Bioinformatics, 32(19), 2016, 2903–2910
doi: 10.1093/bioinformatics/btw347
Advance Access Publication Date: 13 June 2016
Original Paper

OXFORD

Genome analysis

A time-varying group sparse additive model for genome-wide association studies of dynamic complex traits

Micol Marchetti-Bowick[1], Junming Yin[2], Judie A. Howrylak[3] and Eric P. Xing[1,*]

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA, [2]Department of Management Information Systems, University of Arizona, Tucson, AZ, USA and [3]Division of Pulmonary and Critical Care Medicine, Penn State University, Milton S. Hershey Medical Center, Hershey, PA, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 22, 2015; revised on May 24, 2016; accepted on May 27, 2016

Abstract

Motivation: Despite the widespread popularity of genome-wide association studies (GWAS) for genetic mapping of complex traits, most existing GWAS methodologies are still limited to the use of static phenotypes measured at a single time point. In this work, we propose a new method for association mapping that considers dynamic phenotypes measured at a sequence of time points. Our approach relies on the use of Time-Varying Group Sparse Additive Models (TV-GroupSpAM) for high-dimensional, functional regression.

Results: This new model detects a sparse set of genomic loci that are associated with trait dynamics, and demonstrates increased statistical power over existing methods. We evaluate our method via experiments on synthetic data and perform a proof-of-concept analysis for detecting single nucleotide polymorphisms associated with two phenotypes used to assess asthma severity: forced vital capacity, a sensitive measure of airway obstruction and bronchodilator response, which measures lung response to bronchodilator drugs.

Availability and Implementation: Source code for TV-GroupSpAM freely available for download at http://www.cs.cmu.edu/~mmarchet/projects/tv_group_spam, implemented in MATLAB.

Contact: epxing@cs.cmu.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The goal of genome-wide association studies (GWAS) is to analyze a large set of genetic markers that span the entire genome in order to identify loci that are associated with a phenotype of interest. Over the past decade, GWAS has been used to successfully identify genetic variants that are associated with numerous diseases and complex traits, ranging from breast cancer to blood pressure (Hindorff *et al.*, 2015). However, a significant challenge in performing GWAS is that the studies are often vastly under-powered due to the high dimensionality of the feature set relative to the small number of human samples available.

Traditional GWAS methodologies test each variant independently for association with the phenotype, and use a stringent significance threshold to adjust for multiple hypothesis testing (Clarke *et al.*, 2011). Although this approach works well for traits that depend on strong effects from a few loci, it is less suitable for complex, polygenic traits that are influenced by weak effects from many different genetic variants. More recently, a significant body of work has emerged on penalized regression approaches for GWAS that capture the joint effects of all markers (Li *et al.*, 2011; Wu *et al.*, 2009). The majority of these methods model the phenotype as a weighted sum of the genotype values at each locus, and use a

regularization penalty such as the ℓ_1 norm to identify a sparse set of single nucleotide polymorphisms (SNPs) that are predictive of the trait. Although this technique helps to reduce overfitting and detect fewer spurious SNP-trait associations, the lack of statistical power to identify true associations persists.

Here we aim to further boost the statistical power of GWAS by proposing a new model that leverages dynamic trait data, in which a particular trait is measured in each individual repeatedly over time, as depicted in Figure 1a. Such datasets are often generated by longitudinal studies that follow participants over the course of months, years, or even decades. Though broadly available, dynamic trait datasets are frequently underutilized by practitioners who ignore the temporal information. We believe that leveraging time-sequential trait measurements in GWAS can lead to greater statistical power for association mapping.

To illustrate this concept, consider the hypothetical patterns of SNP influence on the phenotype shown in Figure 1b. As in traditional GWAS, an association between a SNP and the phenotype exists if the three SNP genotypes (which we denote *AA*, *Aa* and *aa*) have differential effects on the trait. In the first example, the effects of the three SNP genotypes only differ in the $t \in [0.5, 1]$ time interval. A static method that uses data from an arbitrarily chosen time point or simply treats the time series as i.i.d. samples could easily miss this association, whereas a dynamic method that considers the entire dataset would detect it. The second example shows a SNP in which the difference between the effects of the three genotypes is small but consistent over time. Although this signal could be too weak to be interpreted as a significant association in the static case, it gets much stronger once evidence from the entire time series is considered.

The longitudinal data setting is challenging because traits are measured at irregularly spaced time points over subject-specific intervals. One approach that has been proposed for performing GWAS of dynamic traits, called functional GWAS, or fGWAS (Das *et al.*, 2011), constructs a separate model to estimate the smooth, time-varying influence of each SNP on the phenotype. Once the mean effects have been estimated for each genotype at each time point, a hypothesis test is performed to determine whether the SNP has any additive or dominant effect on the trait. Although the use of dynamic trait data gives fGWAS more statistical power than a standard hypothesis test on static data, the principal drawback of this method is that it is inappropriate for modeling complex traits that arise from interactions between genetic effects at different loci. A related approach extends the fGWAS framework to model multiple SNPs at once using a Bayesian group lasso framework (Li *et al.*, 2015). Although this approach seems promising, it is severely limited by its very slow MCMC inference procedure. There are a number of other methods that have been developed for dynamic trait GWAS, including Yang *et al.* (2009), Furlotte *et al.* (2012), Das *et al.* (2013), and Li and Sillanpää (2013). However, the majority of them either perform single-locus analysis (as in fGWAS) or fail to learn an explicit, interpretable representation of the dynamic effects of the genetic variants at each locus. The notable exception to this is fGWAS with Bayesian group lasso, which we directly compare to our approach in a later section.

In this work, we introduce a new penalized multivariate regression approach for GWAS of dynamic quantitative traits, in which the phenotype is modeled as a sum of nonparametric, time-varying SNP effects. We call this Time-Varying Group Sparse Additive Models, or TV-GroupSpAM. Our method is based on GroupSpAM (Yin *et al.*, 2012), a non-parametric regression model with a group-structured penalty over the input features, which we extend to capture the dynamic effects of SNPs. This model has three major advantages over existing approaches: (i) we leverage dynamic trait data; (ii) we model the contribution of each SNP to the phenotype as a smooth function of time, and explicitly learn these influence patterns; (iii) we model the combined effects of multiple SNPs on the phenotype and select a sparse subset that participate in the model, thereby identifying meaningful SNP-trait associations. We show that TV-GroupSpAM exhibits desirable empirical advantages over baseline methods on both simulated and real datasets.

2 Approach

In this section, we first introduce a time-varying additive model for dynamic complex traits that captures the underlying patterns of genetic effects. We then apply a group sparse regularization scheme to this model in order to impose bias useful for discovering a sparse set of markers that influence the phenotype in a longitudinal setting. Finally, we provide an efficient algorithm for parameter estimation, and thereby association mapping, under our model.

Notation. Let $X_{ij} \in \{0, 1, 2\} : i = 1, \ldots, n; j = 1, \ldots, p$ denote the genotype of individual i at SNP locus j, where n and p denote the number of individuals and SNPs, respectively. Let $Y_{i\tau} \in \mathbb{R} : i = 1, \ldots, n; \tau = 1, \ldots, m$ denote the phenotype value of individual i at the τth time point. Note that the exact time readings for different individuals at their τth time point may be different, i.e. the measurements are not necessarily time-aligned. We therefore introduce an explicit time variable $T_{i\tau} \in \mathbb{R}^+$ to capture the time reading for individual i at the τth time point, and define $Y_{i\tau} \equiv Y(T_{i\tau})$ as a stochastic process that

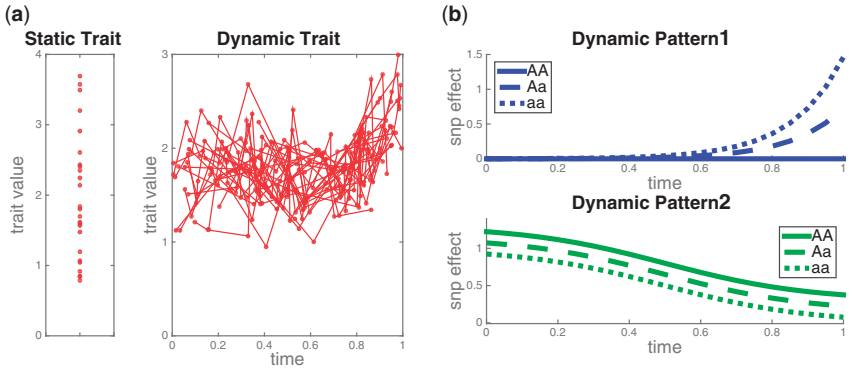

Fig. 1. GWAS has greater statistical power when dynamic traits are used. (a) A toy dataset illustrating the difference between static and dynamic traits. (b) Two synthetic examples of time-dependent patterns of SNP influence on the trait that would be difficult to detect with a static model (Color version of this figure is available at *Bioinformatics* online.)

captures the trait values at each time point. In what follows, we will use uppercase letters X, Y, T to denote random variables and lowercase letters x, y, t to denote their instantiated values.

2.1 Time-varying additive model

We consider the following time-varying additive model with scalar input variables $X_1, \ldots, X_p$ and functional response variable $Y(T)$:

$$Y(T) = f_0(T) + \sum_{j=1}^{p} f_j(T, X_j) + \omega(T) \tag{1}$$

Here $Y(T)$, which represents the trait value at time T, is decomposed into three terms: $f_0(T)$ is an intercept term that represents the non-genetic influence on the phenotype at time T; $f_j(T, X_j)$ represents the genetic effect of marker j with genotype X_j at time T; $\omega(T)$ is the noise term that models the random fluctuation of the underlying process.

Since X_j is a categorical variable, each bivariate component function f_j can be represented more simply as a set of three univariate functions of time, given by $f_j = \{f_j^0, f_j^1, f_j^2\}$. We can subsequently define $f_j(T, X_j) = \sum_g f_j^g(T)\, \mathbb{I}\{X_j = g\}$ where $f_j^g(\cdot) = f_j(\cdot, X_j = g)$. Next we simplify our notation by expanding each X_j into a set of three binary indicator variables such that $X_j^g = 1 \iff X_j = g$. This allows us to rewrite the model in the following form.

$$Y(T) = f_0(T) + \sum_{j=1}^{p} \sum_{g=0}^{2} f_j^g(T)\, X_j^g + \omega(T) \tag{2}$$

Note that the indicator variable X_j^g selects a single function among the set $\{f_j^0, f_j^1, f_j^2\}$ for each SNP.

In the data setting, since each observation is subject to measurement error, we assume $Y_{i\tau} = Y_i(T_{i\tau}) + \epsilon_{i\tau}$ where $\epsilon_{i\tau} \sim \mathcal{N}(0, \sigma^2)$. It follows from the model defined in (2) that the observed phenotypic values satisfy

$$y_{i\tau} = f_0(t_{i\tau}) + \sum_{j=1}^{p} \sum_{g=0}^{2} f_j^g(t_{i\tau})\, x_{ij}^g + \omega(t_{i\tau}) + \epsilon_{i\tau} \tag{3}$$

for subjects $i = 1, \ldots, n$ and measurements $\tau = 1, \ldots, m$. In the remainder of this article, we assume that the residual errors $e_{i\tau} = \omega(t_{i\tau}) + \epsilon_{i\tau}$ are i.i.d. across both subjects and measurements, though an alternative approach would be to impose an autocorrelation structure on $\omega(T)$ to capture the temporal pattern of the underlying longitudinal process (Das *et al.*, 2011; Li and Sillanpää, 2013).

In the model specified above, our only assumption about the genetic effects $\{f_j^0, f_j^1, f_j^2 : j = 1, \ldots, p\}$ is that they are smooth functions of time. A well-established approach to estimate nonparametric functions in additive models (Hastie and Tibshirani, 1990) is to minimize the expected squared error loss:

$$h(f) = \mathbb{E}\left[Y(T) - f_0(T) - \sum_{j=1}^{p} \sum_{g=0}^{2} f_j^g(T)\, X_j^g \right]^2 \tag{4}$$

where the expectation is calculated with respect to the distributions over SNP genotypes $(X_1, \ldots, X_p)$, time T, and phenotypic value Y. In the sample setting, this translates to minimizing

$$\widehat{h}(f) = \sum_{i=1}^{n} \sum_{\tau=1}^{m} \left(y_{i\tau} - f_0(t_{i\tau}) - \sum_{j=1}^{p} \sum_{g=0}^{2} f_j^g(t_{i\tau})\, x_{ij}^g \right)^2 \tag{5}$$

subject to a set of smoothness constraints. We go into detail about how to estimate the parameters of this model in Section 2.3.

2.2 Group sparse regularization

In a typical genome-wide association study, though a large number of markers are assayed, it is believed that only a small subset of them have a real effect on the trait of interest. This assumption motivates us to impose sparsity at the level of the SNPs $X_1, \ldots, X_p$ in the time-varying additive model of (2), such that the effects of many of these variables are zero. To achieve this, we apply a group-sparsity-inducing penalty that leads to shrinkage on the estimated effect of each locus as a whole, including the component functions for all genotypes and their values at all time points. Specifically, we employ a group norm penalty over the component functions in which each group consists of the three functions $\{f_j^0, f_j^1, f_j^2\}$ that correspond to a particular marker X_j.

To construct this group penalty, we use the $\ell_{1,2}$ norm first introduced in the context of the group lasso (Yuan and Lin, 2006). The new empirical objective function for our model is given by

$$\widehat{h}(f) + \lambda \sum_{j=1}^{p} \sqrt{\sum_{g=0}^{2} ||f_j^g||_2^2} \tag{6}$$

and is again subject to a set of smoothness constraints. Here $\lambda > 0$ is a tunable regularization parameter that controls the amount of sparsity in the model, and the squared ℓ_2 norm over f_j^g is defined as

$$||f_j^g||_2^2 = \sum_{i=1}^{n} \sum_{\tau=1}^{m} f_j^g(t_{i\tau})^2 x_{ij}^g \tag{7}$$

The penalty term in (6) induces sparsity at the level of groups by encouraging each set of functions $\{f_j^0, f_j^1, f_j^2\}$ to be set exactly to zero, which implies that the corresponding marker X_j has no effect whatsoever on the phenotype at any time point.

In what follows, we will refer to the model defined by the objective function in (6) as a Time-Varying Group Sparse Additive Model (TV-GroupSpAM). This model is based on both the Group Sparse Additive Model of Yin *et al.* (2012), in which a group sparse regularization penalty is applied to a standard additive model, and the Time-Varying Additive Model of Zhang *et al.* (2013), in which an unpenalized additive model is used to regress a functional response on scalar covariates.

2.3 Optimization algorithm

To estimate the parameters of the TV-GroupSpAM model, we use a block coordinate descent algorithm in which we optimize the objective with respect to a particular group of functions at once while all remaining functions are kept fixed.

Before presenting a complete algorithm for the regularized model, we first describe how to estimate the simpler, unpenalized model introduced in Section 2.1. Given the loss function of (4), some algebra shows that the optimal solution for f_j^g satisfies the following conditional expectation for each genetic marker $j = 1, \ldots, p$ and each genotype value $g \in \{0, 1, 2\}$.

$$f_j^g(T) = \mathbb{E}\left[Y(T) - f_0(T) - \sum_{k \neq j} \sum_{\ell} f_k^\ell(T)\, X_k^\ell \mid T, X_j = g \right] \tag{8}$$

It has been well established in the statistics literature that a scatterplot smoother matrix can be viewed as a natural estimate of the conditional expected value (Hastie and Tibshirani, 1990). To evaluate (8) in the sample setting, we therefore replace the conditional expectation operator $\mathbb{E}[\,\cdot \mid T, X_j = g]$ by left

multiplication with an n-by-n smoother matrix $\mathbf{S}_j^g = \{S_j^g[a,b]\}$, which is defined as

$$S_j^g[a,b] \propto K_h(|t^{(a)} - t^{(b)}|) \qquad \text{if } x_j^{(a)} = g \text{ and } x_j^{(b)} = g$$
$$S_j^g[a,b] = 0 \qquad \text{otherwise}$$

where (a, b) is a pair of data points, each corresponding to a particular individual i and time point τ, and K_h is a smoothing kernel function with bandwidth h. An alternative way to think about $\mathbf{S}_j^g$ is as the element-wise product of a smoother matrix for T, in which entry (a, b) is proportional to $K_h(|t^{(a)} - t^{(b)}|)$, and an indicator matrix for $X_j = g$, in which entry (a, b) is given by $\mathbb{I}\{x_j^{(a)} = x_j^{(b)} = g\}$. This makes intuitive sense because we want to estimate a smooth function over time for each genotype value of each SNP. Thus, to learn each function f_j^g for a particular SNP j and a particular genotype g, we only want to consider data points for which the genotype at SNP j is g and we want to smooth over time.

The empirical estimate of f_j^g will be a vector $\widehat{\mathbf{f}}_j^g \in \mathbb{R}^{nm}$ whose entries correspond to smoothed estimates of the effect of marker j with genotype g on the phenotype at each of the observed time points. Note that the entries of this vector corresponding to samples with genotype $\neq g$ for SNP j will be set to zero because the function is not applicable to those samples. In practice, we drop these dummy entries at the end to obtain our final function estimates. We calculate these estimates using the empirical formula for (8),

$$\widehat{\mathbf{f}}_j^g = \mathbf{S}_j^g \left(\mathbf{y} - \widehat{\mathbf{f}}_0 - \sum_{k \neq j} \sum_{\ell} \widehat{\mathbf{f}}_k^{\ell} \mathbb{I}\{\mathbf{x}_k = \ell\} \right) \tag{9}$$

where $\mathbf{y}$ is the vector of concatenated trait values for each sample, and $\mathbf{x}_k$ is the corresponding vector of genotypes at SNP k for each sample. Here a sample is a measurement for a specific individual i at a specific time point τ. Cycling though SNPs and genotypes and applying the update rule of (9) leads to a variant of the well-known backfitting algorithm. We refer the readers to Hastie and Tibshirani (1990) for details about smoothing and backfitting.

Finally, in order to optimize the penalized objective given in (6), we adapt the block coordinate descent and thresholding algorithms from Yin *et al.* (2012) to our setting. The complete optimization routine is shown in Algorithm 1. After smoothing the partial residual at each iteration, we perform a thresholding step by estimating the group norm $\widehat{w}_j$ and using it to determine whether the group of functions $\widehat{\mathbf{f}}_j$ should be set to zero. If not, we re-estimate the function values by iteratively solving a fixed point equation. We note that Step 9 of our algorithm runs more efficiently than the corresponding step of the thresholding algorithm presented in Yin *et al.* (2012) because we do not need to perform a matrix inversion on each iteration. This property results from the fact that within a particular group of function estimates, each one covers a disjoint set of observations, which simplifies the update equation.

Algorithm 1. Block Coordinate Descent for TV-GroupSpAM

1: **inputs:** genotypes $\mathbf{x}_1, \ldots, \mathbf{x}_p$, time points $\mathbf{t}$, trait values $\mathbf{y}$
2: initialize $\widehat{\mathbf{f}}_0 = \mathbf{0}$ and $\widehat{\mathbf{f}}_j^g = \mathbf{0}$ for $j = 1, \ldots, p$ and $g \in \{0, 1, 2\}$
3: **repeat**
4: update intercept term: $\widehat{\mathbf{f}}_0 = \mathbf{S}_0(\mathbf{y} - \sum_k \sum_\ell \widehat{\mathbf{f}}_k^{\ell} \mathbb{I}\{\mathbf{x}_k = \ell\})$
5: **for** $j = 1, \ldots, p$ **do**
6: compute partial residual:

$$\widehat{\mathbf{R}}_j = \mathbf{y} - \widehat{\mathbf{f}}_0 - \sum_{k \neq j} \sum_{\ell} \widehat{\mathbf{f}}_k^{\ell} \mathbb{I}\{\mathbf{x}_k = \ell\}$$

7: estimate projected residuals by smoothing:

$$\widehat{\mathbf{P}}_j^g = \mathbf{S}_j^g \ \widehat{\mathbf{R}}_j \quad \forall g$$

8: compute group norm:

$$\widehat{w}_j = \sqrt{\sum_{g=0}^{2} ||\widehat{\mathbf{P}}_j^g||_2^2}$$

9: **if** $\widehat{w}_j \leq \lambda$ **then** set $\widehat{\mathbf{f}}_j^g = 0 \quad \forall g$
10: **else** update $\widehat{\mathbf{f}}_j^g \ \forall g$ by iterating until convergence

$$\widehat{\mathbf{f}}_j^{g+} := (1 + \lambda / ||\widehat{\mathbf{f}}_j||_2)^{-1} \widehat{\mathbf{P}}_j^g$$

11: **end if**
12: center each $\widehat{\mathbf{f}}_j$ by subtracting its mean
13: **end for**
14: **until** convergence
15: **outputs:** estimates $\widehat{\mathbf{f}}_0$ and $\widehat{\mathbf{f}}_j = \{\widehat{\mathbf{f}}_j^0, \widehat{\mathbf{f}}_j^1, \widehat{\mathbf{f}}_j^2\}$ for $j = 1, \ldots, p$

3 Simulation study

In order to illustrate the utility of our method, we perform several experiments on synthetic data. We generate data according to the following procedure. First we construct a set of realistic genotypes X_{ij} by randomly subsampling individuals and SNPs from the real asthma dataset that we analyze in the next section. Next we independently sample time points $T_{i\tau} \sim \text{Unif}(0, 1)$ and measurement errors $\epsilon_{i\tau} \sim \mathcal{N}(0, 1)$. We select a subset of SNPs that will have nonzero contribution to the phenotype by placing their functions in an active set $\mathcal{A} \subseteq \{f_1, \ldots, f_p\}$. We then construct the active functions by sampling their values from a diverse set of predefined influence patterns that exhibit a variety of trait penetrance models (including additive, multiplicative, dominant, and recessive) and interact differently with time (including some static patterns for balance). All functions not in the active set, including the intercept term, are defined such that $f(t) = 0 \ \forall t$. Finally, we generate phenotype values $y_{i\tau}$ according to the model defined in (3).

To test the robustness of our model, we generate data according to two slightly different variants of (3). In the first setting, we uphold our original assumption that the residual errors are completely uncorrelated by independently generating $\omega_{i\tau} \sim \mathcal{N}(0, \sigma^2)$. In the second setting, we invalidate this assumption and introduce strong correlation among the errors across time by jointly generating $(\omega_{i1}, \ldots, \omega_{im}) \sim \mathcal{N}(0, \Sigma)$. In all of our experiments, we fix the number of samples at $n = 100$ and the number of time points at $m = 10$. Then, to evaluate our approach in a broad range of settings, we vary the total number of SNPs over $p \in \{50, 100, 200, 500, 1000, 2000, 5000\}$, which covers both the $p \leq n$ and $p > n$ cases, and vary the size of the active set over $|\mathcal{A}| \in \{5, 10, 20\}$.

We compare our method against several baselines, including single-marker hypothesis testing (using the Wald test), group lasso (where each group consists of the 3 genotype indicators for one SNP), fGWAS, and BGL-fGWAS. We used several software packages to run these methods: the PLINK toolkit (Purcell, 2009) for the Wald test, the SLEP Matlab package (Liu *et al.*, 2009) for lasso and group lasso, and the fGWAS2 R package (Wang and Li, 2012) for fGWAS and BGL-fGWAS. To run the static data methods (hypothesis test and group lasso), we summarize the phenotype values by averaging across time.

To evaluate performance, we calculate the maximum power attained by each method at a fixed false discovery rate. In order to

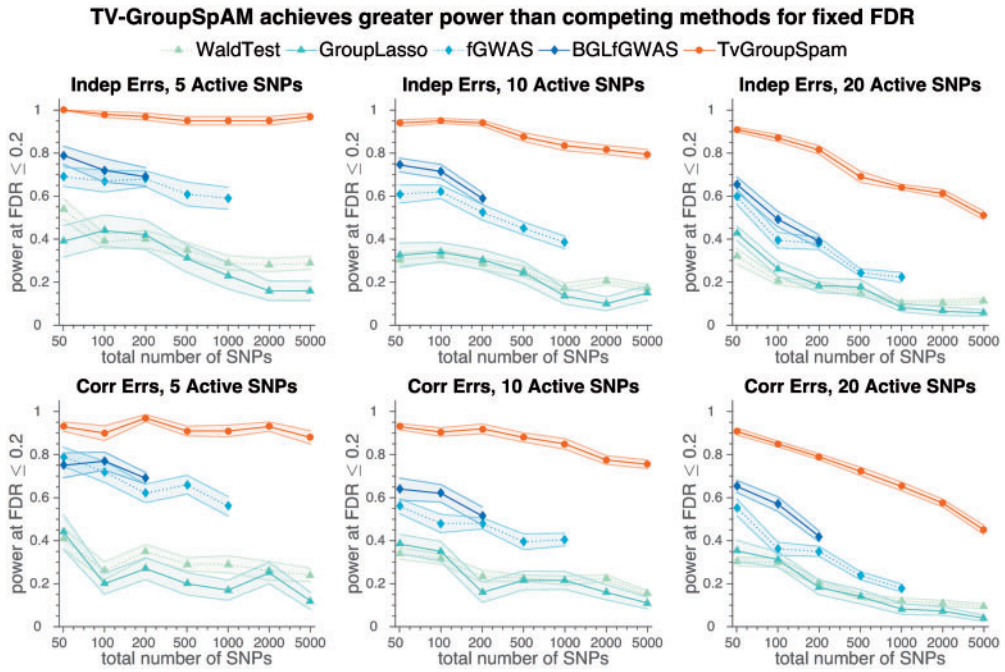

Fig. 2. Comparison of TV-GroupSpAM to baseline methods shows that our approach achieves greater power for a fixed false discovery rate (FDR $\leq$ 0.2). The results are averaged over 20 random synthetic datasets for each setting, and the shaded region denotes the standard error

calculate this metric, we first generate a ranked list of the top $|\mathcal{A}|$ SNPs identified by each method. For the Wald test and fGWAS, this is given by the SNPs with the smallest p-values. For the penalized regression methods, we test a series of values of the regularization parameter, λ, and select the one that yields approximately the desired number of SNPs. We then rank these SNPs according to their fitted model weights or norms. Given this list, we select a cutoff point that yields the largest set of SNPs such that FDR is below 0.2, and we calculate the power at this threshold. The results of our experiments are shown in Figure 2.

Our results indicate that TV-GroupSpAM far outperforms all of the baseline methods in every setting. In many cases, the three dynamic methods are able to detect at least twice as many true associations as the static methods. This underscores the value of leveraging longitudinal data to boost statistical power. The results show that TV-GroupSpAM outperforms fGWAS even when the residual errors are correlated, despite the fact that our model assumes independent errors while fGWAS does not. These results demonstrate that TV-GroupSpAM performs well under many different conditions and is robust to noise.

To obtain a more complete picture of the performance of each method, we plot the precision-recall curves obtained by varying the number of SNPs selected by each method from 0 to p. The average precision-recall curves obtained by averaging results over 20 datasets for the most challenging synthetic data setting ($p = 200$, $|\mathcal{A}| = 20$, correlated errors) are shown in Figure 3. We also report the area under the precision recall curve (AUCPR) for BGL-fGWAS and TV-GroupSpAM. Our approach outperforms the most competitive baseline by a significant margin. Lastly, we compare the run times of the three dynamic trait methods for different values of p, and show the results in Figure 4. For $p = 200$, TV-GroupSpAM ran in 12 minutes, fGWAS ran in 69 minutes, and BGL-fGWAS ran in 20 hours. These results show that our method is by far the most computationally efficient.

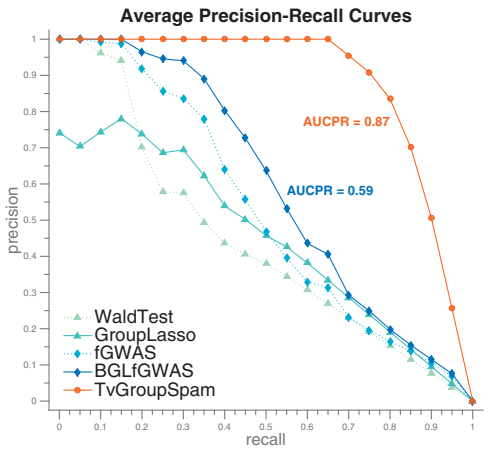

Fig. 3. Comparison of precision-recall curves of TV-GroupSpAM to baseline methods shows that our approach has an average AUCPR of 0.87 $\pm$0.01, which is much higher than the most competitive baseline, BGL-fGWAS, which has an average AUCPR 0.59 $\pm$ 0.02

4 Genome-wide association study of asthma

Next we use TV-GroupSpAM to perform a genome-wide association analysis of asthma traits. We look for associations between SNPs and two quantitative phenotypes frequently used to assess asthma severity: the forced vital capacity (FVC), a sensitive measure of airway obstruction, and bronchodilator response (BDR), which measures lung response to bronchodilator drugs. For this analysis, we use data from the CAMP longitudinal study of childhood asthma (Childhood Asthma Management Program Research Group *et al.*, 1999) with $n = 552$ subjects genotyped at $p = 510\,540$ SNPs from across all 22 autosomal chromosomes. After preprocessing, in which we removed subjects with missing data and SNPs with minor allele frequency below 0.05, we were left with $n = 465$ and $p = 509\,299$.

In order to control for non-genetic effects, we incorporated several static covariates into our model, including: sex, race, the age of onset of asthma, the clinic where the patient's traits were measured, and the treatment or control group to which the patient was assigned in the clinical trial associated with the CAMP study.

For computational efficiency, we first used our approach to filter out a relatively small set of SNPs to include in the final analysis for each phenotype. To do this, we split the dataset into 100 subsets, each containing ∼5000 SNPs, and ran TV-GroupSpAM separately on each set. We regulated the model sparsity by using a binary search procedure to identify a value of λ that selected between 90 and 110 SNPs from each subset, following the example of Wu *et al.* (2009). This yielded a filtered set of 10 118 SNPs for the FVC model and 9621 SNPs for the BDR model. Figure 5 shows the model weight (an indicator of significance) of every SNP that was selected in the filtering step for each phenotype. Next we fit a new global model for each trait using only these selected SNPs, and chose a value of λ that yielded approximately 50 SNPs with nonzero effect on the phenotype (yielding 48 for FVC and 51 for BDR). Finally, we refit the model on just these selected SNPs with no regularization penalty, and use the estimated group functional norms to determine the effect size of each SNP. The complete sets of selected SNPs identified for each trait are listed in Supplementary Tables S1 and S2. Note that the FVC effect sizes are much higher in magnitude than the BDR effect sizes because the FVC phenotype is measured in different units than the BDR phenotype.

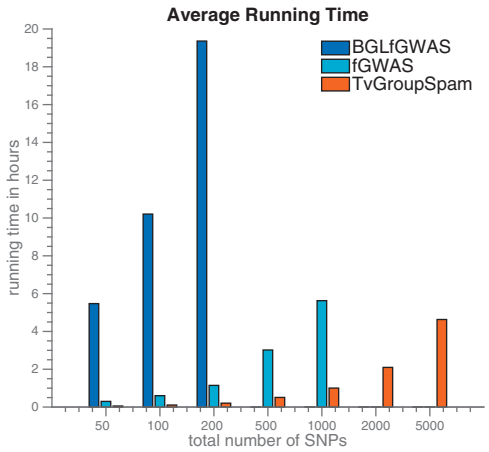

Fig. 4. Comparison of the running time of TV-GroupSpAM to baseline methods shows that our approach runs much faster than both fGWAS and BGL-fGWAS. We were unable to run fGWAS for $p > 1000$ or BGL-fGWAS for $p > 200$ due to time constraints (Color version of this figure is available at *Bioinformatics* online.)

In order to analyze the validity of our results, we identified all genes located within 500 Kb of each SNP and then determined whether any of the genetic loci or nearby genes are known to be associated with asthma or asthma-related functions in the existing literature. Because asthma is a disease characterized by inflammation and constriction of the airways of the lungs, we specifically searched for genes that have been linked to lung function or inflammatory response. Furthermore, since asthma is partly driven by a series of interactions between vascular endothelial cells and leukocytes (Bijanzadeh *et al.*, 2011), we also searched for genes involved in functions of the vascular system or the immune system, particularly those in pathways involving T-helper 2 (Th2) cells, which play a central role in the pathogenesis of asthma (Ober and Yao, 2011).

We list a curated subset of the SNPs selected in the FVC and BDR models in Tables 1 and 2, along with the nearby genes that can be linked to asthma. Our model was able to identify several genetic loci that have a well-established connection to asthma. For example, SNP rs6116189 on chromosome 20 is located near the ADAM33 gene, which has been implicated in asthma by several independent studies (Ober and Hoffjan, 2006). In addition, SNP rs1450118 on chromosome 3 is located near IL1RAP, a gene that produces the Interleukin 1 receptor accessory protein and plays an important role in asthma (Ober and Yao, 2011). Finally, the locus on chromosome 7 at 139.3 Mb is particularly interesting because it was selected in both the FVC and BDR models. This SNP is located near the TBXAS1 gene, which has been linked to asthma (Oh *et al.*, 2011). We plot some examples of the estimated time-varying effects of SNPs selected in our FVC and BDR models in Figure 6.

Finally, in order to evaluate the sensitivity of TV-GroupSpAM to noise in the data, we returned to the two filtered sets of ∼10 000 SNPs each and reran the final selection step on multiple 90% subsamples of the data, then analyzed the stability of the set of selected SNPs. Because the stability naturally varies with the total number of SNPs being selected, we ran our algorithm on each subsample for a fixed set of λ values such that the fraction of selected SNPs ranged from 0.5% to nearly 100%. We then calculated the average stability for a particular value of λ as the average pairwise overlap among the selected SNPs divided by the average number of SNPs selected across all subsamples. We plot the stability as a function of the average percentage of SNPs selected in Figure 7, with the shaded region showing the standard deviation. These results indicate that the stability of the FVC model when selecting 0.47% of SNPs (48 out of 10 118) is 32% and the stability of the BDR model when selecting 0.53% of SNPs (51 out of 9621) is 39%.

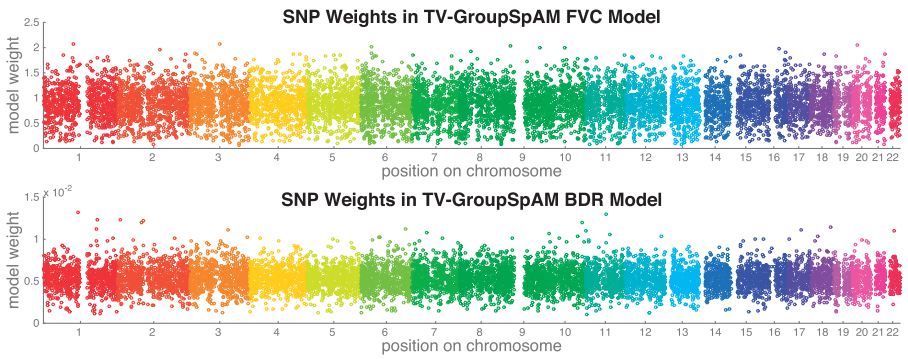

Fig. 5. Manhattan plots of the model weights for each SNP that was selected in the FVC model (left) and BDR model (right) during the filtering stage (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Selected SNPs associated with FVC

SNP	Chrom	Location	Effect size	Nearby genes linked to asthma
rs6442021	3	46.7 Mb	1.5303	*CCR1, CCR2, CCR3, CCR5*—chemokine receptors in the CC family; CCR2 is a receptor for a protein that plays a role in several inflammatory diseases, and has been directly linked to asthma (Batra and Ghosh, 2009); CCR3 may play a role in airway inflammation (NCBI, 2005) *PRSS42, PRSS46, PRSS45, PRSS50*—trypsin-like serine proteases; tryptases are known to cause bronchoconstriction and have been implicated in asthma (Zhang and Timmerman, 1997)
rs2062583	3	56.9 Mb	1.0074	*IL17RD*—interleukin 17 receptor D; IL-17 is a pro-inflammatory cytokine produced by Th17 cells that plays a role in multiple inflammatory diseases, including asthma (Manni *et al.*, 2014)
rs1450118	3	190.4 Mb	0.9027	*IL1RAP*—interleukin 1 receptor accessory protein; enables the binding of IL-33 to its receptor encoded by IL1RL1, which has been repeatedly linked to asthma (Ober and Yao, 2011)
rs3801148	7	139.3 Mb	0.8538	*TBXAS1*—thromboxane A synthase; this enzyme converts prostaglandin H2 to thromboxane A2, a lipid that constricts smooth respiratory muscle (Oh *et al.*, 2011)
rs914978	9	132.3 Mb	1.0631	*PTGES*—prostaglandin E synthase; this enzyme converts prostaglandin H2 to prostaglandin E2, a lipid inflammatory mediator that acts in the lung (Liu *et al.*, 2012)
rs11069178	12	117.9 Mb	0.6869	*NOS1*—nitric oxide synthase 1; nitric oxide levels are elevated in the air exhaled by asthmatics; NOS1 has been linked to a higher risk of asthma (Gao *et al.*, 2000)

Table 2. Selected SNPs associated with BDR

SNP	Chrom	Location	Effect size	Nearby genes linked to asthma
rs7766818	6	46.8 Mb	0.0088	*GPR116*—probable G protein-coupled receptor 116; plays a critical role in lung surfactant homeostasis (Yang *et al.*, 2013) *TNFRSF21*—tumor necrosis factor receptor superfamily member 21; plays a central role in regulating immune response and airway inflammation in mice (Venkataraman *et al.*, 2006)
rs12524603	6	159.8 Mb	0.0075	*SOD2*—superoxide dismutase 2, mitochondrial; plays a role in oxidative stress, and has been linked to bronchial hyperresponsiveness and COPD (Siedlinski *et al.*, 2009)
rs13239058	7	139.3 Mb	0.0079	*TBXAS1*—see Table 1 above
rs10519096	15	59.1 Mb	0.0086	*ADAM10*—disintegrin and metalloproteinase domain-containing protein 10; plays an important role in immunoglobulin E dependent lung inflammation (Mathews *et al.*, 2011)
rs8111845	19	41.6 Mb	0.0066	*TGFB1*—transforming growth factor $\beta 1$; has pro-inflammatory as well as anti-inflammatory properties, and has been linked to asthma and airway remodeling (Nagpal *et al.*, 2005) *CYP2A6, CYP2A7, RAB4B, MIA, EGLND*—genes located in a known COPD locus (Bossé, 2012)
rs6116189	20	4.0 Mb	0.0067	*ADAM33*—disintegrin and metalloproteinase domain-containing protein 33; has been implicated in asthma by several independent studies (Ober and Hoffjan, 2006; Van Eerdewegh *et al.*, 2002)

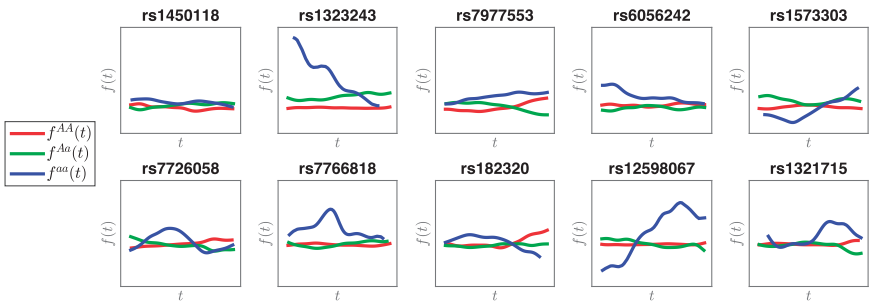

Fig. 6. Examples of estimated dynamic SNP effects. The top row shows four SNPs from the FVC model. The bottom row shows four SNPs from the BDR model (Color version of this figure is available at *Bioinformatics* online.)

5 Discussion

In this work, we propose a new approach to GWAS that bridges the gap between existing penalized regression methods, such as the lasso and group lasso, and dynamic trait methods, such as fGWAS. Our approach uses penalized regression to identify a sparse set of SNPs that jointly influence a dynamic trait. This is a challenging task for several reasons: first, we must contend with high-dimensional data, which necessitates that we regularize the model to perform variable selection; second, we do not know the true underlying model by which each SNP acts on the phenotype, and therefore we must avoid making parametric assumptions about these patterns; third, we assume that SNP effects vary smoothly over time, which means that we cannot apply a standard multi-task regression model that treats the time series as a set of unordered traits.

Although TV-GroupSpAM achieves significantly better performance on synthetic data than existing methods, there are still certain challenging aspects of genome-wide association mapping that are not addressed by this approach. One of these is the task of rare variant detection. Although our method is robust to detecting spurious effects from rare variants, we are also not able to detect true effects from rare variants with high power. This is due to the lack of data available for the *aa* genotype in SNPs with very low minor allele

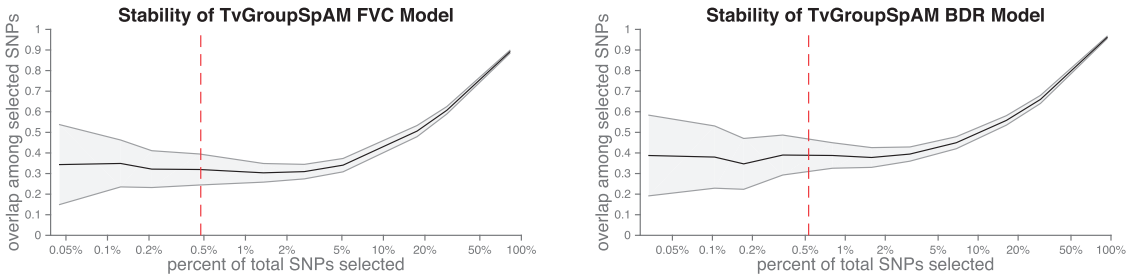

Fig. 7. Average stability of the FVC model (left) and BDR model (right) for different fractions of selected SNPs. Shaded region denotes the standard deviation. Vertical dashed line indicates the fraction of the filtered SNPs that we selected in our final analysis that yielded 48 SNPs for FVC and 51 SNPs for BDR

frequency; because we estimate a separate effect function for each SNP genotype, we are unable to accurately estimate f^{aa} when there are very few data points corresponding to this genotype. Modifying TV-GroupSpAM to more accurately detect the effects of rare variants would be an interesting direction for future work.

Funding

This work was supported by the National Institutes of Health [grant numbers R01-GM093156, P30-DA035778]. Micol Marchetti-Bowick is partly supported by a National Science Foundation Graduate Research Fellowship [under grant number DGE-1252522]. Junming Yin is partly supported by a Ray and Stephanie Lane Research Fellowship from CMU and a research award from the Center for Management Innovations in Health Care at the Eller College of Management.

Conflict of Interest: none declared.

References

Batra,J. and Ghosh,B. (2009) Genetic contribution of chemokine receptor 2 (CCR2) polymorphisms towards increased serum total IgE levels in Indian asthmatics. *Genomics*, **94**, 161–168.

Bijanzadeh,M. *et al.* (2011) An understanding of the genetic basis of asthma. *Indian J. Med. Res.*, **134**, 149.

Bossé,Y. (2012) Updates on the COPD gene list. *Int. J. Chronic Obstruct. Pulmon. Dis.*, **7**, 607.

Childhood Asthma Management Program Research Group. *et al.* (1999) The childhood asthma management program (CAMP) design, rationale, and methods. *Control. Clin. Trials*, **20**, 91–120.

Clarke,G. *et al.* (2011) Basic statistical analysis in genetic case-control studies. *Nat. Protoc.*, **6**, 121–133.

Das,K. *et al.* (2011) A dynamic model for genome-wide association studies. *Hum. Genet.*, **129**, 629–639.

Das,K. *et al.* (2013) Dynamic semiparametric bayesian models for genetic mapping of complex trait with irregular longitudinal data. *Stat. Med.*, **32**, 509–523.

Ferreira,M.A. *et al.* (2005) Robust estimation of experimentwise p values applied to a genome scan of multiple asthma traits identifies a new region of significant linkage on chromosome 20q13. *Am. J. Hum. Genet.*, **77**, 1075–1085.

Furlotte,N. *et al.* (2012) Genome-wide association mapping with longitudinal data. *Genet. Epidemiol.*, **36**, 463–471.

Gao,P. *et al.* (2000) Variants of NOS1, NOS2, and NOS3 genes in asthmatics. *Biochemical and Biophys. Res. Commun.*, **267**, 761–763.

Hastie,T. and Tibshirani,R. (1990). *Generalized Additive Models*. CRC Press.

Hindorff,L. *et al.* (2015). A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies.

Li,J. *et al.* (2011) The bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**, 516–523.

Li,J. *et al.* (2015) Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat.*, **9**, 640–664.

Li,Z. and Sillanpää,M. (2013) A bayesian nonparametric approach for mapping dynamic quantitative traits. *Genetics*, **194**, 997–1016.

Lin,Y.J. *et al.* (2015) Genetic variants in PLCB4/PLCB1 as susceptibility loci for coronary artery aneurysm formation in Kawasaki disease in Han Chinese in Taiwan. *Sci. Rep.*, **5**, 14762.

Liu,J. *et al.* (2009). *SLEP: Sparse Learning with Efficient Projections*. Arizona State University. Available at: http://www.yelab.net/software/SLEP/.

Liu,T. *et al.* (2012) Prostaglandin E2 deficiency uncovers a dominant role for thromboxane A2 in house dust mite-induced allergic pulmonary inflammation. *Proc. Natl. Acad. Sci.*, **109**, 12692–12697.

Manni,M.L. *et al.* (2014) A tale of two cytokines: Il-17 and IL-22 in asthma and infection. *Exp. Rev. Respir. Med.*, **8**, 25–42.

Mathews,J.A. *et al.* (2011) A potential new target for asthma therapy: a disintegrin and metalloprotease 10 (ADAM10) involvement in murine experimental asthma. *Allergy*, **66**, 1193–1200.

Nagpal,K. *et al.* (2005) TGFβ1 haplotypes and asthma in Indian populations. *J. Allergy Clin. Immunol.*, **115**, 527–533.

NCBI. (2005). Entrez Gene Database. Available at: http://www.ncbi.nlm.nih.gov/gene/1232.

Ober,C. and Hoffjan,S. (2006) Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun.*, **7**, 95–100.

Ober,C. and Yao,T.C. (2011) The genetics of asthma and allergic disease: a 21st century perspective. *Immunol. Rev.*, **242**, 10–30.

Oh,S.H. *et al.* (2011) Association analysis of thromboxane A synthase 1 gene polymorphisms with aspirin intolerance in asthmatic patients. *Pharmacogenomics*, **12**, 351–363.

Purcell,S. (2009). PLINK 1.07. Available at: http://pngu.mgh.harvard.edu/~purcell/plink/.

Siedlinski,M. *et al.* (2009) Superoxide dismutases, lung function and bronchial responsiveness in a general population. *Eur. Respir. J.*, **33**, 986–992.

Van Eerdewegh,P. *et al.* (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature*, **418**, 426–430.

Venkataraman,C. *et al.* (2006) Death receptor-6 regulates the development of pulmonary eosinophilia and airway inflammation in a mouse model of asthma. *Immunol. Lett.*, **106**, 42–47.

Wang, Z. and Li, J. (2012). fGWAS2. Available at: http://www.psu.edu/dept/statgen/software/fgwas-r2.html.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yang,J. *et al.* (2009) Nonparametric functional mapping of quantitative trait loci. *Biometrics*, **65**, 30–39.

Yang,M.Y. *et al.* (2013) Essential regulation of lung surfactant homeostasis by the orphan G protein-coupled receptor GPR116. *Cell Reports*, **3**, 1457–1464.

Yin,J. *et al.* (2012) Group sparse additive models. *Proceedings of the 29th International Conference on Machine Learning*, pp. 871–878

You,J. *et al.* (2010) PLC/CAMK IV-NF-κb involved in the receptor for advanced glycation end products mediated signaling pathway in human endothelial cells. *Mol. Cell. Endocrinol.*, **320**, 111–117.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, **68**, 49–67.

Zhang,M. and Timmerman,H. (1997) Mast cell tryptase and asthma. *Mediators Inflamm.*, **6**, 311–317.

Zhang,X. *et al.* (2013) Time-varying additive models for longitudinal data. *J. Am. Stat. Assoc.*, **108**, 983–998.