COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 10, No. 2, pp. 99-114, 2010
© 2010 International Press
003

PREDICTION OF PHENOTYPE INFORMATION FROM GENOTYPE DATA*

NIR YOSEF†, JENS GRAMM‡, QIAN-FEI WANG§, WILLIAM S. NOBLE¶,
RICHARD M. KARP‖, AND RODED SHARAN†**

Abstract. The dissection of complex diseases is one of the greatest challenges of human genetics with important clinical and scientific applications. Traditionally, associations were sought between single genetic markers and disease. The availability of large scale SNP data makes it possible, for the first time, to study the predictive power of genotypes and haplotypes with respect to phenotype data. Here we present a novel method for predicting phenotype information from genotype data. The method is based on a support vector machine that employs new kernel functions for the similarity between genotypes or their underlying haplotypes. We demonstrate our approach on SNP data for the apolipoprotein gene cluster in baboons, predicting plasma lipid levels with significant success rates, and identifying associations that were not detected using extant approaches.

Keywords: Machine learning (Computing Methodologies–Artificial Intelligence–Learning), Parameter learning (Computing Methodologies–Artificial Intelligence–Learning), Classifier design and evaluation (Computing Methodologies–Pattern Recognition–Design Methodology), Biology and genetics (Computer Applications–Life and Medical Sciences).

1. Introduction.

The study of single nucleotide polymorphisms (SNPs) promises to revolutionize the way medical conditions are diagnosed and treated. Key to their successful application is the development of methods that can infer phenotype information from genotype data. Historically, correlations between single genetic markers and phenotype data were exploited to zoom in on regions in the genome that are related to specific traits [16, 31]. Recently, several studies have demonstrated the utility of genotype and haplotype information to mapping complex human traits. Liu *et al.* [14] have developed a Bayesian framework for disease mapping which relies on modeling the evolution of the population haplotypes from a set of founders through mutation and recombination. Rannala & Reeve [25] used a coalescence-based model and an MCMC method to integrate over the unknown gene genealogy and gene coalescence times. Greenspan & Geiger [10] devised a mapping strategy that is based on identifying haplotype blocks and computing a posterior distribution for the association of the disease SNP with each of the blocks. Yosef *et al.* [34] used mining of

*Dedicated to Michael Waterman on the occasion of his 67th birthday.

†Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel.
‡WSI für Informatik, Universität Tübingen, Sand 13, 72076 Tübingen, Germany.
§Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
¶Department of Genome Sciences, Department of Computer Science and Engineering University of Washington Seattle, WA, USA.
‖Computer Science Division, University of California at Berkeley, and International Computer Science Institute, 1947 Center St., Berkeley, CA, 94704, USA.
**To whom correspondence should be addressed. Email: roded@post.tau.ac.il. Tel: 972-3-640-7139. Fax: 972-3-640-9357.

bipartite graphs relating individuals to genotypes in order to detect genotype patterns that discriminate between different phenotypic groups.

Several authors have studied the related diagnostic problem, in which one tries to classify a genotype sample to one of several phenotype classes. The idea is to apply a method that learns from training data a "signature" for each phenotype class, which is subsequently used to determine if a new sample belongs to this class. In [4] the drug-resistance nature of HIV-1 variants was predicted based on their genotypes using decision trees. In [3, 5], a number of regression and classification methods (including decision trees, neural networks and support vector regression) were used to predict the fold change in susceptibility of HIV-1 variants to drugs and co-receptor usage based on their amino acid sequences. Rhee *et al.* [26] used a similar repertoire of methods to relate HIV-1 protease and reverse transcriptase mutations to in vitro susceptibility to antiretroviral drugs. Support vector machine (SVM) [6] analysis based on sets of SNP loci was previously performed to predict susceptibility to breast cancer [13, 27] and coronary heart disease [33], outperforming other classification methods such as decision trees [13, 27].

Here we develop a kernel-based approach for predicting phenotype classes from genotype data. We first introduce a novel kernel function for comparing two genotypes. This kernel was designed to explicitly account for the lack of information which inherently exists in the available genotype data, namely, the different possible pairings between compared heterozygous loci. We then present a novel kernel for comparing haplotype data. Such data is not readily available, but can be derived using computational methods (e.g., [29]) or using pedigree information. We demonstrate our approach on SNP data from [32] on the apolipoprotein gene cluster (*APOA1/C3/A4/A5*) in baboons. This 68 Kb region is orthologous to the human apolipoprotein gene cluster, which is known to be associated with plasma lipid levels, such as HDL-cholesterol and triglyceride concentrations [2, 11, 20]. Our analysis suggests a set of novel associations of genotypes with plasma lipid outcomes which were missed by previous approaches.

Our contribution is three-fold. First, we devise novel kernels for genotype data. In particular, we provide the first kernel that accounts for the underlying haplotype data; this kernel is shown to outperform previous approaches to the problem of predicting phenotypes from genotypes. Second, we apply our method to predict plasma lipid levels in baboons, identifying a set of novel associations. Importantly, the global nature of our approach, which is able to analyze the entire SNP set rather than small subsets at a time, allows us to identify new associations that were missed by a previous analysis of [32]. Third, we demonstrate that a classification framework can be successfully used to analyze quantitative (rather than binary) phenotypes.
Supporting information, including data for download, is available at
http://www.cs.tau.ac.il/∼roded/GP/

2. Results. We analyzed a superset of the genotype data reported in [32]. The original data set contains the genotypes of 621 baboons, over 17 SNPs. The SNPs are located in a gene cluster that is orthologous to the apolipoprotein gene cluster (*APOA1/C2/A4/A5*) in human (http://pga.lbl.gov/SNP/Baboon/APOA1C3A4A5.html). Notably these SNPs were chosen as representatives for the gene cluster and provide a non-redundant representation of that region [32].

The genotyped baboons underwent three different diets [23]—CHOW (basal, low in fat and cholesterol), LCHF (low cholesterol, high fat) and HCHF (high cholesterol, high fat)—and their metabolite levels were measured. In total, we had quantitative information on 10 phenotypes in diet 1, and 9 phenotypes in diets 2 and 3 (Methods). In addition, for each baboon we had covariate information on its sex, age and weight.

Due to inconclusive readings, the genotype vectors contained a total of 843 missing entries. We used the available pedigree data to complete those entries, under the Mendelian assumption (also used in [32]) that no recombinations or mutations have occurred in this region within each pedigree. In total, 154 missing entries were completed in this way, and 219 additional entries were partially resolved, i.e., the state of one of the alleles of the corresponding SNP was derived.

Our goal was to develop methods for classifying the baboons to phenotype classes based on their genotype data. To this end, we partitioned the baboons into three classes with respect to each of the phenotypes: *positives*, corresponding to those baboons attaining the top 15% values for the given phenotype; *negatives*, corresponding to those baboons attaining the bottom 15% values for the given phenotype; and the rest of the baboons, which were considered to be undecided. We experimented with another discretization method for the phenotype data and got similar results in the subsequent analysis (Methods and Table S1).

To limit the influence of relatedness among individuals on our analysis, the classification was based on independent pedigrees. Specifically, we chose a subset of 591 individuals (out of 621) that can be decomposed into 3 independent pedigrees such that no two individuals from different pedigrees have a common ancestor (Supporting information). We trained the classification procedure with baboons from all-except-one pedigrees and tested the procedure's success in predicting phenotypes of baboons from the held-out pedigree.

2.1. Classification by genotypes. Our first set of experiments was aimed at evaluating the predictive power of the genotypes (after the completion of missing entries described above) with respect to the phenotypes. We also examined the influence of the covariate information on the classification process. The classification was performed using the SVM paradigm [6]. The main challenge in applying the SVM algorithm to the data was the development of similarity functions, or *kernels*, to compare the attribute vectors (containing genotype and/or covariate data) of pairs

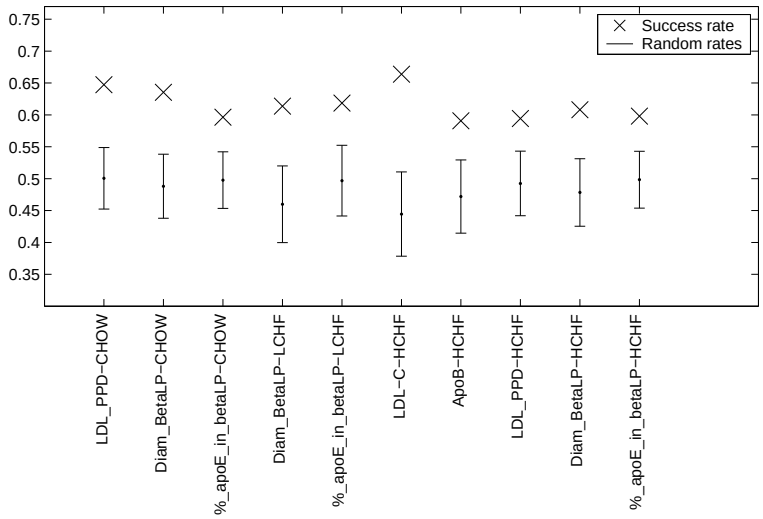

FIG. 1. *Genotype-based classification results using the identity kernel. Shown are the success rates of phenotypes with an empirical p-value lower than 0.05. The error bars indicate the outcome of the permutation tests, denoting the mean and standard deviation values obtained over 200 runs. See Table 1 for phenotype abbreviations.*

of baboons.

First, we applied our classification scheme to the genotype data only. We compared two types of kernels designed for such data. The first kernel, which we call the *identity kernel*, is similar to the one used by [27] for prediction of susceptibility to breast cancer. This kernel computes the maximum number of alleles shared by the two compared individuals. The second kernel, termed the *expectation kernel*, is a novel kernel that accounts for the multiple options for pairing two haplotypes. This kernel computes the expected similarity between the haplotypes corresponding to the two compared genotypes, where the expectation is taken over the space of all possible phasings of the genotype data (Methods).

We embedded these two kernels in a Gaussian function and applied a grid search, varying the width of the Gaussian (reflecting the assumed level of independence between the different SNPs, where a large width resembles a linear kernel which treats each SNP independently) and the regularization coefficient (reflecting the trade off between the empirical error and the width of the classification margin).

We evaluated the different kernels by computing their *success rate*, defined as the average between the percent of positives that were classified correctly and the percent of negatives that were classified correctly. The statistical significance of the success rates was evaluated by an empirical p-value which is based on permutation tests (Methods). Figures 1 and 2 depict the statistically significant cases for the two kernels (empirical p-value < 0.05). The same set of results is also given in Table 1, presenting phenotypes in which at least one of the methods performed well.

TABLE 1

Success rates in predicting plasma lipid levels. The first three columns show the success rate (left) and empirical p-values (right) of our three prediction schemes: identity kernel (genotype data), expectation kernel (genotype data), and all-match kernel (haplotype data). The forth column shows the success rates of a naïve Bayes classifier. The last column presents the sizes of the haplotype blocks for which a significant association was found using the QTDT method of Wang et al. *[32];* NA *denotes phenotypes which were not studied in [32]. The best performance for each phenotype is marked in bold. Cases in which no significant association was found or where the success rate (where applicable) was lower than 55% are marked with "–". Abbreviations: LDL-C (LDL-cholesterol), HDL-C (HDL-cholesterol), ApoB (Apolipoprotein B), ApoE (Apolipoprotein E), LDL_PPD (LDL peak particle diameter), Diam_BetaLP (median diameter of beta lipoproteins), Diam_HDL (diameter of HDL), %_apoE_in_BetaLP (percentage of apoE in beta lipoproteins), TG (triglyceride), TSC (total serum cholesterol).*

Phenotype	Genotype-Identity	Genotype-Expectation	Haplotype	Naive Bayes	Wang *et al.*
TG-CHOW	–	–	–	–	3
HDL-C-CHOW	–	–	–	–	1,2
LDL-C-CHOW	–	–	**0.609, 0.034**	–	–
LDL_PPD-CHOW	**0.647, 0.005**	0.636, 0.005	0.624, 0.014	0.58	–
%_apoE_in_betaLP-CHOW	0.596, 0.020	–	–	–	–
Diam_BetaLP-CHOW	0.635, 0.005	0.647, 0.005	**0.676, 0.007**	0.633	–
LDL_PPD-LCHF	–	0.635, 0.01	**0.651, 0.014**	0.596	–
Diam_BetaLP-LCHF	**0.614, 0.01**	–	–	0.57	–
%_apoE_in_betaLP-LCHF	0.618, 0.02	**0.626, 0.01**	–	0.56	–
TSC-HCHF	–	–	**0.645, 0.007**	–	NA
LDL-C-HCHF	**0.664, 0.005**	0.651, 0.005	0.621, 0.041	–	–
ApoB-HCHF	**0.591, 0.015**	–	–	–	–
LDL_PPD-HCHF	0.594, 0.027	**0.617, 0.01**	0.600, 0.020	0.59	–
Diam_BetaLP-HCHF	0.608, 0.01	0.575, 0.025	**0.634, 0.007**	0.628	–
%_apoE_in_betaLP-HCHF	0.598, 0.020	0.591, 0.015	**0.602, 0.020**	**0.602**	–

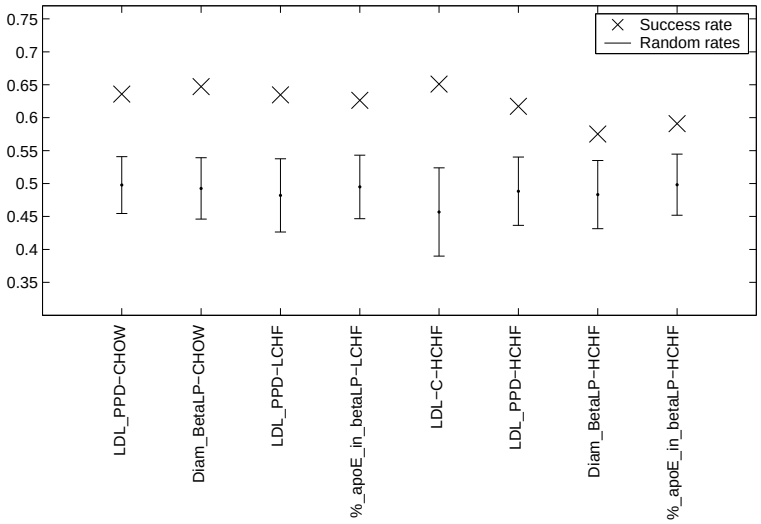

FIG. 2. *Genotype-based classification results using the expectation kernel. Displayed results are as in Figure 1.*

An interesting view into the works of the SVM can be gained from the effect of the Gaussian width on the accuracy of the prediction. In a narrow Gaussian the SVM will typically construct a highly non-linear decision boundary and therefore rely not only on genotypes at individual SNPs, but also on interactions of alleles at different loci. Conversely, in a wide Gaussian the SVM is more likely to resemble a

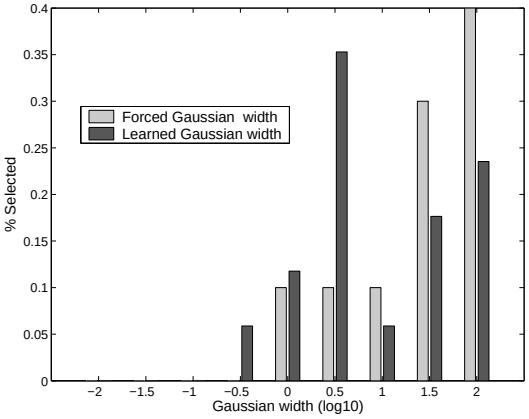

FIG. 3. *The effect of Gaussian width on predicting phenotypes for the different pedigrees. The dark gray bars indicates the percentage of cases each Gaussian width value has been selected during our grid search procedure. The light gray bars indicate the number of cases in which fixing the corresponding width value yielded the best prediction accuracy.*

linear discriminator and rely more heavily on the individual SNPs. Figure 3 shows the distribution of width values selected during our grid search with the expectation kernel (dark gray bars). Evidently, there is an obvious preference for less complex decision boundaries, selecting medium or wide Gaussian values. To validate this we repeated the prediction steps, this time fixing the width value and optimizing only over the regularization parameter. For each experiment (namely, for a given phenotype and a left out pedigree) we record the width value that had the best success rate (also requiring it to be over 55%). The resulting frequencies are displayed in Figure 3 (light gray bars).

Next, we constructed an SVM predictor based on the covariate data (sex, age and weight) alone and on the combined genotype and covariate data (using the expectation kernel). The results are summarized in Figure 4. Evidently, the covariate data are highly predictive for most phenotypes. This predictive ability however is of less interest to us, as the goal of this study is to track down genotype-phenotype relations. In addition, given a combined covariate-genotype data, it is likely that the SVM predictor will tend to concentrate on the more predictive, covariate features, and ignore the genotype data. For these reasons, we limit the discussion in this paper only to genotype data.

For each of the phenotypes that exhibited significant correlation to the SNP data when using the expectation kernel (empirical p-value < 0.05), we also searched for the most predictive subset of SNPs using an SVM-based feature selection method [12]. These results are summarized in the Supporting information.

2.2. Classification by haplotypes. The second set of tests was aimed at evaluating whether the predictions could be improved by using haplotype data in the

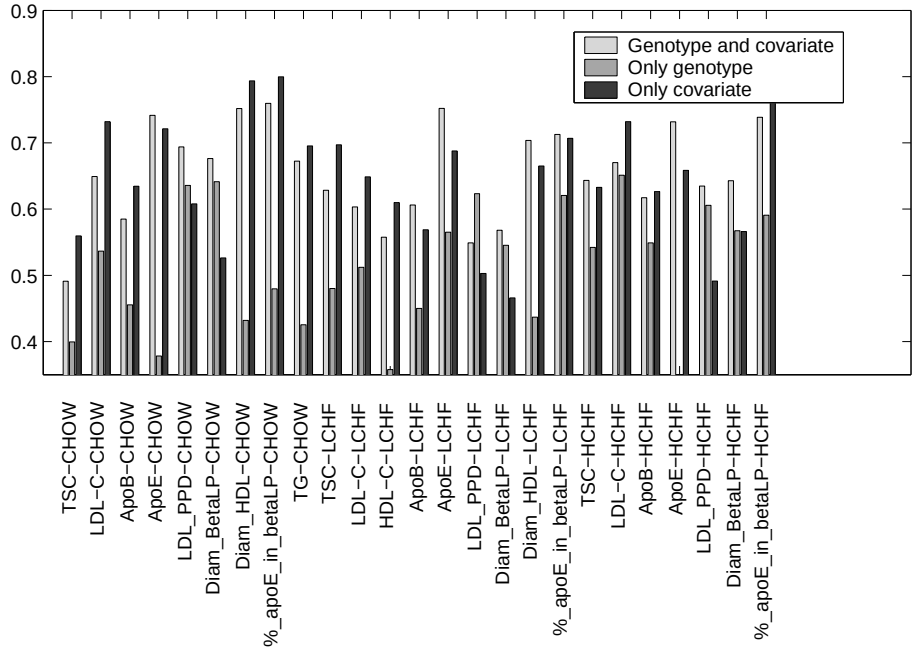

FIG. 4. *Comparison of expectation kernel classification results using genotypes, covariates or both. Shown are the success rates of phenotypes for which at least one of the classifiers scored higher than 55%.*

classification process. Since haplotype data for the genotyped baboons was not available, we inferred the underlying haplotypes, separately for each pedigree, using the popular PHASE method [29].

Given two complete haplotype vectors, each representing two parental haplotypes, the only degree of freedom is which of the two possible pairings to use. Given such a pairing, we can use a simple extension to the genotype kernels (e.g., by concatenating the two parental haplotypes in each vector) as our haplotype kernel. Since the pairing is not available, an ideal approach would be to consider the pairing which yields the best match. However, the resulting similarity function is not known to be a kernel. Hence, we devised an alternative *all-match kernel*. This kernel function is based on the number of matches between the haplotype pairs under the two pairings of haplotypes in one genotype to haplotypes in the other genotype (Methods).

We applied a similar assay to that of the genotype data – the all-match kernel was embedded in a Gaussian function and the best parameter set (width of Gaussian, regularization constant and an additional parameter, specific for the all-match kernel function) were determined using a grid search. The statistical significance of the results was then measured using permutation tests. Figure 5 depicts the statistically significant cases (empirical p-value < 0.05).

Table 1 provides a comparison of the success rates attained by the haplotype-

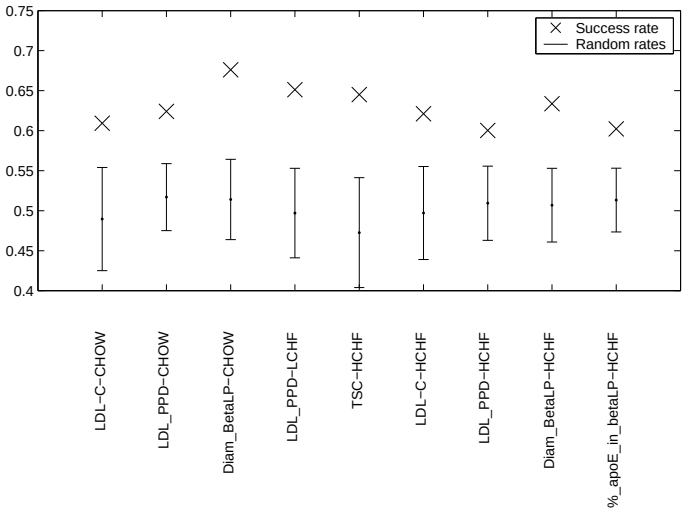

FIG. 5. *Haplotype-based classification results using the all-match kernel. The error bars indicate the outcome of the permutation tests, denoting the mean and standard deviation values obtained over 150 runs. Abbreviations are as in Table 1.*

based kernel, the genotype-based kernels, a boosted naïve Bayes classifier [9], and the QTDT method of Wang *et al.* [32]. Overall, the SVM based methods outperform the latter two methods, finding a larger number of significant associations.

Comparing the SVM based methods we see an overall better performance of the haplotype-based kernel. Specifically, we see that the number of phenotypes (possibly with a different diet) predictable by the genotype-based kernels is equal to the number of phenotypes predictable by the haplotype-based kernel. However, while each of the genotype kernels performed best on 4 or less phenotypes, the haplotype kernel performed best on 6 phenotypes, achieving higher and more significant success rates. In addition, the haplotype-based kernel was the only one to identify LDL-cholesterol particle sizes as a predictable phenotype in all three diets. This association is indeed supported by prior biological knowledge (see below). Notably, we get consistent results, showing the superiority of the haplotype kernel over the genotype based kernels when using the alternative phenotype discretization, presented in SI table 1.

3. Discussion. We presented a methodology for predicting phenotype information from genotype or haplotype data, and for pinpointing genetic markers that are highly predictive of the phenotype data. An application of the approach to SNP data for the apolipoprotein gene cluster in baboons revealed a rich set of associations between the genetic and phenotypic data. The associations that were found in at least two of the three diets in either the genotype-based or the haplotype-based analyses include: Percentage of apoE in beta lipoproteins, LDL-cholesterol concentration, LDL particle size and beta lipoprotein particle size. We note that related phenotypes, such

as apoE in beta lipoprotein/beta lipoprotein particle size and LDL-cholesterol/LDL particle size, came out in our analysis, although each phenotype was treated independently, indicating that our prediction method is robust.

Two of these associations are supported by previous studies in human, serving as a validation of our approach. The apolipoprotein gene cluster was shown to be associated with LDL-cholesterol concentration in [8] and with LDL particle sizes in [15]. The association with beta lipoprotein particle size requires further biological validation.

Wang *et al.* [32] have previously used QTDT (Quantitative Transmission Disequilibrium Tests [1]) to analyze a subset of the reported data. QTDT is a popular family-based statistical method for association tests. It evaluates the association between genotypes and quantitative traits based on a linear model of association and information on nuclear families. Wang *et al.* tested the association between the measured phenotypes and single markers, as well as two-, three- or four-locus haplotypes. In contrast, our SVM-based method focuses on predicting phenotype classes. We note the trade off involved in using a classification method with such discretization of the data: on one hand, we might lose information, while on the other hand a discretization of phenotype values into two classes largely contributes to the robustness against measurement problems. In addition, our tests are based on the entire genotype or haplotype data, and can pinpoint complex associations exhibited by multiple loci, while QTDT requires a prior selection of loci to be tested for association. Our method also has the advantage of not imposing a statistical model on the data as QTDT does. These differences allow us to discover many associations that were missed by QTDT (Table 1). One such example, is the association of LDL cholesterol (LDL-C) levels with the apolipoprotein cluster – a finding supported by previous studies in human [8].

We also compared our method to a Naïve Bayes classifier with boosting [9], which assumes that the SNP values of an individual are independent given the phenotype. As Table 1 shows, the SVM methods clearly outperform the Naïve Bayes approach. A reasonable explanation would be the ability of the SVM to construct non-linear decision boundaries, drawing its predictive power also from interactions of alleles at different loci rather than from individual SNPs alone.

The A1/A5 cluster was shown to be associated with triglyceride level and with HDL-cholesterol concentration by Wang *et al.*. While the association to triglyceride level was not significant in our analysis, we did detect a significant association to the LDL size phenotype, which has been shown to be strongly correlated with triglyceride level, as many groups have reported [17,22,28,30]. The association to HDL-cholesterol was only marginally significant in our analysis ($p \sim 0.058$).

It should be noted that errors in genotype data can play an influencing factor in the results of the ensuing analyzes [21]. However, the loci analyzed in this paper were subjected to further evaluation, ensuring high data quality. 24 baboons were randomly selected from the study population and measured at all SNP locations

using two independent experimental techniques [32]. All SNPs showed at least 95% consistency between the two methods.

Conducting genetic studies of complex phenotypes in baboons rather than humans carries many advantages. Here, the researcher has the option for stringent control over crucial factors such as mating patterns and environmental conditions. Controlling these factors can promise to reduce the impact of environmental side effects and ensure a more accurate detection of influencing genes. One assumption which is generally made in such studies is that polymorphisms in orthologous genes in the two organisms have similar effects on phenotypes. Our results, combined with those of Wang *et al.*, provide further support for this hypothesis, mapping novel associations between the apolipoprotein gene cluster in baboons with plasma lipid levels; findings which match those previously reported in human.

4. Methods.

4.1. Classification method. We consider the set of values of each phenotype as originating from three classes: low phenotype levels (*negatives*), high phenotype levels (*positives*) and medium levels. A natural partition into these classes can be obtained by setting a threshold $1 \leq t \leq 50$ and defining the positives (resp., negatives) to be the top (resp., bottom) t percent of the phenotype values. The results we report here were obtained with $t = 15$. We also experimented with an alternative discretization scheme obtained by fitting to a normal distribution, which yielded similar results in the subsequent analysis (Table S1).

The partition we have defined determines the training examples. Our goal is to use the training data in order to predict the phenotype class of a test sample. We approach the classification problem using a soft margin support vector machine (SVM) [7, 18]. SVM performs the classification task by identifying hyperplanes that "best" separate between data vectors from two classes. We perform the SVM analysis using the *Gist* software package [19].

The application of an SVM requires the specification of a similarity function with certain properties, called *kernel function*, for comparing data vectors. A kernel function can be interpreted as mapping vectors to a high-dimensional *feature space*, where similarity is computed using scalar dot-product. To this end, we developed novel kernels for comparing genotypes and for comparing haplotypes. The challenge was to design a feature space that provides good classification results, while not overfitting the data.

4.2. Kernel functions for genotype data. A given set of m SNPs can be viewed as a vector of length m with entries from $\{0, 1, 2\}$ in the following way: If a position is homozygous, the corresponding entry is set to 0 or 1, depending on the state of the SNP. If a position is heterozygous, the corresponding entry is set to 2. We call the resulting length m vector a *genotype*. In the following we present

two kernel functions. Both functions are based on viewing the genotype vectors as representations for ordered pairs of binary haplotypes.

Given a pair of genotypes of length m, there may exists an exponential number of phasings (representations as binary haplotypes) for each of them and consequently, an exponential number of possible locus-wise pairings. The *expectation kernel* is computed as half the **expected number of identical alleles** over the space of all possible phasings of the two genotypes, assuming that each phasing is equally likely and that the state of a given allele is independent of all other SNP alleles. Thus, for two genotype vectors x and y we define $\kappa(x,y) = \sum_{i=1}^{m} S(x_i, y_i)$, where $S(x_i, y_i)$ denotes the similarity between the genotype entries x_i and y_i. By definition $S(0,0) = S(1,1) = 1$. Similarly, $S(2,2) = 0.5$, since with probability 0.5 two alleles are identical and with probability 0.5 none is. To extend this similarity notion to incomplete vectors, in which some allele states are missing, we treat the missing alleles as drawn at random from the distribution of alleles in the population. Hence, for example, $S(0,?) = p$, where '?' denotes a missing genotype entry and p is the frequency of allele '0' for the corresponding SNP. To show that this definition is indeed a kernel function we provide in Table S4 an explicit mapping of the genotype data to a feature space in which the dot product is equivalent to the expectation kernel.

For the second kernel, we use a formulation closely related to the one used in [27]. This kernel computes the **maximum number of identical alleles** shared by the two compared individuals. We call this kernel the *identity kernel*. Formally, we represent each state using 4 bits, where [1,0,1,0] corresponds to state 0, [0,1,0,1] to state 1 and [1,1,0,0] to state 2. Missing entries are represented by [0,0,0,0]. The identity kernel can be computed as the dot product of the vectors obtained by concatenating these bit strings.

Each of the two kernels was embedded into a *Gaussian* function, which is defined as $exp(\frac{2k(x,y)-k(x,x)-k(y,y)}{2\sigma^2})$, where σ is the width parameter, and $k(\cdot,\cdot)$ is the original kernel.

To use the covariate data in the SVM analysis, we normalized it as follows: continuous data was normalized to have mean zero and variance one; the binary sex character was encoded as 1 or -1. We used the identity mapping of these normalized data to feature space, concatenating these features to the mapped genotype vectors.

4.3. Kernel functions for haplotype data. We also developed kernels for haplotype data. Such data can be derived using phasing approaches (e.g., PHASE [29]) or from pedigree information. Given two complete haplotype vectors (the missing alleles are completed by the PHASE program), each representing two parental haplotypes, the only degree of freedom is which of the two possible pairings to use. Given a pairing, which matches one haplotype in one vector with some haplotype in the other vector, the similarity in each position (for each pair) is defined as 0.5 if the

two corresponding nucleotides match, and 0 otherwise. In practice, a pairing is not available. Ideally, we would want to compute the similarity under the pairing that maximizes it. Since the maximum operator is in general not a kernel function, we present a variant on this ideal formulation.

For two haplotype pairs $x = (x_1, x_2)$ and $y = (y_1, y_2)$, denote by n_{ij} the number of positions in which x_i and y_j agree, for $i, j = 1, 2$. For a parameter k, define $K^k(x, y) = \sum_{i=1}^{2} \sum_{j=1}^{2} \binom{n_{ij}}{k}$. We now show that $K^k(x, y)$ is a kernel by proving that it can be computed as a scalar product in some feature space. Associate with each genotype x a vector of length $\binom{n}{k} 2^{2^k}$ in which each entry is 1 or -1. This vector will be the concatenation of 2^{2^k} subvectors, each corresponding to a different function from $\{0, 1\}^k$ into $\{1, -1\}$ (there are a total of 2^{2^k} such functions). The subvector corresponding to function h will have $\binom{n}{k}$ elements, corresponding to the k-element subsets of $\{1, 2, \ldots, n\}$. The component corresponding to set S will have the entry $h(x_1(S)) + h(x_2(S))$, where $x_i(S)$ is the k-tuple of bits in the positions of x_i corresponding to S. The scalar product of the genotype vectors associated with x and y is $d = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_S \sum_h h(x_i(S)) \cdot h(y_j(S))$. For a set S and $i, j \in \{1, 2\}$, if $x_i(S) \neq y_j(S)$ then $\sum_h h(x_i(S)) \cdot h(y_j(S)) = 0$. If $x_i(S) = y_j(S)$ then $\sum_h h(x_i(S)) \cdot h(y_j(S)) = 2^{2^k}$. Since the number of k-element sets S such that $x_i(S) = y_j(S)$ is $\binom{n_{ij}}{k}$, we have that $d = 2^{2^k} \sum_{i=1}^{2} \sum_{j=1}^{2} \binom{n_{ij}}{k}$. Dividing by 2^{2^k}, the claim is proven.

4.4. Cross validation procedure. To test our approach we generated training data for each phenotype by cross validation. We iterated over the three pedigrees into which the baboons were partitioned. In each iteration baboons from one of the pedigrees were held out, and a classifier was trained based on the remaining two pedigrees and evaluated on the held-out part.

The hyper-parameters of the SVM classifier and the k parameter of the all-match kernel (when classifying by haplotypes) were optimized separately for each left out pedigree by applying a grid search based on an inner cross-validation. The values considered for the hyper-parameters of the SVM span a few orders of magnitude, defined by the following series: $\{10^{-2}, 10^{-1.5}, \ldots, 10^{2}\}$. For the k parameter of the all-match kernel we considered the values $\{2, 3, 4\}$. On each iteration we trained with one of the remaining two pedigrees and tested the success rates with the different parameters on the other one.

The success rate was measured by $\frac{tp}{2(tp+fn)} + \frac{tn}{2(tn+fp)}$, where tp, tn, fp and fn are the numbers of true positives, true negatives, false positives and false negatives, respectively.

To compute the significance of the results we performed a series of permutation tests in which we randomly permuted the values of a certain phenotype among the individuals and applied the prediction procedure to the permuted data.

To preserve the properties of the original data as much as possible, we permuted the phenotype values within each pedigree separately. The p-value that was given to an observation was computed as the fraction of random runs on permuted data that achieved a higher success rate.

4.5. The analyzed phenotypes. The genotyped baboons underwent three different diets—CHOW (basal, low in fat and cholesterol), LCHF (low cholesterol, high fat) and HCHF (high cholesterol, high fat) [23].

Lipid and apolipoprotein phenotypes measured in all three diets include concentrations of HDL and LDL (i.e., non-HDL) cholesterol and the ApoB and ApoE apolipoproteins [23, 24], percentage of apoE in beta lipoprotein, and total serum cholesterol. Lipoprotein size distribution phenotypes measured in all three diets include LDL peak particle diameter, median diameter of beta lipoproteins, and diameter of HDL. Finally, Triglyceride levels were measured only for the basal diet.

5. Acknowledgments. We thank Amir Ben-Dor and Zohar Yakhini for stimulating discussions. Part of this work was done while J.G. and R.S. were at ICSI, Berkeley. N.Y is supported by the Tel-Aviv university president and rector scholarship. J.G. was funded through a postdoc fellowship by the DAAD (German Academic Exchange Association) and by DFG grant NI-369/2. R.S. was supported by an Alon Fellowship and by a grant from the Israel Science Foundation (grant no. 385/06).

6. Web Site References.

http://pga.lbl.gov/SNP/Baboon/APOA1C3A4A5.html;
Details on SNPs in the *APOA1/C3/A4/A5* gene cluster region.
http://www.stat.washington.edu/stephens/software.html; The PHASE program for haplotype inference.
http://svm.sdsc.edu/svm-intro.html; The Gist program for SVM analysis.

REFERENCES

[1] G.R. Abecasis, L.R. Cardon, and W.O. Cookson. *A general test of association for quantitative traits in nuclear families.* Am. J. Hum. Genet., 66(2000), pp. 279–292.

[2] B.E. Aouizerat, M. Kulkarni, D. Heilbron, D. Drown, S. Raskin, C.R. Pullinger, M.J. Malloy, and J.P. Kane. *Genetic analysis of a polymorphism in the human apoA-V gene: effect on plasma lipids.* J. Lipid Res., pages 1167–1173, 2003.

[3] N. Beerenwinkel, M. Dumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter. *Geno2Pheno: estimating phenotypic drug resistance from HIV-1 genotypes.* Nucleic Acids Research, 31:13(2003), pp. 3850–3855.

[4] N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig. *Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype.* Proc. Natl. Acad. Sci. USA, 99(2002), pp. 8271–8276.

[5] N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenfhrer, K. Roomp, I. Savenkov, R. Fischer, D. Hoffmann, J. Selbig, K. Korn, et al. *Computational methods for the design*

of effective therapies against drug resistant hiv strains. Proc. Natl. Acad. Sci. USA, 21:21(2005), pp. 3943–3950.

[6] B.E. Boser, I.M. Guyon, and V.N. Vapnik. *A training algorithm for optimal margin classifiers.* In: D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

[7] N. Cristianini and J. Shawne-Taylor. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[8] G.M. Dallinga-Thie, X.D. Bu, M. van Linde-Sibenius Trip, J.I. Rottera, A.J. Lusis, and T.W. de Bruin. *Apolipoprotein a-i/c-iii/a-iv gene cluster in familial combined hyperlipidemia: effects on ldl-cholesterol and apolipoproteins b and c- iii.* J. Lipid Res., 37(1996), pp. 136–147.

[9] C. Elkan. *Boosting and naive bayesian learning.* (Technical Report). San Diego: Department of Computer Science and Engineering, University of California., 1997.

[10] G. Greenspan and D. Geiger. *High density linkage disequilibrium mapping using models of haplotype block variation.* Bioinformatics, 20:Suppl. 1(2004), pp. I37–I44.

[11] M. Groenendijk, R.M. Cantor, T.W. de Bruin, and G.M. Dallinga-Thie. *The apoAI-CIII-AIV gene cluster.* Atherosclerosis, 157(2001), pp. 1–11.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. *Gene selection for cancer classification using support vector machines.* Machine Learning, 46:1–3(2002), pp. 389–422.

[13] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, J. Mackey, D. Wishart, R. Greiner, and B. Zanke. *Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms.* Clinical Cancer Res, 10(2004), pp. 2725–2737.

[14] J.S. Liu, C. Sabatti, J. Teng, B.J.B. Keats, and N. Risch. *Bayesian analysis of haplotypes for linkage disequilibrium mapping.* Genome Res., 11(2001), pp. 1716–1724.

[15] R. Mar, P. Pajukanta, H. Allayee, M. Groenendijk, G. Dallinga-Thie, R.M. Krauss, J.S. Sinsheimer, R.M. Cantor, T.W.A. de Bruin, and A.J. Lusis. *Association of the apolipoprotein A1/C3/A4/A5 gene cluster with triglyceride levels and LDL particle size in familial combined hyperlipidemia.* Circulation Research, 94(2004), pp. 993—999.

[16] E.R. Martin, W.K. Scott, M.A. Nance, R.L. Watts, J.P. Hubble, W.C. Koller, K. Lyons, R. Pahwa, M.B. Stern, and A. Colcher et al. *Association of single-nucleotide polymorphisms of the tau gene with late-onset parkinson disease.* Journal of American Medical Association, 286:18(2001), pp. 2245–2250.

[17] J.R. McNamara, H. Campos, J.M. Ordovas, J. Peterson, P.W.F. Wilson, and E.J. Schaefer. *Effect of gender, age, and lipid status on low density lipoprotein subfraction distribution.* Arteriosclerosis, 6:5(1987), pp. 483–490.

[18] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. *An introduction to kernel-based learning algorithms.* IEEE Neural Networks, 12:2(2001), pp. 181–201.

[19] P. Pavlidis, I. Wapinski, and W.S. Noble. *Support vector machine classification on the web.* Bioinform, 20:4(2004), pp. 586–587.

[20] L.A. Pennacchio, M. Olivier, J.A. Hubacek, J.C. Cohen, D.R. Cox, J.C. Fruchart, R.M. Krauss, and E.M. Rubin. *An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing.* Science, 294(2001), pp. 169–173.

[21] F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet. *Genotyping errors: causes, consequences and solutions.* Nature Reviews Genetics, 6(2005), pp. 847–846.

[22] D.L. Rainwater. *Lipoprotein correlates of ldl particle size.* Atherosclerosis, 1489(2000), pp. 151–158.

[23] D.L. Rainwater, C.M. Kammerer, K.D. Carey, B. Dyke, J.F. VandeBerg, W.R. Shelledy, P.H. Moore Jr, M.C. Mahaney, H.C. McGill Jr, and J.L. VandeBerg. *Genetic determination of HDL variation and response to diet in baboons.* Atherosclerosis,

161(2002), pp. 335–343.

[24] D.L. Rainwater, C.M. Kammerer, M.C. Mahaney, J. Rogers, L.A. Cox, J.L. Schneider, and J.L. VandeBerg. *Localization of genes that control ldl size fractions in baboons.* Atherosclerosis, 168:1(2003), pp. 15–22.

[25] B. Rannala and J.P. Reeve. *High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence.* Am. J. Hum. Genet., 69(2001), pp. 159–178.

[26] S.Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D.L. Brutlag, and R.W. Shafer. *Genotypic predictors of human immunodeficiency virus type 1 drug resistance.* Proc. Natl. Acad. Sci. USA, 103:46(2006), pp. 17355–17360.

[27] H. Schwender, M. Zucknick, K. Ickstadt, and H. Bolt. *A pilot study on the application of statistical classification procedures to molecular epidemiological data.* Toxicology Letters, 151:1(2004), pp. 291–9.

[28] M.J. Stampfer, R.M. Krauss, J. Ma, P.J. Blanche, L.G. Holl, F.M. Sacks, and C.H. Hennekens. *A prospective study of triglyceride level, low-density lipoprotein particle diameter and risk of myocardial infarction.* JAMA, 276:11(1996), pp. 882–888.

[29] M. Stephens, N. J. Smith, and P. Donnelly. *A new statistical method for haplotype reconstruction from population data.* Am. J. Hum. Genet., 68(2001), pp. 978–989.

[30] A. Tchernof, B. Lamarche, D. Prud'Homme, A. Nadeauand S. Moorjani, F. Labrie, and P.J. Lupienand JP. Despres. *The dense ldl phenotype. association with plasma lipoprotein levels, visceral obesity, and hyperinsulinemia in men.* Diabetes Care, 19:6(1996), pp. 629–637.

[31] D.C. Thomas. *Statistical Methods in Genetic Epidemiology.* Oxford University Press, 2004.

[32] Q.-F. Wang, X. Liu, J. O'Connell, Z. Peng, R.M. Krauss, D.L. Rainwater, J.L. VandeBerg, E.M. Rubin, J.-F. Cheng, and L.A. Pennacchio. *Haplotypes in the APOA1-C3-A4-A5 gene cluster affect plasma lipids in both humans and baboons.* Human Molecular Genetics, 13:10(2004), pp. 1049–1056.

[33] Y. Yoon, J. Song, S. Hong, and J. Kim. *Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines.* Clin Chem Lab Med, 41:1(2003), pp. 529–534.

[34] N. Yosef, Z. Yakhini, A. Tsalenko, V. Kristensen, A.L. Brresen-Dale, E. Ruppin, and R. Sharan. *A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data.* Bioinformatics, 23:2(2007), pp. e91–e80.