, 0.74$ for synthetic and real GOLDN data) and the genomic inflation factor λ, shown in Figures 2(a,b), 3(a,b), 4(a,b) and 5(a,b). These two results suggest that our model adequately models confounding factors and has a low false positive rate in the presence of confounders. For comparison, Figures 2(c,d,e,f), 3(c,d,e,f), 4(c,d,e,f) and 5(c,d,e,f) show p-values obtained from 1) a logistic regression of phenotype onto covariates and SNPs without accounting for confounding factors and 2) from using the PCA-based Eigenstrat method [4]. Here we see that an inflation of the number of significant p-values occurs for these latter two methods, as the distribution of p-values obtained deviates significantly from the uniform distribution ($p < 1 \times 10^{-24}$). One possible explanation for the inflation seen in the p-values produced by the PCA-based method is that it may be biased against due to the relatively small number of markers evaluated in the GAW14 and GOLDN datasets. However, upon additional evaluations on the larger Wellcome Trust Case Control Consortium dataset [16] (Figure 6) containing of 360,657 SNP markers across 3,400 individuals, we observe similar results in that the PCA-based method again produces inflated p-values, whereas our method produced no significant deviation from the uniform distribution of p-values expected under the null hypothesis. We also note that the distributions of p-values obtained are similar for both real data and synthetic data in which the SNP regression weight is set to 0 (Figures 2,3), suggesting that our sampling method has produced synthetic data which is representative of real data.

In addition to testing the calibration of our method, we would also like to test its ability to distinguish spurious associations from real ones, or its statistical power. A method that produces few significant p-values for data where $\beta_{SNP} = 0$ and many significant p-values for data where $\beta_{SNP} > 0$ will have high statistical power, as measured by true and false positive rates. The results of the synthetic experiments are shown in Figure 7 for the GAW14 and GOLDN datasets. The plots are shown as receiver operating characteristic (ROC) curves of the true positive rate as a function of the false positive rate (see Methods). The performance of our model can then be summarized using the area under the ROC curve, or AUC, which is high if our model has high statistical power in discriminating between real and spurious associations. For comparison, we also applied the LMM-based method of [12], which also accounts for confounding factors, to the above synthetic data using the same set of similarities as that used by our method, but interpreted instead as a covariance matrix among individuals under a multivariate Gaussian distribution. As an additional point of comparison, we also applied the Eigenstrat method [4] to the synthetic data. As expected, due to the mismatch between the model used to generate the synthetic test data and our model, there is a modest loss in power as compared to the LMM, whereby the loss in model power decreases as the SNP weight β_{SNP} is increased (Figure 7). The loss in power is partially explained by noting that the data was generated from a GLMM using a Gaussian covariance matrix $\boldsymbol{\theta}$, which corresponds to the same covariance matrix used in the LMM. However, $\boldsymbol{\theta}$ in our model cannot be interpreted as a covariance matrix under a multivariate Gaussian distribution, implying a larger mismatch between our model and the data as compared to that between the LMM and the data. We also see that Eigenstrat, while having low computational cost, does not adequately account for confounding factors and so has significantly lower power as compared to our method.

In addition to assessing the statistical power of our method, we also assessed the runtime of our method as a function of the study size, or number n of individuals. To do this, we synthesized datasets consisting of the phenotypes, SNPs and similarities of the GAW14 dataset replicated several times (up to 35,000 individuals), such that each synthetic dataset generated this way has an increasing number of individuals. We then applied both our method and the LMM to the synthetic datasets and recorded the total time taken to perform a GWAS for each dataset. All experiments were run on a single machine running Windows Server Enterprise with two Intel Xeon E5450 3.0 GHz 64-bit CPUs with 64.0 GB of RAM. Figure 8 shows the runtime of both methods as a function of the study size: as can be seen, the runtime for estimating the parameters of the LMM grows quickly as the number of individuals increases, whereas for our method, the runtime does not grow quickly. In particular, the difference in runtime becomes acute as the study size exceeds 20,000 individuals, resulting in significant runtime speedups (48 mins. for our method as compared to over 33 hours for the LMM for a study with 37,830 individuals). We remark here that although the experiments were carried out on a single machine, the differences in runtime of our method over the LMM-based method would also apply for experiments carried out on computation clusters with multiple compute nodes.

Discussion

We have presented a novel GWAS method that accounts for confounding factors such as population structure, family structure or cryptic relatedness. Similar to LMMs and PCA-based methods for association, our model accounts for confounding factors through the use of pairwise similarities between patients, which allows us to significantly reduce false positive rates when performing associations. In contrast to LMM-based and PCA-based methods, our method retains high statistical power and is relatively inexpensive even as the number of individuals in a study grows. Our experimental results on both real and synthetic genotype data demonstrate that our method can adequately account for confounding factors in order to reduce false positive rates, with a modest loss in statistical power as compared to LMM-based and PCA-based methods for data that is generated from a generalized LMM. We have shown that our method is significantly faster than methods based on LMMs, where significant speedups are obtained as the number of individuals in a study grows. As future studies grow to encompass tens of thousands of individuals [7–9], the speedups afforded by our method over LMM-based methods are expected to be even larger than ones shown here. Although other methods that also have fast runtimes for large datasets could be used, in the regime where the effect of

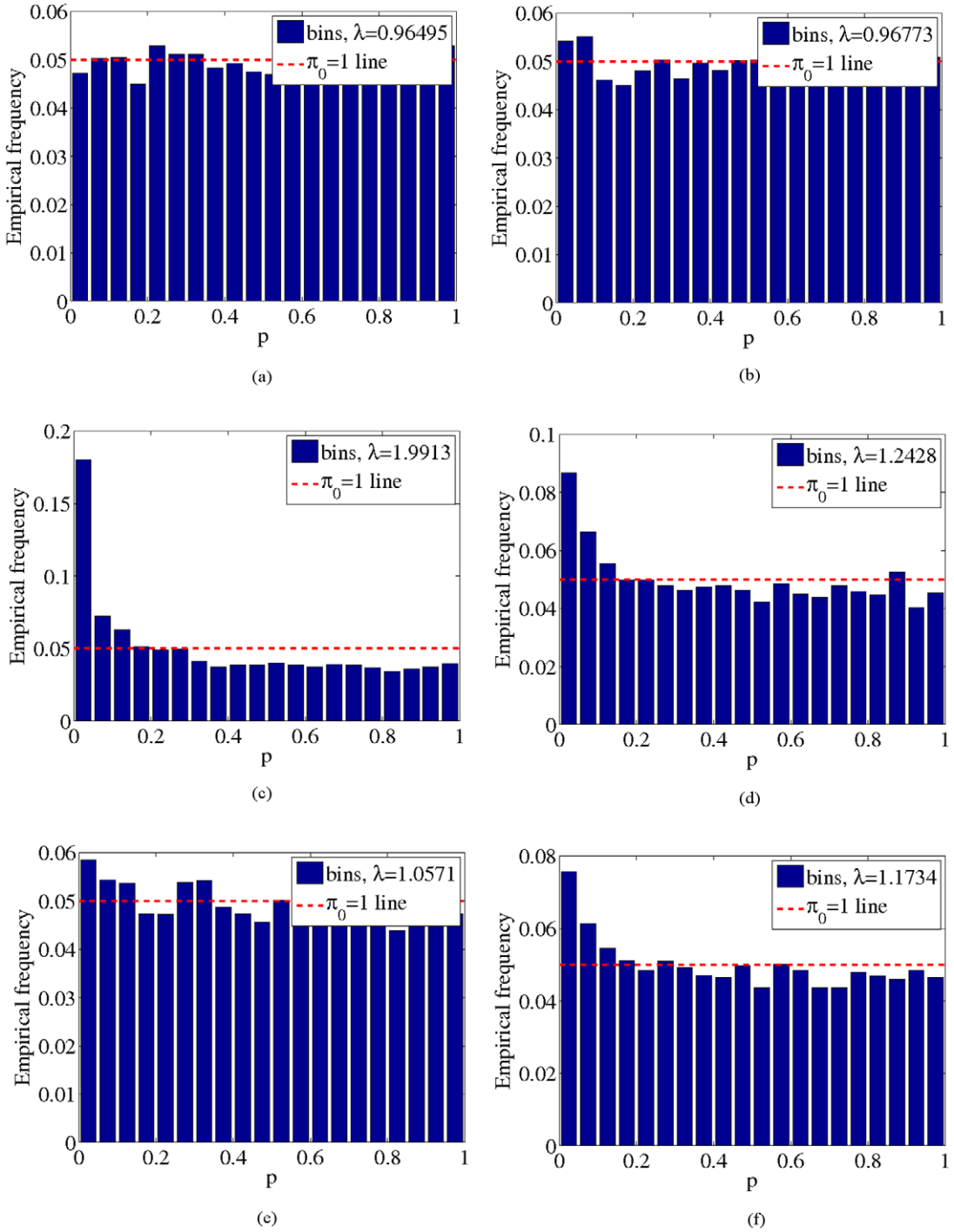

Figure 2. P-value histograms for the GAW14 dataset. a),b) Histograms of p-values obtained from our method for the synthetic (a) and real (b) GAW14 data. For comparison, p-values obtained from a logistic regression that does not account for confounding factors and from Eigenstrat [4] are shown for synthetic (c,e) and real (d,f) GAW14 data. Dotted red lines indicate the expected histogram for the uniform distribution under the null hypothesis $\pi_0 = 1$.
doi:10.1371/journal.pone.0021591.g002

confounders is even stronger than it is for smaller studies, it is expected that these methods will not be able to model confounders adequately so as to reduce false positive associations. Our method presents a reasonable tradeoff between statistical power, low false positive rates and runtime that make it ideally suited for application to larger association studies where other methods either produce too many false positives or incur high computational costs. Future work would involve extending the method to multinomial discrete phenotypes and for modeling multiple phenotypes simultaneously, examining the use of other pairwise similarity measures, or the possibility of incorporating additional covariates into the similarity measures themselves.

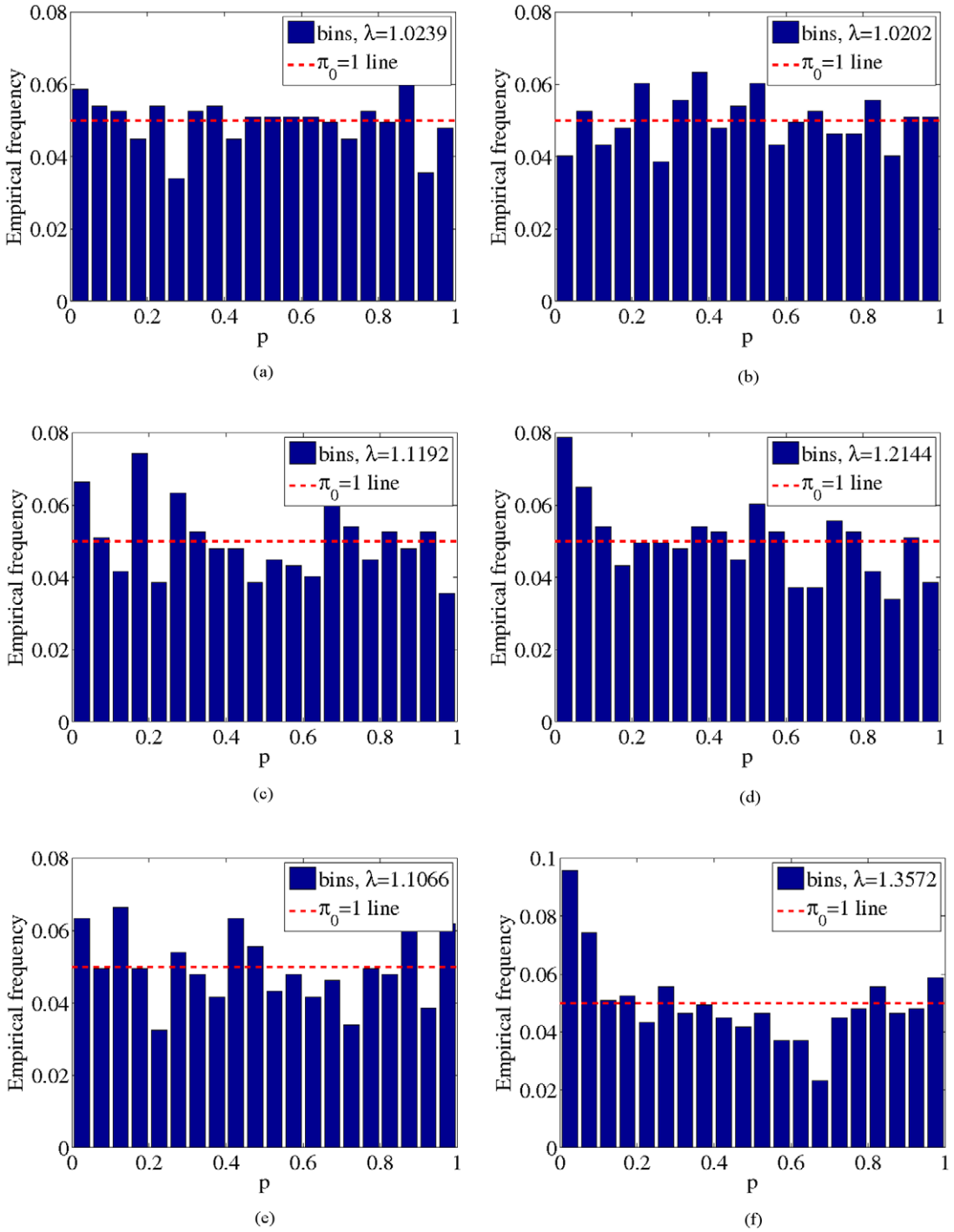

Figure 3. P-value histograms for the GOLDN data. a),b) Histograms of p-values obtained from our method for the synthetic (a) and real (b) GOLDN data. For comparison, p-values obtained from a logistic regression that does not account for confounding factors and from Eigenstrat [4] are shown for synthetic (c,e) and real (d,f) GOLDN data. Dotted red lines indicate the expected histogram for the uniform distribution under the null hypothesis $\pi_0 = 1$.
doi:10.1371/journal.pone.0021591.g003

Methods

Datasets

GAW14 dataset. The GAW14 dataset consisted of a subset of the data provided to the Genetic Analysis Workshop 14 (GAW 14) as part of the Collaborative Study on the Genetics of Alcoholism (U10 AA008401), which is described in detail elsewhere [14]. A total of 1,279 individuals genotyped at 7,579 loci were used from the GAW14 dataset for our analysis. Genotypes are coded using the number of minor alleles,such that the SNP value at a given locus takes on values 0,1,2. Age, sex and ethnic sub-population were recorded for each individual and used as covariates in our analysis. Measured phenotypes included alcohol dependence and smoking activity: the smoking activity phenotype was used for our analysis.

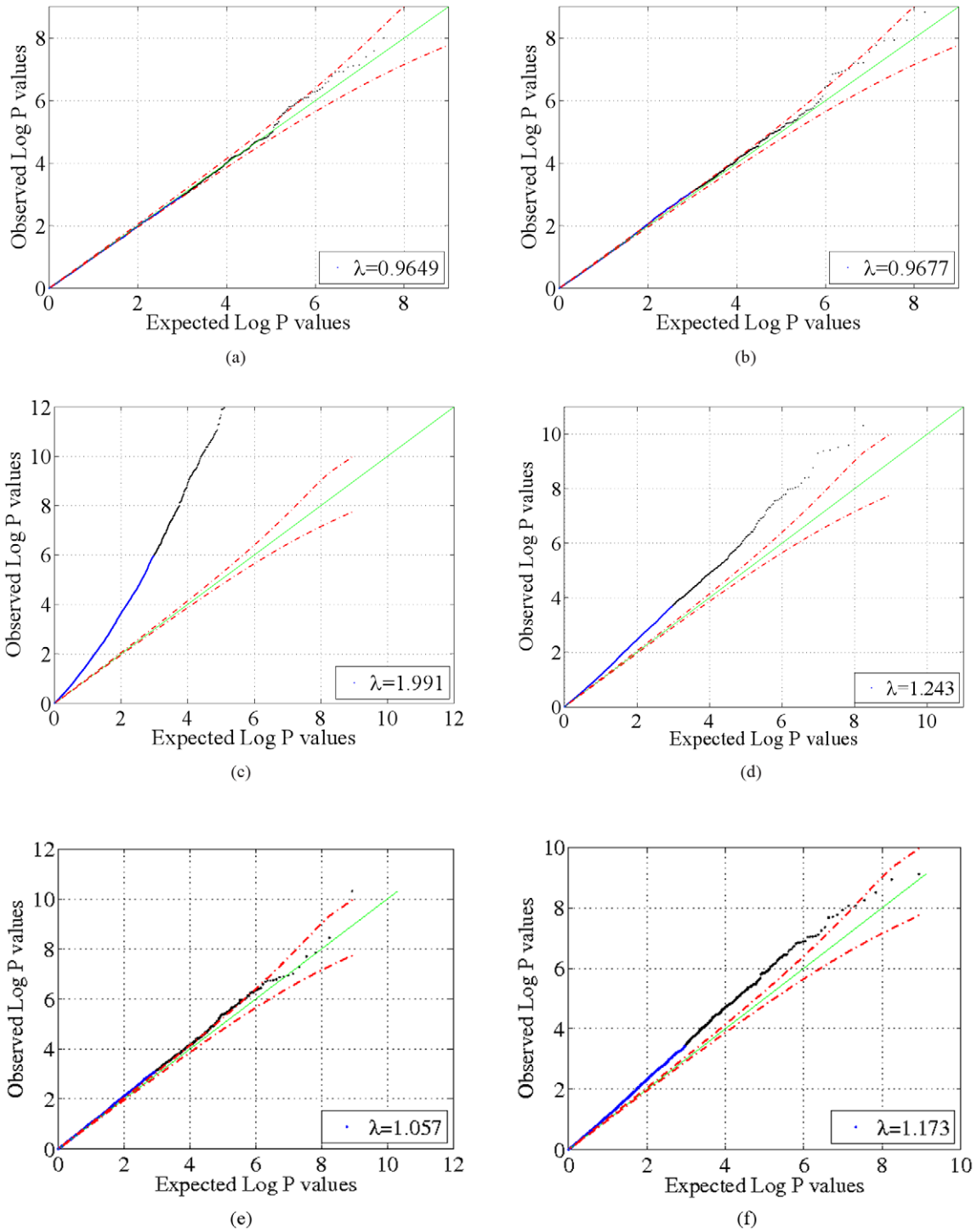

Figure 4. Quantile-quantile (QQ) plots for the GAW14 data. QQ plots of model negative log p-value statistics obtained from our method as a function of expected negative log p-values under the null hypothesis $\pi_0 = 1$ for the synthetic GAW14 data with $\beta_{SNP} = 0$ (a) and real data (b). For comparison, negative log p-value statistics obtained from a logistic regression that does not account for confounding factors and from Eigenstrat [4] are shown for the synthetic data with $\beta_{SNP} = 0$ (c,e) and real data (d,f). Dotted red lines indicate 95% confidence bounds.
doi:10.1371/journal.pone.0021591.g004

GOLDN dataset. Details about the GOLDN study has been described in detail elsewhere [15]. Briefly, the largest three-generation families were recruited from the pool of families that had participated in the National Heart, Lung, and Blood Institute Family Heart Study (FHS) at either the Minnesota or Utah field centers. A total of 1114 individuals from 190 families, genotyped at 647 SNP markers, were included in our analysis. Genotype data was encoded as for the GAW14 dataset. Age and sex was recorded for each individual and used as covariates in our analysis. Measured phenotypes in this dataset included height, physical activity and cholesterol levels: the height phenotype was the one used for our analysis.

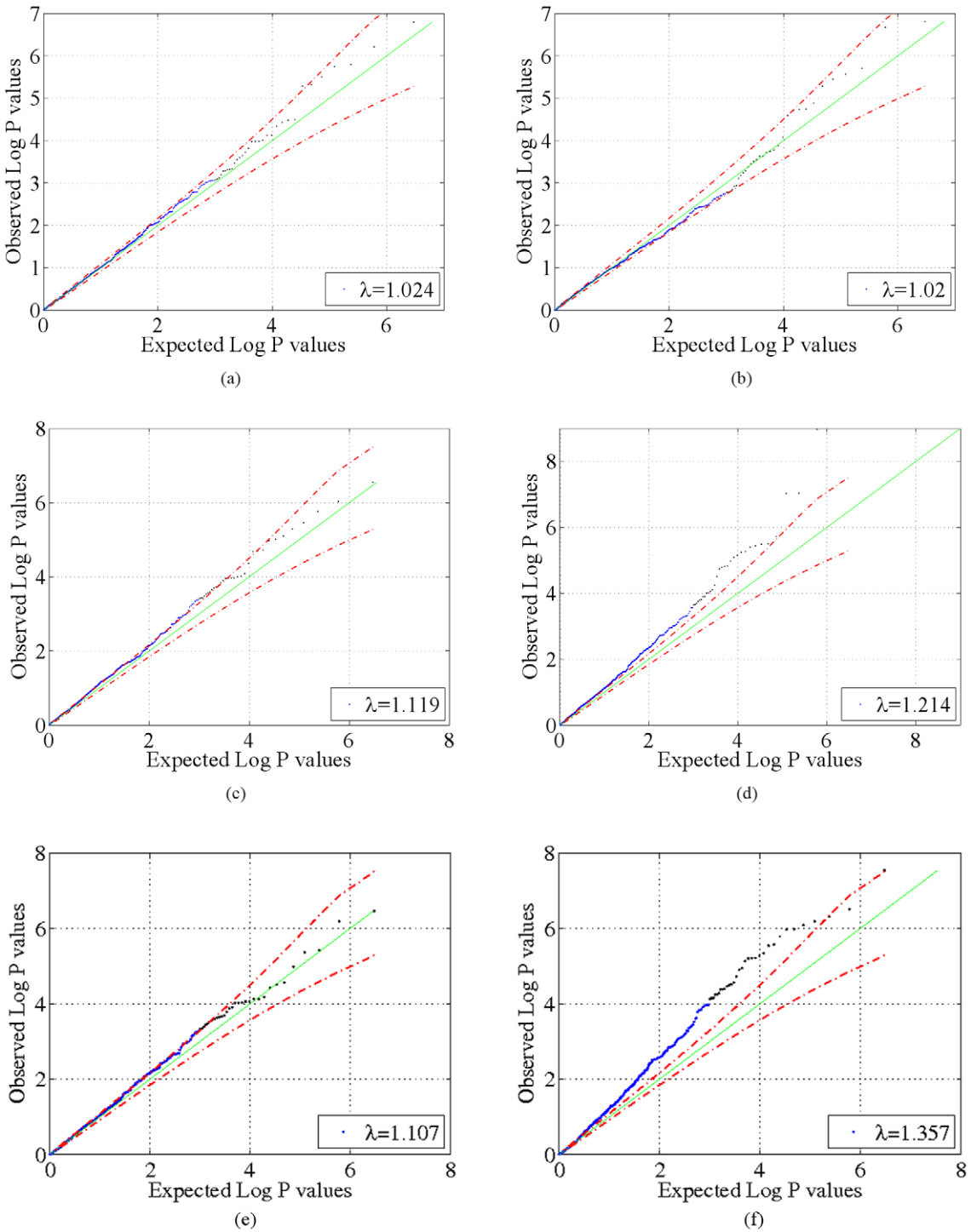

Figure 5. Quantile-quantile (QQ) plots for the GOLDN data. QQ plots of model negative log p-value statistics obtained from our method as a function of expected negative log p-values under the null hypothesis $\pi_0 = 1$ for synthetic GOLDN data with $\beta_{SNP} = 0$ (a) and real data (b). For comparison, negative log p-value statistics obtained from a logistic regression that does not account for confounding factors and from Eigenstrat [4] are shown for the synthetic data with $\beta_{SNP} = 0$ (c,e) and real data (d,f). Dotted red lines indicate 95% confidence bounds.
doi:10.1371/journal.pone.0021591.g005

WTCCC dataset

The Wellcome Trust Case Control Consortium (WTCCC) data consisted of SNP data for about 1,900 individuals with Crohn's disease and about 1,500 controls from the UK Blood Service Control Group (NBS). SNPs were excluded from analysis using the more conservative SNP filter described by the WTCCC in [16], wherein a SNP was excluded if either its minor-allele frequency less than 0.01, it was missing in greater than one percent of individuals, or it was in the extended MHC region. After filtering, 360,657 SNPs remained. Non-white individuals and close family members were not excluded.

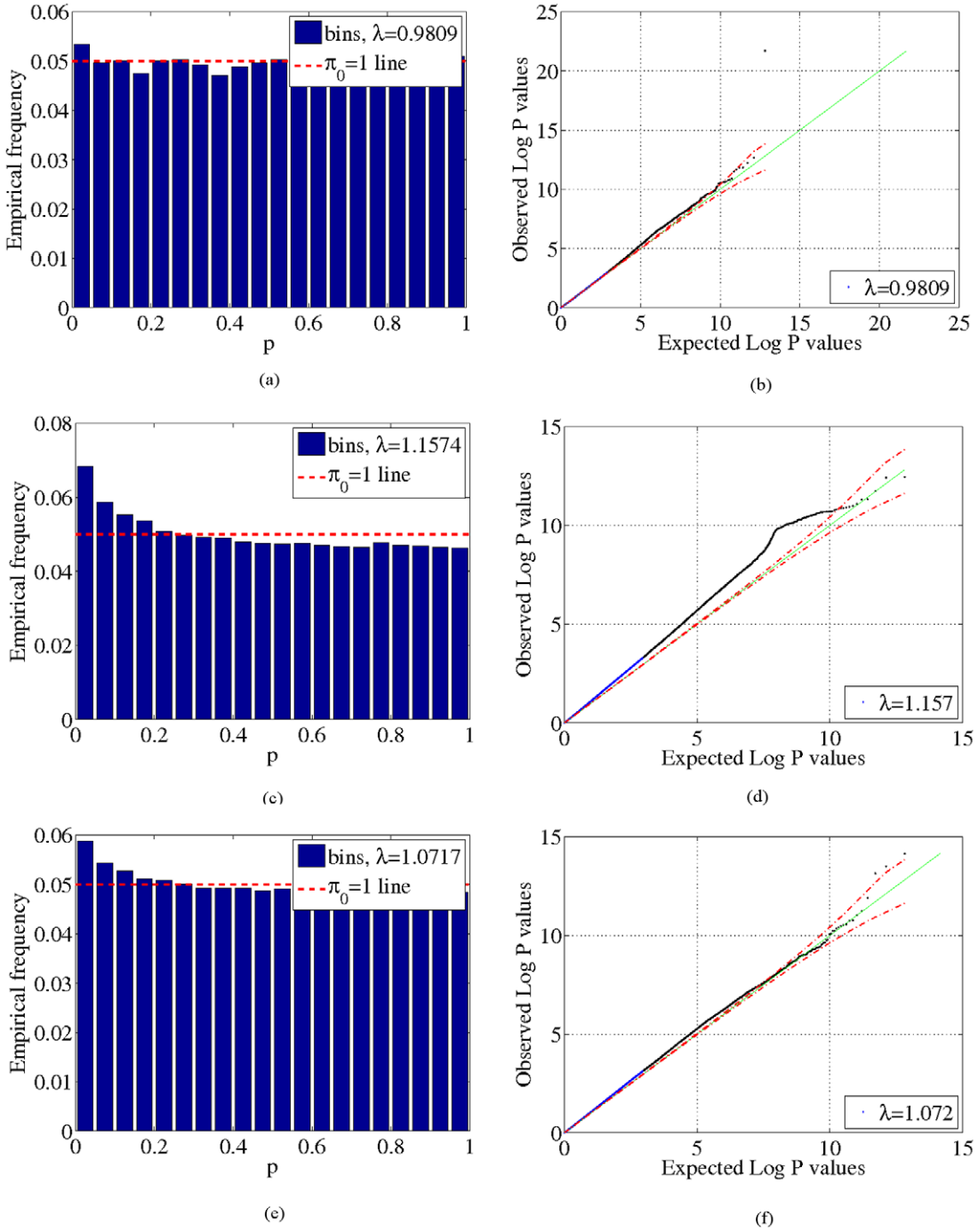

Figure 6. Histograms and quantile-quantile (QQ) plots for the WTCCC data. Negative log p-value statistics obtained from our method (a,b), logistic regression that does not account for confounding factors (c,d) and from Eigenstrat [4] (e,f) for the WTCCC data. Dotted red lines in the QQ plots (right panel) indicate 95% confidence bounds.
doi:10.1371/journal.pone.0021591.g006

Genome-wide association studies using conditional random fields

Given a set of individuals $V=\{1,\cdots,n\}$, we assume that phenotypes consist of binary labels $\{-1,+1\}$ corresponding to the absence/presence of a phenotype in an individual, although the model can easily be generalized to polytomous discrete or continuous phenotypes. Denote by y_i the observed phenotype for the i^{th} individual i and let $\mathbf{y}=(y_1,\cdots,y_n)$ be the vector of observed phenotypes for all individuals in the study. Let $\mathbf{x}_i$ be the vector of covariates for individual i and let $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_n)$ denote the matrix of covariates for the individuals in the study. Here, the covariates for an individual would include that individual's SNP marker at a given loci and possibly labels for age, gender and ethnicity.

For a given locus, our model consists of a probabilistic graphical model over individuals' observed phenotypes, conditioned on each individual's SNP and covariates. A probabilistic graphical model consists of two parts: the first is an graph $G=(V,E)$ in which

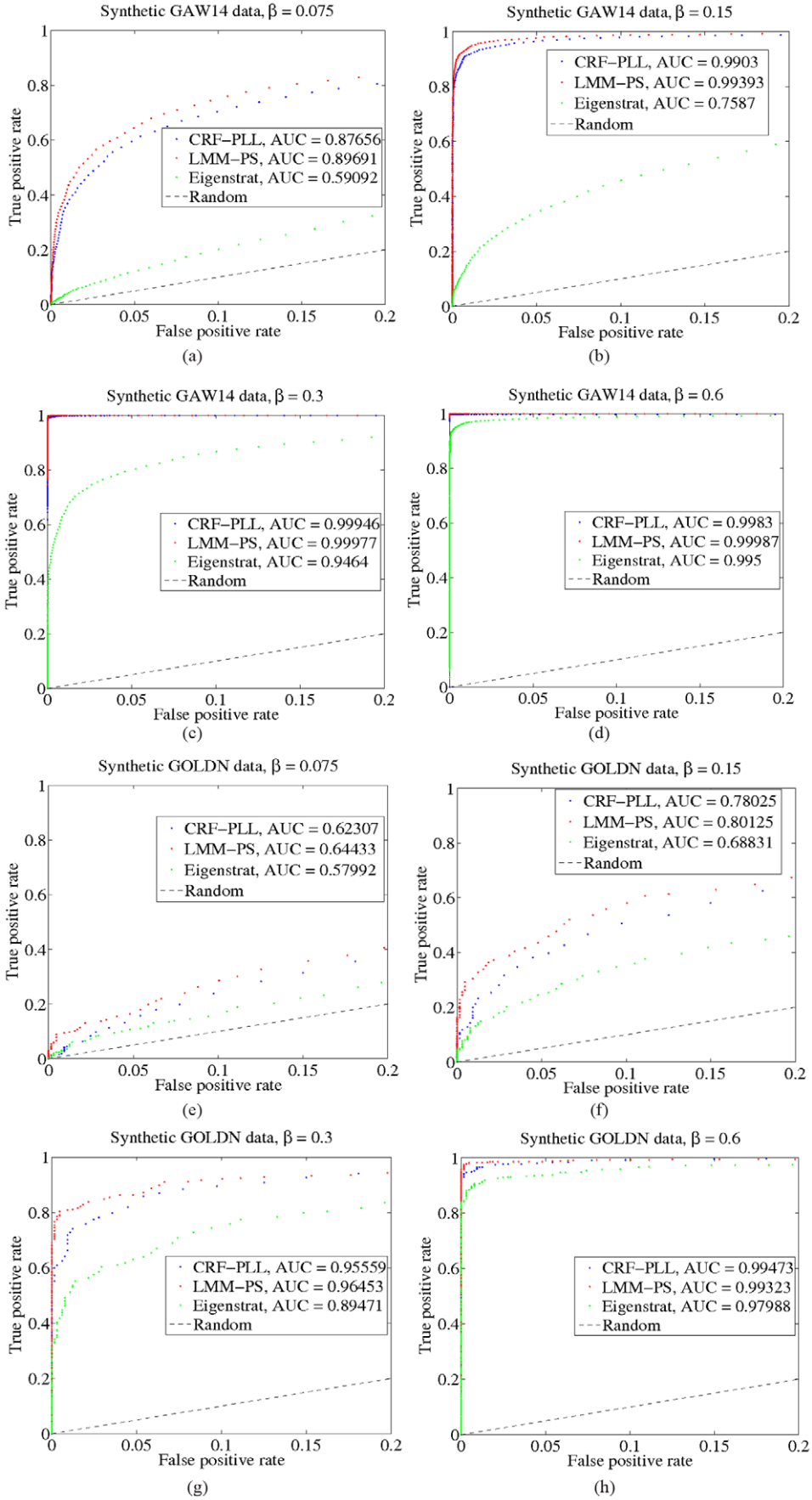
Synthetic GAW14 data, β = 0.075
CRF–PLL, AUC = 0.87656
LMM–PS, AUC = 0.89691
Eigenstrat, AUC = 0.59092
Random
True positive rate
False positive rate
(a)
Synthetic GAW14 data, β = 0.15
CRF–PLL, AUC = 0.9903
LMM–PS, AUC = 0.99393
Eigenstrat, AUC = 0.7587
Random
True positive rate
False positive rate
(b)
Synthetic GAW14 data, β = 0.3
CRF–PLL, AUC = 0.99946
LMM–PS, AUC = 0.99977
Eigenstrat, AUC = 0.9464
Random
True positive rate
False positive rate
(c)
Synthetic GAW14 data, β = 0.6
CRF–PLL, AUC = 0.9983
LMM–PS, AUC = 0.99987
Eigenstrat, AUC = 0.995
Random
True positive rate
False positive rate
(d)
Synthetic GOLDN data, β = 0.075
CRF–PLL, AUC = 0.62307
LMM–PS, AUC = 0.64433
Eigenstrat, AUC = 0.57992
Random
True positive rate
False positive rate
(e)
Synthetic GOLDN data, β = 0.15
CRF–PLL, AUC = 0.78025
LMM–PS, AUC = 0.80125
Eigenstrat, AUC = 0.68831
Random
True positive rate
False positive rate
(f)
Synthetic GOLDN data, β = 0.3
CRF–PLL, AUC = 0.95559
LMM–PS, AUC = 0.96453
Eigenstrat, AUC = 0.89471
Random
True positive rate
False positive rate
(g)
Synthetic GOLDN data, β = 0.6
CRF–PLL, AUC = 0.99473
LMM–PS, AUC = 0.99323
Eigenstrat, AUC = 0.97988
Random
True positive rate
False positive rate
(h)

Figure 7. Assessing statistical power on synthetic data. The plots are shown as receiver operating characteristic (ROC) curves of the true positive rate as a function of the false positive rate for the GAW14 dataset (a,b,c,d) and the GOLDN dataset (e,f,g,h) for various values of the SNP regression weight when using our method (blue), a LMM-based method (red), a PCA-based method (green) and random guessing (black dotted).
doi:10.1371/journal.pone.0021591.g007

nodes in V correspond to individuals and undirected edges in E between pairs of nodes correspond to possible dependencies between the phenotypes of pairs of individuals. The second part of the model is a joint probability distribution on individuals' phenotypes that is a function of all pairs of individuals' phenotypes and each individual's SNP and covariates. Given a graph and the corresponding joint probability distribution, the graphical model captures both the dependencies between individuals due to genetic similarity, in addition to the influence of genetic variation and covariates in generating phenotypes. The influence of genetic variation and covariates is captured using a set of weights $\boldsymbol{\beta}$, where a larger weight magnitude for a given covariate denotes an increased influence of that covariate on determining phenotype. The joint probability of phenotypic labels, conditioned on each individual's genetic variant and covariates is then given by

$$P(\mathbf{y}|\mathbf{X},G,\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{\exp -\left(\sum_{(i,j)\in E}\theta_{ij}y_iy_j - \sum_{i\in V}y_i\boldsymbol{\beta}^T\mathbf{x}_i\right)}{Z(\mathbf{X},G,\boldsymbol{\beta})}, \quad (1)$$

where θ_{ij} is a real-valued genetic similarity for edge (i,j) that models genetic similarity between individuals i and j, and $Z(\mathbf{X},G,\boldsymbol{\beta}) = \sum_{\mathbf{y}}\exp\left(-\sum_{(i,j)\in E}\theta_{ij}y_iy_j - \sum_{i\in V}y_i\boldsymbol{\beta}^T\mathbf{x}_i\right)$ is the partition function that ensures that the probability sums to unity. In the above model, we assume that genetic similarities, denoted collectively as $\boldsymbol{\theta}$, are provided and fixed. Various ways of setting the similarities can be used. Based on their previous use in LMMs [12], we found that using similarities based on Identity-by-State (IBS) worked best, where the IBS value between two individuals is equal to the fraction of SNP marker alleles that are shared between individuals [17] across the entire set of SNPs being studied. The use of the IBS similarity measure here allows us to account for the effects of confounding factors which are encoded in the set of SNPs carried by all individuals.

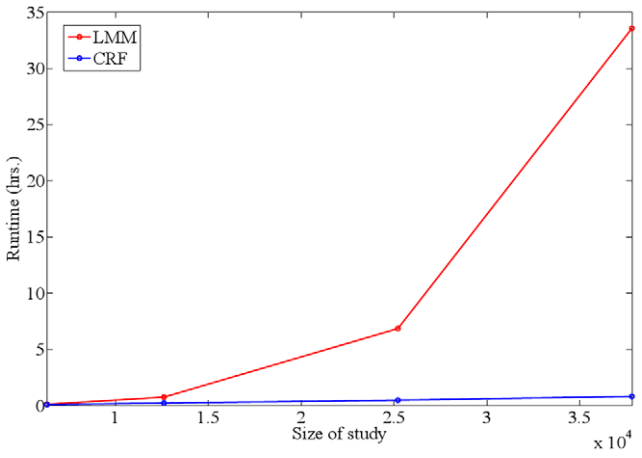

Figure 8. Comparison of runtimes. Runtimes for the CRF and LMM models (in hours) are shown as a function of study size. All experiments were run on a machine with two 3.0 GHz CPUs and 64.0 GB of RAM.
doi:10.1371/journal.pone.0021591.g008

Given individuals' phenotypes $\mathbf{y}$ and covariates $\mathbf{X}$ and a matrix of genetic similarities $\boldsymbol{\theta}$, the goal is to estimate the effect of a particular SNP on the individuals' phenotypes by estimating the weight vector $\boldsymbol{\beta}$. A common criterion that is used consists of maximizing the above probability with respect to the weights for the observed data, or the maximum-likelihood criterion. However, a key difficulty with the above model is that estimating the weights requires that we compute the partition function and its derivatives, which, for even a moderate number of individuals, will be intractable, as it requires summing over all possible joint configurations of the binary vector $\mathbf{y}$. An alternative criterion for parameter estimation that does not require computing $Z(\mathbf{X},G,\boldsymbol{\beta})$ altogether here is to instead optimize the *pseudo-likelihood* [18] function for the above model, which has been previously shown to be asymptotically consistent and here yields fast parameter estimates. We define the negative log-pseudo-likelihood function as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= -\sum_{i\in V}\mathcal{L}_i(\boldsymbol{\beta}) \\ &= -\sum_{i\in V}([y_i=+1]\log p_i + [y_i=-1]\log(1-p_i)),\end{aligned} \quad (2)$$

where the conditional probability of individual i's phenotype given $\mathbf{y}_{-i}$ is denoted as $p_i \equiv p_i(\boldsymbol{\beta}) = P(y_i=+1|\mathbf{y}_{-i},\mathbf{X},G,\boldsymbol{\beta})$. We note that evaluating and differentiating the pseudo-likelihood does not depend on the partition function, as under the above model, the conditional probability for individual i given all other individuals' phenotypes $\mathbf{y}_{-i}$ is given by

$$P(y_i|\mathbf{y}_{-i},\mathbf{X},G,\boldsymbol{\beta}) = \frac{\exp -\left(2y_i\sum_{j\in N(i)}\theta_{ij}y_j - 2y_i\boldsymbol{\beta}^T\mathbf{x}_i\right)}{1+\exp -\left(2y_i\sum_{j\in N(i)}\theta_{ij}y_j - 2y_i\boldsymbol{\beta}^T\mathbf{x}_i\right)}, \quad (3)$$

where $\mathcal{N}(i)$ denotes the set of neighbors of individual i with respect to graph G and we note that the partition function $Z(\mathbf{X},G,\boldsymbol{\beta})$ has dropped out. Thus, to perform genome-wide associations, we optimize the above function with respect to $\boldsymbol{\beta}$: this can be done by using a gradient-based optimization whereby we iteratively update the vector of weights $\boldsymbol{\beta}$ using the gradient of the pseudo-likelihood. The above pseudo-likelihood corresponds to solving a logistic regression problem with covariates $2y_i\mathbf{x}_i$ and an additive term for each individual i, given by $2y_i\sum_{j\in N(i)}\theta_{ij}y_j$, which models the contribution of other individuals' phenotypes in determining the phenotype of i. We remark that computing this additive term need only be done once and requires time that is quadratic in the number of individuals, which contrasts with cubic runtime required by LMM-based methods [12,13]. Furthermore, the time required for parameter estimation per SNP is linear in the number of individuals, as we need only compute a conditional probability p_i for each individual and corresponding derivatives with respect to weight vector $\boldsymbol{\beta}$. The resulting optimization problem is convex, with a unique global optimum, so we are guaranteed to obtain a unique solution $\hat{\boldsymbol{\beta}}$ that maximizes the pseudo-likelihood, in contrast to parameter estimation in LMMs, which may be prone to local minima.

Pseudo-likelihood estimation in the conditional random field

The gradient descent updates for parameter estimation under our method take the form $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \alpha \mathbf{g}$, where $\alpha > 0$ is a learning rate parameter and $\mathbf{g}$ is the gradient of the pseudo-likelihood given by

$$\mathbf{g} = \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = -\sum_{i \in V} \mathbf{x}_i([y_i = +1](1-p_i) - [y_i = -1]p_i).$$

The weight vector $\boldsymbol{\beta}$ is updated until convergence in $\mathcal{L}(\boldsymbol{\beta})$. For our experiments, we used $\alpha = \frac{5}{n}$, which was selected for fast convergence.

Significance testing of SNPs

Given an estimate $\hat{\boldsymbol{\beta}}$ that minimizes the negative log-pseudo-likelihood function, define the robust variance estimator [19] as

$$\sum \equiv \sum(\hat{\boldsymbol{\beta}}) = \mathbf{H}^{-1}\left(\sum_{i \in V} \mathbf{g}_i \mathbf{g}_i^T\right)\mathbf{H}^{-1}, \quad (4)$$

where $\mathbf{H}$ is the Hessian matrix of the pseudo-likelihood objective function, given by

$$\mathbf{H} = \nabla\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = \sum_{i \in V} \mathbf{x}_i \mathbf{x}_i^T p_i(1-p_i), \quad (5)$$

and $\mathbf{g}_i$ is given by

$$\mathbf{g}_i = \nabla_{\boldsymbol{\beta}} \mathcal{L}_i(\boldsymbol{\beta}) = -\mathbf{x}_i([y_i = +1](1-p_i) - [y_i = -1]p_i). \quad (6)$$

The statistic $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ has been shown to be asymptotically distributed according to $N(\mathbf{0}; \sum)$ [20–23]. In particular, it follows that the statistic $\frac{(\hat{\beta}_{SNP} - \beta_0)^2}{\sigma^2_{SNP}}$ is χ^2 with one degree of freedom, where $\hat{\beta}_{SNP}$ is the learned weight for a given SNP, $\sigma^2_{SNP} = \mathbf{S}_{1,1}$ and $\beta_0 = 0$ is the weight for the SNP under the null hypothesis. The above is equivalent to performing a Wald test on $\hat{\beta}_{SNP}$ with a robust variance estimator for the variance of $\hat{\beta}_{SNP}$.

Measuring genomic inflation

Given χ^2 statistics $\chi^2_1, \cdots, \chi^2_p$ for each SNP $j = 1 \cdots, p$ of interest, we can compute a genomic inflation factor λ [2] as

$$\lambda = \frac{median(\chi^2_1, \cdots, \chi^2_p)}{0.4549}. \quad (7)$$

Evaluating model performance

To gauge the calibration and discrimination of our model for both weaker and stronger associations, we generated data with different SNP regression weights using a GLMM. For each SNP, we generated SNP-phenotype associations by setting the SNP regression weight β_{SNP} to 0,0.075,0.15,0.3,0.6, sampling a vector $\mathbf{U}$ from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu} + \mathbf{x}\beta_{SNP}, \boldsymbol{\theta})$ and finally generating the output phenotype for each individual with probability $P(y_i = +1 | u_i, w) = \frac{1}{1 + \exp(w \cdot u_i)}$, where w was chosen in order to obtain synthetic phenotype data with similar phenotype frequencies as those of real phenotype data.

Given a set of model p-values for synthetic data, we define a false positive (FP) to be a SNP that has a significant p-value for some significance level α for synthetic data in which $\beta_{SNP} = 0$. A true negative is defined as a SNP that is not significant at significance level α for synthetic data in which $\beta_{SNP} = 0$. True positives (TP) and false negatives (TN) are defined similarly for synthetic data with $\beta_{SNP} > 0$. By varying the significance level α, we can evaluate the performance of our model using a receiver operating characteristic (ROC) curve, or plotting true positive rate $TP/(TP+FN)$ as a function of the false positive rate $FP/(FP+TN)$ for various synthetic datasets with $\beta_{SNP} > 0$. The performance can then be summarized using the area under the ROC curve, or AUC. Methods that have higher AUC have higher statistical power in discriminating between real and spurious associations.

Acknowledgments

The GAW14 data were provided by the Collaborative Studies on the Genetics of Alcoholism (U10 AA008401). The GOLDN dataset were provided by the Genetics Of Lipid Lowering Drugs and Diet Network (U01 HL72524). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We would like to thank Jennifer Listgarten for providing analysis tools for QQ plots and p-value distributions, and for providing source code for the LMMs.

Author Contributions

Conceived and designed the experiments: JH CM DH. Performed the experiments: JH. Analyzed the data: JH. Wrote the paper: JH CM DH. Contributed source code and computational analysis tools: CK.

References

1. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781–791.
2. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
3. Kang HM, Zaitlen N, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.
4. Price AL, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
5. Yu J, Pressoir G, Briggs W, Vroh BI, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
6. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459–463.
7. Gretarsdottir S, Baas AF, Thorleifsson G, Holm H, den Heijer M, et al. (2010) Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nat Genet 42: 692–697.
8. Sulem P, Gudbjartsson DF, Rafnar T, Holm H, Olafsdottir EJ, et al. (2009) Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche. Nat Genet 41: 734–738.
9. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, et al. (2008) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41: 18–24.

10. Demidenko E (2004) Mixed models: theory and applications. New Jersey: John Wiley and Sons, Inc. 736 p.
11. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Gen 67: 170–181.
12. Listgarten J, Kadie C, Schadt EE, Heckerman D (2010) Correction for hidden confounders in the genetic analysis of gene expression. Proc Nat Acad Sci 107: 16465–16470.
13. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–360.
14. Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, et al. (2005) Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. BMC Genetics 6(Suppl 1): S2.
15. Lai CQ, Arnett DK, Corella D, Straka RJ, Tsai MY, et al. (2007) Fenofibrate effect on triglyceride and postprandial response of apolipoprotein A5 variants: the GOLDN study. Arterioscler. Thromb Vasc Biol 27: 1417–1425.
16. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
17. Chakraborthy R, Lin J (1994) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. Experientia Supp 67: 153–175.
18. Besag J (1975) Statistical analysis of non-lattice data. The Statistician 24: 179–195.
19. Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: 221–233.
20. Arnold BC, Strauss D (1991) Pseudolikelihood estimation: some examples. Ind J Stat, Series B 53: 233–243.
21. Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum likelihood methods: Theory. Econometrica 52: 681–700.
22. Jensen JL, Kunsch HR (1994) On asymptotic normality of pseudo-likelihood estimates for pairwise interaction processes. Ann Inst Statist Math 46: 475–486.
23. Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Berlin: Springer. 687 p.