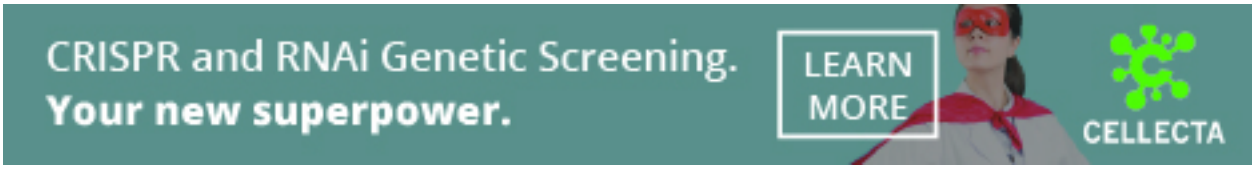
Multikernel linear mixed models for complex phenotype prediction

Omer Weissbrod, Dan Geiger and Saharon Rosset

Genome Res. 2016 26: 969-979 originally published online June 14, 2016
Access the most recent version at doi:10.1101/gr.201996.115



Method

Multikernel linear mixed models for complex phenotype prediction

Omer Weissbrod,[1,2] Dan Geiger,[2] and Saharon Rosset[1]

[1] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 6997801, Israel; [2] Computer Science Department, Technion–Israel Institute of Technology, Haifa 3200003, Israel

Linear mixed models (LMMs) and their extensions have recently become the method of choice in phenotype prediction for complex traits. However, LMM use to date has typically been limited by assuming simple genetic architectures. Here, we present multikernel linear mixed model (MKLMM), a predictive modeling framework that extends the standard LMM using multiple-kernel machine learning approaches. MKLMM can model genetic interactions and is particularly suitable for modeling complex local interactions between nearby variants. We additionally present MKLMM-Adapt, which automatically infers interaction types across multiple genomic regions. In an analysis of eight case-control data sets from the Wellcome Trust Case Control Consortium and more than a hundred mouse phenotypes, MKLMM-Adapt consistently outperforms competing methods in phenotype prediction. MKLMM is as computationally efficient as standard LMMs and does not require storage of genotypes, thus achieving state-of-the-art predictive power without compromising computational feasibility or genomic privacy.

[Supplemental material is available for this article.]

One of the principal aims of genetics research is accurate phenotype prediction. This goal has largely been achieved for Mendelian diseases with a small number of risk variants (Schrodi et al. 2014). However, many genetic traits have a complex genetic architecture that is not well understood (Golan et al. 2014). Phenotype prediction for such traits remains a major challenge.

A key challenge in complex phenotype prediction is accurate modeling of genetic interactions, commonly known as epistatic effects (Cordell 2002). In recent years, there has been mounting evidence that epistatic interactions are widespread throughout biology (Moore and Williams 2009; Lehner 2011; Hemani et al. 2014; Buil et al. 2015). It is well accepted that epistatic interactions are biologically plausible on the one hand (Zuk et al. 2012) and are difficult to detect on the other (Cordell 2009), suggesting that they may be highly influential in our limited success in modeling complex heritable traits.

Linear mixed models (LMMs) have long been considered the method of choice for modeling of complex phenotypes and have gained tremendous interest in recent years (Yu et al. 2006; Kang et al. 2008; Price et al. 2010; Yang et al. 2010; Zhang et al. 2010; Lippert et al. 2011; Zhou and Stephens 2012; de Los Campos et al. 2013). At their core, LMMs encode the assumption that genetically similar individuals are more likely to share similar phenotypes. Despite their popularity, LMM use to date has often been limited by their restriction to relatively simple genetic models that cannot capture interactions. The vast majority of studies to use LMMs assume that genetic variants influence phenotypes in an additive manner. A smaller number of studies investigated LMMs that model dominance effects (Powell et al. 2013; Vitezica et al. 2013; Da et al. 2014; Nishio and Satoh 2014; Zhu et al. 2015), multiplicative interactions between pairs or triples of variants (Henderson 1985; Su et al. 2012; Muñoz et al. 2014; Bloom et al. 2015), gene-environment interactions (Wang et al. 1999; Yang et al. 2007; Ferraudo and Perecin 2014), or LMMs that measure genetic similarity according to identity by state (Wu et al. 2011). Such LMM-based approaches had limited success in complex phenotype prediction, possibly owing to the relatively simple forms of considered interactions.

In recent years, LMMs that can model higher-order interactions have also been investigated, typically under the name, reproducing kernel Hilbert space regression (RKHS) (Liu et al. 2007, 2008; Ober et al. 2011; Gianola et al. 2014; Morota and Gianola 2014; Tusell et al. 2014; Akdemir and Jannink 2015; Jiang and Reif 2015). These works demonstrated improved prediction performance on several plant and animal species compared to simpler methods (Perez-Rodriguez et al. 2012; Rutkoski et al. 2012; Crossa et al. 2013). However, as we demonstrate here, these RKHS-based LMMs are often insufficient to improve prediction performance for complex traits in outbred populations.

Two possible reasons for the limited success of RKHS methods are the extremely high dimensional search space of interacting variants and the assumption of a homogeneous distribution of variant effect sizes across the genome. In recent years, several studies have demonstrated that modeling a heterogeneous effect-size distribution can improve phenotype prediction substantially (Zhou et al. 2013; Moser et al. 2015). Namely, the recently proposed adaptive MultiBLUP (AMB) (Speed and Balding 2014) obtained state-of-the-art phenotype prediction results by grouping variants in close proximity together and inferring different effect-size distributions for different groups. A natural extension of this idea is to model local epistatic interactions within each such group via RKHS, as this restricts the search space for interacting variants considerably. In recent years, there has been an increasing body of evidence for such local genetic interactions between nearby variants (Dimas et al. 2008; Bickel et al. 2011; Haig 2011; Lappalainen et al.

Corresponding author: saharon@post.tau.ac.il
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.201996.115.

2011). However, modeling of a heterogeneous effect-size distribution combined with use of RKHS in a single unified model has not been investigated to date.

Here, we present multikernel linear mixed model (MKLMM), a flexible modeling framework that allows for both global and local high-order interactions modeled via RKHS, as well as modeling of a heterogeneous effect-size distribution. Similarly to AMB, MKLMM first divides the genome into a set of regions and then models the genetic effects within each region via a covariance matrix of genetic similarities between individuals, often called a kernel in the machine learning literature (Rasmussen and Williams 2006). MKLMM differs from AMB by allowing for both additive and nonadditive genetic effects within each region, and differs from previous RKHS methods by dividing the genome into regions and efficiently inferring all model parameters for all kernels jointly via restricted maximum likelihood (Kang et al. 2008). This enables the construction of rich models with a large number of parameters without requiring manual fine tuning, allowing MKLMM to capture complex nonadditive genetic interactions.

In addition to describing the general MKLMM framework, this work presents several specific contributions. First, we evaluate several kernel types in terms of both their underlying assumptions and their predictive performance on real data sets. In particular, we present the saturating pathways kernel, which models interactions via saturation dynamics. The underlying model of this kernel bears similarity to the well-known limiting pathways model (Zuk et al. 2012), but assumes that pathways interact additively rather than competitively. We further demonstrate that this kernel is often superior to other state-of-the-art kernels.

Second, we present MKLMM-Adapt, an adaptive data-driven approach that automatically infers interaction types across different genomic regions and selects appropriate kernels. The resulting model allows different genomic regions to have different interaction patterns and effect-size distributions, leading to state-of-the-art predictive power.

Finally, we demonstrate that MKLMM-based prediction can be carried out without having to store genotypes of training individuals. Although this is a property shared by many predictive methods, it is not trivially carried over to kernel-based methods, because such methods require estimating genetic similarity between training and tested individuals. The storage of personal genomes raises security and privacy concerns that have recently gained considerable attention both within and outside academia (Im et al. 2012; Gymrek et al. 2013; Rodriguez et al. 2013; Erlich and Narayanan 2014; Dove et al. 2015). We demonstrate that MKLMM can approximate genetic similarity well without storing genotypes and phenotypes, thus alleviating privacy and security concerns.

Results

We evaluated MKLMM on synthetic and real data sets. All experiments followed the same procedure of dividing the genome into regions, evaluating models with 0,1,2,…,9 region-kernels, and selecting the best model using cross validation. All models included a genome-wide kernel spanning all genotyped variants.

The evaluated methods included (1) MKLMM-Adapt, which automatically selects the kernel for each region; (2) MKLMM-Poly2, which always uses a weighted combination of a linear and a polynomial kernel of degree 2 for every region, used as a baseline nonlinear method; (3) Adaptive MultiBLUP (AMB) (Speed and Balding 2014), which uses a linear kernel for every region; and (4) Genomic best linear unbiased prediction (GBLUP) (Gianola 2013), which is equivalent to AMB with only one kernel spanning the entire genome. AMB has recently been demonstrated to be superior to other popular methods in several challenging prediction tasks (Speed and Balding 2014). We note that MKLMM-Poly2 is equivalent to models that can capture both additive and multiplicative interaction effects between pairs of variants (Methods; Henderson 1985; Su et al. 2012; Muñoz et al. 2014; Bloom et al. 2015).

The free parameters to infer for each method are the fixed effects (the effects of risk factors such as smoking status or age as well as an overall intercept, which was the only fixed effect used in the simulations), the variance of the environmental component of the phenotype, and the kernel parameters, whose number depends on the number of selected regions. The number of kernel parameters for AMB, MKLMM-Poly2, and MKLMM-Adapt is one per region, two per region, and between one and three parameters per region, respectively (Methods).

Simulation studies

We performed simulation studies by generating synthetic phenotypes based on real genotypes from Chromosome 1 of 2801 individuals from the Wellcome Trust national blood service cohort (The Wellcome Trust Case Control Consortium 2007). The phenotype of each individual was generated by a combination of linear and nonlinear effects distributed across 2, 4, or 6 genomic regions with a mean length of 75 kb, where each region harbored both linear and interaction effects. These numbers were selected based on the number and length of regions typically selected by AMB and MKLMM-Adapt in the analysis of real data sets with similar sample sizes. The effect of these numbers on prediction performance is examined below. Unless otherwise stated, the genomic regions jointly accounted for 25% of the phenotypic variance, another 25% was similarly explained by a global region spanning the entire chromosome, and the remaining 50% was due to an independent normally distributed environmental effect.

The use of Chromosome 1 rather than the entire genome was a choice of convenience and is unlikely to affect the validity of the results, because the average linkage disequilibrium between two variants in different regions for a chromosome of this size is effectively zero, indicating statistical independence between different regions, as would be obtained had the entire genome been utilized. The effect of a chromosome-wide region is also effectively the same as that of a genome-wide region, because in both cases, the number of variants is substantially greater than the sample size, leading to highly similar genetic covariance estimates.

Within each region, the linear effects accounted for 25% of the phenotypic variance explained by the region, and interaction effects accounted for the remaining 75%. Although these settings may exaggerate the magnitude of nonlinear effects expected in real data, they enable investigating the dynamics of nonlinear interactions at greater depth. The simulated interactions included groupwise multiplicative interactions, which generalize pairwise multiplicative interactions to higher orders, and saturating effects, which bound the magnitude of linear effects (Methods). Every simulated interaction involved randomly selected variants from the same region, but this definition includes the chromosome-wide region that spans all variants. A detailed description of the simulation procedure is given in the Supplemental Material.

In the first experiment, we also evaluated two additional MKLMM variants with predetermined kernel types: MKLMM-

Radial Basis Function (MKLMM-RBF) and MKLMM-Saturating Pathways (MKLMM-SP). Each of these variants uses a weighted combination of a linear and a nonlinear kernel of the corresponding type (RBF or SP) for every region (Methods). Briefly, the RBF kernel is a widely used kernel in machine learning and statistical genetics literature (Morota and Gianola 2014), which encodes genetic similarity between individuals k and l in a region with m variants as being proportional to

$$\exp\left(-\frac{1}{2\theta m}\sum_{i=1}^{m}\left(X_k^i - X_l^i\right)^2\right),$$

where θ is a parameter of the model; and X_k^i is the allele of individual k at variant i in the region. The SP kernel encodes genetic similarity between individuals, k and l, as being proportional to

$$\sin^{-1}\left(\frac{\left(\sum_{i=1}^{m} X_k^i X_l^i\right)/m}{\sqrt{\left(\theta+\sum_{i=1}^{m}\left(X_k^i\right)^2/m\right)\left(\theta+\sum_{i=1}^{m}\left(X_l^i\right)^2/m\right)}}\right).$$

Prediction performance was first evaluated using data with various degrees of nonlinear interactions and either two or six genomic regions harboring interacting variants. The advantage of all MKLMM methods over AMB increased with the degree of interactions, and all methods substantially outperformed GBLUP under all settings (Fig. 1). All MKLMM methods performed similarly to AMB when no genetic interactions were present, indicating that the models did not overfit the training data under this setting. The advantage of AMB over GBLUP decreased with the degree of genetic interactions because the high-dimensional effect of the chromosome-wide region can be approximated relatively well by a linear effect (a well-known property of high-dimensional hyperplanes), unlike the lower-dimensional region-specific effects, rendering AMB similar to GBLUP in highly nonlinear settings.

Figure 1 provides additional insight into two aspects of the behavior of the MKLMM methods. First, MKLMM-Poly2 was less powerful than the other MKLMM methods in the presence of two genomic regions and was slightly more powerful under six genomic regions. Second, MKLMM-Adapt performed as well as or better than MKLMM-RBF and MKLMM-SP under all settings. Both results demonstrate that prediction performance depends on a trade-off between model expressiveness and complexity: MKLMM-Poly2 estimates two parameters for every genomic region, whereas the other MKLMM models estimate three parameters for every region. These richer MKLMM models can capture diverse interaction types, but have a greater risk of overfitting the training data when many parameters need to be estimated. Since MKLMM-Adapt performs as well as or better than MKLMM-RBF and MKLMM-SP in most settings, we do not consider these two kernel types in the remainder of this section.

To further investigate the factors affecting prediction performance, we generated data sets with various sample sizes and numbers of genomic regions. Both the advantage of MKLMM methods over AMB and the number of kernels selected by all methods increased with sample size (Fig. 2). The advantage of AMB over GBLUP increased with the number of genomic regions because AMB can capture well the linear portion of the effects of multiple genomic regions. The advantage of MKLMM-Adapt over MKLMM-Poly2 increased with sample size and decreased with the number of regions, indicating that kernel selection improves with sample size but deteriorates with the number of free parameters. Future genetic studies with larger samples will therefore enable constructing richer MKLMM-Adapt models that can capture diverse interaction patterns, which cannot be expressed as simple pairwise multiplicative interactions.

To investigate additional factors affecting prediction performance, we generated data sets with various ratios of explained genetic variance to phenotypic variance (commonly known as heritability when the entire genome is included in the analysis) and various genomic region lengths. The advantage of MKLMM models over AMB increased with the ratio of explained genetic variance to phenotypic variance (Supplemental Fig. S1) and decreased for region lengths larger than 75 kb, indicating that genetic interactions can be better captured over short distances (Supplemental Fig. S2). This result motivates the selection of short regions used by MKLMM-Adapt. We also verified that our results remain consistent in the presence of binary phenotypes (Supplemental Fig. S3). Finally, we verified that our results are highly unlikely to arise due to implicit tagging of genotyped variants (Supplemental Material).

To conclude, our simulations reveal that the advantage of MKLMM-Adapt over AMB increases with the magnitude of genetic interactions within regions, with sample size, with the ratio of genetic to environmental phenotypic variance, and with proximity between interacting variants. We emphasize that these simulations do not attempt to fully mimic realistic genetic studies, because the types

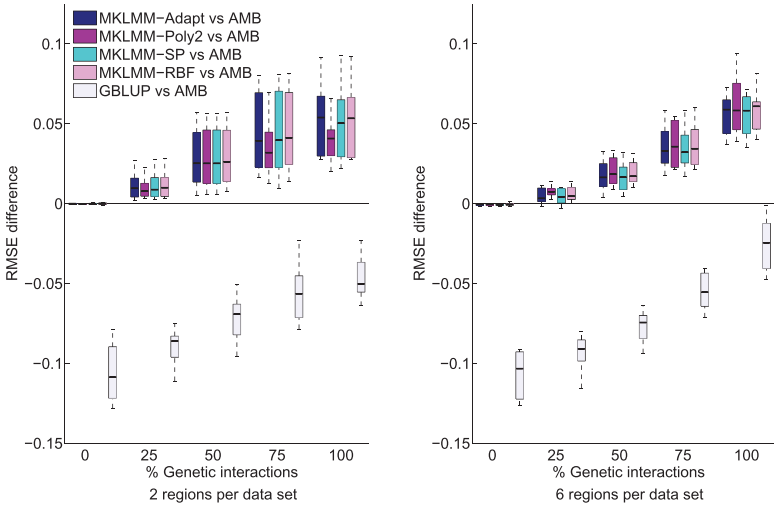


Figure 1. Comparison of the evaluated methods on synthetic data sets with various levels of genetic interactions and numbers of genomic regions harboring interacting variants. The box plots show the difference between the evaluated methods and Adaptive MultiBLUP (AMB), in terms of the root mean square error (RMSE) between the predicted and observed phenotype. Larger values indicate a greater advantage over AMB. The advantage of the MKLMM methods over AMB increased with the percentage of genetic interactions. MKLMM-Poly2 gains a slight advantage over the other MKLMM methods in the presence of six regions, because its simpler model is less prone to overfitting when many parameters need to be estimated from the data.

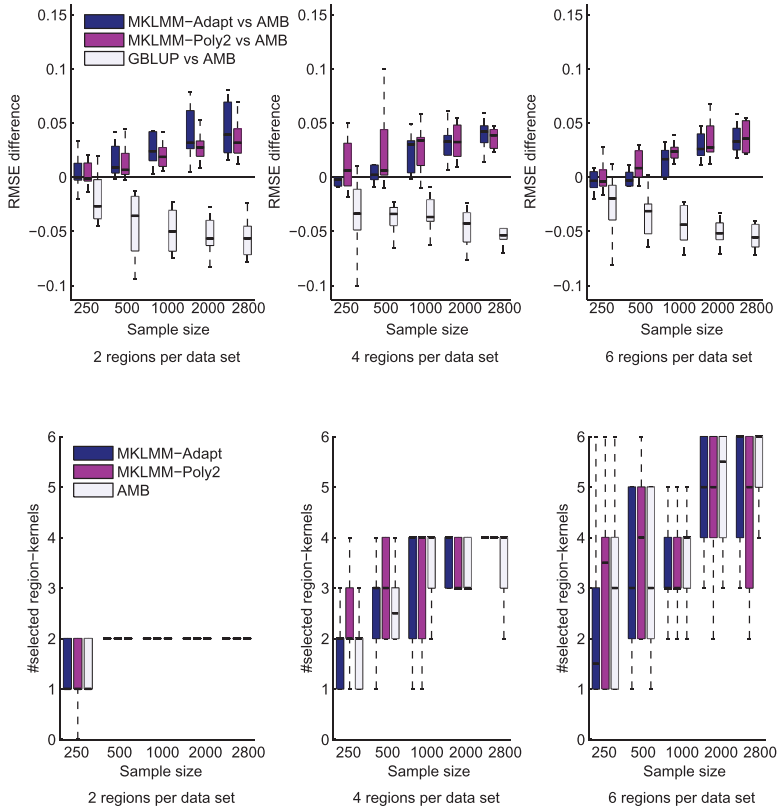


Figure 2. Comparison of the evaluated methods on synthetic data sets with various sample sizes and numbers of genomic regions harboring interacting variants. The values shown are the differences in prediction accuracy compared to AMB (*top*), and the number of region-kernels selected by each method (*bottom*). Larger samples allow using more expressive models with a larger number of kernels. The advantage of MKLMM methods over AMB, and of MKLMM-Adapt over MKLMM-Poly2, increases with sample size and decreases with the number of regions. The median lines in the *bottom* row are sometimes not shown because they intersect with the other quartiles, owing to the discrete nature of these data.

and magnitude of genetic interactions in real data are as yet unknown. Rather, these simulations are intended to shed light on the factors that can affect the prediction performance of MKLMM-Adapt in real genetic studies.

Analysis of mouse phenotypes

We evaluated MKLMM on a data set of 1940 outbred mice measured for 133 quantitative phenotypes spanning several biochemical, behavioral, and disease-related traits (Valdar et al. 2006). The preprocessing procedure is described in the Supplemental Material. Mice from the same cage were always assigned to the same cross-validation fold to prevent leakage, as previously recommended (Speed and Balding 2014). AMB has recently been demonstrated to perform as well as or better than several other state-of-the-art methods on this data set (Speed and Balding 2014) and is thus used as a benchmark method.

MKLMM-Adapt outperformed AMB across 100, 96, and 83 phenotypes in terms of RMSE, out of sample log likelihood (OOS LL), and Pearson correlation, respectively (Fig. 3; Supplemental Figs. S4–S8). After accounting for multiple hypothesis testing, the advantage was statistically significant across 43, 45, and 31 phenotypes, respectively, as computed via permutation testing (Supplemental Material). In contrast, AMB had a significant advantage over MKLMM-Adapt across only four phenotypes, according to RMSE, and across no phenotypes according to the other criteria. A disadvantage of the individual hypothesis tests is that their power decreases with the number of phenotypes due to the multiple testing correction. To assess the global advantage of MKLMM-Adapt, we carried out a one-sided Wilcoxon signed rank test that demonstrated its superiority over AMB in terms of RMSE ($P < 2.09 \times 10^{-12}$), OOS LL ($P < 1.84 \times 10^{-12}$), and Pearson correlation ($P < 6.4 \times 10^{-4}$).

MKLMM-Poly2 results were comparable to those of MKLMM-Adapt, in agreement with our simulation results given the relatively small data set size (Fig. 3; Supplemental Figs. S4–S8). For completeness, we also verified that MKLMM-SP and MKLMM-RBF yield similar results (Supplemental Figs. S9–S11). These results indicate that MKLMM can be routinely exploited to improve prediction performance.

Figure 3 additionally demonstrates that the advantage of MKLMM-Adapt over AMB is inversely correlated with the advantage of GBLUP over AMB (Pearson correlation of –0.26), indicating that MKLMM-Adapt tends to outperform AMB when AMB outperforms GBLUP, possibly indicating traits with more complex and challenging genetic architectures.

Analysis of human diseases

We evaluated MKLMM on seven ascertained human disease data sets from the Wellcome Trust Case Control Consortium 1 (WTCCC1) (The Wellcome Trust Case Control Consortium 2007) and on a large Wellcome Trust Case Control Consortium 2 (WTCCC2) ulcerative colitis data set (Supplemental Material; UK IBD Genetics Consortium et al. 2009). The controls group consisted of individuals from the United Kingdom blood service control group. A second control group was not included in the main experiments to alleviate concerns that it helps the nonlinear methods gain an unfair advantage because of exploitation of population structure. To avoid spurious results owing to population structure, the genotypes of each data set were regressed on the top 10 principal components (PCs), and only the residuals were used for prediction. Each of the WTCCC1 data sets contained approximately 1900 cases and 1600 controls, whereas the ulcerative colitis data set contained 2697 cases and 2801 controls (Supplemental Table S1).

The analysis of binary phenotypes in case-control studies is challenging, because proper modeling requires taking the case-control ascertainment scheme into account (Golan and Rosset 2014), where parameter inference is intractable (Supplemental Material). Nevertheless, several recent works have demonstrated

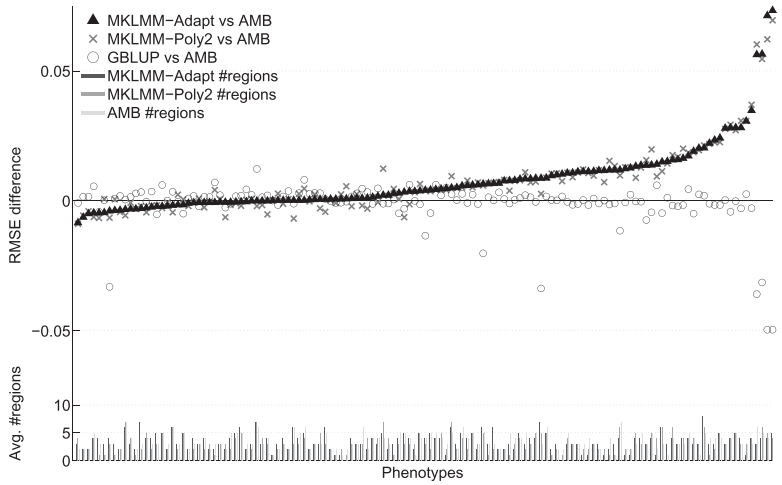


Figure 3. Comparison of the evaluated methods in prediction of mouse phenotypes. Each dot represents the difference between the prediction performance of an evaluated method and AMB across one of the phenotypes, measured according to RMSE. Also shown is the average number of regions selected by each method across the training folds (rounded to the closest integer). MKLMM-Adapt outperformed AMB and GBLUP across 100 and 86 phenotypes, respectively.

that treating binary phenotypes as if they were quantitative in LMMs can lead to effective predictions (Zhou et al. 2013; Speed and Balding 2014; Moser et al. 2015). The methods were compared according to the area under the receiver operating characteristic curve (AUC), which presents the true positive rate (sensitivity) on the *y*-axis plotted against the false positive rate (one minus specificity) on the *x*-axis (Methods). This measure was selected because it is insensitive to ascertainment bias. For completeness, we report additional performance measures in the Supplemental Material. AMB has recently been demonstrated to outperform several other state-of-the-art methods on WTCCC1 data sets (Speed and Balding 2014) and is thus used as a benchmark method.

MKLMM-Adapt held a statistically significant advantage over AMB in prediction of Crohn's disease (CD), type 1 diabetes (T1D), and ulcerative colitis (UC), whereas AMB did not hold a statistically significant advantage over MKLMM-Adapt in any data set (Table 1; Supplemental Tables S2–S4). The large advantage in prediction of UC, which is the largest data set by far, corroborates the finding that the advantage of MKLMM-Adapt increases with sample size. The results suggest that MKLMM-Adapt can consistently be preferred over AMB because its added complexity leads to greater accuracy in the presence of genetic interactions, but is not likely to decrease accuracy when this complexity is not used.

When excluding variants within 5 kb of the major histocompatibility complex (MHC) in the analysis of autoimmune diseases, prediction performance for T1D and rheumatoid arthritis decreased for all methods, whereas there was no decrease for CD, and the advantage of MKLMM-Adapt over AMB in the analysis remained statistically significant (Supplemental Table S2). The results also remained similar when retaining the top 10 principal components, although the linear methods became slightly more powerful, owing to the well-known linear effect of top principal components on phenotypes (Supplemental Table S5; Price et al. 2006). We also evaluated results with a second control group and observed that the advantage of the nonlinear methods over AMB became more pronounced (Supplemental Table S6). For completeness, we also evaluated predictive performance using MKLMM-SP and MKLMM-RBF (Supplemental Table S7). Neither of the two kernel types was superior across all phenotypes, which reiterates the advantage of the data-driven kernel selection of MKLMM-Adapt.

Although MKLMM-Adapt is intended for prediction rather than estimation or hypothesis testing, its estimated parameters can shed some light on the underlying genetic mechanism of heritable traits. For example, in the analysis of T1D, a certain genomic region spanning 119 single-nucleotide polymorphisms (SNPs) in the MHC (Chromosome 6 positions 32,431,292–33,218,180, hg18) was consistently estimated to be very well modeled by a SP kernel across all training folds. Furthermore, the advantage of MKLMM-Adapt over AMB in T1D was significant according to both AUC and OOS LL (Supplemental Table S3). These results indicate that the underlying mechanism of the well-known MHC effect on T1D may be strongly influenced by genetic interactions.

Finally, we verified that the AUC obtained by each method peaks at a certain number of region-kernels and then drops (Fig. 4). This demonstrates that prediction accuracy depends on a trade-off between model complexity and expressiveness. Richer models can capture more complex interaction patterns at the risk

Table 1. Prediction results on WTCCC data sets

Trait	MKLMM-Adapt	MKLMM-Poly2	AMB	GBLUP	*P*-value
CD	**0.667 ± 0.010**	0.650 ± 0.010	0.645 ± 0.011	0.582 ± 0.013	5.00×10^{-5}
T1D	**0.886 ± 0.004**	0.885 ± 0.003	0.883 ± 0.004	0.601 ± 0.008	1.82×10^{-2}
BD	0.563 ± 0.011	0.571 ± 0.012	0.568 ± 0.011	0.578 ± 0.011	9.11×10^{-1}
RA	0.750 ± 0.009	0.749 ± 0.009	0.752 ± 0.010	0.671 ± 0.009	7.30×10^{-1}
T2D	0.634 ± 0.007	0.632 ± 0.008	0.634 ± 0.007	0.598 ± 0.009	3.47×10^{-1}
CAD	0.698 ± 0.014	0.697 ± 0.014	0.699 ± 0.013	0.701 ± 0.015	5.16×10^{-1}
HT	0.611 ± 0.003	0.610 ± 0.003	0.611 ± 0.003	0.576 ± 0.005	9.26×10^{-1}
UC	**0.601 ± 0.007**	0.590 ± 0.003	0.585 ± 0.002	0.583 ± 0.004	1.20×10^{-4}

The reported values are the AUCs obtained across a fivefold cross validation, and the *P*-value of the hypothesis that MKLMM-Adapt obtains a higher AUC than Adaptive MultiBLUP (AMB), as determined via permutation testing. Standard errors were computed by comparing results across the five training folds. Results in bold have a statistically significant advantage over AMB at a 5% false discovery rate.
(CD) Crohn's disease; (T1D) type 1 diabetes; (BD) bipolar disorder; (RA) rheumatoid arthritis; (T2D) type 2 diabetes; (CAD) coronary artery disease; (HT) hypertension; (UC) ulcerative colitis.

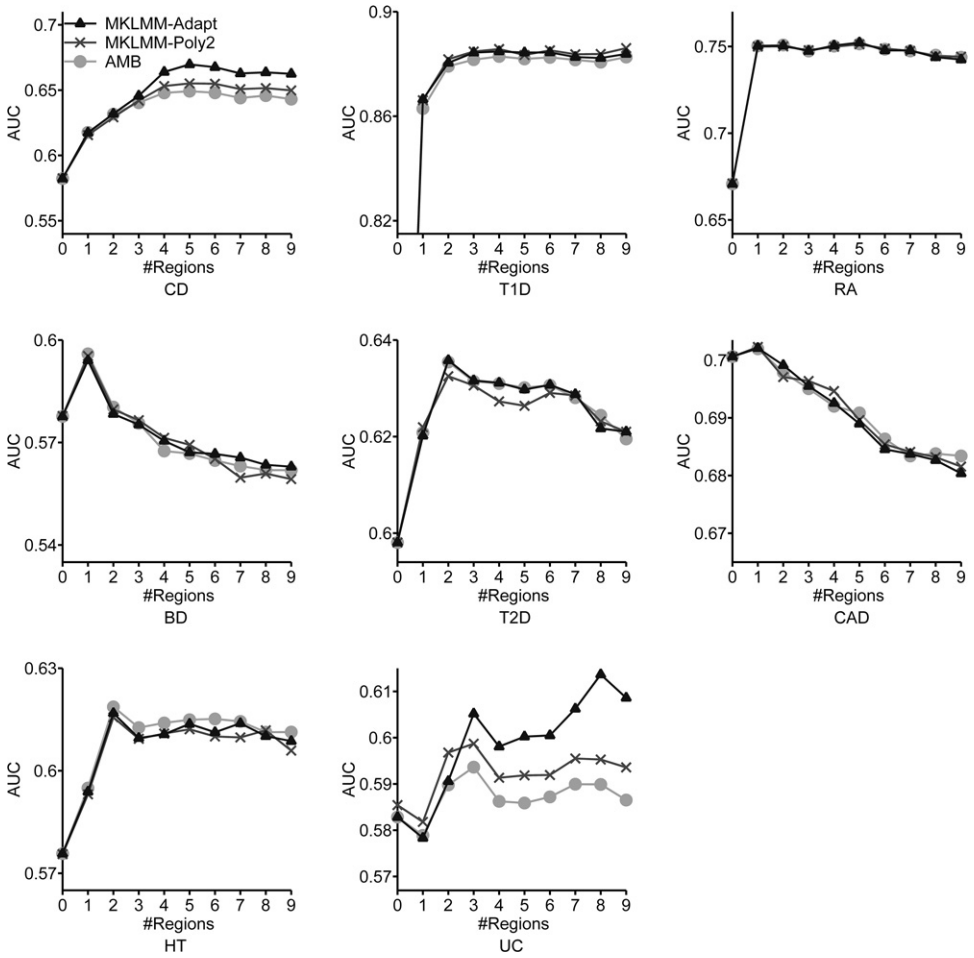


Figure 4. Evaluation of prediction performance in WTCCC data sets as a function of the number of selected regions: (CD) Crohn's disease; (T1D) type 1 diabetes; (BD) bipolar disorder; (RA) rheumatoid arthritis; (T2D) type 2 diabetes; (CAD) coronary artery disease; (HT) hypertension; (UC) ulcerative colitis. The AUC for T1D with zero regions is 0.589 for all methods and is omitted for clarity. GBLUP is not shown explicitly, because AMB with no selected regions is equivalent to GBLUP. MKLMM-Adapt performed as well as or better than AMB across all data sets (evaluated at the number of regions at which prediction performance peaked). Prediction performance always peaked at a certain number of regions and then dropped, indicating that the models may overfit the data when an overly large number of regions is selected. The phenotypes where MKLMM-Adapt performed significantly better than AMB (CD, T1D, UC) appear to be those in which many regions are required for good performance, implying a complex genetic architecture.

of overfitting the training data, suggesting that larger data sets will enable using richer models for improved prediction. Figure 4 additionally demonstrates the aforementioned advantage of MKLMM-Adapt over AMB in prediction of CD, T1D, and UC at the optimal number of regions, which can be found via cross validation. Figure 4 further demonstrates that the additional complexity provided by MKLMM-Adapt does not appear to come at the price of lower accuracies when the extra complexity is not utilized.

Discussion

This work describes the MKLMM phenotype prediction framework that extends the standard LMM to model both linear effects and nonlinear interactions. Our work generalizes and improves upon several recent phenotype prediction approaches and yields state-of-the-art prediction results on several data sets. MKLMM is rich and flexible, yet computationally tractable. The likelihood can be evaluated analytically, and the model only has a single tunable hyperparameter that determines the number of region-specific kernels and serves to balance between model complexity and expressiveness. Our simulations demonstrate that although larger sample sizes lead to greater prediction accuracy, substantial gains in accuracy over alternative methods can be obtained even with samples as small as 250 individuals in the presence of strong genetic interactions between nearby variants.

The MKLMM framework allows incorporating prior knowledge about interaction types via the choice of kernels and their parameters. Alternatively, MKLMM-Adapt can automatically select kernels in a data-driven manner. Compared to AMB, MKLMM-Adapt incurs a small additional computational cost and a small additional statistical cost in estimating more parameters. Our experiments demonstrate that MKLMM-Adapt can be consistently preferred over AMB as it provides greater prediction accuracy in the presence of genetic interactions, but does not appear to harm prediction accuracy when there are no genetic interactions to exploit.

In addition to modeling interactions within genomic regions, MKLMM can potentially use additional sources of information other than SNPs for phenotype prediction, such as gene expression or methylation. MKLMM can naturally be adapted to model

diverse sources of information, by assigning designated kernels for different data sources.

In this study, MKLMM-Adapt is configured to select one of four possible kernel types for each region, one of which is the Poly2 kernel often used in genetic studies to model pairwise multiplicative interactions (Henderson 1985). However, there is a wide range of additional relevant kernels described in both the machine learning literature (Rasmussen and Williams 2006) and statistical genetics literature, such as the dominance kernel (Vitezica et al. 2013) and identity by state kernel (Wu et al. 2011). Such kernels are not considered here because the more complex kernels used here have proven to provide greater prediction accuracy under a wide variety of settings (Morota and Gianola 2014), but they can be incorporated within the MKLMM framework in a straightforward manner.

MKLMM-Adapt is especially suitable for capturing interactions between nearby variants in the same region. Although MKLMM-Adapt can potentially capture interactions between distant variants by using a nonlinear kernel for the genome-wide region, in practice, a linear kernel was always selected for this region in the human disease data sets. A possible reason is that adding an additional genome-wide kernel can improve prediction performance if there are a large number of genome-wide interacting variants, but can also harm prediction performance due to the additional variance introduced into the model. It is likely that the latter alternative was more dominant given the relatively small sample sizes of these data sets. If a researcher is aware of potential interactions between distant variants, it is possible to manually define a nonconsecutive region encompassing these variants, which would enable MKLMM to exploit these interactions for improved prediction.

In addition to presenting MKLMM, we describe a privacy-preserving scheme that enables MKLMM to be used without storing genotypes and phenotypes of training individuals. This property is common to many predictive modeling methods, but is not trivially carried over to kernel-based methods that require genetic similarity between test and training individuals. Although exact recovery of genotypes and phenotypes is impossible in the general case, it may be possible to perform approximate recovery by exploiting domain knowledge, such as specific properties of SNPs. Additional work is required to investigate such approximations.

MKLMM is based on LMM, similarly to many other popular methods for complex trait prediction (Meuwissen et al. 2001; Habier et al. 2011; Zhou et al. 2013; Golan and Rosset 2014; Morota and Gianola 2014; Speed and Balding 2014; Moser et al. 2015). Alternative methods such as decision tree ensembles, support vector machines, and regularized regression techniques (Hastie et al. 2009) are also potential candidates for complex trait prediction. Nevertheless, LMMs share several properties that render them particularly suitable for this task. First, LMMs and their extensions gracefully handle very high-dimensional data, because all data is represented in the covariance matrix, which scales quadratically in the sample size regardless of the data dimensionality. A second advantage is that model parameters such as kernel weights and fixed effects can be inferred analytically because LMMs constitute a full likelihood model. This property is not naturally shared by the aforementioned methods, wherein more expensive grid search methods are typically used (Rasmussen and Williams 2006). This limitation becomes especially severe when dozens of parameters are inferred simultaneously, as in the present study.

A potential concern with kernel-based prediction is that prediction performance may be improved due to better tagging of nongenotyped variants, rather than modeling of true genetic interactions. This can occur when nongenotyped causal variants with a linear effect are accurately imputed in a nonlinear manner by neighboring variants. Our simulation studies indicate that this is highly unlikely to be the reason for the success of MKLMM-Adapt. Nevertheless, improvement in prediction performance due to implicit tagging of nongenotyped variants is legitimate and could remain useful even under whole-genome sequencing due to stringent filtering of low quality variants.

A second potential concern with kernel-based prediction is that population structure may be exploited to improve prediction performance. However, it is generally thought that population structure in the disease data sets investigated here can be accurately captured via principal components (Golan et al. 2014), which were excluded from the analysis. Moreover, in all the disease data sets, the genome-wide kernel was selected to be linear across all folds, indicating that MKLMM-Adapt did not capture global population structure signals that could not be captured by AMB. We further note that exploitation of population structure to improve prediction performance is legitimate, provided that structure in the study reflects the true underlying structure in the population.

One limitation of MKLMM-Adapt is that the null distribution of the statistical test used for kernel selection may differ from the expected theoretical distribution (Lippert et al. 2014). Additionally, MKLMM-Adapt uses a greedy data-driven scheme to construct a composite kernel. It may be possible to circumvent the kernel selection problem by adopting a fully Bayesian framework, wherein the kernel is a mixture of region-specific kernels whose parameters have a prior distribution. Similar formulations have recently been explored in the machine learning literature (Lázaro-Gredilla and Titsias 2011; Gönen 2012). Importantly, such a Bayesian model could encapsulate MultiBLUP and several recent Bayesian LMM extensions (Zhou et al. 2013; Moser et al. 2015) in a unified framework.

In this work, we treat all phenotypes as if they were normally distributed. Quantitative non-normal phenotypes can potentially be transformed to follow a normal distribution via a warping function (Fusi et al. 2014). However, binary phenotypes present a greater challenge. Although MKLMM can be adapted to model binary distributions, accurate parameter inference is intractable in such cases (Supplemental Material). Our extensive simulations demonstrate that adapting MKLMM for binary phenotypes increases prediction performance when the true model parameters are known, but decreases prediction performance otherwise (results not shown). An efficient parameter inference scheme for ascertained case-control studies therefore remains an open problem.

Finally, our work investigates phenotype prediction. Other common tasks in genetic studies include association testing, heritability estimation, and modeling of population structure. However, such statistical analyses intimately depend on the validity of the assumptions underlying the model, unlike prediction, in which performance can be measured empirically. Adapting the MKLMM framework for such tasks remains a potential avenue for future work.

Methods

MKLMM is a probabilistic model that extends LMMs via kernelization of its covariance matrix. Here, we briefly review LMMs,

describe kernelization of the covariance matrix, present a data-driven method for automatic kernel construction, and describe a privacy-preserving scheme for MKLMM-based predictions. Further details are provided in the Supplemental Material.

The linear mixed model

LMMs model the distribution of a normally distributed trait y. Under LMMs, every individual i is associated with a genotype vector X_i and a covariates vector C_i (typically defined as major risk factors, such as smoking status or age). Given a sample of individuals with a genotyped variants matrix $\boldsymbol{X} = [\boldsymbol{X_1}\ \boldsymbol{X_2} \dots\ \boldsymbol{X_n}]^{\boldsymbol{T}}$ and a covariates matrix $\boldsymbol{C} = [\boldsymbol{C_1}\ \boldsymbol{C_2} \dots\ \boldsymbol{C_n}]^{\boldsymbol{T}}$, the phenotypes vector $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$ follows a multivariate normal distribution:

$$\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{C} \sim \mathrm{N}(\boldsymbol{C\beta}, \boldsymbol{G}(\boldsymbol{X}; \boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}). \tag{1}$$

Here, $\boldsymbol{\beta}$ is a vector of covariate coefficients (denoted as *fixed effects*); $\boldsymbol{I}$ is the $n \times n$ identity matrix; σ_e^2 is the variance of the environmental effect; and $\boldsymbol{G(X;\theta)}$ is an $n \times n$ matrix encoding genotypic covariance. This covariance is governed by the parameter vector $\boldsymbol{\theta}$, whose dimension depends on the kernel type used to represent $\boldsymbol{G(X;\theta)}$. Kernelization amounts to defining a functional form for $\boldsymbol{G(X;\theta)}$, as detailed below.

Given a sample of individuals with an observed phenotype vector $\boldsymbol{y}$, the posterior (conditional) phenotype distribution for a tested individual with genotype $\boldsymbol{X}_*$ and covariates $\boldsymbol{C}_*$ is a normal distribution, $y_*|\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{y}, \boldsymbol{C}_*, \boldsymbol{X}_* \sim N(\mu_*, \sigma_*^2 + \sigma_e^2)$, whose parameters are given by (Rasmussen and Williams 2006):

$$\begin{aligned} \mu_* &= \boldsymbol{C}_*^T\boldsymbol{\beta} + \boldsymbol{g}_*^T(\boldsymbol{\theta})(\boldsymbol{G}(\boldsymbol{X}; \boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{C\beta}) \\ {\sigma_*}^2 &= g_{**}(\boldsymbol{\theta}) - \boldsymbol{g}_*^T(\boldsymbol{\theta})(\boldsymbol{G}(\boldsymbol{X}; \boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I})^{-1}\boldsymbol{g}_*(\boldsymbol{\theta}), \end{aligned} \tag{2}$$

where $\boldsymbol{g}_*(\boldsymbol{\theta})$ is a vector of genotypic covariances between the tested individual and all training individuals; and $g_{**}(\boldsymbol{\theta})$ is the prior genotypic variance of the tested individual. The quantities $\boldsymbol{g}_*(\boldsymbol{\theta})$ and $g_{**}(\boldsymbol{\theta})$ are computed according to the selected kernels, as detailed below. Parameter inference can be carried out efficiently via conjugate gradient ascent (Supplemental Material).

LMMs can be extended to handle binary phenotypes by assuming the existence of a latent normally distributed variable and treating the observed phenotype as a threshold indicator for the latent variable (Nickisch and Rasmussen 2008; Golan and Rosset 2014). However, parameter inference under such models is intractable (Supplemental Material).

LMM kernelization

The LMM covariance matrix $\boldsymbol{G}(\boldsymbol{X};\ \boldsymbol{\theta})$ encodes assumptions about the effects of the variants on the phenotype. Standard genomic best linear unbiased prediction (GBLUP) (Gianola 2013) uses a scaled kinship matrix, wherein the entry for individuals k and l is given by the normalized dot product of their (standardized and centered) genotype vectors, scaled by a constant θ:

$$\boldsymbol{G}(\boldsymbol{X}; \theta)_{k,l} = \theta\frac{1}{m}\sum_i X_k^i X_l^i = \theta\left(\frac{1}{\sqrt{m}}\boldsymbol{X}_k\right)^T\left(\frac{1}{\sqrt{m}}\boldsymbol{X}_l\right), \tag{3}$$

where m is the number of genotyped variants; X_k^i is the genotype of individual k at variant i; and $\boldsymbol{X}_k$ is the genotype vector of individual k. We refer to this covariance matrix as the linear kernel. The scaling parameter θ is often termed σ_g^2 in LMM literature. We now consider more complex types of kernels.

The recently proposed MultiBLUP (Speed and Balding 2014) model can be seen as a type of LMM kernelization, in which the covariance matrix is given by a sum of R region-specific covariance matrices, each using a linear kernel:

$$\boldsymbol{G}(\boldsymbol{X}; \boldsymbol{\theta})_{k,l} = \sum_{r=1}^{R}\theta_r\boldsymbol{G}_r(\boldsymbol{X}^r)_{k,l} = \sum_{r=1}^{R}\theta_r\left(\frac{1}{\sqrt{m_r}}\boldsymbol{X}_k^r\right)^T\left(\frac{1}{\sqrt{m_r}}\boldsymbol{X}_l^r\right). \tag{4}$$

Here, θ_r is the magnitude of the linear effect of region r; $\boldsymbol{X}^r$ is a $n \times m_r$ matrix of the m_r variants in a genomic region r; and $\boldsymbol{X}_k^r$, $\boldsymbol{X}_l^r$ are the genotype vectors of individuals k and l in region r, respectively. Equation 4 reflects the assumption that certain genomic regions can have larger effects on the phenotype than others.

The linear kernel encodes the assumption that variants affect phenotypes linearly. Interactions can be encoded via more elaborate kernels. For example, it is well known that pairwise multiplicative interactions can be encoded as follows (Henderson 1985; Su et al. 2012; Bloom et al. 2015):

$$\begin{aligned} \boldsymbol{G}(\boldsymbol{X}; \theta)_{k,l} &= \theta\frac{1}{m^2}\sum_i\sum_j(X_k^iX_k^j)(X_l^iX_l^j) \\ &= \theta\left(\left(\frac{1}{\sqrt{m}}\boldsymbol{X}_k\right)^T\left(\frac{1}{\sqrt{m}}\boldsymbol{X}_l\right)\right)^2. \end{aligned} \tag{5}$$

Equation 5 encodes the assumption that products of pairs of variants have a linear effect on the phenotype, and is known as a polynomial kernel of degree 2 (Poly2 kernel) in the machine learning literature. Equation 5 can also be written as $\boldsymbol{G}(\boldsymbol{X}; \theta)_{k,l} = \theta\pi(\boldsymbol{X}_k)^T\pi(\boldsymbol{X}_l)$, where $\pi(\boldsymbol{X}_k)$ transforms the vector $\boldsymbol{X}_k$ into a new vector with an entry for the product of every pair of variants. However, the transformation π does not need to be computed explicitly, because $\boldsymbol{G}(\boldsymbol{X}; \theta)_{k,l}$ can be more simply computed as $(\theta/m^2)[(\boldsymbol{X}_k)^T\boldsymbol{X}_l]^2$. It turns out that every kernel can be computed implicitly in this manner, as stated by the Mercer theorem that, in informal terms, states that every symmetric positive definite matrix $\boldsymbol{G}(\boldsymbol{X};\boldsymbol{\theta})$ corresponds to a dot product between transformed (possibly infinite-dimensional) genotype matrices $\pi(\boldsymbol{X})$ (Rasmussen and Williams 2006). This allows us to define rich interaction patterns via kernels without explicitly computing the corresponding genotype transformations.

Evaluated kernels

In this work, we consider two kernels corresponding to infinite-dimensional transformations: The first is the well-known radial basis function (RBF) kernel, which is very popular in both machine learning literature and statistical genetics (Gianola et al. 2014; Morota and Gianola 2014) given by

$$\boldsymbol{G}(\boldsymbol{X}; \theta_1, \theta_2)_{k,l} = \theta_1\exp\left(-\frac{1}{2\theta_2 m}\sum_i(X_k^i - X_l^i)^2\right). \tag{6}$$

The RBF kernel is governed by the magnitude parameter θ_1 and the bandwidth parameter θ_2. The magnitude parameter determines the magnitude of the variance explained by this kernel, similarly to the parameter θ in Equation 3. The bandwidth parameter determines the rate of decay of genetic covariance. It can be shown that the RBF kernel generalizes the polynomial kernel, because its underlying transformation $\pi(\boldsymbol{X}_k)$ includes an infinite number of entries, with one entry for every possible polynomial of the vector $\boldsymbol{X}_k$ (Supplemental Material).

The second kernel we consider is the saturating pathways kernel, also known as the neutral network kernel in the machine

learning literature, given by

$$\boldsymbol{G}(\boldsymbol{X};\theta_1,\theta_2)_{k,l}=\frac{\theta_1}{2\pi}\sin^{-1}\left(\frac{\frac{1}{m}(\boldsymbol{X}_k)^T\boldsymbol{X}_l}{\sqrt{\left(\theta_2+\sum_i(X_k^i)^2/m\right)\left(\theta_2+\sum_i(X_l^i)^2/m\right)}}\right). \quad (7)$$

The SP kernel is determined by the magnitude parameter θ_1 and the bandwidth parameter θ_2 that serve roles similar to those in the RBF kernel. It can be shown that the SP kernel corresponds to a neural network with an infinite number of hidden units and a Gaussian prior on the network weights, or alternatively, to a model with an infinite number of interacting biological pathways, as defined in Zuk et al. (2012) (Supplemental Material).

MKLMM-Adapt

MKLMM-Adapt is an MKLMM formulation whose kernel consists of a sum of region-specific kernels, in which the regions and kernels are automatically selected in a data-driven manner. Each region uses a weighted combination of a linear and at most one nonlinear kernel. The resulting model can capture both diverse interaction patterns and heterogeneous effect-size distributions across different genomic regions.

Although models using more region-specific kernels can potentially express richer interaction patterns, they run into risk of overfitting the training data. Our proposed model selection strategy evaluates several models of increasing complexity via a fivefold cross validation and selects the one demonstrating the best out of sample prediction capability. The model selection details follow, and a full algorithmic description is provided in the Supplemental Material.

We begin by dividing the genome into a ranked list of regions, similarly to the approach of adaptive MultiBLUP (AMB) (Speed and Balding 2014). The genome is first divided into many small overlapping subregions spanning ~75 kb. Every subregion is evaluated according to the likelihood obtained when constructing a linear kernel using only its variants. Afterward, subregions whose likelihood is among the bottom 95% are discarded, and every consecutive range of nondiscarded subregions is merged into a region. The regions are ranked in descending order, according to the maximal likelihood obtained by a subregion in each region.

After obtaining a ranked list of regions, we evaluate the performance of models with increasing complexity $M^{(0)}, M^{(1)}, \ldots, M^{(B)}$ via a forward selection scheme, where B is a user-defined parameter. Larger values of B enable evaluating more expressive models at the price of a greater computational cost. All experiments in this paper used $B = 9$, because no improvement was observed for larger values.

Each model $M^{(i)}$ uses a kernel composed of a sum of region-specific kernels for regions 0, …,i, in which region 0 corresponds to a genome-wide region spanning all genotyped variants. Every region-specific kernel is a weighted combination of a linear kernel and at most one out of three additional kernel types: a Poly2 kernel, an RBF kernel, and an SP kernel. At each step i, the candidate models for region i are ranked according to a likelihood ratio test in which the null hypothesis uses only a linear kernel for region i, and the model with the smallest P-value is selected if it is significant at the 5% level after accounting for multiple hypothesis testing. The selected kernel type is also used in models $M^{(i+1)}, \ldots, M^{(B)}$, but its parameters are re-estimated in subsequent models.

The computational complexity of this procedure scales cubically with the sample size, as in standard LMMs. The space complexity scales quadratically with the sample size, owing to the size of the kernel matrices. In our implementation, at most four covariance matrices need to be held in memory simultaneously (two genome-wide kernels and two region-specific kernels), because region-specific kernels can be computed efficiently only when needed and then be discarded.

Privacy-preserving phenotype prediction

MKLMM can perform genetic similarity–based prediction without having to store the genotypes and phenotypes of the training sample. This can be accomplished by using the Bayesian interpretation of MKLMM, which shows that MKLMM is equivalent to a linear regression model wherein all effect sizes have an independent and identically distributed (iid) normal prior distribution (Supplemental Material). It is therefore possible to perform phenotype prediction by storing the parameters of the posterior distribution of the effect sizes, thus alleviating the need to store genotypes and phenotypes of training individuals (Supplemental Material).

Simulations and experiments procedure

We evaluated the performance of MKLMM and AMB on synthetic and real data sets using a fivefold cross validation procedure. In each fold, 80% of the individuals were used for training, and the remaining 20% were used for evaluating prediction performance. The evaluated measures included RMSE, out of sample log likelihood (OOS LL), and Pearson correlation in experiments with quantitative phenotypes, and the area under the receiving operator characteristic curve (AUC) in experiments with binary phenotypes. The AUC was computed by iterating over all test individuals and computing the false and true positive rates obtained when treating the estimated posterior mean of the phenotype of each individual as the affection threshold, such that individuals with estimated posterior mean greater than this value are considered cases, and the others are considered controls. The phenotype was standardized to have a unit variance when it was quantitative to obtain comparable results in different experiments. Each result was computed separately under each fold and then averaged.

All methods were evaluated with varying numbers of kernels. In each fold, one-third of the 20% held-out individuals were used to select the best number of kernels, and the evaluation measure was then computed using the remaining two-thirds. We note that this kernel selection procedure is different from the one used by the authors of AMB (Speed and Balding 2014), that was more conservative and would thus place AMB at a disadvantage relative to the other methods. The statistical significance of the results was evaluated via permutation testing, which takes this procedure into account (Supplemental Material).

Synthetic data sets were created by generating synthetic phenotypes based on real genotypes from Chromosome 1 of 2801 individuals from the WTCCC national blood service cohort. Unless otherwise stated, 25% of the phenotypic variance was explained by either two or six randomly selected genomic regions with a mean length of 75 kb, another 25% was explained by a polygenic term spanning the entire chromosome, and the remaining 50% was explained by an iid normally distributed environmental effect.

All regions (including the chromosome-wide region) exerted both an additive and a nonadditive interaction effect on the phenotype. Nonlinear effects consisted of either a groupwise multiplicative effect, which generalizes pairwise multiplicative effects to also include higher-order interactions, or a saturating effect, which bounds the magnitude of linear effects (Supplemental Material). Each simulated interaction consisted of randomly selected variants in the same region, but we note that one of the regions was a

chromosome-wide region spanning all variants. Ten phenotypes were generated for each unique combination of settings. A detailed description is given in the Supplemental Material.

To evaluate the methods in the presence of binary phenotypes, we created quantitative phenotypes as described above and then dichotomized them at the empirical median. To evaluate performance for ascertained binary phenotypes, we generated synthetic ascertained genotypes and phenotypes by first generating a large number of individuals and then selecting the ones with the most extreme phenotypes as cases. Parameter inference and prediction for binary phenotypes were carried out by treating the phenotype as a normally distributed variable, as is commonly practiced (Zhou et al. 2013; Speed and Balding 2014; Moser et al. 2015). The Supplemental Material contains a discussion of alternative approaches.

Software and code availability

The MKLMM source code is available as Supplemental Material. Updated versions will be available at https://github.com/omerwe/MKLMM. MKLMM is computationally efficient. On a Linux workstation with a single 2-GHz CPU, it can fit an MKLMM model for a data set of 2800 individuals with five linear and five SP kernels in ~1 h.

Acknowledgments

This work was supported by Grant 1487/12 from the Israel Science Foundation. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

References

Akdemir D, Jannink JL. 2015. Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics* **199:** 857–871.

Bickel RD, Kopp A, Nuzhdin SV. 2011. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. *PLoS Genet* **7:** e1001275.

Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. 2015. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nat Commun* **6:** 8712.

Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, Zheng HF, Richards JB, Glass D, Small KS, Durbin R, et al. 2015. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet* **47:** 88–91.

Cordell HJ. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11:** 2463–2468.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10:** 392–404.

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, et al. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 (Bethesda)* **3:** 1903–1926.

Da Y, Wang C, Wang S, Hu G. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* **9:** e87666.

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193:** 327–345.

Dimas AS, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavaré S, Dermitzakis ET. 2008. Modifier effects between regulatory and protein-coding variation. *PLoS Genet* **4:** e1000244.

Dove ES, Joly Y, Tassé AM, Public Population Project in Genomics and Society (P3G) International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, Knoppers BM. 2015. Genomic cloud computing: legal and ethical points to consider. *Eur J Hum Genet* **23:** 1271–1278.

Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy. *Nat Rev Genet* **15:** 409–421.

Ferraudo GM, Perecin D. 2014. Mixed model, AMMI and Eberhart-Russel comparison via simulation on genotype × environment interaction study in sugarcane. *Appl Math* **5:** 2107–2119.

Fusi N, Lippert C, Lawrence ND, Stegle O. 2014. Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat Commun* **5:** 4890.

Gianola D. 2013. Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* **194:** 573–596.

Gianola D, Morota G, Crossa J. 2014. Genome-enabled prediction of complex traits with kernel methods: What have we learned? In *Proceedings, 10th World Congress of Genetics Applied to Livestock Production*, Vancouver, BC, Canada.

Golan D, Rosset S. 2014. Effective genetic-risk prediction using mixed models. *Am J Hum Genet* **95:** 383–393.

Golan D, Lander ES, Rosset S. 2014. Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci* **111:** E5272–E5281.

Gönen M. 2012. Bayesian efficient multiple kernel learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (ed. Langford J, Pineau J), Edinburgh, UK.

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* **339:** 321–324.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* **12:** 186.

Haig D. 2011. Does heritability hide in epistasis between linked SNPs? *Eur J Hum Genet* **19:** 123.

Hastie T, Tibshirani R, Friedman J. 2009. *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. Springer, New York.

Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, Yang J, Gibson G, Martin NG, Metspalu A, et al. 2014. Detection and replication of epistasis influencing transcription in humans. *Nature* **508:** 249–253.

Henderson C. 1985. Best linear unbiased prediction of nonadditive genetic merits. *J Anim Sci* **60:** 111–117.

Im HK, Gamazon ER, Nicolae DL, Cox NJ. 2012. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am J Hum Genet* **90:** 591–598.

Jiang Y, Reif JC. 2015. Modeling epistasis in genomic selection. *Genetics* **201:** 759–768.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178:** 1709–1723.

Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. 2011. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet* **89:** 459–463.

Lázaro-Gredilla M, Titsias MK. 2011. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in neural information processing systems* (ed. Shawe-Taylor J, et al.), pp. 2339–2347, Granada, Spain.

Lehner B. 2011. Molecular mechanisms of epistasis within and between genes. *Trends Genet* **27:** 323–331.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nat Methods* **8:** 833–835.

Lippert C, Xiang J, Horta D, Widmer C, Kadie C, Heckerman D, Listgarten J. 2014. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* **30:** 3206–3214.

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

Liu D, Ghosh D, Lin X. 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics* **9:** 292.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157:** 1819–1829.

Moore JH, Williams SM. 2009. Epistasis and its implications for personal genetics. *Am J Hum Genet* **85:** 309–320.

Morota G, Gianola D. 2014. Kernel-based whole-genome prediction of complex traits: a review. *Front Genet* **5:** 363.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet* **11:** e1004969.

Muñoz PR, Resende MF Jr, Gezan SA, Resende MD, de Los Campos G, Kirst M, Huber D, Peter GF. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* **198:** 1759–1768.

Nickisch H, Rasmussen CE. 2008. Approximations for binary Gaussian process classification. *J Mach Learn Res* **9:** 2035–2078.

Nishio M, Satoh M. 2014. Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One* **9:** e85792.

Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H. 2011. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics* **188:** 695–708.

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S. 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* **2:** 1595–1605.

Powell JE, Henders AK, McRae AF, Kim J, Hemani G, Martin NG, Dermitzakis ET, Gibson G, Montgomery GW, Visscher PM. 2013. Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. *PLoS Genet* **9:** e1003502.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11:** 459–463.

Rasmussen CE, Williams CKI. 2006. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

Rodriguez LL, Brooks LD, Greenberg JH, Green ED. 2013. Research ethics. The complexities of genomic identifiability. *Science* **339:** 275–276.

Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink JL, Sorrells M. 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. *Plant Genome J* **5:** 51.

Schrodi SJ, Mukherjee S, Shan Y, Tromp G, Sninsky JJ, Callear AP, Carter TC, Ye Z, Haines JL, Brilliant MH, et al. 2014. Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future. *Front Genet* **5:** 162.

Speed D, Balding DJ. 2014. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res* **24:** 1550–1557.

Su G, Christensen OF, Ostersen T, Henryon M, Lund MS. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* **7:** e45293.

Tusell L, Pérez-Rodríguez P, Forni S, Gianola D. 2014. Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. *J Anim Breed Genet* **131:** 105–115.

UK IBD Genetics Consortium, Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, et al. 2009. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nat Genet* **41:** 1330–1334.

Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JN, Mott R, Flint J. 2006. Genetic and environmental effects on complex traits in mice. *Genetics* **174:** 959–984.

Vitezica ZG, Varona L, Legarra A. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195:** 1223–1230.

Wang D, Zhu J, Li Z, Paterson A. 1999. Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches. *Theor Appl Genet* **99:** 1255–1264.

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:** 661–678.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89:** 82–93.

Yang J, Zhu J, Williams RW. 2007. Mapping the genetic architecture of complex traits in experimental populations. *Bioinformatics* **23:** 1527–1536.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42:** 565–569.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38:** 203–208.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42:** 355–360.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44:** 821–824.

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

Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, LifeLines Cohort Study, Esko T. 2015. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96:** 377–385.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci* **109:** 1193–1198.

Received November 16, 2015; accepted in revised form May 2, 2016.