

# Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms

Jennifer Listgarten,<sup>4</sup> Sambasivarao Damaraju,<sup>1,4</sup> Brett Poulin,<sup>2</sup> Lillian Cook,<sup>4</sup> Jennifer Dufour,<sup>4</sup> Adrian Driga,<sup>4</sup> John Mackey,<sup>1,4</sup> David Wishart,<sup>3</sup> Russ Greiner,<sup>2</sup> and Brent Zanke<sup>1,4</sup>

University of Alberta Faculties of <sup>1</sup>Medicine, <sup>2</sup>Science, <sup>3</sup>Pharmaceutical Sciences, and the <sup>4</sup>Cross Cancer Institute of the Alberta Cancer Board, Edmonton, Alberta, Canada

## ABSTRACT

Hereditary predisposition and causative environmental exposures have long been recognized in human malignancies. In most instances, cancer cases occur sporadically, suggesting that environmental influences are critical in determining cancer risk. To test the influence of genetic polymorphisms on breast cancer risk, we have measured 98 single nucleotide polymorphisms (SNPs) distributed over 45 genes of potential relevance to breast cancer etiology in 174 patients and have compared these with matched normal controls. Using machine learning techniques such as support vector machines (SVMs), decision trees, and naïve Bayes, we identified a subset of three SNPs as key discriminators between breast cancer and controls. The SVMs performed maximally among predictive models, achieving 69% predictive power in distinguishing between the two groups, compared with a 50% baseline predictive power obtained from the data after repeated random permutation of class labels (individuals with cancer or controls). However, the simpler naïve Bayes model as well as the decision tree model performed quite similarly to the SVM. The three SNP sites most useful in this model were (a) the +4536T/C site of the aldosterone synthase gene *CYP11B2* at amino acid residue 386 Val/Ala (T/C) (rs4541); (b) the +4328C/G site of the aryl hydrocarbon hydroxylase *CYP1B1* at amino acid residue 293 Leu/Val (C/G) (rs5292); and (c) the +4449C/T site of the transcription factor *BCL6* at amino acid 387 Asp/Asp (rs1056932). No single SNP site on its own could achieve more than 60% in predictive accuracy. We have shown that multiple SNP sites from different genes over distant parts of

the genome are better at identifying breast cancer patients than any one SNP alone. As high-throughput technology for SNPs improves and as more SNPs are identified, it is likely that much higher predictive accuracy will be achieved and a useful clinical tool developed.

## INTRODUCTION

Malignant transformation occurs through the accumulation of mutations in genes regulating cell division, apoptosis, invasiveness, or metastasis. These can occur as primary events or as a consequence of defects in “caretaker” genes that function in the maintenance of genomic stability (1). Inherited cancer predisposition from the inheritance of single genes almost exclusively results from abnormalities in DNA maintenance genes such as DNA double-strand break repair factors *BRCA1* or *BRCA2*, which are abnormal in familial breast cancer (2); the check point kinase *ATM*, which is mutated in ataxia telangiectasia (3); the double-strand break repair gene *MRE11*, which is abnormal in a variant of ataxia telangiectasia (4); the helicase *BLM*, which is mutated in Bloom’s syndrome (5); *NBS1*, implicated in the Nijmegen breakage syndrome (6); the XP excision repair enzymes in Xeroderma pigmentosum (7); the mismatch repair enzymes *MSH2* and *MLH1* in hereditary nonpolyposis colon cancer (8, 9); and the transcription regulator *p53* in the Li Fraumeni syndrome (10).

Whereas mutations that render DNA repair enzymes completely inactive can lead to obvious clinical consequences, polymorphisms in these genes that produce subtle alterations in their effectiveness may result in environmental sensitivities, resulting in cancer. The consequence of mutagen exposure may vary between individuals depending on the effectiveness of intrinsic detoxification and repair of induced DNA damage. For instance, procarcinogens such as *N*-nitrosoamines are metabolized into intermediate carcinogenic metabolites by the Phase I cytochrome P450 enzyme 2E1 and are excreted with enhanced solubility through the actions of Phase II enzymes such as glutathione *S*-transferase M1 (11). Increasingly the relationship between the mutagenic potential of genotoxins and inherited allelic variability in carcinogen metabolizing and DNA repair genes is becoming recognized (12–14). The consequence of the “gene-environment” interaction is likely to differ between individuals because of the inheritance of polymorphic alleles and various environmental exposures (15).

With ongoing high-throughput human gene sequencing efforts, human genome variability can now be measured. As many as 3 million sites of “single nucleotide polymorphism” (SNP) have been identified, thus defining the allelic complexity of the human gene pool. Many epidemiological studies have attempted to attribute single alleles to cancer risk. Typically, prior knowledge of tumor pathophysiology permits selection of a candidate gene for which allelic variability has been described. A classic case-control study may be performed after the meas-

Received 7/30/03; revised 1/13/04; accepted 1/13/04.

**Grant support:** This work was sponsored by the Government of Alberta, Ministry of Health and Wellness, Health Strategies Division, and the Alberta Cancer Board.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

**Requests for reprints:** Brent Zanke, Cancer Care Ontario, 1324–620 University Avenue, Toronto, Ontario, M5G 2L7 Canada. Phone: 416-971-9800, extension 2229; Fax: 416-217-1281; E-mail: Brent.Zanke@cancercare.on.ca.

urement of specific alleles in tumors and age-matched control groups. Using such techniques, investigators have linked *CYP3A4* and *hOGG1* alleles to prostate cancer risk (16, 17), a *RET* allele to papillary thyroid carcinoma (18), a P2X7 allele to chronic lymphocytic leukemia (19), a *kallikrein 10* allele to gonadal tumors (20), a *cyclin D1* allele to bladder tumors (21), *p53* and *MMP-1* alleles to lung cancer (22, 23), and *CDKN2A* to melanoma (24).

Such association studies are dependent on prior knowledge of cancer pathogenesis and fortuitous selection of specific polymorphisms for study. Large-scale SNP analytical tools now exist, allowing the simultaneous measurement of many alleles. Interpretation of significant differences in allele distribution between affected individuals and normal controls is difficult because of the hazards of multiple testing (25). When hundreds of alleles are measured and related to even a single clinical patient characteristic, spurious, statistically significant associations may be identified by chance alone. With many clinical patient characteristics, the problem is exacerbated.

Risk for the development of sporadic breast cancer may have a significant inherited component, with as many as 10% of cases having a significant familial component (26, 27). Of these, as few as 13% of cases may be attributable to known BRCA1 or BRCA2 mutations (28). The proportion of breast cancer in the general population that can be explained by these high penetrance genes is relatively small. Variant genotypes in genes that may be involved in the molecular etiology of cancer may confer a relatively smaller degree of cancer risk when considered individually but, when considered collectively, may explain a large component of inherited and sporadic breast cancer (29). Because these genes may be carried by a larger proportion of the general population, the proportion of breast cancer that could be explained by these genes may be relatively large.

To identify polymorphisms in unrecognized breast cancer-associated genes we have measured 98 SNPs distributed over 45 genes in 174 patients with breast cancer and compared these with 158 normal controls. We have compared a variety of machine learning techniques: support vector machines (SVMs), decision trees, and naïve Bayes, and have identified a subset of SNPs that have predictive power in distinguishing breast cancer patients from controls. Many of the genes containing these SNPs are implicated in DNA transcription and repair or in steroid metabolism, suggesting a genetic predisposition to breast cancer in some "nonfamilial" sporadic breast cancers. In this study, the SNP site most able to discriminate between populations, as measured by information gain (described later), was the +4536C/T polymorphism in the aldosterone synthase gene *CYP11B2* at amino acid position 386 (Val/Ala). Alone, evaluation at this site resulted in a naïve Bayes prediction accuracy of 56% as compared with a baseline of 50%. Accuracy was increased to 69% with two additional SNP-based allele determinations in conjunction with a quadratic kernel SVM. Thus, we have shown that machine learning techniques may be used to successfully model relationships between inherited genetic polymorphisms and clinical disease. As high-throughput technology for SNPs improves and, as more SNPs are identified, it is likely that much higher predictive accuracy could be achieved and useful clinical tools be developed with this methodology.

## MATERIALS AND METHODS

**Patient Identification.** The PolyomX Program<sup>5</sup> of the Alberta Cancer Board systematically archives peripheral blood and tumor samples with informed consent from patients and with local institutional review board approval. For this study, 174 local sequentially registered patients with banked breast cancer who were not known to have BRCA1 or BRCA2 abnormalities, were enrolled between January 2001 and June 2002. Blood samples from local age-matched persons not known to have breast cancer were used as controls.

**Tissue Accrual.** Breast tumors removed at the time of primary surgery were identified by gross appearance and placed into liquid nitrogen within 20 min of devitalization. Breast cancer was confirmed histologically on adjacent tissue by two independent pathologists. Peripheral blood was collected into EDTA. Buffy coat cells were isolated by centrifugation and were immediately stored in liquid nitrogen.

**Clinical Informatics.** Clinical parameters were prospectively collected on all patients by multidisciplinary review of imaging studies, histology and by patient interviews conducted by members of the Northern Alberta Breast Cancer Program. Categorical clinical information was entered via web-based information forms and included a detailed family history, disease risk factors, presentation details, pathology, treatment administered, and outcome.<sup>6</sup>

**SNP Measurement.** Polymorphism analysis for various gene SNPs was carried out by the Qiagen genomics service.<sup>7</sup> The assay reproducibility was more than 95% (30). QIAmp DNA blood kit (Qiagen) was used for DNA isolation. DNA was quantitated using the Pico green fluorescence assay (31). The SNPs selected from Human Genome Variability Database were validated using control panel of DNA obtained from Coriell Cell Repositories. From a total of 245 SNPs selected from this public domain database, polymorphisms at 98 sites were reproducibly measured in one or all of the ethnic groups tested from the above panel of DNA, as selected for study in our study subjects. These include 45 well-characterized genes from tumor suppressors, receptors, transcription factors, DNA metabolism enzymes, oncogenes, and other signal transduction pathways.

**Data Analysis.** Correlation of SNPs with presence of cancer was assessed through use of information gain (32), with statistical significance calculated through use of random permutation simulations followed by multiple comparison corrections (33–36). Two-class discriminative models for patients with breast cancer and controls were built and tested using 20-fold cross-validation in conjunction with several machine learning algorithms: naïve Bayes (37), SVM (38), and decision tree (39). The prior in naïve Bayes and decision tree was always set to 50:50. A variety of kernels were used with the SVM, with the quadratic kernel performing maximally. Data analysis was performed with Matlab and SVMLight (40). Relative risk associated with particular genotypes and allele

<sup>5</sup> Internet address: <http://www.polyomx.org/>.

<sup>6</sup> The complete clinical data template can be found at <http://www.cancerboard.ab.ca/polyomx/breastCancerSnpStudy/breastCancerTemplate.html> (best viewed with Internet Explorer).

<sup>7</sup> Internet address for the Qiagen genomics service: <http://www.qiagen.com>.

frequencies were estimated by calculating odds ratios with 95% and 99% confidence intervals (CIs). Because odds ratios could not be computed with any genotype or allele frequencies that were zero, a “pseudo-count” of 0.5 was added to these genotype or allele counts to make the calculation feasible (and biased); this is a typical “Laplacian correction.” Multiple comparisons were not taken into account for the odds ratio CIs.

SNP calls at each site were converted into numeric values assigned according to control population frequencies in the present study: homozygous major allele, 1; heterozygous, 2; homozygous minor allele, 3; ambiguous. Data analysis using this coding convention makes certain assumptions. For models that treat the SNPs as continuous variables, such as SVMs, it makes an additive assumption: heterozygotes are half-way between the homozygotes. Also the two alleles are not treated symmetrically by such models. For models such as naïve Bayes and decision trees, which consider the SNPs to be nominal data, the coding is unimportant. Unknown values refer to data points with poor signal:noise ratio in the genotyping assays. These missing values were ignored in all of the calculations and, thus, were not used as informative. The naïve Bayes algorithm naturally adapts to missing values. It was used with all of the data, as well as with a smaller data set consisting only of patients with all SNP measurements present. SVM and decision tree algorithms were only used with this latter, smaller data set.

## RESULTS

### Description of Breast Cancer and Control Populations.

The 158 control bloods were anonymous, nonduplicated discarded samples obtained from patients attending the University of Alberta Hospital in Edmonton. We selected this tertiary-referral center to obtain control samples because (a) breast cancer patients are not included in the clinical population, and (b) the control and test participants were derived from the same geographical region and referral area. The mean age of the controls was 57.9 years. The 174 samples from patients were derived from women with newly diagnosed invasive breast cancers who consented to primary tumor and blood banking and analysis and attended the Cross Cancer Institute in Edmonton, Canada. All of the tumor samples were independently reviewed to confirm malignancy and histological features. Mean age was 55 years; the mean tumor diameter was 2.2 cm; 74% of tumors were hormone receptor positive (either estrogen receptor and/or progesterone receptor positive) by centralized immunohistochemical analysis, and 59% had node positive disease. Thirty percent of patients were premenopausal, 11% were perimenopausal, and 59% were postmenopausal. American Joint Committee on Cancer stage (fifth edition) was stage II in 89%, stage III in 10%, and stage IV in 1% of patients.

**Predictive SNPs.** Correlation of individual SNPs with occurrence of cancer was computed using information gain (32).<sup>8</sup> Information gain is based on the entropy,  $H$ , of a distribution  $\{p_i\}$ :  $H(p_1, \dots, p_n) = -[\text{summ}]_i p_i \log p_i$ . In this case,  $p_i$  is

the probability of one genotype (e.g., heterozygote) in one population,  $i$ , (e.g., breast cancer patients), and  $n = 2$ , because there are two classes (breast cancer patients and controls). The entropy of a distribution represents the amount of uncertainty in the distribution. In the present context, a high entropy value for a particular genotype for a single SNP would indicate that this genotype is providing information about whether a person has cancer or not. Information gain combines the entropy of each feature value (common homozygous, heterozygous, variant) to form a single number representing the informativeness of the feature (SNP) with respect to the class (cancer patients/controls). Information gain is a measure of the “purity” of the split that a particular feature creates in the data set. For example, if SNP\_1 is present 100% of the time as the minor allele in the breast cancer population and 0% of the time in the normal population, then SNP\_1 creates a perfectly pure split; it is very informative. Conversely, if SNP\_2 is present 30% of the time as the minor allele in breast cancer patients and likewise at 30% in a normal population, then SNP\_2 creates a very impure split; it is completely uninformative. Formally, information gain is calculated by summing the entropy of the split distribution for each possible value of the feature (common homozygous, heterozygous, homozygous variant), weighted by the proportion of values that fall into each possible feature value. This value is then subtracted from the entropy of the split created by the labels alone. The higher the information gain, the more informative the feature and, thus, the more predictive power it has.

Statistical significance was assigned to the information gain values by modeling the null distribution of each SNP with random permutation tests. The significance of each SNP as a predictor for breast cancer *versus* normal was assessed by randomly permuting the labels of the breast cancer and normal SNP data, and then calculating the resulting information gain of each SNP with respect to this random partition. This type of random permutation technique has gained prominence in the microarray community, in which an overabundance of features and feature scoring methods are present (33–36). Ten thousand permutations were performed producing a simulated probability distribution over information gain values for the null hypothesis that the two groups are the same. From this distribution, it was inferred that each of 13 SNPs was individually significant at the  $P \leq 0.05$  level (Table 1; see Table 2 for full SNP information). Because the number of tests was high, a correction for multiple testing was applied so that the overall family of hypotheses has a reasonable false discovery rate. The most conservative such correction is Bonferroni. This correction showed two SNPs to be significant ( $P \leq 0.05$ ; Table 1, SNPs 1–2). Less conservative step-down Bonferroni and Sidak corrections arrived at the same result, with two significant SNPs (Table 1, SNPs 1–2). A less conservative adjustment, the Benjamini-Hochberg step-up false discovery rate indicated that 11 SNPs were significant (Table 1, SNPs 1–11). All of these adjustments, except for Benjamini-Hochberg false discovery rate are known to be highly conservative to preserve the Type I error rate at the expense of increasing the Type II error rate. Benjamini-Hochberg false discovery rate assumes that the  $P$ s across SNPs are independent and uniformly distributed under their respective null hypotheses. In generic association studies, significant differences between populations for a given SNP are often measured using a  $\chi^2$  test on

<sup>8</sup> A complete listing of all SNPs studied in this experiment can be found at <http://www.cancerboard.ab.ca/polyomx/breastCancerSnpStudy/snpData.html>.

**Table 1** The significance of 13 single nucleotide polymorphisms (SNPs)

SNPs found to have significant information gain values (relative to breast cancer patients *versus* controls) as determined by permutations tests. SNPs 1–13 are significant at a  $P \leq 0.05$  level. With adjustments for multiple hypothesis testing through use of Bonferroni, step-down Bonferroni, or Sidak, SNPs 1–2 are significant at a  $P \leq 0.05$  level. With the Benjamin-Hochberg false discovery rate step-up adjustments, SNPs 1–11 are significant at a  $P \leq 0.05$  level. Full information on SNPs is provided in Table 2.

	dbSNP <sup>a</sup>	SNP designation
1	rs4541	<i>CYP11B2</i> (+)4536T/C
2	rs1056836	<i>CYP1B1</i> (+)4328C/G
3	rs1056932	<i>BCL6</i> (+)4449C/T
4	rs10046	<i>CYP19A1</i> (+)32123 (3'UT)
5	rs4545	<i>CYP11B2</i> (+)5215G/A
6	rs1799977	<i>MLH1</i> (+)18529A/G
7	rs1800935	<i>MSH6</i> (+)12742T/C
8	rs5182	<i>AGTR1</i> (+)572C/T
9	rs1799939	<i>RET</i> (+)37412G/A
10	rs17607	<i>CD68</i> (+)1786G/A
11	rs6405	<i>CYP11B1</i> (+)28G/A
12	rs6163	<i>CYP17</i> (+)194G/T
13	rs1800051	<i>CD38</i> (+)55806A/C

<sup>a</sup> dbSNP, double-strand SNP; UT, untranslated.

the  $2 \times 3$  SNP table with subsequent look-up in a  $\chi^2$  distribution table. Use of the  $\chi^2$  distribution makes more stringent assumptions about the structure of the underlying data than use of permutation tests. However, for comparison, we here also applied a  $\chi^2$  analysis. Uncorrected  $P$ s resulting from the  $\chi^2$  test were of the same order of magnitude as those from the information gain tests. Furthermore, application of multiple correction testing to the  $\chi^2$   $P$ s provided almost identical results, with the only exception being the Benjamini-Hochberg step-up false discovery rate, which indicated that only SNPs 1–9 in Table 1

were significant, rather than SNP 1-11 which the information gain provided (data not shown).

**Diagnostic Classifiers.** Machine learning techniques seek to semi-automatically build and validate mathematical models of data. Once a model has been built and validated, the model can then be used for classification or regression or for examining which parts of the data were relevant and in what way. Application of machine learning techniques to a data set involves four steps: (a) positing a class of mathematical or statistical models appropriate for the data; (b) “learning” which particular model in the class is most suitable for the data (this typically involves a numerical optimization of some objective function to produce a fixed set of parameters identifying a specific model within the model class; and (c) validation of the model by use of a test set or cross-validation (explained below). At this point, one has a model, and no longer needs the training data. The final and fourth step can be performed: (4) application of the final model to new data.

Cross-validation is a way to make the most use of a data set for both learning and validation. Rather than separating the data into a single learning set (called the “training” set) and a single test set,  $n$ -fold cross-validation separates the data into  $n$  training sets and  $n$  test sets. If  $n$  were equal to five, cross-validation would work as follows: The entire data set would be divided into five equal-sized groups. The first four groups would be used as training data, and the fifth as test data. The second through to fifth groups would then be used as training data and the first group as test data. This procedure is continued until each group has been used as test data. The aggregate test results from all  $n = 5$  phases of the cross-validation would be used to obtain a final estimate of the predictive accuracy. Cross-validation provides an estimate of how a particular model might do on a new, unseen data set drawn from the same statistical distribution. If the cross validation process produces an estimated accuracy that

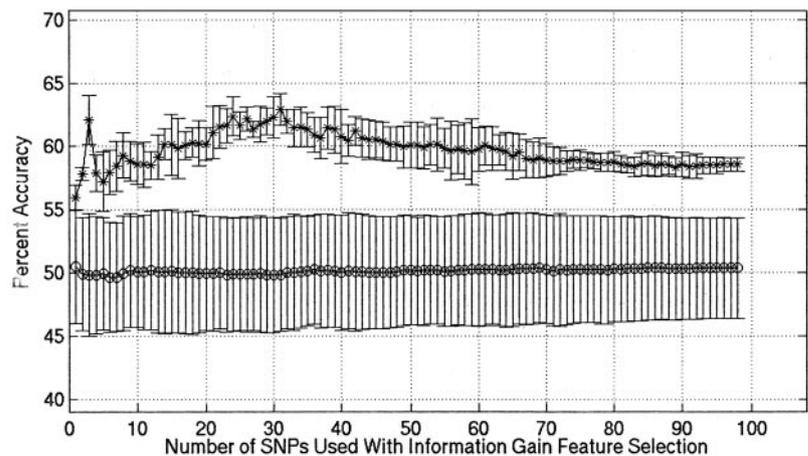
**Table 2** Information on all single nucleotide polymorphisms (SNPs) reported by name in this paper

In the present study, in the control population, SNPs shown in bold were found to have the minor and major alleles opposite from what was reported in the database. References to genotypes in this paper use minor and major alleles as determined by the control population in the present study. For example, *BCL6* homozygous variant refers to *CC*.

	Gene name	SNP designation (as in dbSNP) <sup>a</sup>	Common allele in control population	dbSNP identification	Chromosome	Codon
1	<i>CYP11B2</i>	(+)4536T/C	<i>T</i>	rs 4541	8	386Val/Ala
2	<i>CYP1B1</i>	(+)4328C/G	<i>C</i>	rs 5292	8	293 Leu/Val
3	<b><i>BCL6</i></b>	<b>(+)4449C/T</b>	<b><i>T</i></b>	<b>rs 1056932</b>	<b>3</b>	<b>387 Asp/Asp</b>
4	<i>CYP19A1</i>	(+)32123 (3'UT)T/C	<i>C</i>	<b>rs 10046</b>	<b>15</b>	<b>NA</b>
5	<i>CYP11B2</i>	(+)5215G/A	<i>G</i>	rs 4545	8	435 Gly/Ser
6	<i>MLH1</i>	(+)18529A/G	<i>A</i>	rs 1799977	3	219 Ile/Val
7	<i>MSH6</i>	(+)12742T/C	<i>T</i>	rs 1800935	2	180 Asp/Asp
8	<i>AGTR1</i>	(+)572C/T	<i>C</i>	rs 5182	3	191 Leu/Leu
9	<i>RET</i>	(+)37412G/A	<i>G</i>	rs 1799939	10	691 Gly/Ser
10	<i>CD68</i>	(+)1786G/A	<i>G</i>	rs 17607	17	340 Ala/Thr
11	<i>CYP11B1</i>	(+)28G/A	<i>G</i>	rs 6405	8	10 Cys/Tyr
12	<i>CYP17</i>	(+)194G/T	<i>G</i>	rs 6163	10	65 Ser/Ser
13	<i>CD38</i>	(+)55806A/C	<i>A</i>	rs 1800051	4	168 Ile/Ile
14	<i>ADPRT</i>	(+)22266T/C	<i>T</i>	rs1805414	1	284Ala/Ala
15	<i>ERCC2</i>	(+)17966C/T	<i>C</i>	rs1052555	19	50Asp/Asp
16	<i>CYP11B2</i>	(+)2703C/T	<i>C</i>	rs4546	8	168 Phe/Phe
17	<i>CYP11B2</i>	(-)344UT T/C	<i>T</i>	rs1799998	8	5Flank
18	<i>Tp53</i>	(+)35946G/T	<i>G</i>	rs1802434	15	693 Leu/Leu

<sup>a</sup> dbSNP, double-strand SNP; NA, not applicable.

**Fig. 1** Incremental discriminating power of 98 single nucleotide polymorphisms (SNPs) using a naïve Bayes prediction algorithm with 174 breast cancer patients and 158 controls. This is the larger data set, in which roughly 1% of the SNP measurements were missing. *Permuted Label Prediction* shows the mean and SD of the performance of the naïve Bayes model on the real SNP data, but with the labels (breast cancer patient/control) permuted at random (see “Results”). —, 2 SDs; \*, naïve Bayes prediction; ○, permuted label prediction.



is sufficiently high to warrant the construction of an actual clinical model, one would then use *all* of the available data to train a final, usable model.

It is impossible to determine, *a priori*, which class of models is most appropriate for a data set. For the current study, three machine learning models, naïve Bayes, SVMs, and decision trees were applied to the SNP data to discriminate normal controls from female breast cancer patient samples. Naïve Bayes is one of the simplest classes of models; it assumes independence of each of the features (SNPs). SVM and decision trees can both create extremely rich, complex models that allow many interactions between the features. Each class of model can work well or perform poorly in different contexts. The models used are described in the “Discussion” section.

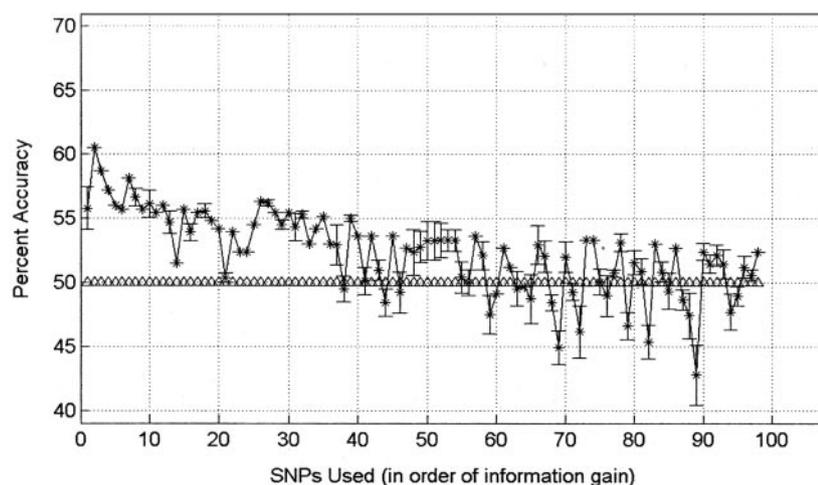
**Entire Data Set.** In the entire data set consisting of 174 breast cancer patients and 158 controls, 1.6% of breast cancer patient calls and 0.9% of control calls were missing because of poor signal:noise ratios in the genotyping assays. Because naïve Bayes naturally handles missing data, we first ran naïve Bayes on this entire data set. This allowed us to use all of our data and

to see how well we could do in the presence of missing data. Later we modified this data set to eliminate missing values.

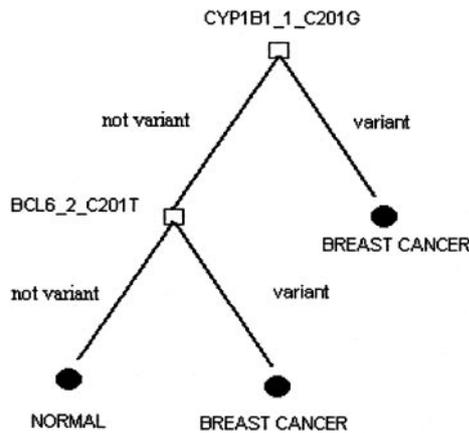
Twenty-fold cross-validation was used. In each fold, SNPs were incrementally selected based on their information gain values. Feature selection was performed once for each fold of the cross-validation rather than once for the whole data set so as not to bias the learner. Feature selection is part of training and, hence, must be performed inside the cross-validation loop. Because creation of cross-validation groups has a stochastic element, the 20-fold cross-validation was repeated five times. Results are reported as mean  $\pm$  SD. Results are shown graphically in Fig. 1.

Maximal performance was achieved using both 3 and 31 SNPs. The former led to a cross-validation accuracy of  $63 \pm 2\%$ , with  $67 \pm 2\%$  sensitivity and  $59 \pm 4\%$  specificity, whereas the latter led to a cross-validation accuracy of  $63 \pm 2\%$ , with  $58 \pm 2\%$  sensitivity and  $66 \pm 2\%$  specificity.

Feature selection was performed inside of each fold of the cross-validation and was, thus, performed 100 times (5 trials  $\times$  20 folds). Feature selection was stable across different folds and



**Fig. 2** Predictive accuracy for individual single nucleotide polymorphisms (SNPs), one at a time, using 174 breast cancer patients, 158 controls, and a naïve Bayes algorithm. —, Naïve Bayes prediction; \*, 2 SDs;  $\Delta$ , baseline.



**Fig. 3** Optimal decision tree as determined by 20-fold cross-validation over five trials. One can think of the decision tree as a series of ordered tests that one performs on a person to predict whether or not that person has cancer. The first test performed is the test at the root (*top*) of the tree, in this case, the single nucleotide polymorphisms (SNP)-type for *CYP11B1* + 4328C/G. If this SNP is variant, then one traverses the right side of the tree to a leaf node, which denotes what category the person falls into. In this case, if a person is variant for *CYP11B1* + 4328C/G, then the leaf node indicates that the model predicts the presence of cancer. Alternatively, if the first tests shows that the person is common homozygous or heterozygous for *CYP11B1* + 4328C/G, then one traverses the left side of the tree and finds that another test is needed before making a classification, namely, the SNP-type for *BCL6* + 4449C/T. The *BCL6* + 4449C/T test, in turn, leads to two leaf nodes, one predicting normal tissue, and the other, breast cancer, for the common homozygous/heterozygous (*left*) and variant (*right*) branches, respectively. In summary, this small decision tree leads to a very simple rule: if a person is variant for *CYP11B1* + 4328C/G or *BCL6* + 4449C/T, then predict that she has breast cancer; otherwise, predict that she does not.

trials. In 96 of 100 feature selections performed, the top three SNPs were *CYP11B2* + 4536T/C, *CYP11B1* + 4328C/G, and *BCL6* + 4449C/T, indicating a robust selection process. These three polymorphisms were also identified when the entire data set was used to rank the SNPs by information gain.

Naïve Bayes was also used on each individual SNP, one at a time, with 20-fold cross-validation and five trials. The maximum predictive accuracy reported was for *CYP11B1* + 4328C/G at  $61 \pm 4\%$ , with sensitivity  $71 \pm 1$  and specificity  $49 \pm 1$ . Results for each individual SNP are shown in Fig. 2.

To determine further whether our results were observed by chance, we also conducted a random permutation test for the naïve Bayes classifier. That is, we conducted 100 random trials in which each trial consisted of the following: (a) random permutation of the labels of the data (cancer/control) so that the labels no longer match the real data in any meaningful way; (b) running of the naïve Bayes classifier algorithm on the data with these random labels; and (c) assessment of the predictive performance. The results are shown in Fig. 1 and labeled “Permutated Label Predictions.” We see that these random data sets have predictive accuracy that is centered on the 50% line and that they are clearly well separated and below the results from the true label partition. Thus it is highly unlikely that the predictive results from the true labels could have arisen by chance alone. In the particular case of three SNPs, which pro-

duces our maximal predictive accuracy, only a single randomly permuted data set, of the 100 such sets, matches the mean value of 63% that the true data partition obtains.

**Smaller Data Set.** Whereas some algorithms such as naïve Bayes and decision trees are amenable to missing values, the missing values can have an adverse effect on the performance of the predictive model. Because SVMs do not naturally handle missing data, it was necessary either to impute missing values or to remove subjects with any missing data before comparing other algorithms to SVMs. We chose the latter so as not to depend on unknown characteristics of the missing data, such as whether or not the missing data are missing completely at random (as opposed, say, to being the result of some experimental bias). This removal of all persons with any missing data resulted in 63 breast cancer patients and 74 controls.

The data partitioning procedure used in the previous section for training and testing was also used with naïve Bayes and SVM (*i.e.*, 20-fold cross-validation, with incremental information gain feature selection, and five separate cross-validation trials). Because SVMs are computationally very intensive, rather than adding a single SNP at a time throughout, we added one SNP at a time until 15 SNPs, and then we increased the number by 5 SNPs at a time (still adding SNPs according to their individual information gain). In the earlier analysis, the critical number of SNPs was approximately three, justifying this approach. For decision trees, feature selection is an inherent part of the algorithm (39). As the tree is being built, features are chosen one at a time on the basis of information content relative to the target classes and the previous features that were selected. This is similar to ranking of features except that interactions between features are considered and can, therefore, be more powerful. SVMs are often touted as doing feature selection as an inherent part of the SVM algorithm. However, in our study, we found that adding an extra layer of feature selection on top of the SVM training algorithm was advantageous (*i.e.*, using the incremental addition of SNPs on the basis of information gain).

We recall that the naïve Bayes model with maximal performance used three SNPs and produced  $67 \pm 2\%$  accuracy, with  $54 \pm 2\%$  sensitivity and  $79 \pm 2\%$  specificity.

The SVMs with quadratic kernel performed better than the other kernels tried. It had maximal performance with the use of

**Table 3** Discrimination of breast cancer patients from normal controls using machine learning techniques. The mean and SD of five 20-fold cross-validation trials.

Algorithm	Maximal accuracy (%)	Sensitivity	Specificity	Number of SNPs <sup>a</sup> used for maximal accuracy
Naïve Bayes	$67 \pm 2$	$54 \pm 2\%$	$79 \pm 2\%$	3
Decision tree	$68 \pm 1$	$64 \pm 2\%$	$70 \pm 4\%$	2
SVM linear kernel	$62 \pm 2$	$57 \pm 2\%$	$57 \pm 2\%$	60
SVM quadratic kernel	$69 \pm 4$	$53 \pm 2\%$	$83 \pm 7\%$	3
SVM cubic kernel	$67 \pm 4$	$47 \pm 2\%$	$84 \pm 4\%$	3

<sup>a</sup> SNP, single nucleotide polymorphism; SVM, support vector machine.

Table 4 Single nucleotide polymorphisms (SNPs) with significant (95 or 99%) genotype odds ratio (OR)<sup>a</sup>  
 The “Sig” column indicates whether the particular genotype OR was significant. Significant results are shown in bold.

	SNP	Genotype	Control	Breast cancer	OR	95% CI <sup>b</sup>	Sig	99% CI	Sig
1	<i>CYP11B2</i> (+) <i>4536T/C</i> (+) <i>4536T/C</i>	1	114	99	1.00	(reference)		(reference)	
		2	42	48	1.32	0.80–2.16		0.69–2.52	
		3	0	19	<b>44.88</b>	<b>2.68–752.89</b>	Yes	<b>1.10–1826.23</b>	Yes
2	<i>CYP11B1</i> (+) <i>4328C/G</i>	1	77	50	1.00	(reference)		(reference)	
		2	56	78	<b>2.15</b>	<b>1.31–3.52</b>	Yes	<b>1.12–4.11</b>	Yes
		3	21	45	<b>3.30</b>	<b>1.76–6.19</b>	Yes	<b>1.44–7.54</b>	Yes
3	<i>BCL6</i> (+) <i>4449C/T</i>	1	67	82	1.00	(reference)		(reference)	
		2	81	60	<b>0.61</b>	<b>0.38–0.96</b>	Yes	0.33–1.11	
		3	10	28	<b>2.29</b>	<b>1.04–5.05</b>	Yes	0.89–6.47	
4	<i>CYP19A1</i> (+) <i>32123</i> (3'UT)	1	49	43	1.00	(reference)		(reference)	
		2	77	67	0.99	0.59–1.68		0.50–1.98	
		3	31	59	<b>2.17</b>	<b>1.19–3.94</b>	Yes	0.99–4.75	
5	<i>MLH1</i> (+) <i>18529A/G</i>	1	76	89	1.00	(reference)		(reference)	
		2	75	64	0.73	0.46–1.15		0.40–1.32	
		3	5	17	<b>2.90</b>	<b>1.02–8.24</b>	Yes	0.74–11.44	
6	<i>MSH6</i> (+) <i>12742T/C</i>	1	90	77	1.00	(reference)		(reference)	
		2	55	82	<b>1.74</b>	<b>1.10–2.75</b>	Yes	0.96–3.18	
		3	13	7	0.63	0.24–1.66		0.18–2.25	
7	<i>AGTR1</i> (+) <i>572C/T</i>	1	51	36	1.00	(reference)		(reference)	
		2	72	84	1.65	0.97–2.81		0.82–3.32	
		3	33	53	<b>2.28</b>	<b>1.24–4.18</b>	Yes	<b>1.02–5.07</b>	Yes
8	<i>RET</i> (+) <i>37412G/A</i>	1	116	109	1.00	(reference)		(reference)	
		2	32	54	<b>1.80</b>	<b>1.08–2.99</b>	Yes	0.92–3.51	
		3	9	5	0.59	0.19–1.82		0.13–2.59	
9	<i>CYP17</i> (+) <i>194G/T</i>	1	68	54	1.00	(reference)		(reference)	
		2	73	89	1.54	0.96–2.46		0.82–2.86	
		3	17	30	<b>2.22</b>	<b>1.11–4.45</b>	Yes	0.89–5.53	
10	<i>CD38</i> (+) <i>55806A/C</i>	1	138	163	1.00	(reference)		(reference)	
		2	19	8	<b>0.36</b>	<b>0.15–0.84</b>	Yes	0.12–1.10	
		3	1	1	0.85	0.05–13.66		0.02–32.73	
11	<i>ADPRT</i> (+) <i>22266T/C</i>	1	48	73	1.00	(reference)		(reference)	
		2	82	77	<b>0.62</b>	<b>0.38–0.99</b>	Yes	0.33–1.16	
		3	27	20	<b>0.49</b>	<b>0.25–0.96</b>	Yes		
12	<i>ERCC2</i> (+) <i>17966C/T</i>	1	90	77	1.00	(reference)		(reference)	
		2	53	80	<b>1.76</b>	<b>1.11–2.80</b>	Yes	0.96–3.24	
		3	14	17	1.42	0.66–3.07		0.52–3.90	
13	<i>CD68</i> (+) <i>1786G/A</i>	1	148	152	1.00	(reference)		(reference)	
		2	7	18	<b>2.50</b>	<b>1.02–6.17</b>	Yes	0.77–8.19	
		3	1	0	0.32	0.01–8.03		0.00–22.01	
14	<i>CYP11B1</i> (+) <i>28G/A</i>	1	134	161	1.00	(reference)		(reference)	
		2	23	13	<b>0.47</b>	<b>0.23–0.96</b>	Yes	0.18–1.21	
		3	1	0	0.28	0.01–6.87		0.00–18.83	
15	<i>CYP11B2</i> (+) <i>2703C/T</i>	1	34	57	1.00	(reference)		(reference)	
		2	95	87	<b>0.55</b>	<b>0.33–0.91</b>	Yes	0.28–1.07	
		3	29	29	0.60	0.31–1.16		0.25–1.43	
16	<i>CYP11B2</i> (–) <i>344 UT</i>	1	34	56	1.00	(reference)		(reference)	
		2	94	86	<b>0.56</b>	<b>0.33–0.93</b>	Yes	0.28–1.10	
		3	30	28	0.57	0.29–1.11		0.24–1.36	
17	<i>Tp53</i> (+) <i>35946G/T</i>	1	102	128	1.00	(reference)		(reference)	
		2	50	35	<b>0.56</b>	<b>0.34–0.92</b>	Yes	0.29–1.08	
		3	6	6	0.80	0.25–2.54		0.17–3.67	

<sup>a</sup> 1, common homozygous; 2, heterozygous; 3, variant.

<sup>b</sup> CI, confidence interval; Sig, significant.

Table 5 Single nucleotide polymorphisms (SNPs) with significant allele (95 or 99%) odds ratio (OR)  
The "Sig" column indicates whether the particular allele OR was significant. Significant results are shown in bold.

	SNP	Allele	Control	Breast cancer	OR	95% CI <sup>a</sup>	Sig	99% CI	Sig
1	<i>CYP11B2</i> (+)4536T/C	N	270	246	1.00	(reference)		(reference)	
		V	42	86	<b>2.25</b>	<b>1.50–3.38</b>	Yes	<b>1.32–3.84</b>	Yes
2	<i>CYP1B1</i> (+)4328C/G	N	210	178	1.00	(reference)		(reference)	
		V	98	168	<b>2.02</b>	<b>1.47–2.78</b>	Yes	<b>1.33–3.08</b>	Yes
3	<i>CYP19A1</i> (+)32123 (3'UT)	N	175	153	1.00	(reference)		(reference)	
		V	139	185	<b>1.52</b>	<b>1.12–2.07</b>	Yes	<b>1.01–2.28</b>	Yes
4	<i>CYP11B2</i> (+)5215G/A	N	286	331	1.00	(reference)		(reference)	
		V	28	13	<b>0.40</b>	<b>0.20–0.79</b>	Yes	<b>0.16–0.98</b>	Yes
5	<i>AGTR1</i> (+)572C/T	N	174	156	1.00	(reference)		(reference)	
		V	138	190	<b>1.54</b>	<b>1.13–2.09</b>	Yes	<b>1.02–2.30</b>	Yes
6	<i>CYP17</i> (+)194G/T	N	209	197	1.00	(reference)		(reference)	
		V	107	149	<b>1.48</b>	<b>1.08–2.03</b>	Yes	0.98–2.24	
7	<i>CD38</i> (+)55806A/C	N	295	334	1.00	(reference)		(reference)	
		V	21	10	<b>0.42</b>	<b>0.19–0.91</b>	Yes	0.15–1.16	
8	<i>ADPRT</i> (+)22266T/C	N	178	223	1.00	(reference)		(reference)	
		V	136	117	<b>0.69</b>	<b>0.50–0.94</b>	Yes	0.45–1.04	
9	<i>CYP11B1</i> (+)28G/A	N	291	335	1.00	(reference)		(reference)	
		V	25	13	<b>0.45</b>	<b>0.23–0.90</b>	Yes	0.18–1.12	

<sup>a</sup> CI, confidence interval; Sig, significant; N, common; V, variant; UT, untranslated.

three SNPs and produced  $69 \pm 4\%$  accuracy, with  $53 \pm 2\%$  sensitivity and  $83 \pm 7\%$  specificity. The use of a linear kernel resulted in maximal performance using 60 SNPs with  $62 \pm 2\%$  accuracy, with  $57 \pm 2\%$  sensitivity and  $67 \pm 2\%$  specificity. The use of a cubic kernel had maximal performance using three SNPs and produced  $67 \pm 4\%$  accuracy, with  $47 \pm 2\%$  sensitivity and  $84 \pm 4\%$  specificity.

For both naïve Bayes and SVMs, the same feature selection method was used (ranking with information gain). In more than 90 of 100 of the feature selections performed, the top three SNPs identified using each of the algorithms were the same as in the previous section in which the entire data set was used: *CYP11B2* + 4536T/C, *CYP1B1* + 4328C/G, and *BCL6* + 4449C/T.

The decision tree with maximal performance used two SNPs (*CYP1B1* + 4328C/G and *BCL6* + 4449C/T), achieving  $68 \pm 1\%$  accuracy, with  $64 \pm 2\%$  sensitivity and  $70 \pm 4\%$  specificity. A graphical picture of the tree is shown in Fig. 3. Results for all algorithms are shown in Table 3.

As an added measure of rigor, permutation tests were applied to the quadratic kernel SVM classifier with the use of three SNPs. The labels of the data (cancer or normal) were randomly permuted, then the three-SNP, quadratic kernel classifier algorithm was run and a model was built in an identical manner to that used with the real data labels. This was repeated 100 times. No random permutation of the labels was able to tie or outperform the mean accuracy of 69% reported above (for three SNPs, quadratic SVM). Average prediction accuracy over 100 trials was 50% with SD of 6.6%.

#### Genotype Odds Ratio and Frequency of Genotypes.

SNP studies often report results in the form of odds ratios for individual SNPs in relation to the presence or absence of a

disease (41, 42). Whereas information gain provides a summary statistic of all genotypes for a particular SNP, odds ratios break this information down into individual genotypes. Table 4 shows odds ratios for all SNPs with at least one genotype (heterozygous or variant) the odds ratio of which, relative to the common homozygous genotype, deviates from unity at a minimum of a 95% significance. Both 95% and 99% confidence intervals, not adjusted for multiple comparisons, are also shown. Table 5 is the same as Table 4 but shows odds ratios for allele frequencies rather than genotype frequencies.

In Table 6 we report the frequency and odds ratio of all occurring genotypes specified by the three SNPs found to be most important for classification in the machine learning section, *CYP11B2* + 4536T/C, *CYP1B1* + 4328C/G, and *BCL6* + 4449C/T. The odds ratio is reported relative to the homozygous common genotype as defined by the control population in this study.

## DISCUSSION

Human genome analysis and high-throughput techniques have spawned a mass of complex, biological data. Analysis of these data creates the bottleneck of many studies at present. Whereas these data are unwieldy, seemingly intractable, and not amenable to traditional methods of statistical analysis, the data are well suited to the application of machine learning algorithms. These algorithms are designed to tease out a variety of patterns, both linear and nonlinear, from large, noisy, and complex data sets that may also contain a great deal of irrelevant information. Traditionally seen in the context of microarray analysis, DNA sequence analysis, protein function, and structure

**Table 6** Frequency of genotypes resulting from single nucleotide polymorphisms CYP11B2 +4536 T/C, CYP1B1 +4328C/G and BCL6 +4449C/T<sup>a</sup>

Total of 161 Breast Cancer and 152 Control (genotypes containing a “no call” were omitted). Odds ratios (ORs) are reported relative to the “normal” genotype of “111.”

Genotype	Control	Breast cancer	OR	95% CI <sup>b</sup>	Sig	99% CI	Sig
113	4	3	1.45	0.29–7.34		0.17–12.21	
213	1	2	3.87	0.32–46.18		0.15–100.66	
313	0	2	9.52	0.43–210.81		0.16–558.05	
<b>123</b>	<b>1</b>	<b>9</b>	<b>17.40</b>	<b>2.01–150.57</b>	Yes	<b>1.02–296.64</b>	Yes
223	1	4	7.73	0.79–75.47		0.39–154.42	
133	2	1	0.97	0.08–11.54		0.04–25.17	
233	1	4	7.73	0.79–75.47		0.39–154.42	
333	0	2	9.52	0.43–210.81		0.16–558.05	
112	21	10	0.92	0.35–2.45		0.25–3.33	
212	11	5	0.88	0.26–3.00		0.18–4.41	
312	0	1	5.71	0.22–148.61		0.08–413.80	
122	23	18	1.51	0.63–3.64		0.48–4.79	
222	10	9	1.74	0.58–5.20		0.41–7.34	
322	0	4	17.13	0.87–339.17		0.34–866.71	
132	9	4	0.86	0.23–3.26		0.15–4.95	
232	4	4	1.93	0.42–8.84		0.26–14.24	
332	0	1	5.71	0.22–148.61		0.08–413.80	
<i>111</i>	29	15	1.00	(reference)			
211	9	7	1.50	0.47–4.84		0.32–6.98	
311	0	3	13.32	0.65–274.72		0.25–711.02	
121	18	17	1.83	0.74–4.54		0.55–6.04	
<b>221</b>	<b>3</b>	<b>7</b>	<b>4.51</b>	<b>1.02–20.00</b>	Yes	0.64–31.94	
321	0	3	13.32	0.65–274.72		0.25–711.02	
<b>131</b>	<b>5</b>	<b>18</b>	<b>6.96</b>	<b>2.16–22.44</b>	Yes	<b>1.49–32.41</b>	Yes
231	0	5	20.94	1.09–403.86		0.43–1023.58	
331	0	3	13.32	0.65–274.72		0.25–711.02	

<sup>a</sup> 1, common homozygous; 2, heterozygous; 3, variant. Genotype = “123” means that CYP11B2 +4536T/C = 1, CYP1B1 +4328C/G = 2, and BCL6 +4449C/T = 3. Genotype = “323” means that CYP11B2 +4536T/C = 3, CYP1B1 +4328C/G = 2, and BCL6 +4449C/T = 3.

<sup>b</sup> CI, confidence interval; Sig, significant.

prediction, the machine learning algorithms have now been applied to SNP data.

**Description of Algorithms.** Naïve Bayes is a simple model that uses the frequencies of different values of each feature, within known classes, to predict the class of a new sample with specified features but no label. It provides a probabilistic framework that assumes that each feature is independent from every other feature, given the class. Although this assumption is typically false, naïve Bayes has been found to work well in practice. Naïve Bayes is generally used as a first pass “naïve” attempt at solving a classification problem. Very simply, naïve Bayes tabulates the number of times a particular SNP occurs as common homozygous, heterozygous, or variant within one population (say, cancer). This directly provides probabilities of the form  $p(\text{SNP} = \text{heterozygous} | \text{class} = \text{cancer})$ , called the class conditional probabilities. To classify a new example, one uses Bayes Rule:

$$p(\text{class} = Y | \text{data} = X) = \frac{p(\text{data} = X | \text{class} = Y)p(Y)}{p(X)}$$

with the assumption that the SNPs are independent,

$$\begin{aligned} p(\text{SNP}_1 = x, \text{SNP}_2 = y, \dots, \text{SNP}_n = z | \text{class} = Y) \\ = p(\text{SNP}_1 = x | \text{class} = Y)p(\text{SNP}_2 = y | \text{class} = Y)p(\text{SNP}_3 \\ = z | \text{class} = Y) \end{aligned}$$

to obtain class probabilities. The class with the higher probability is the one to which the new example is classified.  $p(X)$  need never be computed because it maintains the same value as we change the class,  $Y$ .  $p(Y)$  is simply the probability that a sample came from a particular class, say cancer and can be computed from the relative proportion of samples in the data, or directly set to some known value (e.g., it may be known that in the general population that 5% of persons have cancer).

The decision tree models patterns by examining a single feature at a time in a hierarchical manner, typically including features on the basis of information content related to the desired classification. For example, in the given context, the building of the decision tree (using only training data) would start by finding the single SNP that was most discriminative for classifying cancer *versus* control. This would be at the “root” of the tree (see, e.g., Fig. 3). Next, for each of the possible results of ‘traversing’ this ‘root’ (e.g., go right if the SNP for the given example is variant; to left, otherwise), the same idea is applied again: find the SNP that is the most discriminative for the examples that have traversed to this part of the tree. This criterion is repeatedly applied, each time adding a new “node” (SNP) to the tree. A decision tree also has “leaf nodes,” which, in the present context, would be SNPs for which no tree exists below them. Once an example has traversed to a leaf node, the example is classified as belonging to the class for which the majority of the examples that end up at that leaf node belong.

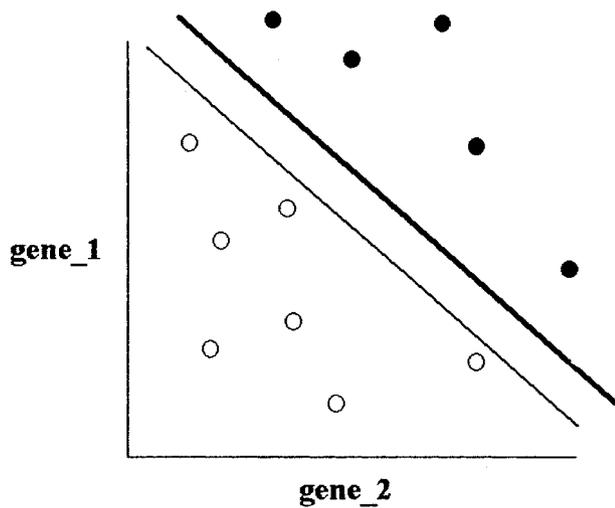


Fig. 4 Representation of a support vector machine (SVM) analysis with a linear kernel using only two features (e.g., transcript levels for two genes, each plotted on one axis), for which the data are immediately separable by a line. The thicker separating line is the one that lies farthest from the two classes (i.e., has the largest margin). The other (thinner) line is a smaller margin and, thus, likely has a weaker ability to predict the class of new persons. ●, cancer; ○, normal.

When building a decision tree model, the building phase of the tree can be stopped using a variety of criteria, such as that a certain maximum number of leaf nodes exist, or that each leaf node must contain at least some minimum number of examples. Additionally, with some algorithms, the tree is pruned back after construction to make sure that the model is not overfitting to noise in the data set. Because the decision tree chooses only one SNP at a time, starting with the root, and never changes any nodes, the optimal sequence of SNPs for prediction may not be chosen.

SVMs extend the notion of a simple linear classifier (e.g., Fisher's linear discriminant) to more complex classifiers by projecting the input data into a user-selected, higher-dimensional space (the space is determined by the choice of 'kernel'). SVMs treat the input data (e.g., SNP values for one person) as continuous values rather than ordinal or discrete. Although this may not always make intuitive sense (e.g., is a common homozygote really a specific amount "larger" than a variant homozygote, or *vice versa*?), it can nevertheless prove powerful in practice. The simplest SVM is one with a linear kernel. Suppose the data had only two features (e.g., transcript levels for two genes; we use this example at this point for illustrative purposes because transcript level are naturally continuous valued variables), measured over many controls and many cancer patients. Then one could plot the data in two dimensions (an example of how this might look is shown in Fig. 4). For this example (Fig. 4), the data can be separated by a straight line, and hence a linear kernel, implying no transformation of the data, is appropriate. In circumstances in which there is no straight line that can separate the two classes, such as illustrated in Fig. 5, a more powerful model is required. With SVMs, this more powerful model is created by modifying the input space. For example, a quadratic

kernel would convert the two-dimensional data points to a three-dimensional space as follows:  $\{gene_1, gene_2\} \rightarrow \{gene_1 \times gene_1, gene_1 \times gene_2, gene_2 \times gene_2\}$ . The SVM would attempt to partition the cancer and control data points in this new space using a hyperplane (a line in more than two dimensions). Clearly the choice of kernel is very important with SVMs. Changing the kernel changes the data transformation, which, in turn, dictates whether a line can be used to separate the data in this new space. With the data shown in Fig. 5, a quadratic transformation turns out to be a suitable one, whereby the data in the new quadratic space can be perfectly separated with a line. In addition to their ability to model complex patterns by changing the input space, SVMs are said to have good generalization bounds because of the principle of "margin maximization," which is at the core of their theoretical development. Generalization refers to the ability of a learned model to generalize to new data (i.e., will it work well on unseen data). The principle of margin maximization states that of all of the linear classifiers that can separate the input data, one should choose the one which lies farthest from all of the training points. For example, in Fig. 4, two lines are shown that separate the data, but one is very close to the boundary of one of the classes. The line that is very close to one of the classes will likely have a weaker ability to predict new examples according to the theory of SVMs.

All three of these algorithms use supervised learning in which the algorithm is told the actual outcome (e.g., whether this patient had cancer or not) during construction of the model. The learned system then predicts the outcome of a sample, given only the feature values and not the target class. Many machine learning methods, including those used in the present study, are related to more traditional statistical methods, such as Fisher's linear discriminant analysis, quadratic discriminant analysis, and logistic regression.

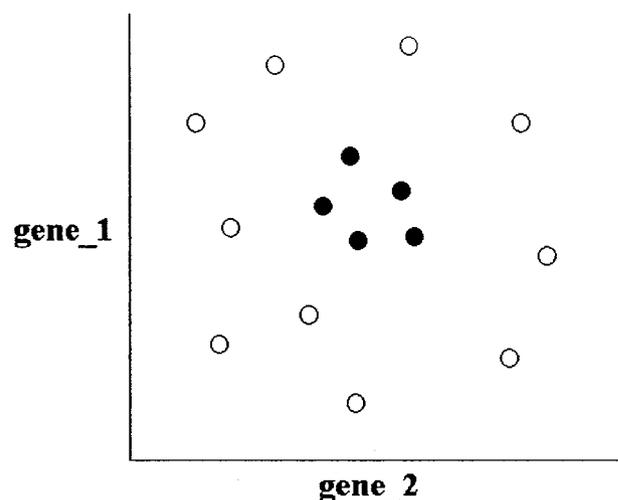


Fig. 5 An example of illustrative data points with only two features (e.g., transcript levels for two genes, each plotted on one axis), for which the data are not immediately separable by a line. To fix this problem, the input data must be transformed into a different space in which it will be linearly separable. ●, cancer; ○, normal.

**Comparison of Algorithm Results.** With the predictive models, we found that the use of the whole data set, including patients with some missing SNP calls, provided a naïve Bayes predictive power of 63%, compared with a baseline of 50%. By pruning the data set down to only complete patient genotypes, this naïve Bayes accuracy was increased to 67%, and further to 69% by using a quadratic kernel SVM. Overall, the three learning algorithms of naïve Bayes, SVM, and decision tree all performed quite similarly. The decision tree had more balanced errors than the other models in that errors occurred more evenly in the prediction of both cancer and noncancerous persons (*i.e.*, the disparity between sensitivity and specificity was less than for other models). The best predictive accuracy from a single SNP using naïve Bayes provided only 61% accuracy. These results illustrate the value of predictive models of breast cancer built from multiple SNP determinations over the whole genome. We anticipate that this may ultimately lead to a useful clinical tool.

**Discussion of Individual SNPs.** About 10% of breast cancers cluster in families, with approximately one-fifth associated with heterozygous germ-line mutations in either the *BRCA1* or the *BRCA2* gene (27, 28, 43). Much smaller proportions are due to germ-line abnormalities in other genes such as the check point kinase *CHEK2* (44), *p53* (45), and the *PTEN* phosphatase gene mutated in Cowden disease (41, 41, 46). Other genetic determinants of familial breast cancer are thought to exist, although they are yet elusive (47).

We have shown that polymorphisms in *CYP11B2* and *CYP11B1*, which are important regulators of steroid metabolism, identify patients with breast cancer. *CYP11B2* steroid hydroxylase catalyzes the final step in aldosterone synthesis. Although cytosine at a polymorphic site within the promoter region at position -344 is associated with essential hypertension (48), coding region variants have not yet been shown to have medical relevance. A polymorphic site at position +1157 (C/T) has been described within the second position of codon 386 that specifies Ala or Val (49). We have shown that the homozygous variant allele at position +4536C/T was the strongest discriminator, as defined by information gain, among 98 SNPs studied in breast cancer and normal cases.

The *CYP11B1*:1A1 activity ratio is a critical determinant of the metabolism and toxicity of estradiol in mammary cells (50). Xenooestrogens, such as the environmental contaminant dioxin alter this ratio, upsetting the metabolism and detoxification of 17  $\beta$ -estradiol (50). We show that Val at position +4328 in *CYP11B1* rather than Leu, is more often observed in breast cancer cases compared with controls, with an odds ratio of 3.3 (99% CI, 1.44–7.54) for the G/G genotype versus the C/C. Other studies have shown that polymorphisms at position +354G/T in codon 119 Ala/Ser of this gene can predict prostate cancer risk with an odds ratio of 4.02 observed in those men having the T/T genotype versus G/G (51). These observations suggest that allelic variation in enzymes metabolizing xenobiotics can affect the carcinogenic effects of endogenous and exogenous sex hormones, affecting cancer risk.

Cytochrome P450 19A1 catalyzes the aromatization of androgenic steroids into estrogens and is etiologically important to postmenopausal breast cancer (52). Aromatase inhibitors are important therapies for postmenopausal breast cancer (53). We

have identified a polymorphism within the first noncoding exon of *CYP19A1* that is predictive of breast cancer risk (double-break SNP rs10046). In our study the presence of T rather than C provides an OR of 1.52 (95% CI, 1.12–2.07). This suggests that, in combination with other steroid hormone metabolizing enzymes, *CYP19A1* may be an important determinant of breast cancer risk.

Hereditary cancer can be caused by mutations in DNA repair enzymes. For instance, breast cancer susceptibility can be caused by mutations in the DNA repair enzymes *BRCA1* and *BRCA2*, whereas abnormalities in the human mismatch repair genes *MSH2* and *MLH1* are linked to hereditary nonpolyposis colorectal cancer (HNPCC). Mutations in *MSH6*, which is found in a complex with *MSH2* and the proliferating cell nuclear antigen, may be implicated in HNPCC of early onset (54–57). We show here that the *MLH1* polymorphism +18529A/G (double-break SNP ID rs1799977), which alters codon 219 to Val from Ile, is associated with breast cancer. The variant homozygous genotype of *MLH1* + 18529A/G is associated with breast cancer with an odds ratio of 2.90 (95% CI, 1.02–8.24). *MLH1* codon 219 is found within the DNA binding region of this mismatch repair enzyme.

*BCL6* is a pox virus and zinc fingers-domain containing transcriptional repressor often rearranged in B cell lymphoma (58). Through repression of gene expression it can control differentiation leading to malignancies of germinal center lymphocytes. There are no reported associations of *BCL6* with breast cancer, although, mechanistically, gene expression in breast tissue may contribute to disease in combination with other risk factors. We demonstrate that the +4449C/T polymorphic site can discriminate between women with breast cancer and those without the disease. The CC genotype specifies a 2.29 odds ratio compared with the TT genotype (95% CI, 1.04–5.05).

Through large scale measurement of SNPs, we have shown that the use of multiple SNPs together, through the use of machine learning algorithms, can achieve significantly better predictive power than any one SNP alone. This is a crucial step away from the traditional methods of looking at single SNP associations, thereby allowing incorporation of disparate biological mechanisms into a single classifier, as well as multifactorial combinations of SNPs that, together, form a single biological mechanism. We have also identified statistically significant differences between women with breast cancer and normal controls. Identified differences are found in genes known to increase the risk for hereditary cancers and an enzyme known to function in estrogen metabolism. If validated, these results indicate the feasibility of premorbid genetic predictive testing and guide the development of rational targeted intervention to interfere with the process of carcinogenesis. For example, the data suggest that aromatase enzyme inhibitors might be most effective for breast cancer chemoprevention in women with risk-associated *CYP19A1* alleles. PolyomX is currently undertaking an assembly of SNP data from a large, independent population to validate the results presented in this report.

## ACKNOWLEDGMENTS

We thank Kathryn Calder and Edith Pituskin for cancer informatics assistance and Drs. Carol Cass and Stephan Gabos for helpful discussions.

## REFERENCES

- Kinzler KW, Vogelstein B. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature (Lond)* 1997;386:761, 763.
- Kerr P, Ashworth A. New complexities for BRCA1 and BRCA2. *Curr Biol* 2001;11:R668–76.
- Savitsky K, Bar-Shira A, Gilad S, et al. A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science (Wash DC)* 1995; 268:1749–53.
- Stewart GS, Maser RS, Stankovic T, et al. The DNA double-strand break repair gene hMRE11 is mutated in individuals with an ataxia-telangiectasia-like disorder. *Cell* 1999;99:577–87.
- Ellis NA, Groden J, Ye TZ, et al. The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* 1995;83:655–66.
- Carney JP, Maser RS, Olivares H, et al. The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. *Cell* 1998;93: 477–86.
- Weeda G, van Ham RC, Vermeulen W, Bootsma D, van der Eb AJ, Hoeijmakers JH. A presumed DNA helicase encoded by ERCC-3 is involved in the human repair disorders xeroderma pigmentosum and Cockayne's syndrome. *Cell* 1990;62:777–91.
- Weber TK, Conlon W, Petrelli NJ, et al. Genomic DNA-based hMSH2 and hMLH1 mutation screening in 32 Eastern United States hereditary nonpolyposis colorectal cancer pedigrees. *Cancer Res* 1997; 57:3798–803.
- Shin KH, Shin JH, Kim JH, Park JG. Mutational analysis of promoters of mismatch repair genes hMSH2 and hMLH1 in hereditary non-polyposis colorectal cancer and early onset colorectal cancer patients: identification of three novel germ-line mutations in promoter of the hMSH2 gene. *Cancer Res* 2002;62:38–42.
- Malkin D, Li FP, Strong LC, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science (Wash DC)* 1990;250:1233–8.
- Shewetta SA. Drug-metabolizing enzymes: mechanisms and functions. *Curr Drug Metab* 2000;1:107–32.
- da Fonte de Amorim L, Rossini A, Mendonca G, et al. CYP1A1, GSTM1, and GSTT1 polymorphisms and breast cancer risk in Brazilian women. *Cancer Lett* 2002;181:179–86.
- Wu MS, Chen CJ, Lin MT, et al. Genetic polymorphisms of cytochrome P450 2E1, glutathione S-transferase M1 and T1, and susceptibility to gastric carcinoma in Taiwan. *Int J Colorectal Dis* 2002; 17:338–43.
- Goode EL, Dunning AM, Kuschel B, et al. Effect of germ-line genetic variation on breast cancer survival in a population-based study. *Cancer Res* 2002;62:3052–7.
- Brennan P. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis (Lond)* 2002;23:381–7.
- Tayeb MT, Clark C, Sharp L, et al. CYP3A4 promoter variant is associated with prostate cancer risk in men with benign prostate hyperplasia. *Oncol Rep* 2002;9:653–5.
- Xu J, Zheng SL, Turner A, et al. Associations between hOGG1 sequence variants and prostate cancer susceptibility. *Cancer Res* 2002; 62:2253–7.
- Lesueur F, Corbex M, McKay JD, et al. Specific haplotypes of the RET proto-oncogene are over-represented in patients with sporadic papillary thyroid carcinoma. *J Med Genet* 2002;39:260–5.
- Wiley JS, Dao-Ung LP, Gu BJ, et al. A loss-of-function polymorphic mutation in the cytolytic P2X7 receptor gene and chronic lymphocytic leukaemia: a molecular study. *Lancet* 2002;359:1114–9.
- Bharaj BB, Luo LY, Jung K, Stephan C, Diamandis EP. Identification of single nucleotide polymorphisms in the human kallikrein 10 (KLK10) gene and their association with prostate, breast, testicular, and ovarian cancers. *Prostate* 2002;51:35–41.
- Wang L, Habuchi T, Takahashi T, et al. Cyclin D1 gene polymorphism is associated with an increased risk of urinary bladder cancer. *Carcinogenesis (Lond)* 2002;23:257–64.
- Biros E, Kalina I, Biros I, et al. Polymorphism of the p53 gene within the codon 72 in lung cancer patients. *Neoplasma* 2001;48: 407–11.
- Zhu Y, Spitz MR, Lei L, Mills GB, Wu X. A single nucleotide polymorphism in the matrix metalloproteinase-1 promoter enhances lung cancer susceptibility. *Cancer Res* 2001;61:7825–9.
- Kumar R, Smeds J, Berggren P, et al. A single nucleotide polymorphism in the 3' untranslated region of the CDKN2A gene is common in sporadic primary melanomas but mutations in the CDKN2B, CDKN2C, CDK4 and p53 genes are rare. *Int J Cancer* 2001;95:388–93.
- Hemminki K, Shields PG. Skilled use of DNA polymorphisms as a tool for polygenic cancers. *Carcinogenesis (Lond)* 2002;23:379–80.
- Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
- Rebbeck TR. The contribution of inherited genotype to breast cancer. *Breast Cancer Res* 2002;4:85–9.
- Turchetti D, Cortesi L, Federico M, Romagnoli R, Silingardi V. Hereditary risk of breast cancer: not only BRCA. *J Exp Clin Cancer Res* 2002;21:17–21.
- Rebbeck TR. Inherited genetic predisposition in breast cancer. a population-based perspective. *Cancer (Phila)* 1999;86:2493–501.
- Kokoris M, Dix K, Moynihan K, et al. High-throughput SNP genotyping with the Masscode system. *Mol Diagn* 2000;5:329–40.
- Breen G, Harold D, Ralston S, Shaw D, St Clair D. Determining SNP allele frequencies in DNA pools. *Biotechniques* 2000;28:464–6, 468, 470.
- Cover TM, Thomas JA. Elements of information theory. New York: John Wiley; 1991.
- Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 2001;344:539–48.
- Ben-Dor, Amir, Friedman N, Yakhini Z. Scoring genes for relevance. Technical report AGL-2000, Agilent Technologies. Palo Alto, CA: Agilent Technologies; 2000.
- van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature (Lond)* 2002;415:530–6.
- Olshen AB, Jain AN. Deriving quantitative conclusions from microarray expression data. *Bioinformatics* 2002;18:961–70.
- Duda RO, Hart PE. Pattern classification and scene analysis. New York: John Wiley and Sons; 1973.
- Cristianini N, Shawe-Taylor J. An introduction to support vector machines (and other kernel-based learning methods). Cambridge: Cambridge University Press; 2000.
- Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton, FL: CRC Press; 1995.
- Joachims T. Making large-scale SVM learning practical. Cambridge: Massachusetts Institute of Technology Press; 1999.
- Becker N, Nieters A, Rittgen W. Single nucleotide polymorphism—disease relationships: statistical issues for the performance of association studies. *Mutat Res* 2003;525:11–8.
- Tanaka Y, Sasaki M, Kaneuchi M, Shiina H, Igawa M, Dahiya R. Polymorphisms of the CYP1B1 gene have higher risk for prostate cancer. *Biochem Biophys Res Commun* 2002;296:820–6.
- Schwab M, Claas A, Savelyeva L. BRCA2: a genetic risk factor for breast cancer. *Cancer Lett* 2002;175:1–8.
- Meijers-Heijboer H, van den Ouweland A, Klijn J, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(\*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* 2002;31:55–9.
- Lehman TA, Haffty BG, Carbone CJ, et al. Elevated frequency and functional activity of a specific germ-line p53 intron mutation in familial breast cancer. *Cancer Res* 2000;60:1062–9.
- Carroll BT, Couch FJ, Rebbeck TR, Weber BL. Polymorphisms in PTEN in breast cancer families. *J Med Genet* 1999;36:94–6.
- Peto J. Breast cancer susceptibility—a new look at an old model. *Cancer Cell* 2002;1:411–2.

48. Tsukada K, Ishimitsu T, Teranishi M, et al. Positive association of CYP11B2 gene polymorphism with genetic predisposition to essential hypertension. *J Hum Hypertens* 2002;16:789–93.
49. Halushka MK, Fan JB, Bentley K, et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 1999;22:239–47.
50. Coumoul X, Diry M, Robillot C, Barouki R. Differential regulation of cytochrome P450 1A1 and 1B1 by a combination of dioxin and pesticides in the breast tumor cell line MCF-7. *Cancer Res* 2001;61:3942–8.
51. Tanaka Y, Sasaki M, Kaneuchi M, Shiina H, Igawa M, Dahiya R. Polymorphisms of the CYP1B1 gene have higher risk for prostate cancer. *Biochem Biophys Res Commun* 2002;296:820–6.
52. Meinhardt U, Mullis PE. The essential role of the aromatase/p450arom. *Semin Reprod Med* 2002;20:277–84.
53. Haiman CA, Hankinson SE, De Vivo I, et al. Polymorphisms in steroid hormone pathway genes and mammographic density. *Breast Cancer Res Treat* 2003;77:27–36.
54. Kariola R, Raevaara TE, Lonnqvist KE, Nystrom-Lahti M. Functional analysis of MSH6 mutations linked to kindreds with putative hereditary non-polyposis colorectal cancer syndrome. *Hum Mol Genet* 2002;11:1303–10.
55. Charames GS, Millar AL, Pal T, Narod S, Bapat B. Do MSH6 mutations contribute to double primary cancers of the colorectum and endometrium? *Hum Genet* 2000;107:623–9.
56. Verma L, Kane MF, Brassett C, et al. Mononucleotide microsatellite instability and germline MSH6 mutation analysis in early onset colorectal cancer. *J Med Genet* 1999;36:678–82.
57. Flores-Rozas H, Clark D, Kolodner RD. Proliferating cell nuclear antigen and Msh2p-Msh6p interact to form an active mismatch recognition complex. *Nat Genet* 2000;26:375–8.
58. Staudt LM, Dent AL, Shaffer AL, Yu X. Regulation of lymphocyte cell fate decisions and lymphomagenesis by BCL-6. *Int Rev Immunol* 1999;18:381–403.