BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i508–i518
doi:10.1093/bioinformatics/btm207

From protein microarrays to diagnostic antigen discovery: a study of the pathogen *Francisella tularensis*

Suman Sundaresh[1,2], Arlo Randall[1,2], Berkay Unal[3], Jeannine M. Petersen[4], John T. Belisle[5], M. Gill Hartley[6], Melanie Duffield[6], Richard W. Titball[6], D. Huw Davies[3], Philip L. Felgner[3] and Pierre Baldi[1,2,*]

[1]School of Information and Computer Sciences, [2]Institute for Genomics and Bioinformatics, [3]Center for Virus Research, University of California, Irvine, CA, [4]Centers for Disease Control and Prevention, [5]Mycobacteria Research Laboratories, Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO, USA and [6]Defence Science and Technology Laboratory, Porton Down, UK

ABSTRACT

Motivation: An important application of protein microarray data analysis is identifying a serodiagnostic antigen set that can reliably detect patterns and classify antigen expression profiles. This work addresses this problem using antibody responses to protein markers measured by a novel high-throughput microarray technology. The findings from this study have direct relevance to rapid, broad-based diagnostic and vaccine development.

Results: Protein microarray chips are probed with sera from individuals infected with the bacteria *Francisella tularensis*, a category A biodefense pathogen. A two-step approach to the diagnostic process is presented (1) feature (antigen) selection and (2) classification using antigen response measurements obtained from *F.tularensis* microarrays (244 antigens, 46 infected and 54 healthy human sera measurements). To select antigens, a ranking scheme based on the identification of significant immune responses and differential expression analysis is described. Classification methods including *k*-nearest neighbors, support vector machines (SVM) and *k*-Means clustering are applied to training data using selected antigen sets of various sizes. SVM based models yield prediction accuracy rates in the range of ~90% on validation data, when antigen set sizes are between 25 and 50. These results strongly indicate that the top-ranked antigens can be considered high-priority candidates for diagnostic development.

Availability: All software programs are written in R and available at http://www.igb.uci.edu/index.php?page=tools and at http://www.r-project.org

Contact: pfbaldi@uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A fundamental problem in disease diagnostics is to identify a serodiagnostic antigen set whose expression profiles can be used to reliably diagnose infectious disease and potentially form the basis for the development of vaccines against that disease. The need to identify these antigens is further heightened by the urgency for more rapidly assessing the prevalence and spread of infectious diseases due to the emergence of new strains of infectious organisms and the recent concerns related to bioterrorism.

In response to these needs, a novel high-throughput approach has been developed to rapidly convert genome sequence information from infectious bacteria and viruses into the proteins that are encoded by each of the genes (Davies *et al.*, 2005a; Sundaresh *et al.*, 2006). This approach enables fast, comprehensive, and high-throughput analysis of immune responses to infectious disease antigens that can be applied to the discovery and development of serodiagnostic tests. All of the individual proteins from an infectious microorganism are printed onto a microarray chip and the chip is probed with sera from vaccinated or infected humans and animals and the antibody reactivity against each antigen can be quantified to obtain immunodominant antigen profiles. Thus, the proteome microarrays can be used to interrogate the *entire* proteome of any infectious microorganism, potentially comprising thousands of antigens each. Once the microarrays are fabricated they can be produced in large numbers, enabling large numbers of sera to be conveniently probed while consuming small quantities of individual sera ($<2\,\mu$l/test). To the best of our knowledge, no other method of serodiagnostic antigen discovery can quantitatively and comprehensively interrogate the humoral immune response on an antigen specific basis against bacteria, parasites and viral proteomes with large numbers of individual patients' sera and with comparable accuracy, efficiency and speed.

In particular, this study analyzes the antigen expression profiles for *Francisella tularensis*, which is the etiological agent of tularemia, a serious and sometimes fatal disease of humans and animals (Ellis *et al.*, 2002; Isherwood *et al.*, 2005; Larsson *et al.*, 2005). It is a category A biodefense pathogen and there is concern over its illegitimate use as an agent of bioterrorism or biological warfare (Dennis *et al.*, 2001). Until recently, little was known of the genetic makeup of this bacterium. This was resolved by the determination of the genome sequence of

*To whom correspondence should be addressed.

F.tularensis strain Schu S4 (Karlsson *et al*., 2000; Prior *et al*., 2001).

Among the National Institutes of Health Biodefense research goals is the objective to identify new *F.tularensis* vaccine candidates that can prevent or modulate infection both before and after exposure. In addition, the NIAID Biodefense Effort has called for initiatives to identify and characterize adaptive immune responses that occur after initial exposure to *F.tularensis* and to develop rapid, inexpensive and broad-based clinical diagnostics approaches for tularemia.

The main focus of this article is to (1) identify serodiagnostic subsets of antigens for a given pathogen whose expression profiles reliably support classification and diagnosis of healthy and disease samples; (2) build classification models and investigate the effects of varying the sizes of serodiagnostic antigen subsets on prediction accuracy; (3) compare the performance of different classification models; and finally (4) validate and generalize their predictive power in the presence of new, unseen cases.

2 MATERIALS AND METHODS

2.1 Immunoblots and microarrays

Protein microarray chips consisting of 1741 *F.tularensis* antigens are fabricated as described previously in Davies *et al*. (2005a). These large chips are probed using a representative set of infected human samples to generate a smaller chip comprising 244 of the most reactive proteins determined by average signal intensity. Briefly, this is a three step process involving: (1) PCR amplification of each ORF, (2) *in vivo* recombination cloning and (3) in vitro transcription/translation and microarray chip printing.

Custom PCR primers comprising 20 bp of gene-specific sequence with 33 bp of 'adapter' sequences are used in PCRs with *F.tularensis*, SchuS4 strain, genomic DNA as template. The adapter sequences, which become incorporated into the termini flanking the amplified gene, are homologous to the cloning site of the linearized T7 expression vector pXT7 (Davies *et al*., 2005a) and allow the PCR products to be cloned by *in vivo* homologous recombination in competent DH5α cells. The resulting fusion protein also incorporates a 5′ polyhistidine epitope, an ATG translation start codon and a 3′ hemagglutinin epitope and T7 terminator. Sequence-confirmed plasmids are expressed in 5 h *in vitro* transcription–translation reactions (RTS 100 kits from Roche) according to the manufacturer's instructions. Protein expression is monitored either by dot blot or microarray using monoclonal antipolyhistidine (clone His-1 from Sigma) and antihemagglutinin (clone 3F10, Roche). Microarrays are printed onto nitrocellulose coated glass slides FAST slides (Whatman) using an Omni Grid 100 microarray printer (Genomic Solutions). Prior to array staining, the sera are diluted to 1/200 in protein array blocking buffer (Whatman) containing *Escherichia coli* lysate at a final concentration of 30% (final concentration 4–5 mg/ml) and incubated at room temperature for 30 min with constant mixing. The arrays are rehydrated in blocking buffer for 30 min and probed with the pretreated sera overnight at 4°C with constant agitation. The slides are then washed five times in tris(hydroxymethyl)aminomethane (Tris) buffer containing 0.05% (v/v) tween 20, and incubated in biotin-conjugated goat antihuman immunoglobulin diluted 1/200 in blocking buffer; the secondary antibodies are obtained from Jackson Immuno Research and are anti-IgG, Fc-γ chain-specific. After washing, bound antibodies are detected by incubation with streptavidin-conjugated PBXL-3 (Martek). The slides are then washed three times in Tris buffer containing 0.05% (v/v) tween 20 and three times in Tris buffer without tween followed by a final water wash. The slides are air dried under brief centrifugation and examined in a Perkin Elmer ScanArray Express HT microarray scanner. Intensities are quantified using QuantArray software. All signal intensities are corrected for spot-specific background.

Table 1. *F.tularensis* data: number of sera in each category

Diagnostic group	Number of sera	Training set	Validation set
Infected	46	34	12
Healthy	54	41	13
Total	**100**	**75**	**25**

2.2 *F.tularensis* sera measurements

Sera are acquired from 46 individuals in the US diagnosed with tularemia (Table 1). All sera are banked diagnostic samples submitted to Centers for Disease Control and Prevention (CDC) and tested for *F.tularensis* specific antibodies using a standard microagglutination assay, with titres > 1:128 positive. The clinical form of the disease [ulceroglandular ($n=20$) or pneumonic ($n=21$) tularemia] and specimen timing with respect to symptom onset is known from submission forms accompanying the diagnostic specimen. Sera are drawn from the majority of the 46 individuals within 1 month of symptom onset. Subspecies responsible for infection [type A ($n=10$) or type B ($n=5$)] is identified for those cases of tularemia, where a culture is also obtained from the patient and the *F.tularensis* subspecies identified by biochemical subtyping (glycerol fermentation). Control sera are obtained from healthy blood donors in the US of which, 54 sera probed using protein microarrays, are used in this study.

The *F.tularensis* data set thus comprises measurements of 244 antigens in 100 human sera. In addition, seven internal controls (cell-free expression reactions lacking template gene) are spotted on the array.

2.3 Diagnostic engine

A two-step approach is adopted for building a diagnostic engine that reliably classifies healthy and infected samples—feature selection i.e. selecting the most relevant antigens that determine the diagnosis, followed by classification i.e. determining a pattern linking the selected antigens' profiles to the diagnosis.

There are several important reasons for narrowing down a small set of antigens from an entire proteome. First, identification of the most immunodominant antigens in a given disease is a vital step toward understanding the biology of the pathogen and the disease and consequently boosting research efforts in both diagnostic antigen discovery and subunit vaccine development. Second, it becomes possible to obtain significant antigen profiles for several pathogens on a single chip simultaneously, vastly reducing the cost of diagnosis for a single sample. Third, computational diagnostic models typically perform better when irrelevant variables or features are removed during parameter optimization, especially when data are limited.

2.3.1 Data preprocessing and normalization Each serum is measured in duplicate. If missing values are present in any one of the replicated measurements, it is replaced with the other value. Data are normalized using a log-variant (asinh) transformation called 'vsn' (Durbin *et al*., 2002; Huber *et al*., 2002) so that experimental variations are minimized and the measurements are in the same range and scale. The measurement error model (Rocke and Durbin, 2001) that is

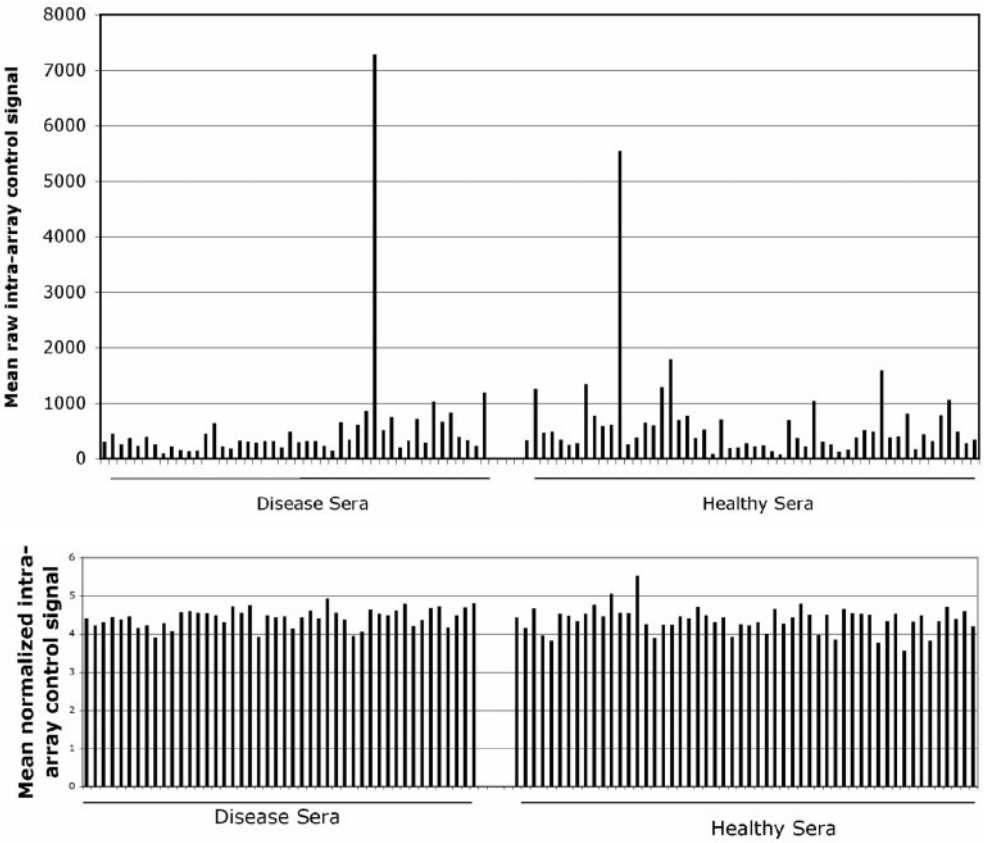

Fig. 1. Normalizing the data set using known true-negative signals (cell-free expression reactions lacking template gene). The upper panel shows the variation in the average raw intra-array control signal across the 100 measurements. The lower panel shows the normalized and 'vsn' transformed average control signal.

assumed by this transformation is appropriate for signals from protein microarrays (Sundaresh *et al*., 2006).

Since the dataset contains expression profiles of 244 of the 1741 *F.tularensis* antigens that generated some immune response, only the seven known true-negative intra-array control signals (cell-free expression reactions lacking template gene) are used as 'house-keeping' probes to obtain the scale and offset parameters. The transformation function 'vsn' is then applied to the whole dataset using these parameters. This method calibrates the measurements and renders the variance relatively independent of the mean signal. Figure 1 shows the effects of applying the 'vsn' (asinh) transformation on the intra-array controls. The 'vsn' transformation resembles the log transformation for large signal intensities and is defined for zero and negative intensities that may occur after background subtraction.

If experiments to obtain measurements for each serum sample are replicated, they are averaged, as is the case in the *F.tularensis* data set. In addition, the seven intra-array control signals are averaged to obtain an estimate for the array-specific background signal.

2.3.2 Differential expression analysis After normalization, significant immune responses within diagnostic subgroups as well as differential immune responses across subgroups are determined by conducting a series of statistical tests (*t*-tests). In these tests, the sample variance of the measurements of each protein is substituted with a Bayes-regularized estimate (Baldi and Hatfield, 2002; Baldi and Long, 2001) that takes into account the variance of neighboring proteins, i.e. proteins with similar expression levels. A web implementation called Cyber-T is available at http://www.igb.uci.edu/index.php?page=tools. Statistical tests that use the Bayes-regularized estimate of the variance have been shown to effectively determine differential expression in both DNA (Hatfield *et al*., 2002; Hung *et al*., 2002; Long *et al*., 2001) and protein microarray expression measurements (Sundaresh *et al*., 2006) to correct unrealistically inflated or deflated sample variances, especially when replications are low. The Bayes-regularized *t*-test generates a list of antigens, ranked by their *P*-values, which are differentially expressed between two groups. An independent study (Choe *et al*., 2005) that analyzed a spiked microarray data set with known concentrations reported that Cyber-T outperforms other differential expression estimation methods including standard *t*-tests and SAM (Tusher *et al*., 2001). *P*-values pw_d *and* pw_h associated with the strength of immune responses of an antigen *within* each group disease (*d*) and healthy (*h*), respectively, are obtained by comparing expressions of the antigens in that group with the corresponding array-specific background signal. *P*-values representing the strength of differential expression for each antigen *between* groups (*pb*) are also computed.

2.3.3 Feature selection Feature selection has been applied to the feature space prior to classification in the context of DNA microarray data (Ding and Peng, 2005; Golub *et al*., 1999) to filter out variables that are irrelevant with respect to determining the class and studies have shown that removing them can improve prediction accuracy. Let us now assume that the data are randomly partitioned into a training set and validation set, each of which includes measurements from both diagnostic groups (Table 1). Using the training set, the values pw_d, pw_h and *pb* are computed for each antigen. Features that are to be

used in the classifier are selected based on a ranking measure computed for each antigen as follows:

(1) Presence of response in disease sera: rank antigens in the disease group based on the strength of the within array response compared with intra-array controls (i.e. lowest $pw_d \rightarrow$ highest $rank_d$).
(2) Significant differential expression: rank antigens based on significantly greater expression in the disease group compared with the healthy group as determined by the Bayes-regularized *t*-test (i.e. lowest $pb \rightarrow$ highest $rank_{d>h}$).

Intuitively, both ranking measurements aim to pick up antigens that specifically show a significant immune response in the disease group. The final rank for each antigen is computed as the sum of the two ranking measures. The top x antigens (x is provided by the user) with the highest rank are included in the set (S) to be used in the classifier. If two or more antigens have the same rank, they are both added. Further refinements of this ranking scheme include assigning weights (indicating relative importance) to each of the measures prior to combining them or using *P*-values and corresponding false discovery rate estimates as cutoffs.

2.3.4 Classification and clustering For DNA microarray data analysis, there has been extensive work done to demonstrate the reliability of automated machine learning techniques for class prediction, notably in cancer classification (Golub *et al*., 1999; Lee and Lee, 2003; Nguyen and Rocke, 2002; Tibshirani *et al*., 2002). Techniques for class prediction and clustering have also been applied to other protein microarray technologies. Belov *et al*. (2006) applied discriminant functions analysis to CD (cluster of differentiation) antibody microarrays to diagnose various forms of leukemias and lymphomas. Groathouse *et al*. (2006) used unsupervised classification techniques such as hierarchical clustering and self-organizing maps (SOM) to identify disease (leprosy) state specific patterns in native-based protein arrays. Binder *et al*. (2006) demonstrated the effectiveness of the *k*-nearest neighbor method applied to profiles from microarrays containing protein-coated beads for the identification of autoimmune diseases. We tried different machine learning methods and report the comparative results for three of them—two supervised methods, *k*-nearest neighbor (*k*-NN) and support vector machines (SVM) (Vapnik, 1995) and one unsupervised method, *K*-means clustering. In addition, their performance is assessed across a range of feature set sizes.

k-NN determines the k nearest measurements (with known class) to the unknown measurement that needs to be classified. Typically, Euclidean distance is used to compute distance between two measurement vectors. The majority class is assigned to the unknown case. The algorithm does not explicitly build a classification model. The implicit model comprises the signals of the selected features in the training set. The classifier accuracy on the training set is determined by applying leave-one-out-cross-validation. Each measurement in the validation set is assigned the majority class belonging to its closest 'k' neighbors in the training set. An analysis using $k = \{3,5,9\}$ neighbors showed no notable differences in performance and for the rest of the article, the results of $k = 5$ are presented.

SVM is a classification method that aims to maximize generalization accuracy and minimize classification error by determining a maximum margin hyperplane in feature space separating two classes. Non-linear classification in SVMs is achieved by the use of a kernel. A preliminary investigation of kernel functions indicated that the performance of linear and polynomial kernels was comparable for this data set, while RBF tended to overfit the data. We therefore present results generated using the linear kernel in the following sections.

K-Means clusters data by minimizing intracluster distance between members of a cluster and the cluster's centroid (mean). In the given context, the number of clusters, a required input, is trivially equal to $k = 2$. Once the clusters are determined, the specific diagnostic class label is assigned to the clusters such that the *total prediction accuracy is maximized.* This allows for the computation of classification accuracy measures such as sensitivity and specificity. The primary reason for including *k*-means in this work is to assess the performance of an unsupervised learning approach, which is useful in situations where class labels are not available for this kind of data (e.g. new virus strain).

2.3.5 Validation procedure Repeated random subsampling or repeated holdout, a commonly used technique (Dudoit *et al*., 2002), is employed for generating training and validation/holdout sets and assessing classification accuracy. The training and validation sets are stratified, meaning that they each have approximately the same proportions of classes or diagnostic cases as in the original data set. Unlike in *m*-fold cross-validation, validation sets may contain overlapping samples. However, the advantage of using this approach is that, given limited data, it affords the flexibility of simultaneously generating larger validation set sizes and conducting any desired number of repetitions, thus reducing discreteness of classification error rates and shrinking confidence intervals for the mean estimates (Dudoit *et al*., 2002; Mitchell, 1997). It is important to note that antigen selection is performed on the training set only.

2.3.6 Machine learning procedure A simple algorithm for the feature selection and classification approach is presented below. The input is a table with a set of A antigens, n_d and n_h measurements in each diagnostic group respectively. R is the number of iterations or runs. 'Classifier' refers to a classification or clustering model.

For each run r, r = 1,2,...,R,

(1) *Randomly partition data into subsets T_r and V_r for feature selection/training and validation, respectively.*
(2) *Generate significance measures for each antigen $a \in A$ in T_r for both within group response (compared with intra-array controls), (pw_d, pw_h,) and differential expression between groups (pb) using Bayes-regularized t-tests.*
(3) *Perform feature selection based on the ranking scheme described earlier. Generate top ranking antigen sets $s \in S$, e.g. top 10, top 25 antigens. For each top ranking antigen set, $s \in S$,*
   (i) *Train y classifiers $C_{r,1}, C_{r,2}, \ldots, C_{r,y}$ on training data T_r using the signals of antigens in $s \in S$*
   (ii) *Compute training accuracy of $C_{r,1}, C_{r,2}, \ldots, C_{r,y}$ on T_r using the signals of antigens in $s \in S$*
   (iii) *Compute validation accuracy of $C_{r,1}, C_{r,2}, \ldots, C_{r,y}$ on V_r using the signals of antigens in $s \in S$*

Finally, training and validation accuracy of each classifier are averaged across all R runs.

2.3.7 Parameters and initializations $R = 10$ runs are conducted. In each run, the data are randomly partitioned into T_r and V_r in the ratio of 3:1, ensuring a reasonable number of measurements for training and validation respectively. The actual distribution is shown in Table 1.

2.3.8 Software All software programs used for these analyses are written using the statistical software *R*. Normalization and variance stabilization is performed using the 'vsn' package described in Huber *et al*. (2002). The program to perform Bayes-regularized *t*-tests is available for download at http://www.igb.uci.edu/?page=tools&subPage=dmss. *k*-NN, SVM and *k*-means implementations are available in the packages 'class', 'e1071' and 'stats', respectively at http://www.R-project.org.

3 RESULTS

3.1 Antigen selection

In each run, the top $x = \{10, 25, 50, 100, 150\}$ antigens are selected based on the ranking scheme described. Table 2 presents antigens that are in the top 25 list in all 10 runs. Table A of Supplementary Materials presents antigens in the top 50 list in all 10 runs.

Figure 2 shows the elevated responses of top ranking antigen signals in infected sera. These antigens appear in the top 25 list in all 10 runs. The intra-array control is also included. The difference in the control signal is negligible as is expected when the measurements are calibrated.

3.1.1 The Δh statistic (Huber *et al.*, 2002) The bars in Figure 2 represent the difference (Δh) between mean 'vsn' transformed disease sera signals and healthy sera signals, used in the Bayes-regularized *t*-test to compute differential expression. The *P*-values (p_b) from the *t*-test are used to determine one of the two measures that make up the cumulative ranking score proposed earlier. For 'vsn' transformed data, this $\triangle$h statistic coincides with the log-ratio for high intensities and the difference for near-zero intensities. It should be noted that Figure 2 depicts the mean difference when all sera (both training and validation sets) are used. When performing the differential expression analysis on randomly generated training subsets, antigens are selected based on the mean difference between disease and healthy sera in that particular training set.

Table 2. Antigens that appear in the top 25 list in all 10 runs using the proposed ranking scheme

Antigen	No. of times in top 25 list	No. of times in top 10 list
FTT1116	10	10
FTT0106	10	10
FTT1484	10	10
FTT1314	10	10
FTT1696	10	10
FTT0472	10	10
FTT0956	10	8
FTT0077	10	8
FTT0101	10	7
FTT1163	10	6
FTT0949	10	3
FTT0989	10	1
FTT1540	10	1
FTT0975	10	0
FTT1775	10	0

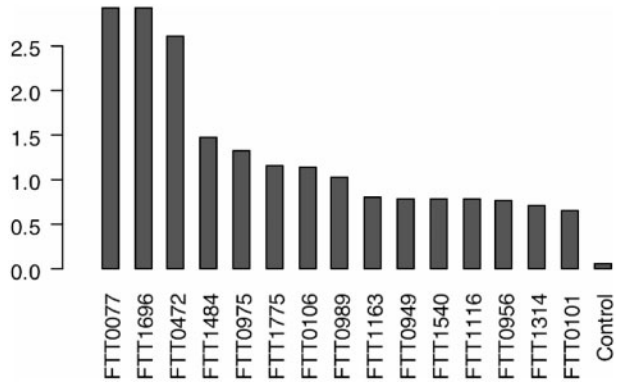

Fig. 2. Differential expression between mean 'vsn' transformed disease and healthy sera signals for antigens appearing in the top 25 list in all 10 runs (Table 2). Bars represent the Δh statistic (see text). 'Control' is the mean intra-array background/true negative signal.

3.1.2 Immune responses in disease subtypes Since the clinical status of some of the patients is known with respect to whether the infection was pneumonic (lung infection) or ulceroglandular (cutaneous), and whether the infection was from the virulent Type A strain or the less virulent Type B strain, the differences in immunoreactivity in these patient groups can also be examined.

Figures 3 and 4 show the immune responses in the specific infection subgroups and enable us to visually appreciate some of the differences. For example, antigen FTT1484 is more reactive in the ulceroglandular group compared to pneumonic group of sera. Similarly, antigen FTT0975 shows a stronger

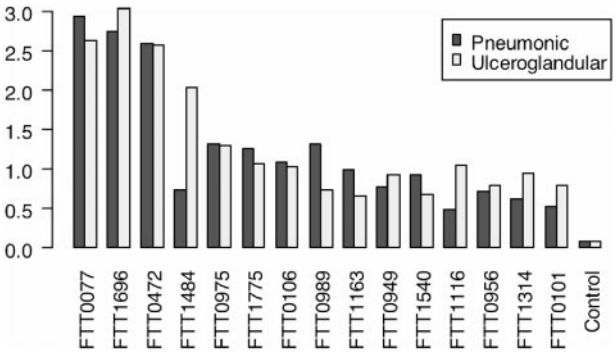

Fig. 3. Immune responses in pneumonic and ulceroglandular disease subgroups compared to mean 'vsn' transformed healthy sera signals for antigens in Table 2. Bars represent the Δh statistic. 'Control' refers to the mean intra-array background/true negative signal.

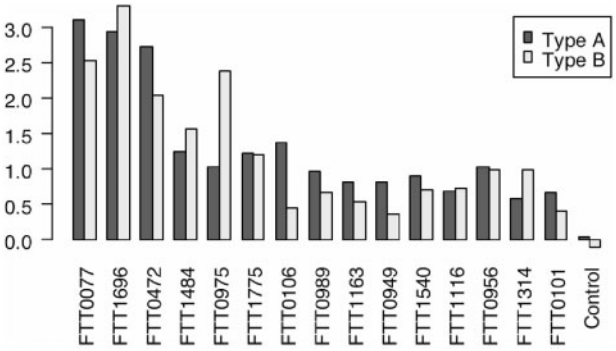

Fig. 4. Immune responses in Type A and Type B disease sub-groups compared to mean 'vsn' transformed healthy sera signals for antigens in Table 2. Bars represent the Δh statistic. 'Control' refers to the mean intra-array background/true negative signal.

response in Type B infected subjects. This can be confirmed by more rigorous assessment, using statistical tests, to determine which responses in each of the subgroups are significantly different (outside the scope of this work). The following section explores how these selected lists of antigens help in improving class prediction. It is of particular interest to observe whether smaller numbers of antigens are effectively able to summarize the information in the proteome for reliable diagnosis of *F.tularensis* disease in human sera.

3.2 Classification/clustering

Figure 5 shows the performance of the classification/clustering techniques on the training and validation sets. Prediction accuracy rates (%) are averaged across the 10 runs and presented in terms of total accuracy, sensitivity and specificity. The primary aim is to observe the expected levels of prediction accuracy using the selected antigens' signals. If smaller antigen sets are able to capture immune response information required for diagnosis as reliably as, if not better than, larger sets, then antigens occurring with high frequency in these smaller sets can be considered high-priority candidates for further evaluation.

Given the available data, a limited analysis is also conducted to compare methods when different feature set sizes are used for training. When comparing methods, the variance corrected resampled paired *t*-test (Nadeau and Bengio, 2003) is applied, to account for the overlapping test samples. Berrar *et al*. (2006) noted that this correction, while drastically improving the Type I error of the *t*-test, might cause a decrease in power, especially when the training set is not ~5–10 times larger than the test set. *P*-values are therefore reported when both borderline and significant differences in accuracy measures are observed. Holm–Bonferroni correction is applied because $c = 3$ methods are compared across six feature set sizes, resulting in $\aleph = 1/2 \times c(c-1) \times 6 = 18$ multiple comparisons.

3.2.1 Total accuracy Total accuracy refers to the percentage of cases in the validation set that are correctly classified as either infected or healthy sera. Figure 5f shows a decline in accuracy as the feature set size increases, particularly beyond 50 antigens. The decline is apparent for *k*-NN and even more so for *k*-means since both use distance measures that can be sensitive to noise in the feature set. The validation accuracy of *k*-means (~60%) is lower than SVM (~85%), when no feature selection is performed (borderline significant $P = 0.003$, Fig. 6d). As expected, the performance of *k*-means, an unsupervised learning method, steadily improves with feature selection, which is based on criteria that uses class label information. SVM is fairly robust to changes in feature set size and also appears to be resistant to overfitting when linear and polynomial (data not shown) kernels are used.

3.2.2 Sensitivity and specificity Both SVM and *k*-NN methods diagnose cases with higher average specificity (>95%) than sensitivity (80–90%) (Fig. 5b and d) on validation data, using the top 10–50 antigens. One reason for this could be the variation in immune responses in infected samples due to genetic factors such as MHC haplotype, disease subtype, specimen timing with respect to onset of symptoms and *F.tularensis* strain. Milder infections may not show significantly different responses from uninfected samples. High sensitivity is observed when ~25–50 antigens are selected for model building (Fig. 5b). In addition to the noise removed when feature selection is applied, the increased sensitivity when feature set size is small is probably because the antigens are selected based on disease-specific responses. The sensitivity of *k*-NN decreases rapidly (Fig. 6e) as the feature set size increases but the difference is not statistically significant when compared with other methods. We also note a decrease in specificity on the validation set as feature set size increases, (Fig. 5d) with *k*-means performing worse than both *k*-NN ($P < 0.002$) and SVM ($P < 0.003$) when all features are used (Fig. 6f).

3.2.3 Diagnostic antigen discovery Given the high prediction accuracy rates in the 90% range with the available data, when the feature set size is ~25–50 antigens (Fig. 6-upper panel), the *F.tularensis* antigens occurring with the highest frequency in the top 25 list can be regarded as a high-priority serodiagnostic antigen set. Figure 2 presents these antigens ordered by the $\triangle$h statistic. The mechanisms responsible for particularly elevated responses to FTT0077, FTT1696 and FTT0472, or to other antigens in specific disease subgroups, require further investigation. Here, all antigens in Figure 2 are further validated through proteomic analyses aimed at predicting localization as shown in Table 3. An extended table containing antigens appearing in top 50 list in all 10 runs is available in Table A (Supplementary Material). Some of the immunodominant antigens shown in Table 3 are not immunologically unique to *Francisella*. For example, GroEL from other bacterial species is regularly identified as being reactive with human sera (Chen *et al*., 2004; Cole *et al*., 2005). A potential extension of this study would be to focus on the proteins which are either unique to *Francisella* or investigate whether the proteins have significant sequence diversity from other bacterial homologs so they are more likely to be immunologically distinct.

Table 3 also discusses the likely cellular location of the proteins. As a gram-negative bacterium, *F.tularensis* contains both helical transmembrane (TMH) proteins and transmembrane beta-barrel proteins (TMBs). TMH proteins are found in the bacterial inner membrane, while TMBs localize in the outer membrane. To predict TMH proteins we look at the localization predictions in conjunction with the number of predicted transmembrane segments from TMHMM (Krogh *et al*., 2001). To predict TMBs, we consider the localization predictions, combined with the output from three publicly available TMB screening tools: PRED-TMBB (Bagos *et al*., 2004), TMB-HUNT (Garrow *et al*., 2005) and ProfTMB (Bigelow *et al*., 2004). All three TMB screening tools provide a numerical confidence in their output. For the antigens submitted, scores generated by each tool are normalized to 0–1 intervals. Two independent computational tools are used to predict localization: PSORTb (Gardy *et al*., 2005) and a SVM based predictor (Wang *et al*., 2005). Both predictors are designed specifically for gram-negative bacteria. It is not altogether surprising that many top antigens are predicted to be surface or membrane associated proteins, which are highly visible to the host immune

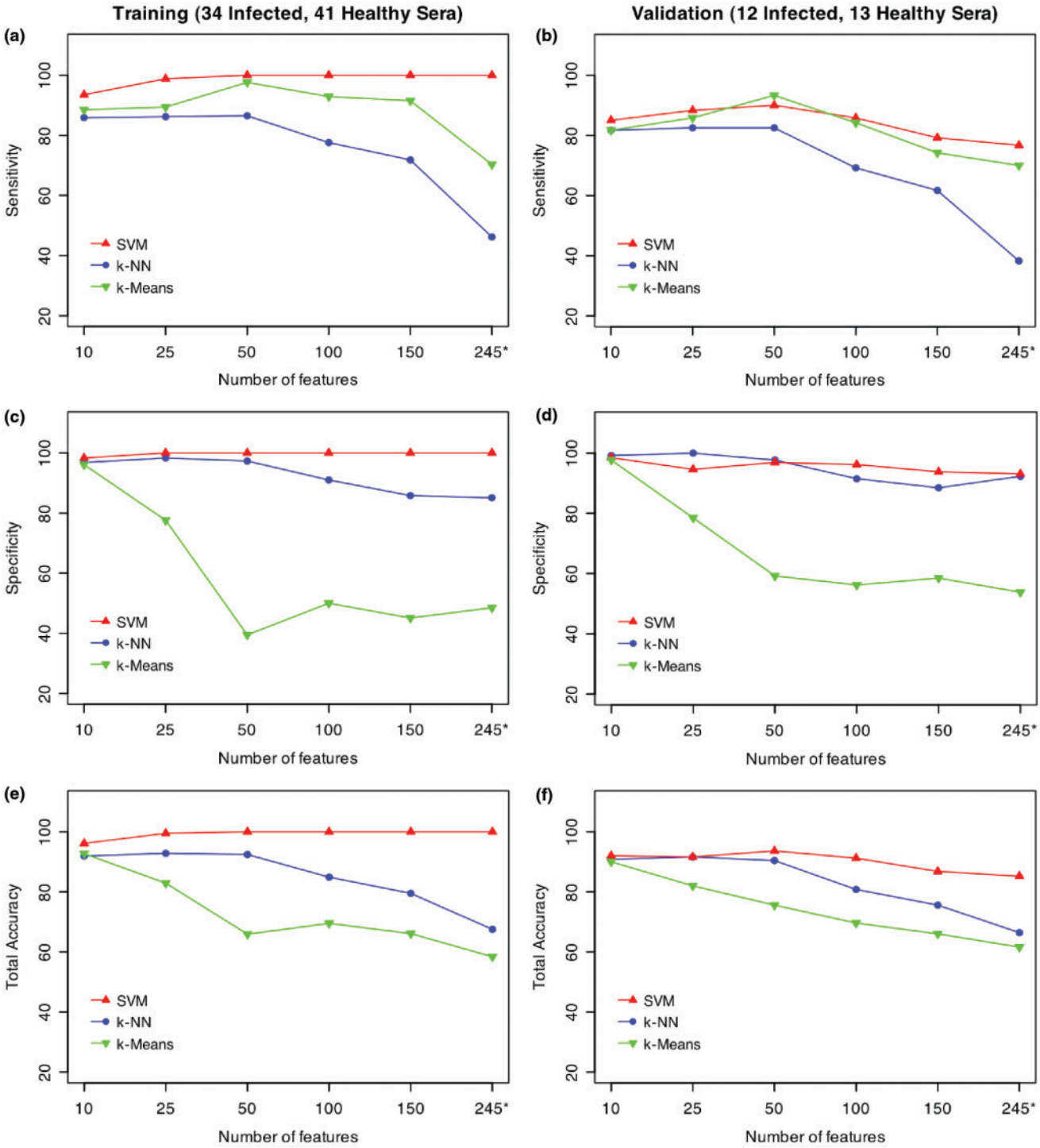

Fig. 5. Prediction accuracy of classification (%) models averaged over 10 runs for different feature subset sizes. *Without antigen/feature selection i.e. all 244 antigens +1 averaged control used by classifier for training and validation.

system and hence more likely to trigger an initial immune response. In several pathogens that have been studied using the same techniques, it has been observed that a large number of top antigens are predicted to be membrane proteins (~50%) (data not shown). It must be noted that PSORT relies partly on the presence of signal sequences to identify likely exported proteins. However, there are many proteins exported via other systems, which may not be identified using PSORT. For example, FTT1314 is almost certainly on the surface since this is the location of type IV pili. Predicting subcellular localization

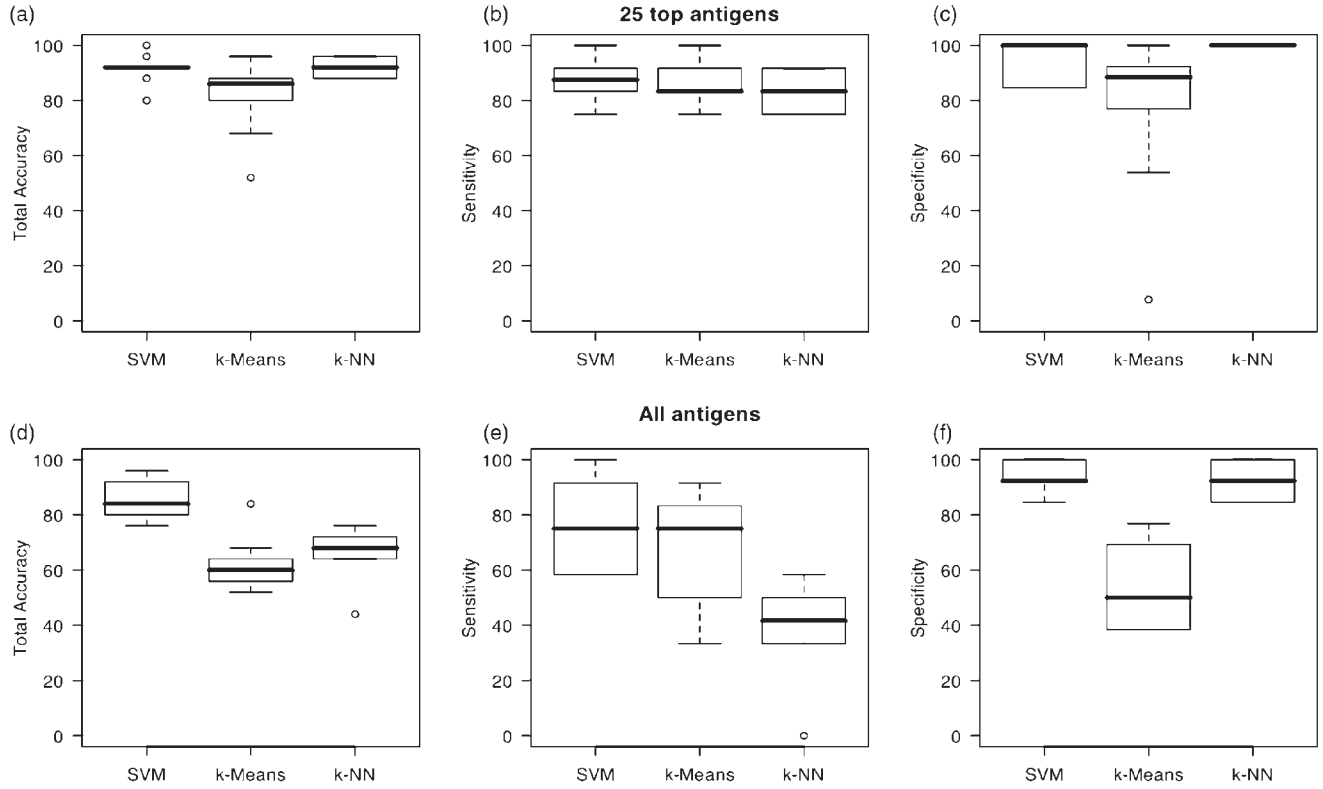

Fig. 6. Box plots summarizing validation accuracy (total accuracy, sensitivity and specificity) of classification methods. Upper panels show validation accuracy (%) when 25 top-antigens are selected for training. Lower panels show accuracy (%) when no feature selection is performed.

is also complex because some proteins, like GroEL, which are clearly predominately cytoplasmic, can be found on the surface under some conditions (Bergonzelli *et al*., 2006; Garduno *et al*., 1998). One protein, FTT0949, is predicted to be encoded by a pseudogene in *F.tularensis* subspecies *holarctica* OSU18 (type B) (Petrosino *et al*., 2006) and *tularensis* SchuS4 (type A). This is interesting since evidence of its immunoreactivity suggests that it is expressed *in vivo* in the infected patient. The immunoprotemics data does not only indicate possible protective antigens, but also genes that are expressed *in vivo*. Therefore, some of the proteins identified might well play roles in virulence. The most obvious example of this is FTT1314. Type IV pili are strongly implicated in virulence of *F.tularensis* (Forslund *et al*., 2006).

4 DISCUSSION AND CONCLUSIONS

The foundation of the array platform is a high-throughput PCR and homologous-recombination cloning methodology that enables the genes of any sequenced pathogen to be cloned quickly and efficiently. Genomes of pathogens comprising several thousand genes can be cloned in relatively short time. This gives the capability of screening the whole proteome for immunoreactive targets. It also means that the platform is not restricted to screening just known antigens. Conventional proteomic methodologies use 2D gels to separate bacterial cell components followed by the identification of immunoreactive spots by mass spectrometry. These methods are only able to sample proteins which are produced by the bacterium in broth culture in the laboratory. Different proteins that are only expressed in an animal or human host will not be available for analysis. In contrast, the protein microarray technology studied here samples each protein equally and allows the entire proteome to be interrogated in an unbiased manner. Completely novel antigens are discovered leading to a more comprehensive data set. Moreover, the arrays are fabricated from proteins expressed in cell-free transcription/translation reactions that are printed directly without further purification. This considerably alleviates the bottleneck associated with protein purification, particularly when thousands of antigens are required. More than 12 500 proteins from at least 14 microorganisms including vaccinia virus (Crotty *et al*., 2003; Davies *et al*., 2005a, b), *Leptospira interrogans, Mycobacterium tuberculosis, Plasmodium falciparum* (Doolan *et al*., 2003; Sundaresh *et al*., 2006), *Plasmodium vivax, Burkholderia pseudomallei, Coxiella burnetii, Borrelia burgdorferi and C.trachomatis,* orthopox viruses, herpes viruses and papilloma viruses have already been cloned, expressed and printed.

Using this array technology, it has been demonstrated that signals measuring strong antigen-antibody binding activity can be computationally analyzed to determine infection. The screening process for identifying immunodominant antigens occurs in stages. Initially, chips containing 1741 *F.tularensis* antigens are probed with sera and the top 244 antigens with the

Table 3. Summary of the immunoreactive antigens identified in this study that are part of the serodiagnostic antigen set. The probability that a signal peptide is present in the amino acid sequence is determined by SignalP 3.0 (Bendtsen *et al*., 2004) using Hidden Markov Models (HMM) where the prediction is given as a probability. Predicted TMH count comes from the TMHMM predictor (Krogh *et al*., 2001). The TMB screening values come from normalizing the numeric confidence outputs for each predictor: TMB-HUNT (Garrow *et al*., 2005), ProfTMBB (Bigelow *et al*., 2004) and PRED-TMBB (Bagos *et al*., 2004). **The most likely TMB or TMH proteins in this set, i.e. FTT0106 and FTT1775, respectively**, are highlighted in bold. The intracellular localization is predicted by PSORTb (Gardy *et al*., 2005) and an SVM based predictor (Wang *et al*., 2005)

FTT#	Description	SignalP 3.0	TMHMM helix count	TMB screening			Localization		
				TMB-HUNT	Prof-TMBB	PRED-TMBB	PSORTb	SVM	By function*
FTT1116	Preprotein translocase family protein	1.00	2	0.33	0.00	0.72	unknown	inner mem	mem
FTT0106	**Efflux protein, RND family, MFP subunit**	**0.67**	**2**	**0.87**	**0.81**	**0.47**	**cyto mem**	**outer mem**	**mem**
FTT1484	Pyruvate dehydrogenase, E2 cmpt.	0.00	0	0.03	0.96	0.74	cyto mem	cyto	cyto
FTT1314	Type IV pili fiber building block protein	0.86	1	0.17	0.79	0.37	unknown	extracellular	surface
FTT1696	Chaperone protein, GroEL	0.00	0	0.00	0.79	1.00	cyto	cyto	cyto
FTT0472	Acetyl-CoA carboxylase, biotin carboxyl carrier protein subunit	0.09	0	0.12	0.78	0.80	unknown	cyto	cyto
FTT0956	Hypothetical membrane protein	0.99	1	0.44	0.59	0.29	unknown	periplasmic	mem
FTT0077	Dihydrolipoamide succinyltransferase cmpt.	0.00	0	0.01	0.79	0.54	cyto	cyto	cyto
FTT0101	Conserved membrane hypothetical protein	1.00	1	0.08	0.63	0.18	cyto mem	periplasmic	mem
FTT1163	Hypothetical membrane protein	0.27	2	0.53	0.00	0.26	cyto mem	inner mem	mem
FTT0949	Pseudogene	1.00	1	0.30	0.48	0.28	unknown	cyto	mem
FTT0989	Hypothetical protein	0.94	1	0.19	0.49	0.26	unknown	outer mem	mem
FTT1540	Hypothetical protein	1.00	0	0.43	0.82	0.18	unknown	extracellular	mem
FTT0975	Hypothetical protein	1.00	0	0.34	0.71	0.45	unknown	extracellular	mem
FTT1775	**Chloride channel protein**	**0.79**	**11**	**0.17**	**0.82**	**0.48**	**cyto mem**	**inner mem**	**mem**

*In addition, antigens are identified as likely membrane proteins based on similarity with other proteins of membrane localization, or presence of a signal sequence, presence of membrane spanning domains predicted by TMHMM, presence of a lipid attachment (prosite domain ps0013). Abbreviations: cyto,cytoplasmic; mem,membrane.

highest signal intensity are used to build a second chip so that more specific measurements can be obtained. The next stage of analysis employs more stringent criteria based on robust statistical techniques, which enable identifying candidate serodiagnostic antigen sets. A comparative study of different computational techniques for classification highlights the effects of varying antigen set sizes on predictive power. When signals of 25–50 antigens are used for classifier training, SVM based models perform well overall yielding prediction rates of ~90%. For a validation set containing $n = 25$ cases, this accuracy level corresponds to a misdiagnosis of ~2–3 cases. An investigation of support vectors identified in the models did not yield any association with available patient information such as disease subtype and specimen timing with respect to symptom onset. SVM models are found to be on average more specific than sensitive, making this an important consideration in the specific application that this test is used for. A large population-wide study will provide better estimates of the cost-effectiveness of these diagnostic tests and enable a more rigorous comparison of different computational techniques for disease diagnosis. Finally, the immunodominant antigens, presented as high-priority candidates for diagnostic development, can be further evaluated using proteomic analysis techniques.

ACKNOWLEDGEMENTS

The bioinformatics and primer design components in this work were supported primarily by National Institutes of Health Biomedical Informatics Training Program Grant 5T15LM007743 and National Science Foundation Grant MRI EIA-0321390 to P.B. and the Institute for Genomics and Bioinformatics at UCI. The protein array component was supported primarily by National Institute of Allergy and Infectious Diseases Grants U01AI056464 and 1U01AI061363-01 to P.F.

Conflict of Interest: none declared.

REFERENCES

Bagos,P.G. *et al.* (2004) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res.*, **32**, W400–W404.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Baldi,P. and Hatfield,G.W. (2002) *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. Cambridge University Press.

Belov *et al.* (2006) Analysis of human leukaemias and lymphomas using extensive immunophenotypes from an antibody microarray. *Br. J. Haematol.*, **135**, 184–197.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Bergonzelli,G.E. *et al.* (2006) GroEL of *Lactobacillus johnsonii* La1 (NCC 533) is cell surface associated: potential role in interactions with the host and the gastric pathogen *Helicobacter pylori*. *Infect. Immun*, **74**, 425–434.

Berrar,D. *et al.* (2006) Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics*, **22**, 1245–1250.

Bigelow,H.R. *et al.* (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.*, **32**, 2566–2577.

Binder,S.R. *et al.* (2006) Protein arrays and pattern recognition: new tools to assist in the identification and management of autoimmune disease. *Autoimmun. Rev.*, **5**, 234–241.

Chen,Z. *et al.* (2004) Rapid screening of highly efficient vaccine candidates by immunoproteomics. *Proteomics*, **4**, 3203–3213.

Choe,S.E. *et al.* (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.*, **6**, R16.

Cole,J.N. *et al.* (2005) Surface analyses and immune reactivities of major cell wall-associated proteins of group A *Streptococcus*. *Infect Immun.*, **73**, 3137–3146.

Crotty,S. *et al.* (2003) Cutting edge: long-term B cell memory in humans after smallpox vaccination. *J. Immunol.*, **171**, 4969–4973.

Davies,D.H. *et al.* (2005a) Profiling the humoral immune response to infection by using proteome microarrays: high-throughput vaccine and diagnostic antigen discovery. *Proc. Natl Acad. Sci. USA*, **102**, 547–552.

Davies,D.H. *et al.* (2005b) Vaccinia virus H3L envelope protein is a major target of neutralizing antibodies in humans and elicits protection against lethal challenge in mice. *J. Virol.*, **79**, 11724–11733.

Dennis,D.T. *et al.* (2001) Tularemia as a biological weapon: medical and public health management. *Jama*, **285**, 2763–2773.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 185–205.

Doolan,D.L. *et al.* (2003) Utilization of genomic sequence information to develop malaria vaccines. *J. Exp. Biol.*, **206**, 3789–3802.

Durbin,B. *et al.* (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105–S110.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Ellis,J. *et al.* (2002) Tularemia. *Clin. Microbiol. Rev.*, **15**, 631–646.

Forslund,A.-L. *et al.* (2006) Direct repeat mediated deletion of a Type IV pilin gene results in major virulence attenuation of *Francisella tularensis*. *Mol. Microbiol.*, **59**, 1818–1830.

Garduno,R.A. *et al.* (1998) Surface-associated hsp60 chaperonin of *Legionella pneumophila* mediates invasion in a HeLa cell model. *Infect. Immun.*, **66**, 4602–4610.

Gardy,J.L. *et al.* (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623.

Garrow,A.G. *et al.* (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. *Nucleic Acids Res.*, **33**, W188–W192.

Groathouse *et al.* (2006) Use of protein microarrays to define the humoral immune response in leprosy patients and identification of disease-state-specific antigenic profiles. *Infect. Immun.*, **74**, 6458–6466.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hatfield,G.W. *et al.* (2002) Differential analysis of DNA microarray gene expression data. *Mol. Microb.*, **47**, 871–877.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl.), S96–S104(ISMB 2002).

Hung,S.-P. *et al.* (2002) Global gene expression profiling in *Escherichia coli* K12: The effects of leucine-responsive regulatory protein. *J. Biol. Chem.*, **277**, 40309–40323.

Isherwood,K.E. *et al.* (2005) Vaccination strategies for *Francisella tularensis*. *Adv. Drug Deliv. Rev.*, **57**, 1403–1414.

Karlsson,J. *et al.* (2000) Sequencing of the *Francisella tularensis* strain Schu 4 genome reveals the shikimate and purine metabolic pathways, targets for the construction of a rationally attenuated auxotrophic vaccine. *Microb. Comp. Genomics*, **5**, 25–39.

Krogh,A., Larsson,B., von Heijne,G., and Sonnhammer,E.L.L. (2001), Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes, J. Mol. Biol., 305, 567–580.

Larsson,P. *et al.* (2005) The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. *Nat. Genet.*, **37**, 153–159.

Lee,Y. and Lee,C.K. (2003) Classification of multiple cancer types by multi-category support vector machines using gene expression data. *Bioinformatics*, **19**, 1132–1139.

Long,A.D. *et al.* (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *J. Biol. Chem.*, **276**, 19937–19944.

Mitchell,T. (1997) *Machine Learning*. McGraw Hill.

Nadeau,C. and Bengio,Y. (2003) Inference for generalization error. *Mach. Learn.*, **52**, 239–281.

Nguyen,D.V. and Rocke,D.M. (2002) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.

Petrosino,J.F. *et al*. (2006) Chromosome rearrangement and diversification of *Francisella tularensis* revealed by the type B (OSU18) genome sequence. *J. Bacteriol*., **188**, 6977–6985.

Prior,R.G. *et al*. (2001) Preliminary analysis and annotation of the partial genome sequence of *Francisella tularensis* strain Schu 4. *J. Appl. Microbiol*., **91**, 1–7.

Rocke,D.M. and Durbin,B. (2001) A model for measurement errors for gene expression arrays. *J. Comput. Biol*., **8**, 557–569.

Sundaresh,S. *et al*. (2006) Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics*, **22**, 1760–1766.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci*, **98**, 5116–5121.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Wang,J. *et al*. (2005) Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinform*., **6**, 174.