# Supporting Information: Text

## BAKR Binary Classification Model

Here, we detail an extension for the BAKR modeling framework when carrying out binary classification. Assume $\mathbf{y} = [\mathrm{y}_1, \ldots, \mathrm{y}_n]^\intercal$ is a vector of binary response variables. We specify the following generalized kernel model (Mallick et al., 2005; Chakraborty et al., 2007; Chakraborty, 2009; Zhang et al., 2011; Chakraborty et al., 2012) in matrix notation

$$\mathbf{y} \sim p(\mathbf{y} \,|\, \boldsymbol{\mu}) \quad \text{with} \quad g^{-1}(\boldsymbol{\mu}) = \widetilde{\mathbf{K}}\boldsymbol{\alpha}.$$

In the case of binary responses, typical link functions used for $g$ are the probit or logit functions. In the current work, we use the probit link due to its tractability for calculating marginal likelihood estimates. We now specify the hierarchical classification model using the empirical kernel factor representation of BAKR where $\widetilde{\mathbf{K}} = \widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{U}}^\intercal$:

$$\begin{aligned} \mathrm{y}_i &= \begin{cases} 1 & \text{if} \quad \mathrm{s}_i > 0 \\ 0 & \text{if} \quad \mathrm{s}_i \leq 0, \end{cases} \\ \mathbf{s} &= \widetilde{\mathbf{U}}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}), \\ \boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}, \sigma^2\widetilde{\boldsymbol{\Lambda}}), \\ \sigma^2 &\sim \text{Scale-inv-}\chi^2(\nu, \phi). \end{aligned} \tag{S1}$$

Under this specification, the responses in the kernel model are latent variables with standard normal errors. We denote this vector of latent responses as $\mathbf{s} = [\mathrm{s}_1, \ldots, \mathrm{s}_n]^\intercal$.

The Gibbs sampler for the model specified in Equation (S1) is a slight adaptation of the sampler specified for the approximate nonlinear regression specified in the main text. The MCMC procedure detailed here is similar to the standard posterior sampling scheme for probit regression (Albert and Chib, 1993). Posterior samples are generated by iterating the following conditional densities:

(1) For $i = 1, .., n$

$$\mathrm{s}_i^{(t+1)} \,|\, \boldsymbol{\theta}, \sigma^2, \mathbf{s}^{(t)}, \mathbf{y} \sim \begin{cases} \mathrm{N}(\tilde{\mathbf{u}}_i^\intercal\boldsymbol{\theta}, 1)\mathbb{1}(\mathrm{s}_i^{(t)} \leq 0) & \text{if} \quad \mathrm{s}_i^{(t)} \leq 0 \\ \mathrm{N}(\tilde{\mathbf{u}}_i^\intercal\boldsymbol{\theta}, 1)\mathbb{1}(\mathrm{s}_i^{(t)} > 0) & \text{if} \quad \mathrm{s}_i^{(t)} > 0; \end{cases}$$

(2) $\boldsymbol{\theta} \,|\, \mathbf{s}, \sigma^2, \mathbf{y} \sim \mathrm{MVN}(\mathbf{m}^*, \mathbf{V}^*)$ where $\mathbf{V}^* = \sigma^2(\widetilde{\boldsymbol{\Lambda}}^{-1} + \sigma^2\mathbf{I})^{-1}$ and $\mathbf{m}^* = \mathbf{V}^*\widetilde{\mathbf{U}}^\intercal\mathbf{s}$;

(3) $\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\widetilde{\boldsymbol{\Psi}}^\intercal(\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{U}}^\intercal\widetilde{\mathbf{K}}^{-1}\widetilde{\boldsymbol{\Psi}}^\intercal)^{-1}\boldsymbol{\theta}$;

(4) $\sigma^2 \,|\, \mathbf{s}, \boldsymbol{\theta}, \mathbf{y} \sim \text{Scale-inv-}\chi^2(\nu^*, \phi^*)$ where $\nu^* = \nu + q$ and $\phi^* = \nu^{*-1}(\nu\phi + \boldsymbol{\theta}^\intercal\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\theta})$.

Again, the third step is deterministic and allows for posterior inferences to be made in the original covariate space. Iterating the above procedure $T$ times will result in the following set of posterior draws: $\left\{\boldsymbol{\theta}^{(t)}, \sigma^{2(t)}, \widetilde{\boldsymbol{\beta}}^{(t)}\right\}_{t=1}^{T}$.

## BAKR Mixed Model Extension

There are many applications where a mixed modeling framework is desired. Examples of this include cases where the observations are not independent but related via some genetic population structure or known kinship, or cases where one needs to control for covariates and other fixed effects (e.g. age, sex, or genotype principal components). Here, we detail a mixed model extension of BAKR. The extension to binary classification is straightforward and based on the steps outlined for the BAKR-probit model in the main text. One can adapt the representation of BAKR to include a random component as follows:

$$\mathbf{y} = \widetilde{\mathbf{K}}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \tau^2\mathbf{I}) \tag{S2}$$

where, $\mathbf{Z}$ contains covariates representing additional population structure, and $\boldsymbol{\delta}$ are the corresponding random effects assumed to be normally distributed with mean $\mathbf{0}$ and some covariance structure $\boldsymbol{\Delta}$. In

many statistical genetics applications, $\boldsymbol{\Delta}$ is not assumed to be diagonal or block-diagonal, which implies that the elements in the response vector $\mathbf{y}$ are correlated via the random effects (Liu et al., 2007). The relevance of the random effects is that they capture a key proportion of the phenotypic variance that allows for more accurate statistical inferences. This correction can increase the model's power to detect true causal variants, rather than falsely identifying significant covariates that may have large effect sizes simply due to correlations within the population structure (Kang et al., 2008, 2010; Zhang et al., 2010; Yang et al., 2014). This flexibility of the mixed modeling approach is a major reason why it is used in applications such as genome-wide association studies (GWAS) (Yang et al., 2014).

We now specify the following empirical factor hierarchical mixed model

$$\begin{aligned} \mathbf{y} &= \widetilde{\mathbf{U}}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \tau^2\mathbf{I}), \\ \boldsymbol{\theta} &\sim \text{MVN}(\mathbf{0}, \sigma_\theta^2\widetilde{\boldsymbol{\Lambda}}), \\ \boldsymbol{\delta} &\sim \text{MVN}(\mathbf{0}, \sigma_\delta^2\boldsymbol{\Delta}), \\ \sigma_\theta^2, \sigma_\delta^2, \tau &\sim \text{Scale-Inv-}\chi^2(\nu, \phi), \end{aligned} \tag{S3}$$

Note that the model specification is almost identical to the original BAKR formulation — the difference being the addition of simulating the random effects from the kinship matrix $\boldsymbol{\Delta}$. We call this version of the model the BAKR mixed model (BAKR-MM).

Given the model specification in Equation (S3), we can again use a Gibbs sampler to draw from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\theta^2, \sigma_\delta^2, \tau^2 \,|\, \mathbf{y})$. The Gibbs sampler consists of iterated sampling of the following conditional densities:

(1) $\boldsymbol{\theta} \,|\, \boldsymbol{\delta}, \sigma_\theta^2, \sigma_\delta^2, \tau^2, \mathbf{y} \sim \text{MVN}(\mathbf{m}_\theta^*, \mathbf{V}_\theta^*)$ with $\mathbf{V}_\theta^* = \tau^2\sigma_\theta^2(\tau^2\widetilde{\boldsymbol{\Lambda}}^{-1} + \sigma_\theta^2\mathbf{I})^{-1}$ and $\mathbf{m}_\theta^* = \tau^{-2}\mathbf{V}_\theta^*\widetilde{\mathbf{U}}^\intercal(\mathbf{y} - \mathbf{Z}\boldsymbol{\delta})$;

(2) $\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\widetilde{\boldsymbol{\Psi}}^\intercal(\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{U}}^\intercal\widetilde{\mathbf{K}}^{-1}\widetilde{\boldsymbol{\Psi}}^\intercal)^{-1}\boldsymbol{\theta}$;

(3) $\boldsymbol{\delta} \,|\, \boldsymbol{\theta}, \sigma_\theta^2, \sigma_\delta^2, \tau^2, \mathbf{y} \sim \text{MVN}(\mathbf{m}_\delta^*, \mathbf{V}_\delta^*)$ with $\mathbf{V}_\delta^* = \tau^2\sigma_\delta^2(\tau^2\boldsymbol{\Delta}^{-1} + \sigma_\delta^2\mathbf{I})^{-1}$ and $\mathbf{m}_\delta^* = \tau^{-2}\mathbf{V}_\delta^*\mathbf{Z}^\intercal(\mathbf{y} - \widetilde{\mathbf{U}}\boldsymbol{\theta})$;

(4) $\sigma_\theta^2 \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\delta^2, \tau^2, \mathbf{y} \sim \text{Scale-inv-}\chi^2(\nu_\theta^*, \phi_\theta^*)$ where $\nu_\theta^* = \nu + q$ and $\phi_\theta^* = \nu_\theta^{*-1}(\nu\phi + \boldsymbol{\theta}^\intercal\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\theta})$;

(5) $\sigma_\delta^2 \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\theta^2, \tau^2, \mathbf{y} \sim \text{Scale-inv-}\chi^2(\nu_\delta^*, \phi_\delta^*)$ where $\nu_\delta^* = \nu + m$ and $\phi_\delta^* = \nu_\delta^{*-1}(\nu\phi + \boldsymbol{\delta}^\intercal\boldsymbol{\Delta}^{-1}\boldsymbol{\delta})$;

(6) $\tau^2 \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\theta^2, \sigma_\delta^2, \mathbf{y} \sim \text{Scale-inv-}\chi^2(\nu_\tau^*, \phi_\tau^*)$ where $\nu_\tau^* = \nu + n$ and $\phi_\tau^* = \nu_\tau^{*-1}(\nu\phi + \mathbf{e}^\intercal\mathbf{e})$, with $\mathbf{e} = \mathbf{y} - \widetilde{\mathbf{U}}\boldsymbol{\theta} - \mathbf{Z}\boldsymbol{\delta}$.

Once again, the second step is deterministic and maps back to the effect size analogs. Iterating the above procedure $T$ times results in a set of samples $\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T$. Prediction under this mixed modeling extension is similar to the procedure we presented in the main text. More specifically, given an out of sample test set $\mathbf{X}^*$ with corresponding covariates $\mathbf{Z}^*$, draws from the conditional posterior predictive distribution are given as

$$\left\{\mathbf{y}^{*(t)} = \mathbf{X}^*\widetilde{\boldsymbol{\beta}}^{(t)} + \mathbf{Z}^*\boldsymbol{\delta}^{(t)}\right\}_{t=1}^T. \tag{S4}$$

Note that this is also similar to other mixed effect modeling strategies (Skrondal and RabeHesketh, 2009).

**Alternative Specifications.** In the event that specific covariates are not known, posterior inference of BAKR-MM may easily be adapted to mirror that of a Bayesian Gaussian process or any other standard Bayesian nonparametric statistical method (Mallick et al., 2005; Chakraborty et al., 2007; Chakraborty, 2009; Liang et al., 2009; de los Campos et al., 2010; Zhang et al., 2011; Chakraborty et al., 2012). In these cases, the response variables to be predicted are simply treated as missing random variables that we will impute. The MCMC algorithm above can be easily adapted to allow for the sampling of the missing response variables. Briefly, we partition the vector of response variables $\mathbf{y}$ into a set of training $\mathbf{y}_r$ and validation samples $\mathbf{y}_v$. The design matrix can be similarly partitioned into $[\mathbf{X}_r; \mathbf{X}_v]$. Under a randomized feature map, the approximate kernel matrix $\widetilde{\mathbf{K}}$, and its eigenvalue decomposition $\widetilde{\mathbf{U}}$, are computed based on the full design matrix $\mathbf{X}$. The matrix $\mathbf{X}_v$ implicitly forms part of the kernel model prior structure, even though the

corresponding responses are missing. We now add an additional step to the MCMC procedure where $\mathbf{y}_v$ is imputed from the implied conditional posterior, which will be a draw from multivariate normal distribution for this model.

There are some issues to consider with this model specification and inference procedure. Here, posterior inferences are made using all the data, including $\mathbf{X}_v$. Therefore, if any new validation samples are introduced, the entire posterior analyses must be repeated (West, 2003; Mallick et al., 2005; Chakraborty et al., 2007; Chakraborty, 2009; Liang et al., 2007, 2009). Furthermore, posterior inferences on the original covariate effect sizes begin to lose meaning and interpretability when the sample size of the training set is smaller than that of the validation set (i.e. $n_r < n_v$). Often the objective is to make inferences on a set of explanatory variables, while correcting for population structure — meaning, there is no testing set to be considered.

## Identifiability of the Effect Size Analog

The space of nonlinear functions we consider in this paper is the subspace of the RKHS realized by the representer theorem. Namely,

$$\mathcal{H}_{\mathbf{x}} = \left\{ f \,|\, f(\mathbf{x}) = \boldsymbol{\Psi}_{\mathbf{x}}^{\intercal}\mathbf{c} \text{ and } \|f\|_{\mathrm{K}}^2 < \infty \right\}$$

with $\boldsymbol{\Psi}_{\mathbf{x}} = [\boldsymbol{\psi}(\mathbf{x}_1), \ldots, \boldsymbol{\psi}(\mathbf{x}_n)]$. Here, the coefficients $\mathbf{c}$ determine the nonlinear function. Note that the empirical kernel factor model implicitly models these coefficients via a lower dimensional representation with effect sizes $\boldsymbol{\theta}$. For convenience in this analysis, we consider the full parameter space $\mathbf{c}$, rather than the reduced rank parameter space $\boldsymbol{\theta}$.

A reasonable identifiability requirement for the effect size analog is that two different functions in $\mathcal{H}_{\mathbf{x}}$ will result in two different vectors for $\widetilde{\boldsymbol{\beta}}$. This requirement can be restated as the projection $\widetilde{\mathbf{P}} = \mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}$ should be an injective map from $\mathbf{c}$ to $\widetilde{\boldsymbol{\beta}}$. Since we are working with an approximation of the Gaussian kernel and $p \gg n$, we can assume that the approximate shift-invariant kernel matrix $\widetilde{\mathbf{K}} = \widetilde{\boldsymbol{\Psi}}^{\intercal}\widetilde{\boldsymbol{\Psi}}$ is positive definite. First we consider the classic linear regression setting where

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\dagger}\mathbf{y},$$

where $\mathbf{X}^{\dagger}$ is the Moore-Penrose pseudoinverse. Observe that two vectors $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ that only differ in the null space of $\mathbf{X}$ will give rise to the same response $\mathbf{y}$. This same issue will arise for our nonlinear effect size analog. Hence, the statement we will make about the injectivity of the map $\widetilde{\mathbf{P}}$ will hold modulo the null space of $\mathbf{X}$.

**Claim S1.** *Consider an approximate shift-invariant kernel matrix $\widetilde{\mathbf{K}}$ that is strictly positive definite with random feature map $\tilde{\psi} : \mathbb{R}^p \to \mathbb{R}^p$. The projection $\widetilde{\mathbf{P}}$ is injective for any coefficient vector for which the projection $\widetilde{\mathbf{P}}$ is in the span of the design matrix $\mathbf{X}$. Alternatively, the projection $\widetilde{\mathbf{P}}$ is injective for the span of the design matrix, span($\mathbf{X}$).*

*Proof.* Consider positive definite kernel matrices $\widetilde{\mathbf{K}}$. The assumption that the approximate kernel matrix is positive definite is key as it implies that $\widetilde{\boldsymbol{\Psi}}$ spans the entire $p$-dimensional covariate space. In the case that $\widetilde{\mathbf{K}}$ is positive semi-definite, we would have to understand the composition of the null space of $\widetilde{\mathbf{K}}$ with the null space of the design matrix $\mathbf{X}$.

Let $\mathbf{c}_1$ and $\mathbf{c}_2$ be two different coefficient vectors corresponding to the restricted RKHS subspace. There exists $\boldsymbol{\delta}$ such that $\mathbf{c}_2 = \mathbf{c}_1 + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \neq \mathbf{0}$ and

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_1 &= \mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}\mathbf{c}_1 \\ \widetilde{\boldsymbol{\beta}}_2 &= \mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}\mathbf{c}_2 = \mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}(\mathbf{c}_1 + \boldsymbol{\delta}) = \mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}\mathbf{c}_1 + \mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}\boldsymbol{\delta}. \end{aligned}$$

Since $\widetilde{\mathbf{K}}$ is a positive definite matrix,

$$\mathbf{X}^{\dagger}\widetilde{\boldsymbol{\Psi}}^{\intercal}\boldsymbol{\delta} = \widetilde{\boldsymbol{\delta}}_{\|} + \widetilde{\boldsymbol{\delta}}_{\perp},$$

where $\widetilde{\boldsymbol{\delta}}_{\|}$ is the projection onto the span of $\mathbf{X}$, and $\widetilde{\boldsymbol{\delta}}_{\perp}$ is the projection onto the null space of $\mathbf{X}$. Note that $\mathbf{X}\widetilde{\boldsymbol{\delta}}_{\perp} = \mathbf{0}$, so we cannot separate $\widetilde{\boldsymbol{\beta}}_1 \neq \widetilde{\boldsymbol{\beta}}_2$ if the difference between $\mathbf{c}_1$ and $\mathbf{c}_2$ projects onto the null space of $\mathbf{X}$. By definition, if part of the vector $\boldsymbol{\delta}$ projects onto the span of $\mathbf{X}$, then $\mathbf{P}\boldsymbol{\delta} \neq \mathbf{0}$ and $\widetilde{\boldsymbol{\beta}}_1 \neq \widetilde{\boldsymbol{\beta}}_2$. □

## Variable Selection via Hard Thresholding

In many applications, an important objective is to determine how well a given predictor is associated with a response. There are many approaches to computing marginal statistics for each explanatory variable based on its corresponding effect size estimate. In the frequentist literature, these marginal statistics are typically p-values. In the Bayesian literature, two commonly used quantities are posterior inclusion probabilities (PIPs) (Barbieri and Berger, 2004) and posterior probabilities of association (PPAs) (Stephens and Balding, 2009). In this subsection, we will develop an analog to these selection quantities, which we call the posterior probability of association analog (PPAA).

The BAKR framework does not allow for a direct computation of PIPs or PPAs. However, here we show how to compute an analog to these selection quantities (i.e. the PPAA), which is based on a hard thresholding operation. The main idea of considers a posterior probability of the form $\Pr[|\beta| \geq z \mid \mathbf{y}] > r$ — in which $z$ and $r$ denotes some effect size and posterior probability thresholds — which has been proposed as an alternative to the conventional Bayesian hypothesis testing in a series of previous works (Stephens and Balding, 2009; Shi and Kang, 2015; Pasanen et al., 2015). The PPAA quantity is compatible to the PIP and PPA when the posterior distribution of $|\beta|$ is assumed to be heavy tailed and centered at zero.

We use a variation of an adaptive technique used in Bayesian image analysis to select an explanatory variable specific threshold $z_j$ (Abramovich and Benjamini, 1995; Rajankar and Talbar, 2014; Shi and Kang, 2015). The following procedure to compute PPAAs can be added as a deterministic step to the proposed MCMC that computes effect size analogs. Assume that under the null hypothesis $\mathrm{H}_0^{(j)} : \widetilde{\beta}_j = 0$ with the classical testing significance level $\lambda$:

(1) Calculate the probability of a variable being associated under the null hypothesis

$$\widetilde{\pi}_j = 2\left(1 - \Phi\left(\frac{|\widetilde{\beta}_j|}{\widehat{\sigma}}\right)\right) \quad j = 1, \ldots, p.$$

Here, $\Phi$ is the cumulative distribution function of the standard normal distribution. The procedure is adaptive since $\sigma$ is unkown and needs to be estimated. We use the consistent median absolute deviation (MAD), $\widehat{\sigma} = \mathrm{MAD}/0.674$.

(2) Sort all of the probabilities corresponding to the variant effect sizes in ascending order as $\widetilde{\pi}_1 \leq \widetilde{\pi}_2 \leq \widetilde{\pi}_3 \leq \ldots \leq \widetilde{\pi}_p$.

(3) Compute $j^*$ as the largest $j$ such that $\widetilde{\pi}_j \leq \lambda(j/p)$.

(4) For this $j^*$, calculate $z_{j^*} = \widehat{\sigma}\Phi^{-1}\left(1 - \frac{\widetilde{\pi}_{j^*}}{2}\right)$.

(5) Threshold all $|\widetilde{\beta}_j|$ at the level $z_{j^*}$.

Note that ideal false discovery rates are achieved through $\lambda(j/p)$. This procedure allows allows us to derive a metric of evidence for each predictor variable directly from posterior inference on effect sizes, $\widetilde{\beta}_j$. After the computing each $z_{j^*}$, the PPAA may then be alternatively represented by the following

$$\widetilde{\gamma}_j = \begin{cases} 1 & \text{if} \quad |\widetilde{\beta}_j| \geq z_{j^*} \\ 0 & \text{if} \quad \text{Otherwise} \end{cases} \quad \text{for} \quad j = 1, \ldots, p \tag{S5}$$

where $\widetilde{\gamma}_j$ effectively represents an indicator that predictor variable $j$ is associated with the response.

The PPPA allows for posterior inferences on the relevance of each original covariate. In statistical genetics applications, we can define candidate causal variables as those covariates that satisfy $\{\widetilde{\gamma} : \Pr[\widetilde{\gamma} = 1 \mid \mathbf{y}] > r\}$.

This is analogous to the classical association studies that use p-values, PIPs, and PPAs as measures of covariate relevance. In practice, $r$ may be chosen subjectively (Hoti and Sillanpaa, 2006), or taken to be $r = 0.5$ in order to obtain an equivalence of a Bayesian "median probability model" (Barbieri and Berger, 2004). Another option is to define $r$ through k-fold permutation to find an effective predictor-wide threshold. For any set of significant variables, further analyses may be carried out involving the relative costs of false positives and false negatives to make an explicitly reasoned decision about which predictors to pursue (Stephens and Balding, 2009).

**Simulations: Controlling False Discovery Rates.** To validate BAKR and the proposed posterior probability of association analogue (PPAA), we carried out a simulation study. Specifically, we use a simulated matrix $\mathbf{X}$ with $p = 2000$ covariates to create continuous outcomes using the following polynomial model: $\mathbf{y} = \mathbf{X}^3\mathbf{b} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{I})$ and $\mathbf{X}^3 = \mathbf{X} \circ \mathbf{X} \circ \mathbf{X}$ is the element-wise third power of $\mathbf{X}$. We assume that the first 100 covariates are relevant to the response with $\mathbf{b}_{1:100} \sim \text{MVN}(\mathbf{0}, \mathbf{I})$, while the remainder are assumed to have zero effect. In order to investigate that the PPAA gives the correct control on the false discovery rate (FDR), we first set up the term $\lambda(j/p)$ (i.e. the procedure to compute Equation (S5)) to give a desired FDR = 0.05. Next, for a given sample size $n$, we check that PPAA yields the desired FDR. Here, we consider $n = \{500, 750, 1000\}$ where we analyzed 100 datasets in each case. The results for each $n$ are 0.063 (0.004), 0.057 (0.002), and 0.051 (0.001), respectively, with the numbers in parentheses representing the variability between simulated runs. As expected based on previous notes in the literature, the PPAA controls the FDR for reasonably sized datasets, and can be slightly liberal when the sample size is small. Presumably, the liberal behavior of the PPAA in small samples arises from the fact that the hard thresholding procedure mirrors a frequentist test and does not directly take into account the uncertainty in the estimates of the effect size analog.

## Sparsity Conditions for the Effect Size Analog

In this section, we state conditions for which the projection of the approximate nonlinear kernel function inferred onto the input covariates can be sparse. We will adapt arguments from the compressive sensing literature (Candès et al., 2006; Ben-Haim et al., 2010) to provide conditions under which inference using BAKR, coupled with hard thresholding, can result in sparse effect size analogs. We will first provide rigorous results under which a frequentist adaptation of BAKR can result in sparse effect size analogs.

In the main text, we formulated the following approximation for nonlinear kernel regression

$$\mathbf{K}\boldsymbol{\alpha} \approx \mathbf{X}\widetilde{\mathbf{B}}\boldsymbol{\theta}, \tag{S6}$$

where $\boldsymbol{\theta}$ are the empirical factor regression parameters to be inferred by BAKR, $\widetilde{\mathbf{B}} = \mathbf{X}^\dagger\widetilde{\boldsymbol{\Psi}}^\intercal(\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{U}}^\intercal\widetilde{\mathbf{K}}^{-1}\widetilde{\boldsymbol{\Psi}}^\intercal)^{-1}$, and $\widetilde{\boldsymbol{\beta}} = \widetilde{\mathbf{B}}\boldsymbol{\theta}$ are the effect size analogs. We define the set of $s$-sparse signals as $\boldsymbol{\Theta}_s = \{\widetilde{\boldsymbol{\beta}} : \|\widetilde{\boldsymbol{\beta}}\|_{\ell_0} \leq s\}$. Given equation (S6), we can formally state the following definition of sparsity for nonlinear regression.

**Definition S1** (Sparsity of Effect Size Analog)**.** *The effect size analog $\widetilde{\boldsymbol{\beta}}$ is $(s, \varepsilon)$-sparse if for the following minimization problem*

$$\left[\widehat{\boldsymbol{\beta}} \equiv \widetilde{\mathbf{B}}\widehat{\boldsymbol{\theta}}\right] = \arg\min_{\boldsymbol{\theta}}(\|\widetilde{\mathbf{B}}\boldsymbol{\theta}\|_0 \text{ subject to } \|\mathbf{X}\widetilde{\mathbf{B}}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \leq \epsilon),$$

*the minimizer $\widehat{\boldsymbol{\beta}}$ is $s$-sparse, with $\widehat{\boldsymbol{\beta}} \in \boldsymbol{\Theta}_s$.*

We now state conditions under which one can infer $(s, \varepsilon)$-sparse effect sizes based on sparse regression methods. One of the conditions that will arise is coherence of the design matrix $\mu(\mathbf{X})$, which is defined as

$$\mu(\mathbf{X}) = \max_{1 \leq i < j \leq n} \frac{|\langle \mathbf{x}_i, \mathbf{x}_j \rangle|}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2},$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the $i$-th and $j$-th columns of $\mathbf{X}$, respectively. We assume the following data generation

model

$$\mathbf{y} = \widetilde{\mathbf{K}}\boldsymbol{\alpha}^* + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \tag{S7}$$

where $\widetilde{\boldsymbol{\beta}} = \widetilde{\mathbf{B}}\boldsymbol{\theta}$, $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ are the true parameters, and $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \tau^2\mathbf{I})$. The inference algorithm we consider computes an estimator by minimizing the following functional

$$\left[\widehat{\boldsymbol{\beta}} \equiv \widetilde{\mathbf{B}}\widehat{\boldsymbol{\theta}}\right] = \arg\min_{\boldsymbol{\theta}}(\|\mathbf{X}\widetilde{\mathbf{B}}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda\|\widetilde{\mathbf{B}}\boldsymbol{\theta}\|_1), \quad \lambda > 0 \tag{S8}$$

The following is an immediate result of Corollary 1 in (Ben-Haim et al., 2010).

**Proposition S1.** *Given the generating model (S7) with coherence $\mu(\mathbf{X})$ and $\boldsymbol{\beta}^* \in \boldsymbol{\Theta}_s$ with $s \leq 1/(3\mu)$, and $\widehat{\boldsymbol{\beta}}$ inferred by algorithm (S8), with $\lambda = \sqrt{8\tau^2(1+\alpha)\log(n-s)}$ and $\alpha > 0$ being a small number. Then with probability greater than*

$$\left(1 - \frac{1}{(n-s)^\alpha}\right)\left(1 - \exp\left\{-\frac{s}{7}\right\}\right),$$

*the solution $\widehat{\boldsymbol{\beta}}$ is unique, $supp(\widehat{\boldsymbol{\beta}}) \subset supp(\boldsymbol{\beta}^*)$, and*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \leq (\sqrt{3} + 3\sqrt{2(1+\alpha)\log(n-k)})^2 s\tau^2. \tag{S9}$$

The above proposition states that if coherence assumptions are satisfied and the generating effect analog $\boldsymbol{\beta}^*$ is $(s, \varepsilon)$-sparse then algorithm (S8) will recover an $s$-sparse estimate $\widehat{\boldsymbol{\beta}}$ that is close to $\boldsymbol{\beta}^*$. Furthermore, any nonzero element of $\widehat{\boldsymbol{\beta}}$ is a nonzero element of the true effect analog $\boldsymbol{\beta}^*$.

We have shown that sparse regularization methods can recover this more general notion of an effect size. In its current formulation, BAKR does not implement sparse regularization. However, there are two trivial adaptations of BAKR that do implement sparse regularization. The first class of adaptations involves the replacement of the normal prior on the effect size analogs with a heavy tailed prior that will induce sparsity. An example of this approach is the Bayesian Lasso (Park and Casella, 2008). There is rich literature providing theory on how prior specification in sparse linear models gives rise to sparse posterior effect sizes (Castillo et al., 2015). The second approach is based on the observation that iterating least-squares regression and hard thresholding implements sparse regularization (Monajemi et al., 2013). The Stagewise Orthogonal Matching Pursuit (StOMP) algorithm is a concrete example of this approach (Donoho et al., 2012). BAKR with a diffuse prior on the effect size analog, followed by hard thresholding, is equivalent to one step of an algorithm such as StOMP. If we were to run multiple iterations of BAKR followed by hard-thresholding, then we would be implementing a sparse regularization algorithm.

**Extension to Anisotropic Kernel Functions.** Recall in Section 2.2 of the main text, the anisotropic kernel that has been previously used for variable selection often takes on the following form

$$k_{\boldsymbol{\vartheta}}(\mathbf{u}, \mathbf{v}) = k\Big((\mathbf{u} - \mathbf{v})^\intercal \text{Diag}(\boldsymbol{\vartheta})(\mathbf{u} - \mathbf{v})\Big), \quad \vartheta_j > 0, \quad j = 1, \ldots, p.$$

A very natural variation of the StOMP algorithm is to use the inferred vector $\widetilde{\boldsymbol{\beta}}$ to set the magnitude of the elements in $\boldsymbol{\vartheta}$ at each iteration. From an optimization perspective, this would involve applying an exponential decay to each coordinate rather than performing a hard thresholding at each step of the iteration. A Bayesian procedure for coupling the effect size analog and the anisotropic kernel may be of interest, but we have concerns about how well the sampling from the posterior of $\boldsymbol{\vartheta}$ will mix.

# Note About Collinearity and Interpretation

Collinearity between covariates is important to consider when analyzing effect sizes estimated by a linear regression model. If not dealt with correctly, this issue can cause problems with the interpretation of results

for practical applications. The same concerns arise with respect to the interpretation of the effect sizes estimated by the BAKR. To mitigate some of these concerns we apply the following analogous steps to those prescribed in linear regressions to reduce the effects of collinearity (Gelman and Hill, 2007):

(1) **Centering the Data:** Before performing spectral decomposition and dimensionality reduction, we center the approximate kernel matrix to ensure that the first principal component is proportional to the maximum variance of the multidimensional data. The aim behind this is to reduce collinearity between predictors in the basis space and RKHS—which we expect to implicitly reduce collinearity between the original covariates.

(2) **Orthogonalizing the Data:** The spectral decomposition of the approximate kernel matrix achieves this step on the data.

(3) **Variable Selection:** The g-prior specification on the kernel factor coefficients induces shrinkage on the original covariate effect sizes. Further variable selection may be carried out by enforcing sparsity across the original coefficients via thresholding as detailed in the previous subsection.

## Preprocessing of Real Datasets

The WTCCC data set is from the Wellcome trust case control consortium (WTCCC) 1 study (The Wellcome Trust Case Control Consortium, 2007). The data set consists of about 14,000 cases of seven common diseases, including 1,868 cases of bipolar disorder (BD), 1,926 cases of coronary artery disease (CAD), 1,748 cases of Crohn's disease (CD), 1,952 cases of hypertension (HT), 1,860 cases rheumatoid arthritis (RA), 1,963 cases of type 1 diabetes (T1D) and 1,924 cases of type 2 diabetes (T2D), as well as 2,938 shared controls. We selected a total of 458,868 shared single nucleotide polymorphisms (SNPs) following a previous study (Zhou et al., 2013). In the analysis, we mapped SNPs to the closest neighboring gene(s) using the the databases dbSNP, ImmunoBase, and UCSC Genome Browser, which can be found at the following:

- dbSNP: `http://www.ncbi.nlm.nih.gov/SNP/`
- ImmunoBase: `http://www.immunobase.org/`
- UCSC Genome Browser: `http://ucscbrowser.genap.ca/`

The heterogeneous stock of mice consists of 1,904 individuals from 85 families, all descended from eight inbred progenitor strains (Valdar et al., 2006). The data contains 129 quantitative traits that are classified into 6 broad categories including behavior, diabetes, asthma, immunology, haematology, and biochemistry. A total of 12,226 autosomal SNPs were available for all mice. For individuals with missing genotypes, we imputed missing values by the mean genotype of that SNP in their family. All polymorphic SNPs with minor allele frequency above 1% in the training data were used for prediction.

## Variance Component Analysis of Stock Mice Traits

For the phenotypic decomposition of the 129 quantitative mice traits, we consider a linear mixed model with multiple variance components (Morota et al., 2014; Zhou, 2016). Specifically, this random effect model is formulated as the following:

$$\mathbf{y} = \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3 + \mathbf{g}_c + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \tau^2\mathbf{I}) \tag{S10}$$

where $\mathbf{g}_1 \sim \text{MVN}(\mathbf{0}, \sigma_1^2\mathbf{K})$ is the linear effects component; $\mathbf{g}_2 \sim \text{MVN}(\mathbf{0}, \sigma_2^2\mathbf{K}^2)$ is the pairwise interaction component; $\mathbf{g}_3 \sim \text{MVN}(\mathbf{0}, \sigma_3^2\mathbf{K}^3)$ is the third order interaction component; and $\mathbf{g}_c \sim \text{MVN}(\mathbf{0}, \sigma_c^2\mathbf{C})$ is the common environmental component. Here, we let $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_c^2\}$ be the corresponding random effect variance terms. The matrix $\mathbf{I}$ is an identity matrix. The covariance matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^\intercal/p$ is a linear kernel matrix (Keerthi and Lin, 2003; Jiang and Reif, 2015). The covariance matrix $\mathbf{K}^2 = \mathbf{K} \circ \mathbf{K}$ represents a pairwise interaction relationship matrix and is obtained by using the Hadamard product (i.e. the squaring of each element) of the linear kernel matrix with itself. Similarly, the matrix $\mathbf{K}^3 = \mathbf{K} \circ \mathbf{K} \circ \mathbf{K}$ represents a

third order interaction relationship matrix (i.e. the cubing of each element). Lastly, $\mathbf{C}$ is a matrix of common environmental factors where a given entry $C_{ij} = 1$ if mice $i$ and $j$ are from the same cage. One can think of $\mathbf{g}_c$ as structured noise determined via cage assignment, while $\boldsymbol{\varepsilon}$ is considered as random noise.

The point of this analysis is to directly estimate the contribution of nonlinear genetic effects across an array of different phenotypes and traits — particularly amongst samples that are related through some common environmental structure. We quantify these contributions by examining the proportion of phenotypic variance explained (pPVE) using the following equation (Zhou et al., 2013; Zhou, 2016):

$$\text{pPVE}_j \propto \frac{\widehat{\sigma}_j^2}{n}\text{tr}(\boldsymbol{\Sigma}_j) \quad \text{and} \quad \sum_j \text{pPVE}_j = 1,$$

where $\boldsymbol{\Sigma} = [\mathbf{K}, \mathbf{K}^2, \mathbf{K}^3, \mathbf{C}]$. In the main text, we plot the pPVEs corresponding to the random effect variance terms $\widehat{\boldsymbol{\sigma}}^2 = \{\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\sigma}_3^2, \widehat{\sigma}_c^2\}$. The variance component that explains the greatest portion of the overall PVE then represents the most influential effect onto that particular phenotypic response. We implement this model by using the `-vc` argument in the software, GEMMA (Zhou and Stephens, 2012). This software is publicly available at `http://www.xzlab.org/software.html`. Briefly, GEMMA fits variance component models by using a MQS algorithm (Zhou, 2016), which is based on a combination of a method of moments (MoM) (Hansen, 1982) and minimal norm quadratic unbiased estimation criteria (MINQUE) (Rao, 1971). Note, that each phenotype is quantile normalized before running the analysis in GEMMA.

## Potentially Novel Loci Discovered in WTCCC Study

In the main text, we apply BAKR to an association mapping analysis of all seven diseases from the Wellcome Trust Case Control Consortium (WTCCC) 1 study (The Wellcome Trust Case Control Consortium, 2007). Overall, BAKR identified 29 significantly associated genomic regions — 14 of which were highlighted in the original WTCCC study as having strong associations, and 3 others that were highlighted in other studies which analyzed the same dataset. BAKR missed 6 genomic regions that were identified as strongly associated in the original WTCCC study, but was able to discover 12 new loci in five of the seven diseases: CD, HT, RA, T1D, and T2D. We detail these potentially novel findings here:

**Crohn's Disease (CD).** Variants spanning from 70.20Mb-70.29Mb on chromosome 10 were detected by BAKR as being associated with Crohn's disease. The leading significant SNP in this region with the highest PPAA is rs2579176. This variant, in particular, has been reported as being upstream of *DLG5*, a gene which has been found to be associated with perianal Crohn's disease (de Ridder et al., 2007). This gene was also validated (Zhang, 2012) as a member of a pairwise genetic interaction that is very influential in the cause of the trait and hard to detect. Complete details of the potentially novel loci discovered by BAKR can be found in the Supporting Information.

**Hypertension (HT).** The original WTCCC study did not report any regions of the genome as strongly associated with hypertension. Moreover, across all other compared studies, there does not appear to be much cohesion or obvious patterns in the regions determined to be significant. This is most likely due to a few reasons. First, many of the studies we compare (including our own) deal with different variations of the same dataset (e.g. depending on data origin, preprocessing measures, etc). Second, hypertension may be more susceptible to misclassification bias due to the presence of hypertensive individuals within the control samples (The Wellcome Trust Case Control Consortium, 2007). Hence, we could lessen the chance for false positives if we excluded controls with elevated blood pressure. Nonetheless, BAKR detected one locus as being moderately associated with the trait, marked by rs762015. This locus is near the gene *LOC100506412* — which is the same gene that is reported to be near a moderately significant locus found in the original WTCCC study. However, because of the reasons we just mentioned, we do not feel confident in speculating that this is a true associated region, and conclude that the locus reported by BAKR is possibly a false positive.

**Rheumatoid Arthritis (RA).** BAKR identified 3.85Mb-4.16Mb on chromosome 17 as one potentially novel region associated with rheumatoid arthritis. The leading significant SNP in this genomic region with the highest PPAA is rs9913077. This variant is upstream from *ANKFY1*, whose specific mechanistic function is unknown, but is a notable member of the ankyrin repeat gene family. This particular family of genes has been indicated to play a role in rheumatoid arthritis (Stahl et al., 2010; Zhernakova et al., 2011; Okada et al., 2014), and thus suggests evidence that this finding by BAKR may be a true positive.

**Type 1 Diabetes (T1D).** BAKR identified seven new associated regions. On chromosome 2, BAKR identifies two regions marked by SNPs rs4147713 and rs6737675, which are upstream of genes *NDUFS1* and *ABCA12*, respectively. *NDUFS1* has been reported as being responsible for transferring electrons from the NADH to the respiratory chain, while *ABCA12* plays a role in lipid and ATP transfers (Akiyama et al., 2005). We note that both of these cellular mitochondrial functions have been cited as becoming dysfunctional in the presence of type 1 diabetes (Sivitz and Yorek, 2010). The closest gene to rs1618545, on chromosome 3, is *TSEN2*. Nothing is known about the specific influence of *TSEN2* on type 1 diabetes. There is more indication that rs9302151 on chromosome 15 might be associated with the disease, as it is known that the corresponding gene *ATP8B4* is involved with ATP transfer and phospholipid transport in the cell membrane (Harris and Arias, 2003). Nothing is cited about the other SNPs rs1097157 and rs10934261 on chromosome 3, or rs12660882 on chromosome 6, as it pertains to type 1 diabetes. We note that they might be false positives, or the product of dependencies between SNPs that have yet to be detected by previous methods based on linear assumptions. Type 1 diabetes appears to be a more complex disease than hypertension, therefore we believe the speculation of nonlinear and relatedness influence here to be more valid.

**Type 2 Diabetes (T2D).** BAKR identifies two previously unmentioned regions on chromosomes 4 and 5, marked by SNPs rs7698608 and rs11167666, respectively. The former is upstream of *CISD2* — which is an iron/sulfur cluster gene and is said to be involved in calcium homeostasis in the liver — has been suggested to have an association with type 2 diabetes (Kang et al., 2012). The latter SNP, on the other hand, is most likely to be connected to more relevant and insightful biology. This variant is near the gene *GALNT10*, which has been previously been linked to the fluctuation of body mass index (BMI), obesity, and the cause of type 2 diabetes (Schwenk et al., 2013).

# Supporting Information: Algorithmic Overview

**Algorithm S1** Bayesian Approximate Kernel Regression

Select a desired positive definite shift-invariant kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i - \mathbf{x}_j)$

Compute the Fourier transform of the kernel function: $f(\boldsymbol{\omega}) = (2\pi)^{-1} \int \exp\{-\iota \boldsymbol{\omega}^\intercal (\mathbf{x}_i - \mathbf{x}_j)\} k(\mathbf{x}_i - \mathbf{x}_j) \mathrm{d}(\mathbf{x}_i - \mathbf{x}_j)$

Draw $\boldsymbol{\omega}_\ell \overset{iid}{\sim} f(\boldsymbol{\omega})$ and $\mathrm{b}_\ell \overset{iid}{\sim} U[0, 2\pi]$ for $\ell = 1, \ldots, p$

Configure $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p] \in \mathbb{R}^{p \times p}$ and $\mathbf{b} = [\mathrm{b}_1, \ldots, \mathrm{b}_p] \in \mathbb{R}^p$

Construct $\widetilde{\mathbf{K}} = \widetilde{\boldsymbol{\Psi}}^\intercal \widetilde{\boldsymbol{\Psi}}$ where $\widetilde{\boldsymbol{\Psi}} = [\widetilde{\boldsymbol{\psi}}(\mathbf{x}_1), \ldots, \widetilde{\boldsymbol{\psi}}(\mathbf{x}_n)]$ and $\widetilde{\boldsymbol{\psi}}(\mathbf{x}_i)^\intercal = \sqrt{\frac{2}{d}} \cos(\mathbf{x}_i \boldsymbol{\Omega} + \mathbf{b})$

Factorize $\widetilde{\mathbf{K}} = \widetilde{\mathbf{U}} \widetilde{\boldsymbol{\Lambda}} \widetilde{\mathbf{U}}^\intercal$, where $\widetilde{\mathbf{U}} \in \mathbb{R}^{n \times q}$ and $\widetilde{\boldsymbol{\Lambda}} \in \mathbb{R}^{q \times q}$ for chosen $q$

Run the Gibbs Sampler ($T$ Iterations)

**for** $t = 1 \to T$ **do**

$\quad \boldsymbol{\theta} \mid \sigma^2, \tau^2, \mathbf{y} \sim \mathrm{MVN}(\mathbf{m}^*, \mathbf{V}^*)$ with $\mathbf{V}^* = \tau^2 \sigma^2 (\tau^2 \widetilde{\boldsymbol{\Lambda}}^{-1} + \sigma^2 \mathbf{I}_q)^{-1}$ and $\mathbf{m}^* = \tau^{-2} \mathbf{V}^* \widetilde{\mathbf{U}}^\intercal \mathbf{y}$

$\quad \widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger \widetilde{\boldsymbol{\Psi}}^\intercal (\widetilde{\boldsymbol{\Lambda}} \widetilde{\mathbf{U}}^\intercal \widetilde{\mathbf{K}}^{-1} \widetilde{\boldsymbol{\Psi}}^\intercal)^{-1} \boldsymbol{\theta}$;

$\quad \sigma^2 \mid \boldsymbol{\theta}, \tau^2, \mathbf{y} \sim \text{Scale-inv} - \chi^2(\nu_\sigma^*, \phi_\sigma^*)$ where $\nu_\sigma^* = \nu + q$ and $\phi_\sigma^* = \nu_\sigma^{*-1} (\nu\phi + \boldsymbol{\theta}^\intercal \widetilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{\theta})$

$\quad \tau^2 \mid \boldsymbol{\theta}, \sigma^2, \mathbf{y} \sim \text{Scale-inv} - \chi^2(\nu_\tau^*, \phi_\tau^*)$ where $\nu_\tau^* = \nu + n$ and $\phi_\tau^* = \nu_\tau^{*-1} (\nu\phi + \boldsymbol{\varepsilon}^\intercal \boldsymbol{\varepsilon})$

**end for**

**Algorithm S2** Bayesian Approximate Kernel Probit Regression

Select a desired positive definite shift-invariant kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i - \mathbf{x}_j)$

Compute the Fourier transform of the kernel function: $f(\boldsymbol{\omega}) = (2\pi)^{-1} \int \exp\{-\iota \boldsymbol{\omega}^\intercal (\mathbf{x}_i - \mathbf{x}_j)\} k(\mathbf{x}_i - \mathbf{x}_j) \mathrm{d}(\mathbf{x}_i - \mathbf{x}_j)$

Draw $\boldsymbol{\omega}_\ell \overset{iid}{\sim} f(\boldsymbol{\omega})$ and $\mathrm{b}_\ell \overset{iid}{\sim} U[0, 2\pi]$ for $\ell = 1, \ldots, p$

Configure $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p] \in \mathbb{R}^{p \times p}$ and $\mathbf{b} = [\mathrm{b}_1, \ldots, \mathrm{b}_p] \in \mathbb{R}^p$

Construct $\widetilde{\mathbf{K}} = \widetilde{\boldsymbol{\Psi}}^\intercal \widetilde{\boldsymbol{\Psi}}$ where $\widetilde{\boldsymbol{\Psi}} = [\widetilde{\boldsymbol{\psi}}(\mathbf{x}_1), \ldots, \widetilde{\boldsymbol{\psi}}(\mathbf{x}_n)]$ and $\widetilde{\boldsymbol{\psi}}(\mathbf{x}_i)^\intercal = \sqrt{\frac{2}{d}} \cos(\mathbf{x}_i \boldsymbol{\Omega} + \mathbf{b})$

Factorize $\widetilde{\mathbf{K}} = \widetilde{\mathbf{U}} \widetilde{\boldsymbol{\Lambda}} \widetilde{\mathbf{U}}^\intercal$, where $\widetilde{\mathbf{U}} \in \mathbb{R}^{n \times q}$ and $\widetilde{\boldsymbol{\Lambda}} \in \mathbb{R}^{q \times q}$ for chosen $q$

Run the Gibbs Sampler ($T$ Iterations)

**for** $t = 1 \to T$ **do**

$\quad$ **for** $i = 1 \to n$ **do**

$$\quad \mathrm{s}_i^{(t+1)} \mid \boldsymbol{\theta}, \sigma^2, \mathbf{s}^{(t)}, \mathbf{y} \sim \begin{cases} \mathrm{N}(\tilde{\mathbf{u}}_i^\intercal \boldsymbol{\theta}, 1) \mathbb{1}(\mathrm{s}_i^{(t)} \leq 0) & \text{if} \quad \mathrm{s}_i^{(t)} \leq 0 \\ \mathrm{N}(\tilde{\mathbf{u}}_i^\intercal \boldsymbol{\theta}, 1) \mathbb{1}(\mathrm{s}_i^{(t)} > 0) & \text{if} \quad \mathrm{s}_i^{(t)} > 0; \end{cases}$$

$\quad$ **end for**

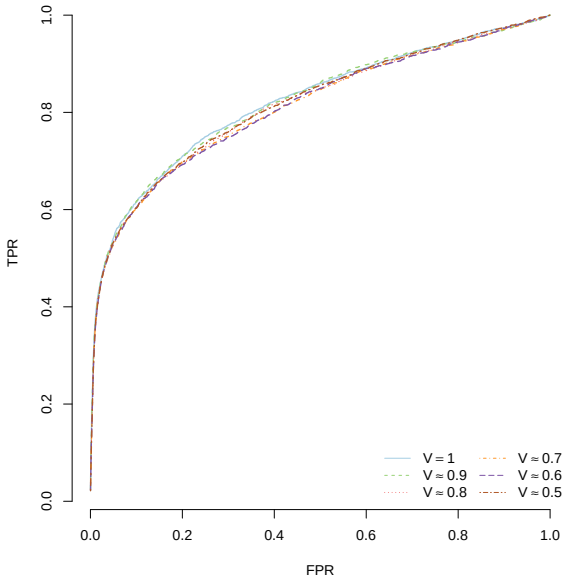
$\quad \boldsymbol{\theta} \mid \mathbf{s}, \sigma^2, \mathbf{y} \sim \mathrm{MVN}(\mathbf{m}^*, \mathbf{V}^*)$ where $\mathbf{V}^* = \sigma^2 (\widetilde{\boldsymbol{\Lambda}}^{-1} + \sigma^2 \mathbf{I})^{-1}$ and $\mathbf{m}^* = \mathbf{V}^* \widetilde{\mathbf{U}}^\intercal \mathbf{s}$

$\quad \widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger \widetilde{\boldsymbol{\Psi}}^\intercal (\widetilde{\boldsymbol{\Lambda}} \widetilde{\mathbf{U}}^\intercal \widetilde{\mathbf{K}}^{-1} \widetilde{\boldsymbol{\Psi}}^\intercal)^{-1} \boldsymbol{\theta}$;

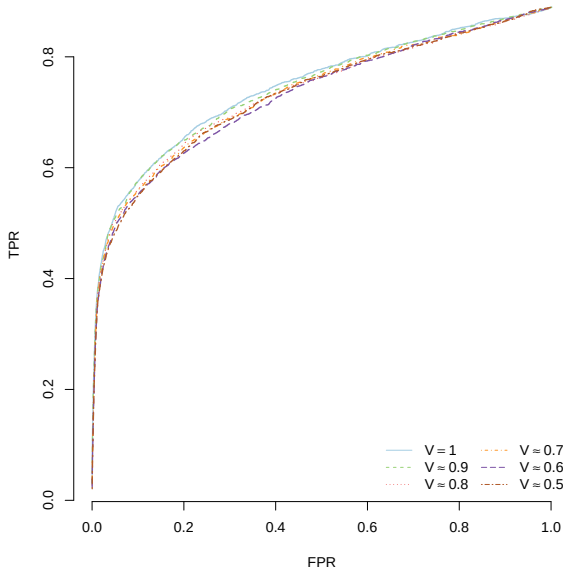
$\quad \sigma^2 \mid \mathbf{s}, \boldsymbol{\theta}, \mathbf{y} \sim \text{Scale-inv-}\chi^2(\nu^*, \phi^*)$ where $\nu^* = \nu + q$ and $\phi^* = \nu^{*-1} (\nu\phi + \boldsymbol{\theta}^\intercal \widetilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{\theta})$

**end for**

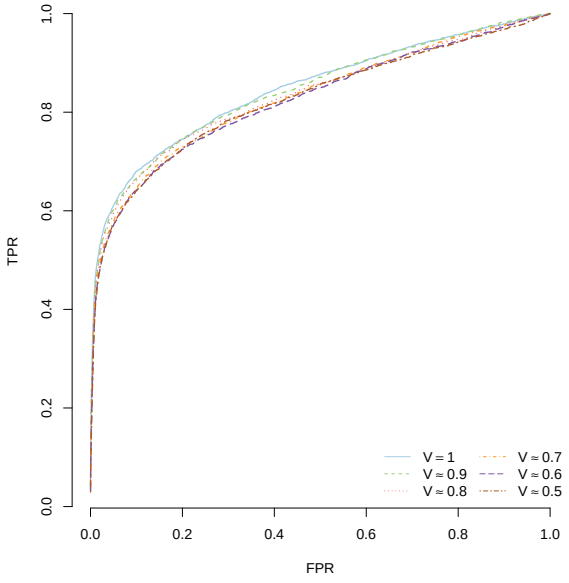
# Supporting Information: Figures



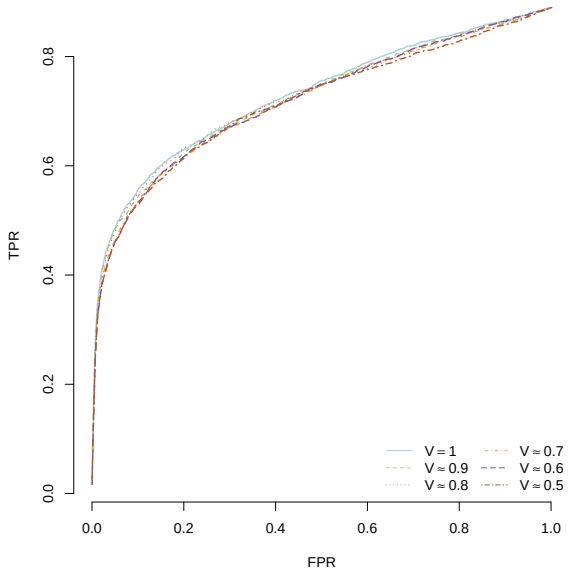

(a) Scenario I ($\rho = 0.2$): Group 1 SNPs




(b) Scenario I ($\rho = 0.2$): Group 2 SNPs




(c) Scenario II ($\rho = 0.8$): Group 1 SNPs




(d) Scenario II ($\rho = 0.8$): Group 2 SNPs

Figure S1: Sensitivity analysis comparing the power of BAKR under its empirical factor representation when explaining different amounts of the cumulative variance (V) in the approximate kernel matrix $\widetilde{\mathbf{K}}$. Values considered here are V $= \{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$. Group 1 SNPs are those that exhibit additive effects, while the SNPs in group 2 are those involved in interactions. The x-axis shows the false positive rate, while the y-axis gives the rate at which true causal predictor variables were identified. Results are based on 100 different simulated datasets in each scenario.

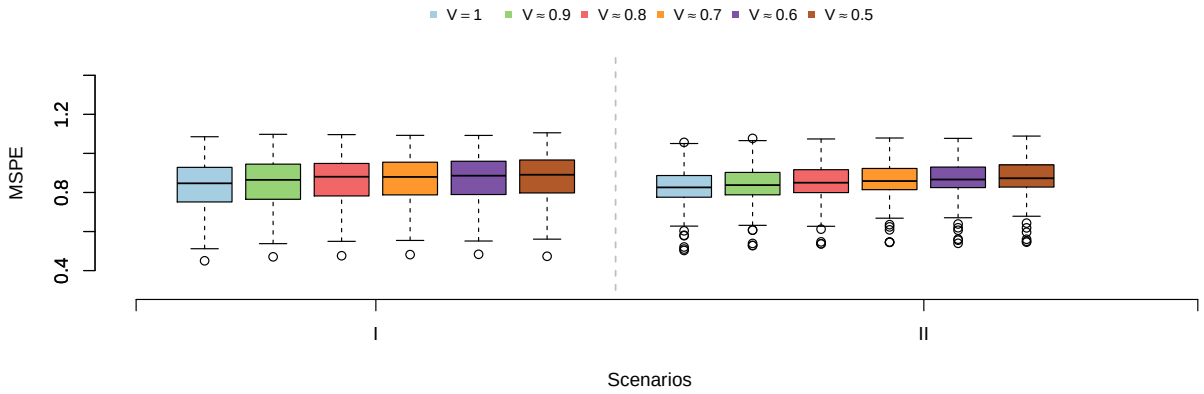
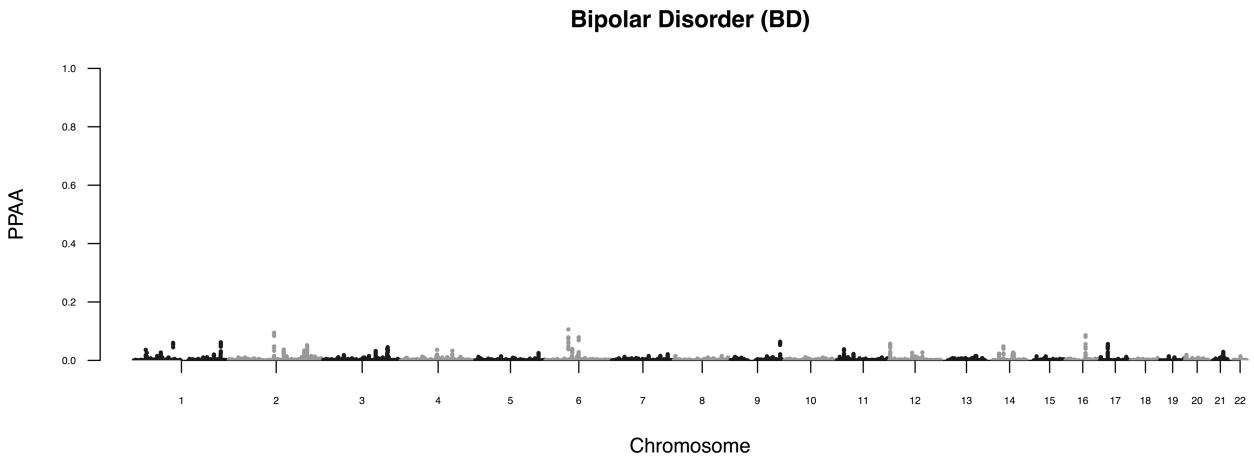
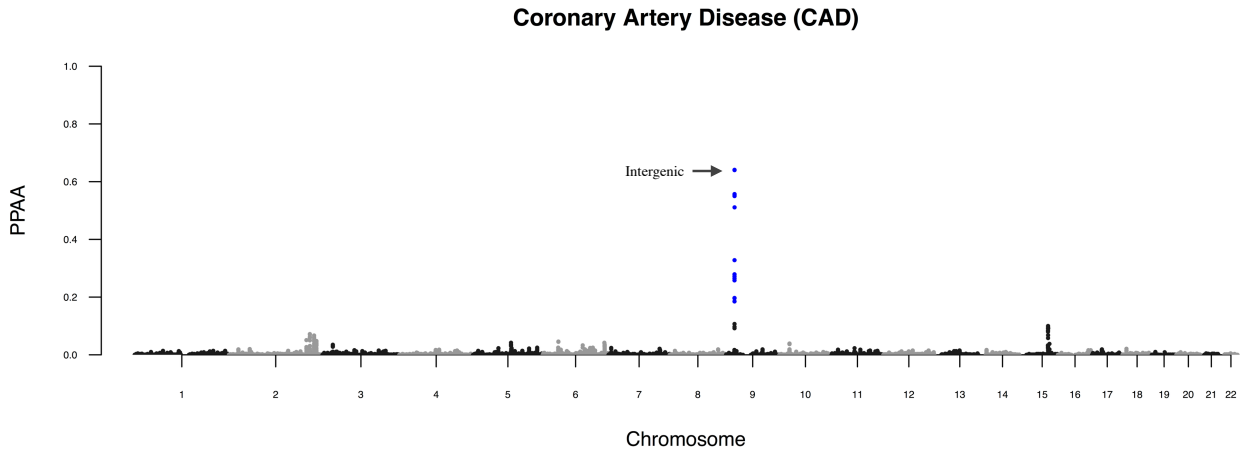
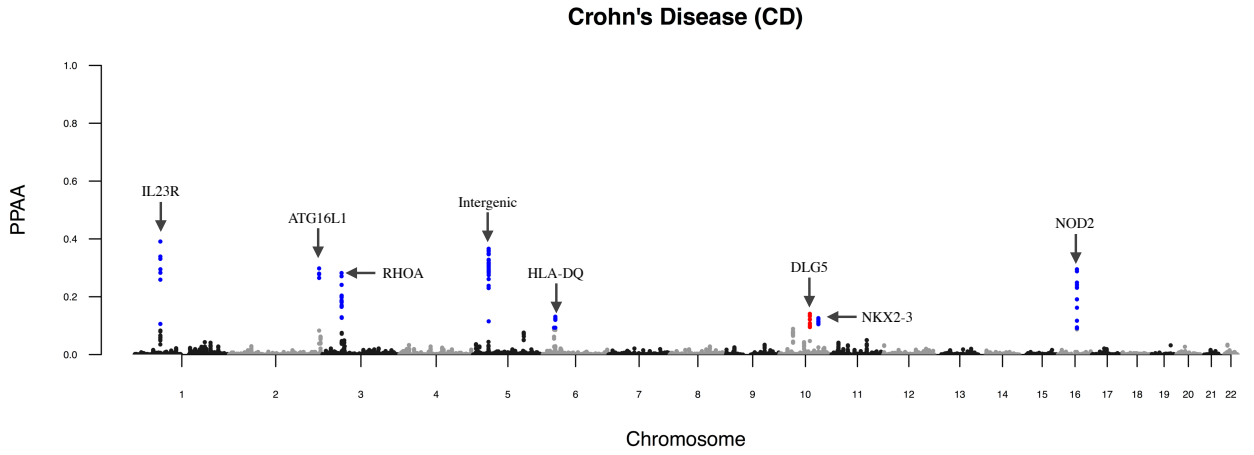


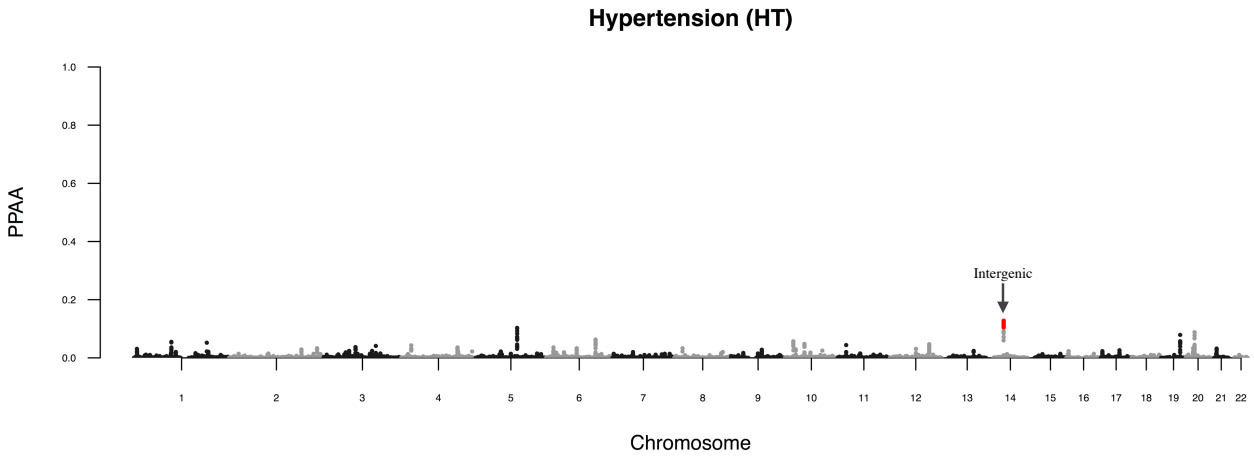

Figure S2: Sensitivity analysis comparing the mean square prediction error (MSPE) for BAKR under its empirical factor representation when explaining different amounts of the cumulative variance (V) in the approximate kernel matrix $\widetilde{\mathbf{K}}$. Values considered here are V $= \{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$. In Scenario I, pairwise interactions make up 80% of the broad-sense heritability (i.e. $\rho = 0.2$). In Scenario II, additive effects dominate 80% of the broad-sense heritability (i.e. $\rho = 0.8$). Values in bold represent the method with the lowest MSPE. These results are based on 100 different simulated datasets in both scenarios.

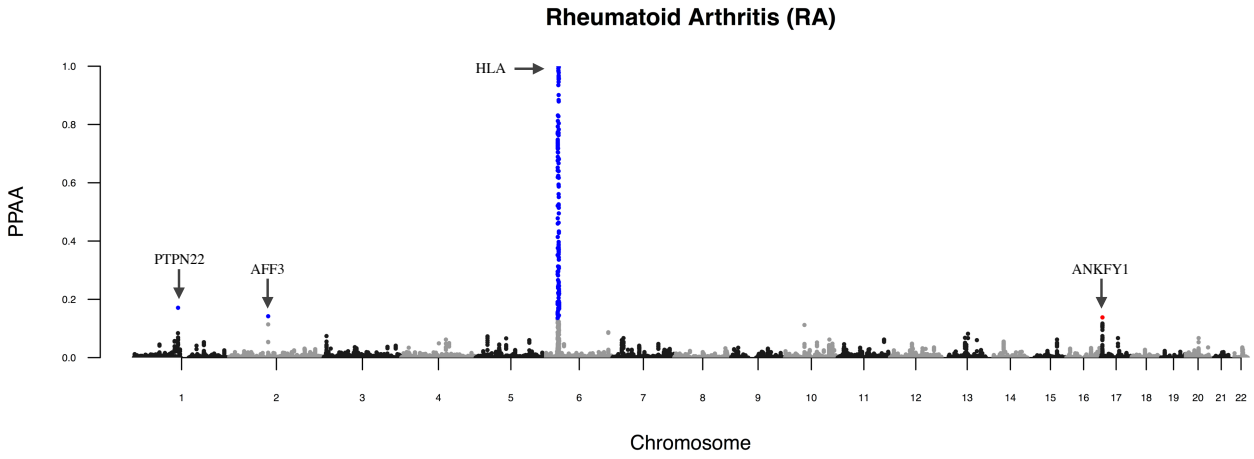

Figure S3: The portion of phenotypic variance explained (pPVE) by four different genetic effects of interest: (1) additive effects, (2) pairwise interactions, (3) third order interactions, and (4) common environmental (i.e. cage specific) effects.

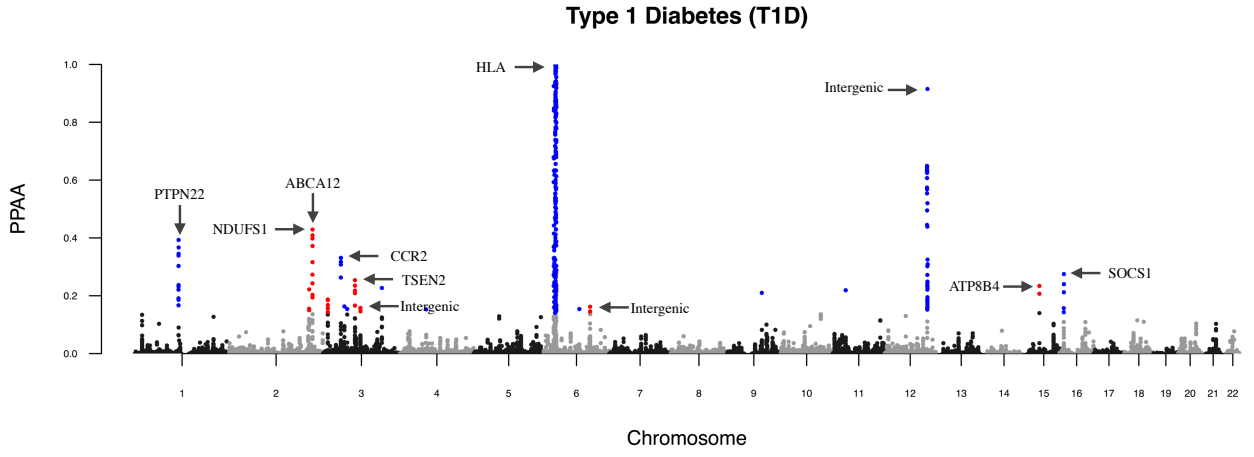

Bipolar Disorder (BD)
PPAA
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


Coronary Artery Disease (CAD)
PPAA
1.0
0.8
0.6
0.4
0.2
0.0
Intergenic
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


Crohn's Disease (CD)
PPAA
1.0
0.8
0.6
0.4
0.2
0.0
IL23R
ATG16L1
RHOA
Intergenic
HLA-DQ
DLG5
NKX2-3
NOD2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


Hypertension (HT)
PPAA
1.0
0.8
0.6
0.4
0.2
0.0
Intergenic
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


Rheumatoid Arthritis (RA)
PPAA
1.0
0.8
0.6
0.4
0.2
0.0
HLA
PTPN22
AFF3
ANKFY1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


Type 1 Diabetes (T1D)
PPAA
1.0
0.8
0.6
0.4
0.2
0.0
HLA
Intergenic
PTPN22
ABCA12
NDUFS1
CCR2
TSEN2
Intergenic
Intergenic
ATP8B4
SOCS1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

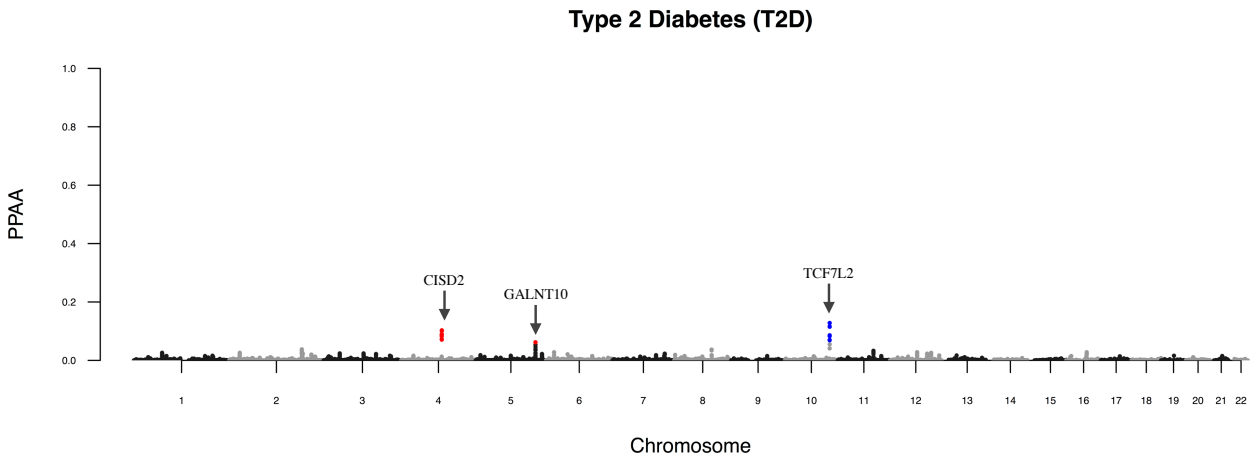



Figure S4: Genome-wide scan for seven diseases in the WTCCC dataset. For each of the seven diseases, the posterior probabilities of association analogs (PPAA) for quality-control-positive SNPs are plotted against position on each chromosome. Chromosomes are shown in alternating colors for clarity, with corresponding PPAAs exceeding a trait-specific 5% FWER threshold highlighted in blue. The SNPs highlighted in red are those that exceed the FWER threshold and lie in a potentially novel locus associated with the disease. Labeled are the genes located near the enriched positions. The gene symbols are from Entrez Gene (Maglott et al., 2011).

# Supporting Information: Tables

## Table S1

| | Sample Size | | |
|---|---|---|---|
| | $n = 500$ | $n = 750$ | $n = 1000$ |
| $\mathrm{V}(\mathrm{R}^2)$ | $7.02 \times 10^{-6}$ | $2.26 \times 10^{-6}$ | $1.04 \times 10^{-6}$ |
| $\mathrm{Prop}(\mathrm{R}^2)_{\mathrm{MCMC}}$ | 0.92 (0.001) | 0.87 (0.001) | 0.85 (0.001) |
| $\mathrm{Prop}(\mathrm{R}^2)_{\widetilde{\mathbf{K}}}$ | 0.08 (0.001) | 0.13 (0.001) | 0.15 (0.001) |

Table S1: **Assessment of how much of the variation between runs of BAKR is due to posterior estimation via the Gibbs sampler, versus how much is due to the sampling of the approximate kernel function.** Here, we consider sample sizes $n = \{500, 750, 1000\}$, where we analyzed 100 different datasets in each case. Within each individual dataset, we run BAKR 100 different times in order to get a clear illustration of the variation in performance between runs of the model. We treat the variance of the computed $\mathrm{R}^2$ (i.e. $\mathrm{V}(\mathrm{R}^2)$) across runs as a quantity to measure error. We then decompose the variance of $\mathrm{R}^2$ into the proportion due to MCMC (i.e. $\mathrm{Prop}(\mathrm{R}^2)_{\mathrm{MCMC}}$) and the proportion due to the approximate kernel (i.e. $\mathrm{Prop}(\mathrm{R}^2)_{\widetilde{\mathbf{K}}}$). Reported here are the averages across all simulated datasets. Standard errors across replicates are given the parentheses.

## Table S2

**Computational time (in seconds) to run different Bayesian variable selection methods for 100 MCMC iterations, as a function of sample size and the number of covariates.** Compared here are Bayes Ridge (BRR), Bayes Lasso (BL), Bayes LMM (BLMM), and Bayes C$\pi$. BAKR is assessed under both its full model specification (V = 1), as well as under its empirical factor representation (V $\approx$ 0.9 and 0.8, respectively). Sample sizes considered were $n$ = 100, 500, 1000, and 2000. The number of covariates considered were $p = 1 \times 10^3$, $1 \times 10^4$, $5 \times 10^4$, and $1 \times 10^5$, respectively. Note that only cases in which $p > n$ were timed. Results are based on 100 different simulated datasets. Standard errors across these replicates for each model are given the parentheses. (PDF)

## Table S3

| | | Cumulative Variance Explained in $\widetilde{\mathbf{K}}$ | | | | | |
|---|---|---|---|---|---|---|---|
| | Scenario | V = 1 | V $\approx$ 0.9 | V $\approx$ 0.8 | V $\approx$ 0.7 | V $\approx$ 0.6 | V $\approx$ 0.5 |
| MSPE (SD) | I | **0.708** | 0.729 | 0.741 | 0.750 | 0.758 | 0.766 |
| | | **(0.20)** | (0.20) | (0.20) | (0.20) | (0.21) | (0.21) |
| | II | **0.687** | 0.714 | 0.733 | 0.746 | 0.758 | 0.768 |
| | | **(0.18)** | (0.18) | (0.18) | (0.19) | (0.19) | (0.19) |

Table S3: **Sensitivity analysis comparing the mean square prediction error (MSPE) for BAKR under its empirical factor representation when explaining different amounts of the cumulative variance (V) in the approximate kernel matrix $\widetilde{\mathbf{K}}$.** Values considered here are V = $\{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$. In Scenario I, pairwise interactions make up 80% of the broad-sense heritability (i.e. $\rho = 0.2$). In Scenario II, additive effects dominate 80% of the broad-sense heritability (i.e. $\rho = 0.8$). Values in bold represent the method with the lowest MSPE. These results are based on 100 different simulated datasets in both scenarios. Standard errors across these replicates for each model are given the parentheses.

## Table S4

**A table that lists the 129 quantitative mice phenotypes which are classified into the 6 categories: behavior, diabetes, asthma, immunology, haematology, and biochemistry.** (XLSX)

## Table S5

**Table of all significant SNPs, discovered by BAKR according to the 0.05 FWER threshold, for each of the seven diseases in the WTCCC dataset.** Listed are the PPAAs for each variant, along with their marginal p-value which was computed using a single-SNP linear model. The phenotype specific FWER thresholds are given on page 2. (XLSX)

## Table S6

**Table of regions with at least two SNPs having PPAAs satisfying the 5% FWER threshold in the analysis of the WTCCC Data.** Listed for all regions are the SNPs with the highest PPAA and their corresponding marginal p-values. The marginal p-values reported are found via linear regression and used as a direct comparison. The reference column gives literature sources that have previously suggested some level of association between a given region and disease. Rows listed in bold are those for which we did not find any sources that previously suggested association with that disease. These regions could potentially be novel. Note that some of the listed references are works that utilize methods that consider pairwise interactions between SNPs. *Multiple SNPs in the HLA region are significant, so we choose the SNP with the lowest marginal p-value and report that as the most extreme. (PDF)

# References


Abramovich, F. and Y. Benjamini (1995). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis 22*, 351–361.

Akiyama, M., Y. Sugiyama-Nakagiri, K. Sakai, J. R. McMillan, M. Goto, K. Arita, Y. Tsuji-Abe, N. Tabata, K. Matsuoka, R. Sasaki, D. Sawamura, and H. Shimizu (2005). Mutations in lipid transporter ABCA12 in harlequin ichthyosis and functional recovery by corrective gene transfer. *Journal of Clinical Investigation 115*(7), 1777–1784.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *The Annals of Statistics 32*(3), 870–897.

Ben-Haim, Z., Y. C. Eldar, and M. Elad (2010). Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Transactions on Signal Processing 58*(10), 5030–5043.

Candès, E. J., J. K. Romberg, and T. Tao (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math. 59*(8), 1207–1223.

Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics 43*(5), 1986–2018.

Chakraborty, S. (2009). Bayesian binary kernel probit model for microarray based cancer classification and gene selection. *Computational Statistics & Data Analysis 53*(12), 4198–4209.

Chakraborty, S., M. Ghosh, and B. K. Mallick (2012). Bayesian non-linear regression for large $p$ small $n$ problems. *Journal of Multivariate Analysis 108*, 28–40.

Chakraborty, S., B. K. Mallick, D. Ghosh, M. Ghosh, and E. Dougherty (2007). Gene expression-based glioma classification using hierarchical bayesian vector machines. *Sankhyā: The Indian Journal of Statistics 69*(3), 514–547.

de los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research (Cambridge) 92*(4), 295–308.

de Ridder, L., R. K. Weersma, G. Dijkstra, G. van der Steege, M. A. Benninga, I. M. Nolte, J. A. Taminiau, D. W. Hommes, and P. C. Stokkers (2007). Genetic susceptibility has a more important role in pediatric-onset Crohn's disease than in adult-onset Crohn's disease. *Inflammatory Bowel Diseases 13*(9), 1083–1092.

Donoho, D. L., Y. Tsaig, I. Drori, and J.-L. Starck (2012). Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Transactions on Information Theory 58*(2), 1094–1121.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Volume Analytical Methods for Social Research. New York: Cambridge University Press.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Harris, M. J. and I. M. Arias (2003). FIC1, a P-type ATPase linked to cholestatic liver disease, has homologues (ATP8B2 and ATP8B3) expressed throughout the body. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1633*(2), 127–131.

Hoti, F. and M. J. Sillanpaa (2006). Bayesian mapping of genotype expression interactions in quantitative and qualitative traits. *Heredity 97*, 4–18.

Jiang, Y. and J. C. Reif (2015). Modeling epistasis in genomic selection. *Genetics 201*, 759–768.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. yee Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics 42*(4), 348–354.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008). Efficient control of population structure in model organism association mapping. *Genetics 178*(3), 1709–1723.

Kang, H. P., A. A. Morgan, R. Chen, E. E. Schadt, and A. J. Butte (2012). Coanalysis of gwas with eqtls reveals disease-tissue associations. *AMIA Summits on Translational Science Proceedings 2012*, 35–41.

Keerthi, S. S. and C.-J. Lin (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation 15*(7), 1667–1689.

Liang, F., K. Mao, S. Mukherjee, and M. West (2009). Nonparametric Bayesian kernel models. *Department of Statistical Science, Duke University, Discussion Paper*, 7–10.

Liang, F., S. Mukherjee, and M. West (2007). The use of unlabeled data in predictive modeling. *Statistical Science 22*(2), 189–205.

Liu, D., D. Ghosh, and X. Lin (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machinesand linear mixed models. *Biometrics 63*, 1079–1088.

Maglott, D., J. Ostell, K. D. Pruitt, and T. Tatusova (2011). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research 39*(Database issue), D52–D57.

Mallick, B. K., D. Ghosh, and M. Ghosh (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 219–234.

Monajemi, H., S. Jafarpour, M. Gavish, S. C. . Collaboration, and D. L. Donoho (2013). Deterministic matrices matching the compressed sensing phase transitions of gaussian random matrices. *Proceedings of the National Academy of Sciences 110*(4), 1181–1186.

Morota, G., P. Boddhireddy, N. Vukasinovic, D. Gianola, and S. DeNise (2014). Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Frontiers in Genetics 5*, 56.

Okada, Y., D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. A. F. J. van de Laar, H.-J. Guchelaar, T. W. J. Huizinga, P. Dieude, X. Mariette, S. Louis Bridges Jr, A. Zhernakova, R. E. M. Toes, P. P. Tak, C. Miceli-Richard, S.-Y. Bang, H.-S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapaa-Dahlqvist, L. Arlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, the RACI consortium, the GARNET consortium, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S.-C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature 506*(7488), 376–381.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Pasanen, L., L. Holmström, and M. J. Sillanpää (2015). Bayesian lasso, scale space and decision making in association genetics. *PLoS ONE 10*(4), e0120017.

Rajankar, S. and S. Talbar (2014). Generalized false discovery rate based adaptive thresholding approach for ecg signal compression. *International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 266–271.

Rao, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis 1*(4), 445–456.

Schwenk, R. W., H. Vogel, and A. Schürmann (2013). Genetic and epigenetic control of metabolic health. *Molecular Metabolism 2*(4), 337–347.

Shi, R. and J. Kang (2015). Thresholded multiscale Gaussian processes with application to Bayesian feature selection for massive neuroimaging data. arXiv:1504.06074.

Sivitz, W. I. and M. A. Yorek (2010). Mitochondrial dysfunction in diabetes: From molecular mechanisms to functional significance and therapeutic opportunities. *Antioxidants & Redox Signaling 12*(4), 537–577.

Skrondal, A. and S. RabeHesketh (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(3), 659–687.

Stahl, E. A., S. Raychaudhuri, E. F. Remmers, G. Xie, S. Eyre, B. P. Thomson, Y. Li, F. A. S. Kurreeman, A. Zhernakova, A. Hinks, C. Guiducci, R. Chen, L. Alfredsson, C. I. Amos, K. G. Ardlie, A. Barton, J. Bowes, E. Brouwer, N. P. Burtt, J. J. Catanese, J. Coblyn, M. J. H. Coenen, K. H. Costenbader, L. A. Criswell, J. B. A. Crusius, J. Cui, P. I. W. de Bakker, P. L. De Jager, B. Ding, P. Emery, E. Flynn, P. Harrison, L. J. Hocking, T. W. J. Huizinga, D. L. Kastner, X. Ke, A. T. Lee, X. Liu, P. Martin, A. W. Morgan, L. Padyukov, M. D. Posthumus, T. R. D. J. Radstake, D. M. Reid, M. Seielstad, M. F. Seldin, N. A. Shadick, S. Steer, P. P. Tak, W. Thomson, A. H. M. van der Helm-van Mil, I. E. van der Horst-Bruinsma, C. E. van der Schoot, P. L. C. M. van Riel, M. E. Weinblatt, A. G. Wilson, G. J. Wolbink, B. P. Wordsworth, C. Wijmenga, E. W. Karlson, R. E. M. Toes, N. de Vries, A. B. Begovich, J. Worthington, K. A. Siminovitch, P. K. Gregersen, L. Klareskog, and R. M. Plenge (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics 42*(6), 508–514.

Stephens, M. and D. J. Balding (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics 10*, 681–690.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*(7145), 661–678.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics 38*(8), 879–887.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics 7*.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics 46*(2), 100–106.

Zhang, Y. (2012). A novel Bayesian graphical model for genome-wide multi-SNP association mapping. *Genetic Epidemiology 36*(1), 36–47.

Zhang, Z., G. Dai, and M. I. Jordan (2011). Bayesian generalized kernel mixed models. *Journal of Machine Learning Research 12*, 111–139.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics 42*(4), 355–360.

Zhernakova, A., E. A. Stahl, G. Trynka, S. Raychaudhuri, E. A. Festen, L. Franke, H.-J. Westra, R. S. N. Fehrmann, F. A. S. Kurreeman, B. Thomson, N. Gupta, J. Romanos, R. McManus, A. W. Ryan, G. Turner, E. Brouwer, M. D. Posthumus, E. F. Remmers, F. Tucci, R. Toes, E. Grandone, M. C. Mazzilli, A. Rybak,

B. Cukrowska, M. J. H. Coenen, T. R. D. J. Radstake, P. L. C. M. van Riel, Y. Li, P. I. W. de Bakker, P. K. Gregersen, J. Worthington, K. A. Siminovitch, L. Klareskog, T. W. J. Huizinga, C. Wijmenga, and R. M. Plenge (2011). Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet 7*(2), e1002004.

Zhou, X. (2016). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *bioRxiv*.

Zhou, X., P. Carbonetto, and M. Stephens (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet 9*(2), e1003264.

Zhou, X. and M. Stephens (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics 44*(7), 821–825.