Supporting Information

1 Preprocessing of Real Datasets

The WTCCC data set is from the Wellcome trust case control consortium (WTCCC) 1 study [1]. 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 [2]. 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 [3]. 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.

2 Variance Component Analysis

For the variance component analysis, we consider a linear mixed model with multiple variance components [4, 5]. Specifically, this random effect model is formulated as the following:

$$\mathbf{y} = \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3 + \mathbf{g}_c + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n) \tag{S2.1}$$

where $\mathbf{g}_1 \sim \mathrm{MVN}_n(\mathbf{0}, \sigma_1^2 \mathbf{G})$ is the linear effects component; $\mathbf{g}_2 \sim \mathrm{MVN}_n(\mathbf{0}, \sigma_2^2 \mathbf{G}^2)$ is the pairwise interaction component; $\mathbf{g}_3 \sim \mathrm{MVN}_n(\mathbf{0}, \sigma_3^2 \mathbf{G}^3)$ is the third order interaction component; and $\mathbf{g}_c \sim \mathrm{MVN}_n(\mathbf{0}, \sigma_c^2 \mathbf{C})$ is the common environmental component. One can think of $\mathbf{g}_c$ as structured noise and $\boldsymbol{\varepsilon}$ as random noise. Here, we let $\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_c^2\}$ be corresponding random effect variance terms. The matrix $\mathbf{I}_n$ is an $n \times n$ identity matrix. The matrix $\mathbf{G} = \frac{1}{p}\mathbf{X}\mathbf{X}'$ is a linear kernel (Gram) matrix [6, 7]. The matrix $\mathbf{G}^2 = \mathbf{G} \circ \mathbf{G}$ 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. The matrix $\mathbf{G}^3 = \mathbf{G} \circ \mathbf{G} \circ \mathbf{G}$ represents a third order interaction relationship matrix (i.e. the cubing of each element), and $\mathbf{C}$ is a matrix of common environmental factors where: $C_{ij} = 1$ if mouse i and j are from the same cage.

The point of this analysis is to directly estimate the contribution of nonlinear 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 portion of phenotypic variance explained (pPVE) using the following equation defined in [2, 5]:

$$\text{pPVE}_j \propto \frac{\sigma_j^2}{n}\text{tr}(\text{GSM}_j) \quad \text{and} \quad \sum_j \text{pPVE}_j = 1,$$

where under (S2.1), $j = 1, \ldots, 4$. We specifically plot the pPVEs corresponding to the random effect variance terms $\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_c^2\}$. The variance component that explains the greatest portion of the overall PVE then represents the most influential effect onto that particular phenotypic response. We run this analysis using the GEMMA software [8] (publicly available at `http://www.xzlab.org/software.html`), which is designed for multiple variance component models and performs by using a MQS algorithm based on a method of moments and a minimal norm quadratic unbiased estimation criteria [5]. Each phenotype is quantile normalized before running the analysis in GEMMA.

3 BAKR Mixed Model Extension

There are applications where a nonlinear mixed model is desired. Examples of this include cases where the observations are not independent but related via some population structure or known kinship, or cases where one needs to control for confounders such as batch effects. Here, we detail a nonlinear mixed regression model. The extension to binary classification is straightforward based on the steps outlined for the BAKR-probit model in the main text. One can adapt the empirical factor representation of BAKR to include a random component as follows:

$$\mathrm{y}_i = \tilde{\mathbf{u}}_i'\boldsymbol{\theta} + \varphi_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \quad \varphi_i \perp\!\!\!\perp \varepsilon_i \tag{S3.1}$$

with $\mathbb{E}[\mathrm{y}_i \,|\, \varphi_i] = \tilde{\mathbf{u}}_i'\boldsymbol{\theta} + \varphi_i$ and φ_i is independent of ε_i. Jointly, $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_n]'$ are assumed to be normally distributed with zero mean and covariance structure $\boldsymbol{\Delta}$. In our applications, $\boldsymbol{\Delta}$ is not diagonal or block-diagonal, which implies that the elements in the response vector $\mathbf{y}$ are correlated via the random effects [10]. In the statistical genetics context, the relevance of the random effect is that the fixed and random effects capture a larger portion of the total covariance structure and allow for more accurate posterior summaries of quantities of interest, such as effect sizes. This correction increases 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 with the population structure [11–14]. A standard approach in quantitative and statistical genetics is to define the covariance of $\boldsymbol{\varphi}$ as a known kinship matrix $\boldsymbol{\Delta}$ which can model either direct family relations between individuals or population structure across individuals, and is estimated from SNP data [13, 14]. This flexibility of the linear mixed model is a major reason it is used in applications such as genome-wide association studies (GWAS) [13].

We specify the following hierarchical model

$$\begin{aligned}
\mathrm{y}_i &= \tilde{\mathbf{u}}_i'\boldsymbol{\theta} + \varphi_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \quad \varphi_i \perp\!\!\!\perp \varepsilon_i, \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \widetilde{\boldsymbol{\Lambda}}), \\
\sigma_\theta^2 &\sim \mathrm{Scale-inv-}\chi^2(\nu_\theta, \phi_\theta), \\
\sigma_\varepsilon^2 &\sim \mathrm{Scale-inv-}\chi^2(\nu_\varepsilon, \phi_\varepsilon), \\
\boldsymbol{\varphi} &\sim \mathrm{MVN}(\mathbf{0}_n, \boldsymbol{\Delta}).
\end{aligned} \tag{S3.2}$$

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

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

(1) $\boldsymbol{\theta} \mid \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{\varphi}, \mathbf{y}, \boldsymbol{\Delta} \sim \mathrm{MVN}(\mathbf{m}_\theta^*, \mathbf{V}_\theta^*)$ with $\mathbf{V}_\theta^* = \sigma_\varepsilon^2\sigma_\theta^2(\sigma_\varepsilon^2\widetilde{\boldsymbol{\Lambda}}^{-1} + \sigma_\theta^2\mathbf{I}_q)^{-1}$ and $\mathbf{m}_\theta^* = \frac{1}{\sigma_\varepsilon^2}\mathbf{V}_\theta^*\widetilde{\mathbf{U}}'(\mathbf{y} - \boldsymbol{\varphi})$;

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

(3) $\sigma_\theta^2 \mid \boldsymbol{\theta}, \sigma_\varepsilon^2, \boldsymbol{\varphi}, \mathbf{y}, \boldsymbol{\Delta} \sim \mathrm{Scale\text{-}inv} - \chi^2(\nu_\theta^*, \phi_\theta^*)$ where $\nu_\theta^* = \nu_\theta + q$ and $\phi_\theta^* = \frac{1}{\nu_\theta^*}(\nu_\theta\phi_\theta + \boldsymbol{\theta}'\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\theta})$;

(4) $\sigma_\varepsilon^2 \mid \boldsymbol{\theta}, \sigma_\theta^2, \boldsymbol{\varphi}, \mathbf{y}, \boldsymbol{\Delta} \sim \mathrm{Scale\text{-}inv} - \chi^2(\nu_\varepsilon^*, \phi_\varepsilon^*)$ where $\nu_\varepsilon^* = \nu_\varepsilon + n$ and $\phi_\varepsilon^* = \frac{1}{\nu_\varepsilon^*}(\nu_\varepsilon\phi_\varepsilon + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})$, with $\boldsymbol{\varepsilon} = \mathbf{y} - \widetilde{\mathbf{U}}\boldsymbol{\theta} - \boldsymbol{\varphi}$;

(5) $\boldsymbol{\varphi} \mid \boldsymbol{\theta}, \sigma_\theta^2, \sigma_\varepsilon^2, \mathbf{y}, \boldsymbol{\Delta} \sim \mathrm{MVN}(\mathbf{m}_\varphi^*, \mathbf{V}_\varphi^*)$ where $\mathbf{V}_\varphi^* = \sigma_\varepsilon^2(\sigma_\varepsilon^2\boldsymbol{\Delta}^{-1} + \mathbf{I}_n)^{-1}$ and $\mathbf{m}_\varphi^* = \frac{1}{\sigma_\varepsilon^2}\mathbf{V}_\varphi^*(\mathbf{y} - \widetilde{\mathbf{U}}\boldsymbol{\theta})$.

Once again, the second step is deterministic and maps back to the effect size analogues $\hat{\boldsymbol{\beta}}$. Iterating the above procedure T times results in a set of samples $\{\hat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}$.

Prediction under this mixed modeling extension is similar to that of a Gaussian process or any other standard nonparametric statistical methods [15]. The response variables to be predicted are simply 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. Partition the vector of response variables $\mathbf{y}$ into a set of training $\mathbf{y}_t$ and validation samples $\mathbf{y}_v$. The design matrix can be similarly partitioned $[\mathbf{X}_t; \mathbf{X}_v]$. Under the randomized feature map $\tilde{\psi}$, the approximate kernel matrix $\widetilde{\mathbf{K}}$ and its eigenvalue decomposition $\widetilde{\mathbf{U}}$ are formulated based on the full design matrix $\mathbf{X}$. The matrix $\mathbf{X}_v$ implicitly forms part of the model and the kernel factor 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. The inferences are made using all the data, including $\mathbf{X}_v$. Therefore, if any new validation samples are introduced, the entire analysis must be repeated [16]. 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. $n_t < n_v$). 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.

4 Description of other Supporting Tables

Table S1

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 S2

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)

5 WTCCC Supporting Result Table

Table S3: Notable regions of the genome showing the strong association

Disease	Chr.	Region (Mb)	Reference	SNP	PPAA	P-Value
CAD	9	22.01-22.12	[1, 17–21]	rs9632884	0.64	2.53E-13
CD	1	67.38-67.46	[1, 19, 21, 22]	rs10489629	0.39	3.71E-12
CD	2	233.94-233.97	[1, 19, 21, 22]	rs6431654	0.30	7.37E-14
CD	3	49.43-49.87	[1, 19–22]	rs6784820	0.28	2.93E-05
CD	5	40.43-40.64	[1, 19–22]	rs10213846	0.37	3.84E-12
CD	6	32.82-32.84	[1, 19, 22]	rs7768538	0.13	2.24E-06
CD	**10**	**79.20-79.29**	**N/A**	**rs2579176**	**0.14**	**2.76E-04**
CD	10	101.26-101.28	[1, 19, 21, 22]	rs7081330	0.13	1.85E-06
CD	16	49.30-49.36	[19–22]	rs17221417	0.29	8.06E-12
HT	**14**	**45.46-45.66**	**N/A**	**rs762015**	**0.12**	**1.96E-03**
RA	1	114.02	[1, 19–21, 23, 24]	rs6679677	0.17	1.55E-26
RA	2	100.19	[25]	rs11694875	0.14	3.15E-04
RA	6	HLA	[1, 19–21, 23, 24]	rs6457617*	1.00	6.22E-79
RA	**17**	**4.10**	**N/A**	**rs9913077**	**0.14**	**1.29E-04**
T1D	1	113.80-114.15	[1, 19–21, 24, 26, 27]	rs1217396	0.39	1.62E-10
T1D	**2**	**206.67-206.85**	**N/A**	**rs4147713**	**0.22**	**1.82E-03**
T1D	**2**	**215.52-215.65**	**N/A**	**rs6737675**	**0.43**	**3.49E-04**
T1D	**3**	**12.51-12.58**	**N/A**	**rs1618545**	**0.19**	**3.11E-04**
T1D	3	46.26-46.37	[27]	rs1799865	0.33	4.89E-05
T1D	**3**	**82.74-82.82**	**N/A**	**rs1097157**	**0.25**	**2.33E-04**
T1D	**3**	**97.03-97.09**	**N/A**	**rs10934261**	**0.16**	**1.16E-04**
T1D	6	HLA	[1, 19–21, 24, 26, 27]	rs9273363*	1.00	0.00E+00

Notable regions of the genome showing the strong association (Continued)

Disease	Chr.	Region (Mb)	Reference	SNP	PPAA	P-Value
T1D	**6**	**120.74-120.84**	**N/A**	**rs12660882**	**0.16**	**3.50E-04**
T1D	12	109.82-111.40	[1, 19–21, 26, 27]	rs17696736	0.92	2.10E-15
T1D	**15**	**48.08-48.11**	**N/A**	**rs9302151**	**0.23**	**3.10E-03**
			N/A	**rs2414005**	**0.21**	**2.60E-03**
T1D	16	10.96-11.34	[1, 19, 21, 24, 26, 27]	rs243327	0.28	1.87E-04
T2D	**4**	**104.04-104.30**	**N/A**	**rs7698608**	**0.10**	**5.02E-04**
T2D	**5**	**153.62-153.63**	**N/A**	**rs11167666**	**0.06**	**3.99E-03**
T2D	10	114.74-114.80	[1, 19, 21, 26]	rs11196205	0.13	5.10E-11

Table of regions with at least two SNPs having PPAAs satisfying the 5% FWER threshold. Listed for all regions are the SNPs with the highest PPAA and its corresponding marginal p-value. The marginal p-values reported are found via linear regression. The reference column gives literature 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 the listed references [19, 26, 28] 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.

References

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

[2] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, February 2013.

[3] William Valdar, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet*, 38(8):879–887, August 2006.

[4] Gota Morota, Prashanth Boddhireddy, Natascha Vukasinovic, Daniel Gianola, and Sue DeNise. 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, 2014.

[5] Xiang Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *bioRxiv*, March 2016.

[6] Yong Jiang and Jochen C. Reif. Modeling epistasis in genomic selection. *Genetics*, 201:759–768, October 2015.

[7] S. Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines with gaussian kernel.

[8] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 44(7):821–825, July 2012.

[9] James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[10] Dawei Liu, Debashis Ghosh, and Xihong Lin. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machinesand linear mixed models. *Biometrics*, 63:1079–1088, 2007.

[11] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, April 2010.

[12] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, March 2008.

[13] Jian Yang, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, February 2014.

[14] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, Jianming Yu, Donna K. Arnett, Jose M. Ordovas, and Edward S. Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, April 2010.

[15] Fiang Liang, Kai Mao, Sayan Mukherjee, and Mike West. Nonparametric Bayesian kernel models. *Department of Statistical Science, Duke University, Discussion Paper*, pages 7–10, 2009.

[16] Mike West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 7, 2003.

[17] Daniela Zanetti, Robert Carreras-Torres, Esther Esteban, Marc Via, and Pedro Moral. Potential signals of natural selection in the top risk loci for coronary artery disease: 9p21 and 10q11. *PLoS ONE*, 10(8):e0134840, 2015.

[18] Sonny Dandona, Alexandre F. R. Stewart, Li Chen, Kathryn Williams, Derek So, Ed O'Brien, Christopher Glover, Michel LeMay, Olivia Assogba, Lan Vo, Yan Qing Wang, Marino Labinaz, George A. Wells, Ruth McPherson, and Robert Roberts. Gene dosage of the common variant 9p21 predicts severity of coronary artery disease. *Journal of the American College of Cardiology*, 56(6):479–486, August 2010.

[19] Christoph Lippert, Jennifer Listgarten, Robert I. Davidson, Jeff Baxter, Hoifung Poon, Carl M. Kadie, and David Heckerman. An exhaustive epistatic SNP association analysis on expanded wellcome trust data. *Scientific Reports*, 3:1099 EP, January 2013.

[20] Tao Feng and Xiaofeng Zhu. Genome-wide searching of rare genetic variants in WTCCC data. *Human Genetics*, 128(3):269–280, September 2010.

[21] Erich Dolejsi, Bernhard Bodenstorfer, and Florian Frommlet. Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian information criterion. *PLoS ONE*, 9(7):e103322, July 2014.

[22] Andre Franke, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A Anderson, Joshua C Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G Mathew, Grant W Montgomery, Natalie J Prescott, Soumya Raychaudhuri, Jerome I Rotter, Philip Schumm, Yashoda Sharma, Lisa A Simms, Kent D Taylor, David Whiteman, Cisca Wijmenga, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Buning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jurgen Glas, Andre Van Gossum, Stephen L Guthery, Jonas Halfvarson, Hein W Verspaget,

Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lemann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julian Panes, Anne Phillips, Deborah D Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A Hillary Steinhart, Pieter C F Stokkers, Leif Torkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R Targan, Steven R Brant, John D Rioux, Mauro D'Amato, Rinse K Weersma, Subra Kugathasan, Anne M Griffiths, John C Mansfield, Severine Vermeire, Richard H Duerr, Mark S Silverberg, Jack Satsangi, Stefan Schreiber, Judy H Cho, Vito Annese, Hakon Hakonarson, Mark J Daly, and Miles Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–1125, December 2010.

[23] Ian C. Scott, Seth D. Seegobin, Sophia Steer, Rachael Tan, Paola Forabosco, Anne Hinks, Stephen Eyre, Ann W. Morgan, Anthony G. Wilson, Lynne J. Hocking, Paul Wordsworth, Anne Barton, Jane Worthington, Andrew P. Cope, and Cathryn M. Lewis. Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking. *PLoS Genet*, 9(9):e1003808, September 2013.

[24] Hariklia Eleftherohorinou, Victoria Wright, Clive Hoggart, Anna-Liisa Hartikainen, Marjo-Riitta Jarvelin, David Balding, Lachlan Coin, and Michael Levin. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE*, 4(11):e8068, November 2009.

[25] Eli A Stahl, Soumya Raychaudhuri, Elaine F Remmers, Gang Xie, Stephen Eyre, Brian P Thomson, Yonghong Li, Fina A S Kurreeman, Alexandra Zhernakova, Anne Hinks, Candace Guiducci, Robert Chen, Lars Alfredsson, Christopher I Amos, Kristin G Ardlie, Anne Barton, John Bowes, Elisabeth Brouwer, Noel P Burtt, Joseph J Catanese, Jonathan Coblyn, Marieke J H Coenen, Karen H Costenbader, Lindsey A Criswell, J Bart A Crusius, Jing Cui, Paul I W de Bakker, Philip L De Jager, Bo Ding, Paul Emery, Edward Flynn, Pille Harrison, Lynne J Hocking, Tom W J Huizinga, Daniel L Kastner, Xiayi Ke, Annette T Lee, Xiangdong Liu, Paul Martin, Ann W Morgan, Leonid Padyukov, Marcel D Posthumus, Timothy R D J Radstake, David M Reid, Mark Seielstad, Michael F Seldin, Nancy A Shadick, Sophia Steer, Paul P Tak, Wendy Thomson, Annette H M van der Helm-van Mil, Irene E van der Horst-Bruinsma, C Ellen van der Schoot, Piet L C M van Riel, Michael E Weinblatt, Anthony G Wilson, Gert Jan Wolbink, B Paul Wordsworth, Cisca Wijmenga, Elizabeth W Karlson, Rene E M Toes, Niek de Vries, Ann B Begovich, Jane Worthington, Katherine A Siminovitch, Peter K Gregersen, Lars Klareskog, and Robert M Plenge.

Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*, 42(6):508–514, June 2010.

[26] Yu Zhang, Jing Zhang, and Jun S. Liu. Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. *Annals of Applied Statistics*, 5(3):2052–2077, 2011.

[27] Jonathan P. Bradfield, Hui-Qi Qu, Kai Wang, Haitao Zhang, Patrick M. Sleiman, Cecilia E. Kim, Frank D. Mentch, Haijun Qiu, Joseph T. Glessner, Kelly A. Thomas, Edward C. Frackelton, Rosetta M. Chiavacci, Marcin Imielinski, Dimitri S. Monos, Rahul Pandey, Marina Bakay, Struan F. A. Grant, Constantin Polychronakos, and Hakon Hakonarson. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet*, 7(9):e1002293, 09 2011.

[28] Min-Seok Kwon, Mira Park, and Taesung Park. IGENT: efficient entropy based algorithm for genome-wide gene-gene interaction analysis. *BMC Medical Genomics*, 7, 2014.