Liu *et al. BMC Bioinformatics* (2016) 17:47
DOI 10.1186/s12859-016-0889-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

An individualized predictor of health and disease using paired reference and target samples

Tzu-Yu Liu[1], Thomas Burke[2], Lawrence P. Park[2], Christopher W. Woods[2], Aimee K. Zaas[2], Geoffrey S. Ginsburg[2*] and Alfred O. Hero[3,4*]

Abstract

Background: Consider the problem of designing a panel of complex biomarkers to predict a patient's health or disease state when one can pair his or her current test sample, called a target sample, with the patient's previously acquired healthy sample, called a reference sample. As contrasted to a population averaged reference this reference sample is individualized. Automated predictor algorithms that compare and contrast the paired samples to each other could result in a new generation of test panels that compare to a person's healthy reference to enhance predictive accuracy. This paper develops such an individualized predictor and illustrates the added value of including the healthy reference for design of predictive gene expression panels.

Results: The objective is to predict each subject's state of infection, e.g., neither exposed nor infected, exposed but not infected, pre-acute phase of infection, acute phase of infection, post-acute phase of infection. Using gene microarray data collected in a large scale serially sampled respiratory virus challenge study we quantify the diagnostic advantage of pairing a person's baseline reference with his or her target sample. The full study consists of 2886 microarray chips assaying 12,023 genes of 151 human volunteer subjects under 4 different inoculation regimes (HRV, RSV, H1N1, H3N2). We train (with cross-validation) reference-aided sparse multi-class classifier algorithms on this data to show that inclusion of a subject's reference sample can improve prediction accuracy by as much as 14 %, for the H3N2 cohort, and by at least 6 %, for the H1N1 cohort. Remarkably, these gains in accuracy are achieved by using smaller panels of genes, e.g., 39 % fewer for H3N2 and 31 % fewer for H1N1. The biomarkers selected by the predictors fall into two categories: 1) contrasting genes that tend to differentially express between target and reference samples over the population; 2) reinforcement genes that remain constant over the two samples, which function as housekeeping normalization genes. Many of these genes are common to all 4 viruses and their roles in the predictor elucidate the function that they play in differentiating the different states of host immune response.

Conclusions: If one uses a suitable mathematical prediction algorithm, inclusion of a healthy reference in biomarker diagnostic testing can potentially improve accuracy of disease prediction with fewer biomarkers.

Keywords: Reference-aided prediction, Precision medicine, Automated diagnostics, Biomarker discovery, Sparse multi-block classifier algorithm

*Correspondence: geoffrey.ginsburg@duke.edu; hero@eecs.umich.edu
[2]Center for Applied Genomics and Precision Medicine, Department of Medicine, Duke University, Durham NC, USA
[3]Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor MI, USA
Full list of author information is available at the end of the article

Background

It is evident that that patient history can improve interpretability of diagnostic data such as a panel of assayed biomarkers. When this history includes a previously collected assay, the assay constitutes a reference baseline against which the current assay can be quantitatively compared. However, as the size and complexity of clinical biomarker panels increase, manual cross-assay comparisons become impractical. This motivates the development of automated algorithms that can combine a current target assay and a reference assay with improved prediction or classification performance. In this paper we consider the problem of using a panel of biomarkers to predict a patient's health state when both the target sample and reference sample are available. Two questions are of interest. Can such a reference sample be used to more accurately assess the deviation of the target sample from a previously established patient baseline, potentially translating into improved predictions? Can such predictions be performed accurately with relatively fewer biomarkers, i.e., a smaller test panel, potentially translating into a less expensive test? In this paper we show that the answer to both of these questions is affirmative. Using a state-of-the-art multi-block sparse predictor algorithm, and a large-scale serially sampled data set collected in a human viral challenge study, we present an algorithm for reference-aided health prediction that attains higher predictive accuracy using a smaller panel of biomarkers.

The reader may not find it surprising that automated diagnostics may benefit from pairing a reference sample and a target sample. Indeed, it has been common clinical practice for a physician to manually compare a small number of a patient's analytes to his or her previous test results. However, such manual comparison will become increasingly difficult as we enter the era of precision medicine where whole genome expression or next generation sequencing platforms may play an important clinical role [1–3]. In this era, automated algorithms will be needed not only for accurate prediction but also for selection of a suitably small subset of the thousands of probes generated by these platforms. Such algorithms impose sparsity on the predictor by utilizing only a small fraction of the available probes. The reduction of the number of probes (genes) is relevant to personalized medicine applications since it leads to a more economical (lower complexity) targeted biomarker assay. Previous work has developed such algorithms in the context of prediction of acute respiratory virus infection [4–6]. This paper goes one step further and shows that adding one healthy reference sample can result in improved prediction performance.

The paper is organized as follows. We first present the formulation of our optimization problem in the "Methods" section, including the loss function used as surrogates in reference-based classification, the proper regularization that selects variables relevant simultaneously to all classes and references, and followed by a discussion about the general algorithm we propose to solve the optimization. Then we present the performance of the reference-based classification applied to H3N2, H1N1, HRV, and RSV flu challenge data sets in the results section. Advantages of the methods and biological interpretation are presented in the discussion section. The conclusion section concludes this paper.

Methods

The proposed reference-aided prediction method is based on a state-of-the-art supervised multi-block multi-class classifier algorithm with variable selection [7]. To illustrate the advantages of the proposed predictor, we will demonstrate superior prediction performance on data collected from large scale serially sampled respiratory virus challenge studies. Data from the challenge studies have previously been used by us and others to derive molecular signatures for acute respiratory infection (ARI) [4, 5, 8, 9]. This paper's contribution is the introduction of a new individualized reference-aided predictor that is demonstrated on an extended set of data collected from additional challenge studies (see Table 1). More details on these challenge studies can be found in the aforemementioned references and in the Additional file 1. We describe the challenge studies first and then turn to the automated predictor afterwards.

Viral challenge study model

To demonstrate the advantages of reference-based prediction, we use data from a serially sampled challenge study. The challenge study consists of a total of 151 subjects (human volunteers) that, shortly after enrollment in the study, were inoculated with sham or live virus from one of 4 categories of pathogen (HRV, RSV, H3N2, H1N1). The overall study was conducted over a 4 year time period in 7 stages (see Table 1 for a summary). Research participants in these studies provided informed consent and all research activities were conducted in accordance with the Declaration of Helsinki and local policies and regulations. These studies were approved by the Duke University Health System (DUHS) Institutional Review Board (IRB). Where applicable, additional approval was obtained from a local governing IRB where the study activities occurred: Western Institutional Review Board (WIRB) and the University of Virginia IRB approved the studies that were conducted Retroscreen Virology, London, UK and UVA, respectively.

Each subject in the study was serially sampled for several days quantifying time courses of whole blood gene expression by Affymetrix Human U133A 2.0 GeneChips, self-reported clinical symptom scores over

Table 1 Composition of data collected in the respiratory virus challenge study. The study enrolled a total of 151 subjects challenged with 4 difference viruses over seven different challenge sub-studies and samples at multiple regularly spaced time points over a time period ranging from 3⊠5 days. The first column is the sub-study designation. Second column is the virus used in the challenge. Third and fourth columns are the year and location the sub-study was conducted. Fifth column is the DUHS IRB protocol number. Sixth column is the duration of the sub-study in hours. Last two columns are the number of subjects and the number of time points collected per subject, respectively

Challenge	Virus	Year	Location	IRB protocol	Duration (hrs)	# Subjects	# Time points
DEE1	RSV	2008	Retroscreen	Pro00002796	166	20	21
DEE2	H3N2	2009	Retroscreen	Pro00006750	166	17	21
DEE3	H1N1	2009	Retroscreen	Pro00018132	166	24	20
DEE4	H1N1	2010	Retroscreen	Pro00019238	166	19	21
DEE5	H3N2	2011	Retroscreen	Pro00029521	680	21	23
HRV UVA	HRV	2008	Univ. of Virginia	Pro00003477	120	20	15
HRV Duke	HRV	2010	Duke Univ.	Pro00022448	136	30	19

8-10 symptoms (varied by study), and viral shedding from periodic nasopharyngeal titrations. The Affymetrix gene probes were log transformed and normalized using the RMA package with quantile normalization, median polish and a custom cdf mapping from oligoprobes to gene yielding 12,023 gene probes (see Additional file 1 for details on genechip normalization and symptom symptom score definitions).

Subjects were sampled at least once before the viral inoculum was administered and at least 14 times after inoculation. Each subject was designated as a symptomatic subject (Sx) or an asymptomatic subject (Asx) and as an infected subject (Inf) or uninfected subject (UnInf). The Asx/Sx designation was based on a modified Jackson score computed from the self-reported clinical symptoms [10, 11]. The Inf/UnInf designation was determined from viral shedding data: a subject was declared infected if the viral titers exceed a high threshold at any time point or if they exceed a lower threshold at any tow time points. Further details are provided in the Additional file 1 deposited to the GEO database (accession number GSE73072).

For the prediction analysis, we excluded 44 clinically ambiguous subjects due to inconsistencies between their declared symptomatic status and measured shedding status and 3 subjects that had no Affymetrix gene probes collected. These 44 clinically ambiguous subjects were at some time either acutely infected but asymptomatic or not infected but acutely symptomatic. Thus the results reported below are restricted to the 104 unambiguously healthy (Asx and uninfected) and unambiguously ill (Sx and infected) subjects. Of these 104 unambiguous subjects 41 were infected subjects and 63 were uninfected subjects, and they will be designated as such in the sequel.

For these 104 subjects five time-specific infection states were determined on the basis of symptom scores and the viral shedding measurements. State 1 is “baseline” before inoculation. The other states occur after inoculation. State 2 is “Asx and UnInf” and applies to all post-inoculation samples of the uninfected subjects. States 3, 4 and 5 occur in the infected subjects after inoculation. State 3 is “Sx and pre-acute Inf,” State 4 is “Sx and acute Inf,” and State 5 is “Sx and post-acute Inf.” For each subject a healthy reference genechip sample was taken from baseline (state 1) and paired with one of the post-inoculation genechip samples taken from the subject’s post-inoculation time course (states 2-5). The state predictors, described below, were trained and tested on subsets of these paired samples. The 5 state designations are illustrated in Fig. 2c for the H3N2 DEE2 cohort and in corresponding figures for the HRV, RSV and H1N1 cohorts availabel on GEO (GSE73072). The Additional file 1 also contains details about the data collection and the criteria used for designating the states from shedding and symptom score data.

Mathematically, we denote the i-th microarray sample as the $p = 12,023$ dimensional vector $\mathbf{x}_i$. The i-th sample is labeled as $y_i \in \{1, 2, \ldots, K\}$ corresponding to one of the $K = 4$ possible infected/symptomatic states. The subject from whom the i-th sample was collected is denoted as $s_i \in \{1, \ldots, m\}$, where m is the total number of subjects. Figure 1 illustrates the time vs. subject matrix layout of the challenge study. The time instant labeled 0 (white vertical line) corresponds to the time of inoculation. The location of a hypothetical reference sample and target sample for a given subject s_i is shown in the figure. For illustration, Fig. 2 shows the titration and symptom data collected from subjects in the H3N2 DEE2 study. Similar figures for the other studies, summarized in Table 1, are available on GEO (GSE73072).

Prediction algorithms

To establish and quantify the value of including a subject’s reference sample, we implement a state-of-the-art automated prediction algorithm that performs variable selection and accomodates a reference sample in addition

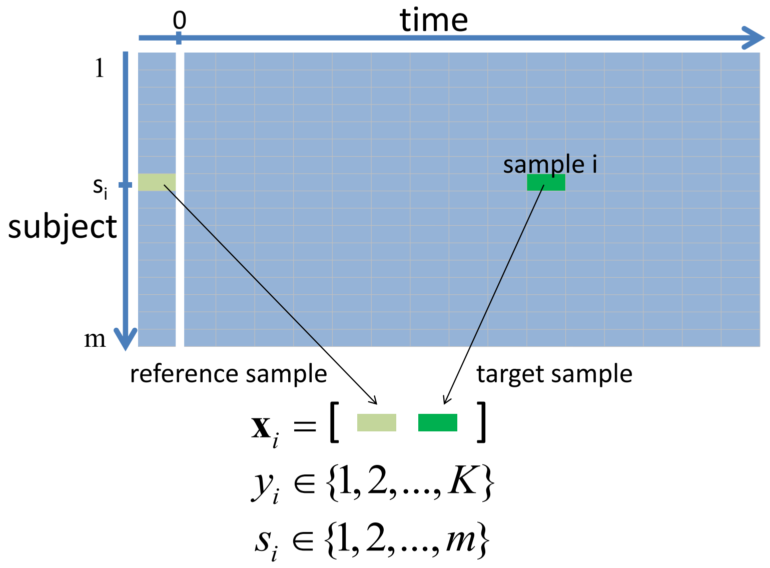

Fig. 1 Sample data layout for standard and reference-aided prediction. Each cell in this matrix corresponds to a sample $\mathbf{x}_i$ of a subject s_i taken at some time during the viral challenge study. The corresponding infection state label y_i for each target sample $\mathbf{x}_i$ is shown in Fig. 2.C for the particular case of the H3N2 challenge study. The standard predictor tries to predict the state using only the target sample. The reference-aided predictor uses both the target sample and the reference sample taken prior to inoculation time, denoted by the white vertical line

to a target sample. The predictor for the state y_i is learned from the biomarker data $\mathbf{x}_i$ using a supervised sparse multi-block multi-class classification algorithm, described in detail below. The different classes classified by the algorithm correspond to the different infection states. Sparsity forces the algorithm to select a small number of biomarkers (genes) from the 12,023 possible biomarkers. The imposition of sparsity is required in order to minimize overfitting error since the number of samples available to train the classifier is much smaller than the total number of biomarkers [12, 13]. The multi-block structure is used to force the reference-aided classifier to use the same subset of biomarkers for the paired reference and target samples in the classifier function. More specifically, as discussed below, for the reference-aided predictor there are two blocks corresponding to, respectively, the gene probe values in the reference sample and the target sample. For the standard predictor there is only one block corresponding to the gene probe values of the target sample.

A classifier is a function that operates on a data point $\mathbf{x}$ (the input) and produces a decision $\hat{y}$ (the output) about the class, where $\hat{y} \in \{1, \ldots, K\}$. In machine learning the classifier function is optimized to achieve the best possible classification accuracy over a set of training samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$, which are typically a subset (the training set) of all the available data. The result of this optimization is often averaged over many different training subsets of the data, e.g., by random resampling or leave-one-out resampling, a process called classifier cross-validation [12]. Among the many different algorithms available for multi-class classification the support vector machine (SVM) is one of the most prevalent. There are two common strategies for multi-class classification that have been proposed: (1) solving the multi-class problem by a series of binary SVM classifiers [14, 15]; (2) formulating a single unified multi-class SVM [16–21]. In this paper we adopt the latter more direct approach to multi-class classification.

As described in [19], the unified K-class classifier is a scoring based algorithm that classifies the input by computing its score for each class and outputs the class label associated with the maximum score. Specifically, given K functions $f_1, \ldots, f_K$ the unified K-class classifier outputs the decision $\hat{y} = \arg\max_k f_k(\mathbf{x})$ These functions assign confidence scores to the input $\mathbf{x}$ and can be chosen as linear functions of the form

$$f_k(\mathbf{x}) = \mathbf{w}_k{}^T \mathbf{x} + b_k, \qquad k = 1, \ldots, K \quad (1)$$

where $\mathbf{w_k} \in \mathbb{R}^\mathbf{P}$ is a (column) vector of p weights $\{w_{ki}\}_{i=1}^p$ and $b_k \in \mathbb{R}$ is a scalar offset. While other forms of the score functions are also common, e.g., kernelized linear, polynomial or sigmoidal functions, we will use the linear function (1) to design both the standard and reference-aided predictor.

Since there are many fewer samples (n) than variables (p) it is desirable to reduce the number of biomarkers used by the classifier in order to minimize overfitting errors [12]. This can be accomplished by constraining the weight vectors $\{\mathbf{w_k}\}_{\mathbf{k=1}}^{\mathbf{K}}$ to have common sparsity, i.e., the $\mathbf{w_k}$'s have many common entries equal to zero. Defining the

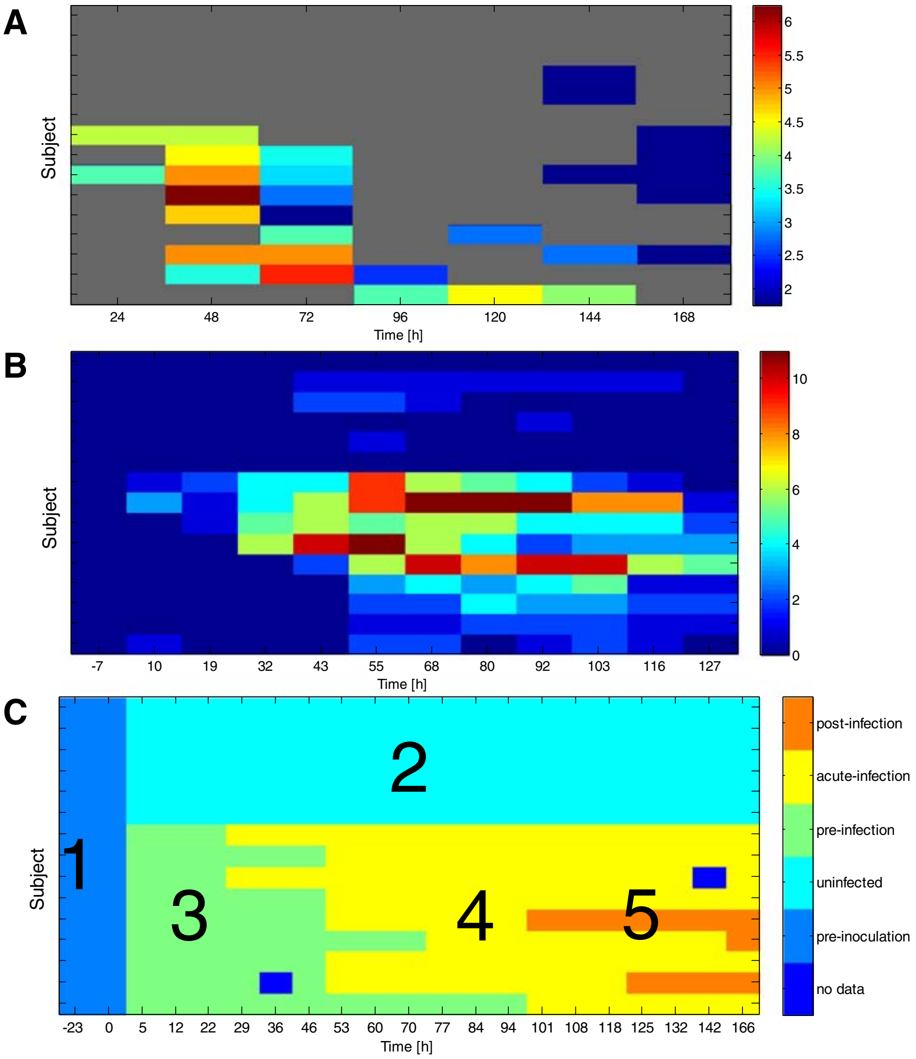

Fig. 2 H3N2 DEE2 challenge study viral shedding map (**a**), symptom score map (**b**), infection state map (**a**). **a** shows viral titration measurements for each subject at each sample time. **b** shows the sum of the 10 self-reported symptom scores. We use the measurements in A and B to designate each subject as infected (Inf), noninfected (UnInf), symptomatic (Sx), or asymptomatic (Asx). The subject dedignations can be found in the Additional file 1. Subjects whose titer scores and symptom scores agree, i.e., those who are either infected and symptomatic or uninfected and asymptomatic, are used for training the predictors. We assign 5 state labels to these subjects, as shown in (**c**), that correspond to baseline reference (state 1), UnInf (state 2), pre-accute Inf (state 3), acute Inf (state 4), and post-acute Inf (state 5). The onset and offset time of detectable titration are used to set the boundaries between class 3 and 4 and class 4 and 5 respectively (see Additional file 1)

$K \times p$ weight matrix $\mathbf{W} = [\mathbf{w_1}, \ldots, \mathbf{w_K}]^\mathbf{T}$ and K-element vector $\mathbf{b} = [\mathbf{b_1}, \ldots, \mathbf{b_K}]^\mathbf{T}$, this common sparsity constraint is expressed as $\mathbf{W}$ having many columns identically equal to zero. This is a form of structured sparsity [22], also called group sparsity, that is mathematically expressed as the "mixed ℓ_1/ℓ_0 norm" constraint on $\mathbf{W}$: $\|\mathbf{W}\|_{\mathbf{1,0}} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \|\mathbf{w_{(j)}}\|_\mathbf{0} \le \mathbf{q}$, where q is much less than p, $\mathbf{w_{(j)}}$ is the j-th column of $\mathbf{W}$ and $\|\mathbf{u}\|_\mathbf{0}$ is a function that counts the number of non-zeros in a K-element vector $\mathbf{u}$. A convex relaxation of this constraint, adopted for the classifier used in this paper, is the mixed ℓ_1/ℓ_2 norm constraint [7, 23]:

$$R(\mathbf{W}) = \|\mathbf{W}\|_{\mathbf{1,2}} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \|\mathbf{w_{(j)}}\|_\mathbf{2} \le \mathbf{q}, \tag{2}$$

where $\|\mathbf{u}\|_\mathbf{2}^\mathbf{2} = \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{u_k^2}$ denotes the ℓ_2 or Euclidean norm of $\mathbf{u}$.

To specify the unified multiclass classifier it therefore suffices to select the sparse weights $\mathbf{W}$ and offsets $\mathbf{b}$ defining (1). These are learned from the data by solving the sparsity penalized empirical risk minimization problem:

$$\min_{\mathbf{W},\mathbf{b}} \frac{1}{n} \sum_{i=1}^{n} V(\mathbf{W}, \mathbf{b}, \mathbf{x}_i, y_i) + \lambda R(\mathbf{W}), \tag{3}$$

where $R(\mathbf{W}) = \|\mathbf{W}\|_{\mathbf{1,2}}$ is the relaxed group sparsity inducing regularization function (2), $\lambda > 0$ is a regularization parameter, and $V(\mathbf{W}, \mathbf{b}, \mathbf{x}_i, y_i)$ is an empirical loss function, depending on the parameters and the training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, that penalizes errors between the classifier output $\hat{y}_i$ and the true class label y_i.

As contrasted to the standard multi-class classifier, developed above, the reference-aided multi-class classifier uses a higher dimensional subject-specific input $\mathbf{x_s}$, which is a $2p$-dimensional vector, constructed by concatenating the paired reference and target samples of subject s, denoted $\mathbf{x}_\mathbf{s}^{\mathbf{ref}}$) and $\mathbf{x}_\mathbf{s}^{\mathbf{target}}$, into a single vector. Note that the j-th and $(j + p)$-th elements of the vector $\mathbf{x_s}$ correspond to the same biomarker (gene probe). With the subject-specific input $\mathbf{x_s}$ the weight matrix $\mathbf{W}$ of the multi-class classifier is $K \times 2p$ dimensional. Figure 3 illustrates the group sparsity constraint that enforces that $\mathbf{W}$ select only a few common biomarker variables from each pair. Such sparsity structure can be induced into the predictor by modifying the penalty function in (3) from the mixed ℓ_1/ℓ_2 norm (2) to the following convex function:

$$\begin{aligned} R(\mathbf{W}) &= \sum_{j=1}^{\mathrm{p}} \left|\left|\tilde{\mathbf{w}}_{(j)}\right|\right|_2 \\ \tilde{\mathbf{w}}_{(j)} &= \left[\mathbf{w}_{(j)}^T, \mathbf{w}_{(j+p)}^T\right]^T. \end{aligned} \tag{4}$$

Both the standard and the reference-aided multi-class classifier are learned by minimizing a risk function of the form (3). For the purposes of this paper, we adopt the multi-class hinge loss function $V(\mathbf{W}, \mathbf{b}, \{\mathbf{x_i}, \mathbf{y_i}\})$ proposed in [19] which, along with the proposed mixed ℓ_1/ℓ_2 norm sparsity penalty R, makes (3) a convex but non-smooth optimization problem. This problem can be solved with iterative optimization methods and we use an optimization algorithm, developed in [7, 23], that is based on variable splitting [24]. The optimization algorithm used in this paper is given as Algorithm 3.

A two-stage adaptive group sparsity method was used to further reduce the danger over-fitting in training the classifier. This method is an extension of the adaptive lasso [25, 26] to the group sparse multi-class classification framework developed above. The method is implemented as follows. Suppose after solving (3) we have an initial estimate $\mathbf{W_{init}}, \mathbf{b}$ of the classifier parameters. Then we refine this estimate by solving (3) once more, except in place of $R(\mathbf{W})$ in (2) we use

$$R_{adapt}(\mathbf{W}, \mathbf{W_{init}}) = \sum_{\mathbf{j=1}}^{\mathbf{p}} \frac{\|\mathbf{w_{(j)}}\|_\mathbf{2}}{\|\mathbf{w_{init,(j)}}\|_\mathbf{2}},$$

for the standard classifier. A two-stage adaptive reference-aided classifier is defined similarly except that in the summand defining R_{adapt} the weights $\mathbf{w}_{(j)}$ and $\mathbf{w}_{init,(j)}$ are respectively replaced by $\tilde{\mathbf{w}}_{(j)}$ and $\tilde{\mathbf{w}}_{init,(j)}$.

Algorithm 1: Reference-aided classification algorithm. The algorithm minimizes the convex mixed ℓ_1/ℓ_2 mixed norm sparsity penalized empirical risk (3) by applying variable splitting in an ADMM framework [24]. The minimization step 3 is solved by sequential dual method discussed in [27, 28] and the minimization step 5 is solved non-iteratively by applying the proximal operator for the mixed ℓ_1/ℓ_2 norm [29]. The parameters λ and μ are user-defined tuning parameters that control the column-sparsity of $\mathbf{W}$ and the convergence rate, respectively. In the experiments conducted in this paper, λ is selected by cross-validation, μ is fixed to be 10

```
1 set $\tau = 0$, choose $\mu > 0, \mathbf{M}_0, \mathbf{W}_0, \mathbf{D}_0$
2 **while** *stopping criterion is not satisfied* **do**
3     $\mathbf{W}_{\tau+1} = \arg\min_{\mathbf{W}} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \frac{\mu}{2}||\mathbf{W} - \mathbf{M}_\tau - \mathbf{D}_\tau||_F^2$
4     s.t. $\forall i, k \ (\mathbf{w}_{y_i}^T \mathbf{x}_i + b_{y_i}) + \delta_{y_i,k} - (\mathbf{w}_k^T \mathbf{x}_i + b_k) \geq 1 - \xi_i$
5     $\mathbf{M}_{\tau+1} = \arg\min_{M} \lambda \sum_{j=1}^{p} ||\tilde{\mathbf{m}}_{(j)}||_2 + \frac{\mu}{2}||\mathbf{W}_{\tau+1} - \mathbf{M} - \mathbf{D}_\tau||_F^2$
6     $\mathbf{D}_{\tau+1} = \mathbf{D}_\tau - \mathbf{W}_{\tau+1} + \mathbf{M}_{\tau+1}$
7     $\tau = \tau + 1$
```

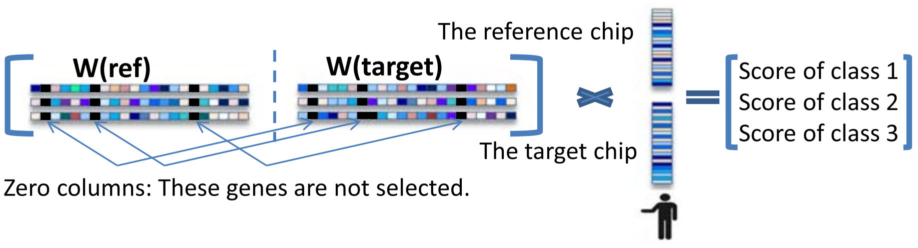

Fig. 3 Multi-block group sparsity structure for proposed reference-based predictor. A multi-block multi-class classifier of $K = 3$ classes applies a $K \times 2p$ matrix $\mathbf{W}$ =[$\mathbf{W}$(**ref**), $\mathbf{W}$(**target**)] to the combined vector of probes on reference and target gene chips in order to form the vector of scores for each of the $K = 3$ states. The two blocks of the classifier correspond to the block of weights $\mathbf{W}$(**ref**), applied to the reference sample, and the block of weights $\mathbf{W}$(**target**), applied to the target sample. The classifier decision rule is to assign the state label that corresonds to the maximum score. The blocks share sparsity, denoted by the black columns in the weight matrix, which designate columns that are identically zero. Variables associated with these zero columns are not selected by the classifier

Results

Here we demonstrate that the reference-based predictor described in the Methods section results in improved state classification accuracy with a smaller panel of biomarkers for the challenge study dataset studied. The standard and reference-aided predictors were trained and tested separately on data from each virus category. These data are denoted H3N2, H1N1, HRV and RSV, respectively, for the pooled data from the two H3N2 studies, the pooled data from the two H1N1 studies, the pooled data from the two HRV studies, and the data from the single RSV study (see Table 1). These four virus-specific datasets consisted of $m = 29, 24, 31, 17$ subjects, respectively. Each of the virus-specific datasets was divided into m training-test partitions containing $m-1$ subjects for training by successively removing subjects one at a time for testing (leave-one-out partitions). For each of these subsets the predictors were trained by minimization of the empirical risk (3), using 2-fold cross-validation to first select the regularization parameter λ with the mixed norm sparsity constraint R, and an additional 2-fold cross-validation to select the regularization parameter with the adaptive sparsity inducing regularizers R_{adapt} discussed at the end of the Methods section on Prediction algorithms. The prediction performance and variable selection frequencies were assessed by averaging the predictor's state misclassification errors over the m training-test partitions.

Furthermore, each of the variables in each training set was standardized to z-scores by subtracting the sample mean and dividing by the sample standard deviation, where these sample statistics were computed over the samples in the training set. The biomarkers of each subject in the each test set were standardized using the sample mean and standard deviation computed from the associated training set. To reduce possible bias due to imbalance in the numbers of samples across classes (states), at each training iteration we applied uneven cost to each sample such that the average sampling proportions among the classes were identical.

The accuracy of the reference-aided predictor is presented in row 1 of Table 2 for each of the four virus-specific datasets. For comparison the accuracy of three other predictors is shown in the remaining rows of Table 2. The proposed reference-aided predictor achieves better performance in terms of average error rates. This improvement is achieved using biomarker panels with significantly fewer genes, as compared to the standard predictor in row 2 of Table 2. The number of genes selected by the standard predictor was optimized by the two-stage adaptive cross-validation procedure described in Section Methods. Row 3 shows the performance of a constrained standard predictor when the regularization parameter is selected so that it uses approximately the same number of genes as the proposed reference-aided predictor. The actual genes selected by this constrained predictor differ from those slected by the reference-aided predictor and the performance is significantly worse than the unconstrained predictor in row 2. Row 4 shows the performance of a differential predictor that is implemented by applying the standard predictor to the difference between target and reference sample $\mathbf{x}_s^{\mathbf{target}} - \mathbf{x}_s^{\mathbf{ref}}$. This differential predictor is implemented by restricting the reference-aided predictor to the case where the reference and target weights have identical magnitudes but opposite signs. The performance in row 4 of Table is better than that of the standard predictors (rows 2 and 3) but worse than that of the proposed predictor (row 1). This indicates that simple differencing of the target and reference samples, corresponding to using all probes as contrast genes and none as reinforcement genes, leads to a less effective predictor than the one proposed in row 1 of Table 2.

Table 2 Average accuracy (error rate) and average size of the gene panel (number of selected genes) selected by automated predictors of infected state (class) for different viral challenges (data from DEE2/DEE5, DEE3/DEE4 and HRV-UVA/HRV-Duke were pooled and designated as H3N2, H1N1, and HRV in table). Shown are the reference-aided predictor (w/ baseline reference), the standard predictor (w/o baseline reference), the standard predictor with constraints to have the similar complexity in terms of the number of genes as the reference-aided predictor (constrained standard predictor), and the differential predictor. Across all viruses the inclusion of the baseline reference results in a decrease in error rate and a reduction in the number of genes used by the predictor. The reported accuracy and size of panel were computed by cross-validation of the predictors using leave-one-subject-out resampling to partition the data into training and target samples

Virus	H3N2	H1N1	HRV	RSV
Classes	2345	2345	2345	234
w/ baseline reference				
Error rate	0.386	0.480	0.483	0.635
Number of selected genes	200.34	287.54	287.55	238.53
w/o baseline reference				
Error rate	0.448	0.508	0.526	0.728
Number of selected genes	327.90	420.25	358.48	593.82
Constrained standard predictor				
Error rate	0.458	0.517	0.544	0.737
Number of selected genes	207.38	271.83	298.00	243.29
Differential predictor				
Error rate	0.415	0.492	0.571	0.678
Number of selected genes	481.48	1209.46	464.39	503.29

The per-chip misclassification error rates, computed by aggregating over the m training-test partitions, are shown in Fig. 4 as heatmaps over times and subjects for each virus-specific dataset. Figures 4 and 5 show that

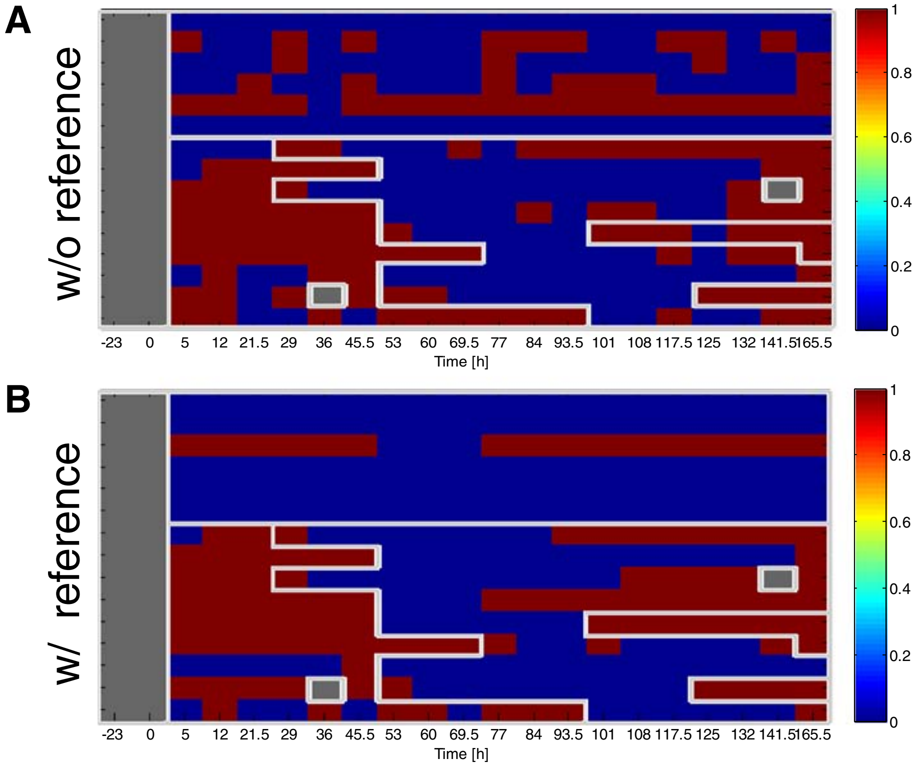

Fig. 4 Heatmaps of predictor error rates for H3N2 DEE2 dataset. The heatmaps of the sample-specific predictor error rate for classifying states 2,3,4 and 5, defined in Fig. 2, for samples in the H3N2 DEE2 dataset. The top figure (**a**) shows the error rates of the standard predictor (no reference). The bottom figure (**b**) show the corresponding results of the reference-aided predictor

the proposed reference-aided predictor achieves the most improvement in predicting states 2 (UnInf) and 3 (pre-acute Inf), which are the most difficult states to classify.

The different roles of the biomarkers that were automatically selected by the reference-aided predictor yields insight into these accuracy improvements. For concreteness, we focus on the H3N2 dataset. See Additional file 1 for analysis of the biomarkers selected in the HRV, RSV and H1N1 datasets. The variables that were frequently selected by the reference-aided predictors are shown in Fig. 6a for H3N2. The genes in this figure were selected with frequency at least 70 %; i.e., they were included in the trained reference-aided classifier in at least 70 % of the training-test partition sets. The genes are ordered according to the cluster order illustrated in panel B. The bars in Fig. 6a represent the average value of the weight $w(ref)$ applied to a specific gene in the reference sample (R), denoted as yellow bar, and the weight $w(target)$ applied to the same gene in the target sample (T), denoted as a green bar. These selected biomarkers can be grouped into two categories: (1) contrasting genes (R and T weights have opposite sign); and (2) reinforcing genes (R and T weights have the same sign). Contrasting genes are selected by the predictor for their differential expression between R and T, while reinforcing genes do not differentially express but rather serve to normalize the other variables (recall that the gene probes were log transformed in the RMA normalization).

An example of a contrasting gene is the interferon induced gene IFI27 which differentially expresses between R and T for states 2, 3, 4 and 5. Interestingly, the signs of the R and T weights for IFI27 are reversed in the score function for state 2 (UnInf) as compared to their signs for the score functions of the other three states (pre-acute Inf, acute-Inf, post-acute Inf): a relative decrease in IFI27 from reference to target sample induces a high UnInf score while a relative increase induces a high Inf score. An example of a reinforcing gene is the immunoglobin lambda variable IGLV3-25 that plays the role of reinforcing the UnInf state (positive contribution to state 2 score) to the detriment of the Inf states (negative contributions to state 3, 4, 5 scores). Another example, that reinforces the Inf states instead of the UnInf state, is the NEDD4 binding protein gene N4BP3. Note that some genes, e.g., PEX13 and LOC26010, take on a contrast role for some states while they take on a normalizing role for other states. Many of the genes that were selected by the reference-aided predictor were not selected by other predictors

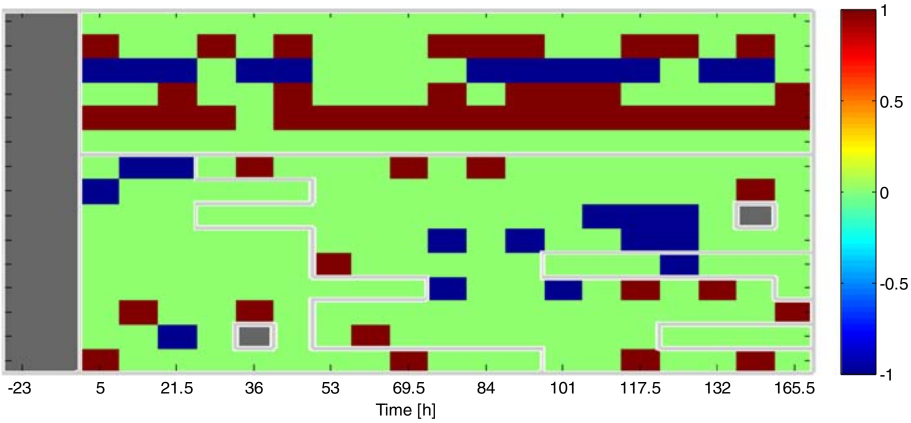

Fig. 5 Heatmaps of predictor error rates for H3N2 DEE2 dataset. The heatmaps of the sample-specific predictor error rate for classifying states 2,3,4 and 5, defined in Fig. 2, for samples in the H3N2 DEE2 dataset. The figure shows the difference between the error rates of the standard predictor (no reference) and the error rates of the reference-aided predictor (with reference) for H3N2 DEE2. The entries in red have higher error rate using the standard predictor than the reference-aided predictor, and vice versa

studied in Table 2. (See Sec. 4.1 in Additional file 1). Since the differential predictor can only form contrasts between reference and target gene probes none of the reinforcing genes. Indeed, we did not find the reinforcing genes, e.g., IGVL3-25, NBP3 and MYOM2 were selected by the differential predictor. The lack of reinforcement genes deprives the differential classifier of potential normalizing variables leading to poorer performance.

The average expression levels over time of the frequently selected H3N2 genes shown in Fig. 6a are shown as heatmaps in Fig. 6b, where the expression levels are averaged over the uninfected and infected subjects, respectively, in the left and right heatmaps. Notice that the reinforcing genes, such as IGVL3-25, NBP3 and MYOM2, appear to be related to susceptibility since the expression levels are substantially higher or lower in the uninfected population than in the infected population, even before viral inoculation. The expression levels of contrast genes such as IFI27, PEX13 and LOC26010, are relatively stable in the uninfected subjects but rapidly increase as an infected subject enters the acute Inf phase (roughly 40 h after inoculation).

The improved accuracy of the reference-aided predictor can be visualized by rendering a scatter plot of the vector of confidence scores, defined in (1), over all of the samples. In Fig. 7 the scatterplot of the $K = 3$ dimensional vector of scores is shown for the H3N2 pooled challenge studies. In the scatterplot each of these vectors has been given a different color depending on the state of the particular target sample at which the score is evaluated. The right panel of Fig. 7 shows the scatterplot of confidence scores computed with the classifier weight matrix $\mathbf{W}$ by reference-aided predictor, averaged over the m training-test partitions. The left panel shows the associated scatterplot for the standard predictor, implemented with the average weights. Notice that the scores are better separated when the references are taken into account. Note that both with and without reference-aided training, among all pairs of states, discrimination between state 3 (pre-infection) and state 4 (acute infection) is the most difficult. This is explained by the fact that states 3 and 4 apply to the same group of subjects, namely, those that develop acute symptomatic infection with significant levels of viral shedding. For these subjects the ARI signature is developing during state 3 and comes into full bloom during state 4, thus making these states more similar to each other.

There are common genes in the four sets of biomarkers selected by the predictors trained separately on the HRV, RSV, H3N2, and H1N1 viral-specific datasets. We call these common biomarkers pan-viral predictive genes. A Venn diagram of all the biomarkers that have been found in each virus study is shown in Fig. 8, and the pan-viral predictive genes in the intersection among all of the studies are listed in Table 3. Heatmaps, analogous to those shown in the H3N2 heatmap Fig. 6b, are shown for these pan-viral predictive genes in Fig. 9 for all 4 viral-specific datasets. Similarly to the H3N2 heatmap, some genes seem to have a reinforcement role, such as C7orf58, and others have a contrasting role, such as IFI27.

Discussion

The proposed reference-aided predictor significantly outperformed the standard predictor that does not use the reference, implemented with a single block multi-class classification algorithm. Specifically, the reference-aided predictor achieved an average (cross-validated) state prediction accuracy improvement of: 14 % for RSV, 13 % for H3N2, 9 % for HRV, and 6 % for H1N1. Remarkably, for all of these viral challenges this gain in accuracy was achieved with a smaller panel of genes: 60 % fewer for

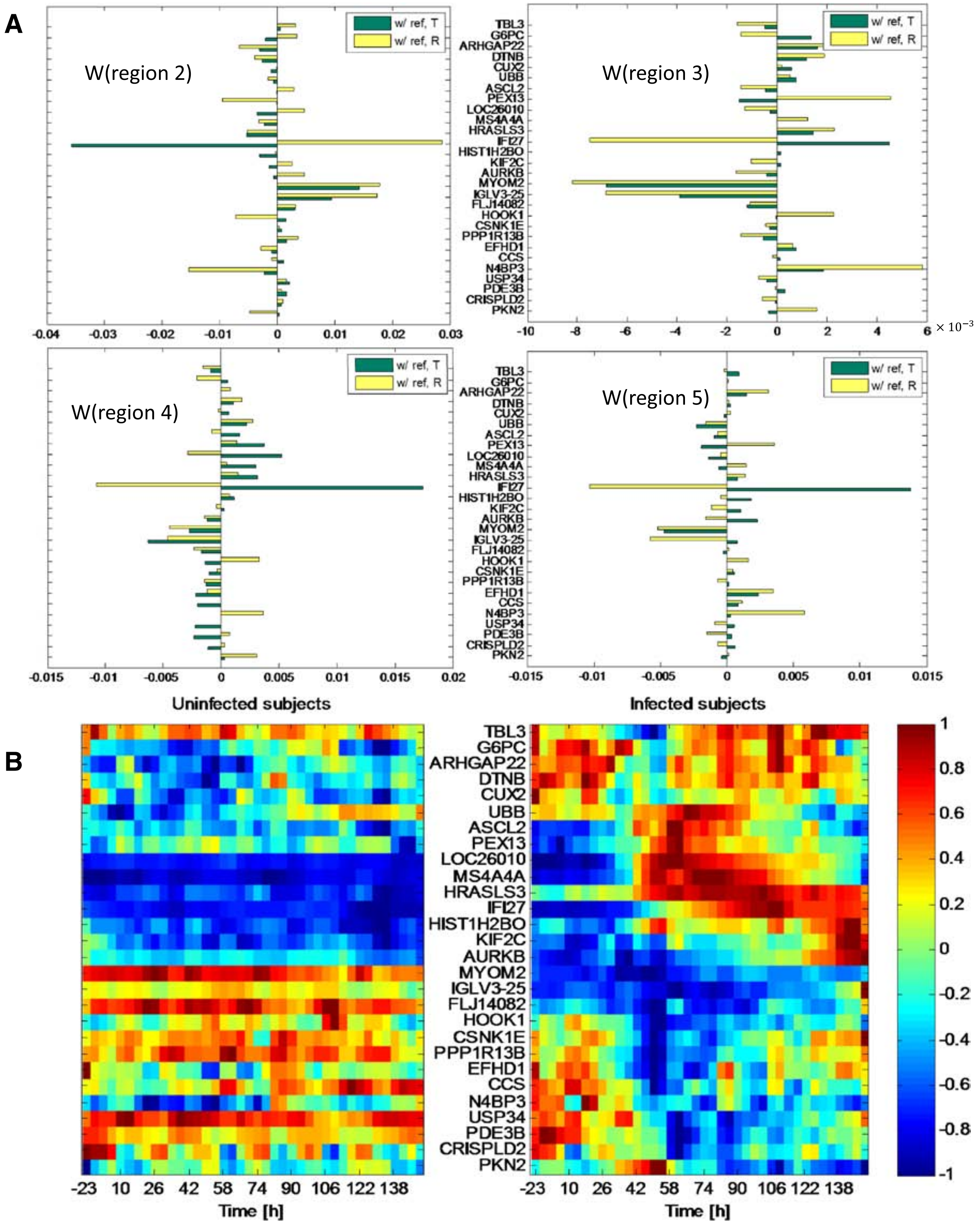

Fig. 6 Biomarkers selected by reference-aided predictor. The top figures in (**a**) show the genes selected by the proposed reference-aided predictor with selection frequency $\geq 70\,\%$ for the 4 different score functions for states 2,3,4,5. The value of the classifier weights for each of the score functions are shown as yellow bars (weights applied to reference sample R) and green bars (weights applied to target sample T). Note that genes having yellow and green bars of opposite sign are contrasting information in R vs T while genes having these bars with the same sign are reinforcing information in R and T. The bottom figures in (**b**) show the expression of the genes shown in (**a**) averaged over the uninfected subjects (*left*) and infected subjects (*right*). The expression levels are normalized such that the maximum and minimum of each gene achieve 1 and -1 respectively. Let the averaged expression at time t be $z(t)$, the maximum of $z(t)$ be z_{max}, and the minimum be z_{max}. The normalized expression levels are computed as $\tilde{z}(t) = 2 \times [z(t) - 0.5 \times (z_{max} + z_{min})] / (z_{max} - z_{min})$

RSV, 39 % fewer form H3N2, 20 % fewer for HRV, and 31 % fewer for H1N1, as shown in Table 3. This suggests that by including such a reference sample with the target sample, the reference-aided predictor can build a more parsimonious description of the infected vs uninfected phenotypes. As seen in the previous section, this better description translates into improved prediction of infection state. There are several possible reasons that the reference-aided predictor achieves improvement in accuracy while using significantly fewer predictor variables. First, by pairing an individual's target sample to his baseline reference sample, the reference-aided predictor turns the standard predictor into an individualized predictor. Second, the availability of a baseline sample allows the predictor to learn genes that display a contrast between an individual's healthy baseline and sick target samples. Third, even though the reference-aided predictor has to learn twice as many coefficients, our predictor sparsity penalty forces most of these to zero, resulting in a more parsimonious predictor that minimizes overfitting errors.

Moreover, many of the genes in the panels selected by the automated reference-aided predictor and the standard predictor were different. The genes in the standard predictor were similar to the acute respiratory infection (ARI) signature reported in [4]. The reference-aided predictor selected a panel of genes that fall into two classes: 1) contrast genes that exploit the fact the the baseline reference differs significantly from the target sample; 2) reinforcement genes that do not differ significantly but are used by the classifier for baseline normalization. Specifically, for a contrast gene, the predictor forms a weighted average of the baseline and target expression levels using two coefficients having opposite sign. For a reinforcement gene these two coefficients have the same sign. We caution that these definitions only make sense when the two coefficients have similar magnitudes.

The reference-aided predictor identified the pan-viral predictive genes as some of the best subject-specific genes that either reinforce or contrast expression in the subject's reference and target samples. Many of these genes are not included in the standard predictor that does not use a reference. Table 3) indicates that the reference-aided predictor found 8 pan-viral predictive genes (C7orf58, CCR1, IFITM3, MTMR12, NUDT13, ORM2, TSPAN8, and TSTAT3) that were not found by the standard predictor. While some of these are orthologs to genes in the standard predictor, others might represent additional pathways that can only be picked up by analysis of paired samples. For example, some studies suggest that IFITM3 is important for intrinsic viral resistance. Specifically, in vitro studies show that many pathogenic viruses' replication can be restricted by genes in the interferon inducible transmembrane (IFITM) protein family, and it has been found that IFITM3 plays an important role in the host's defense against influenza A virus [30]. Furthermore, it has been reported that during RSV infection deletion of CCR1 leads to attenuated pathophysiologic responses [31] and, as reported by the NCBI gene database, an important acute phase plasma protein is encoded by orosomucoid 2 (ORM2), which can be stimulated during acute inflammation and be may an important factor in immunosuppression. Other genes among the 8 genes specific to the reference-aided predictor have no obvious function in immune response but appear to have been selected to serve as normalization genes.

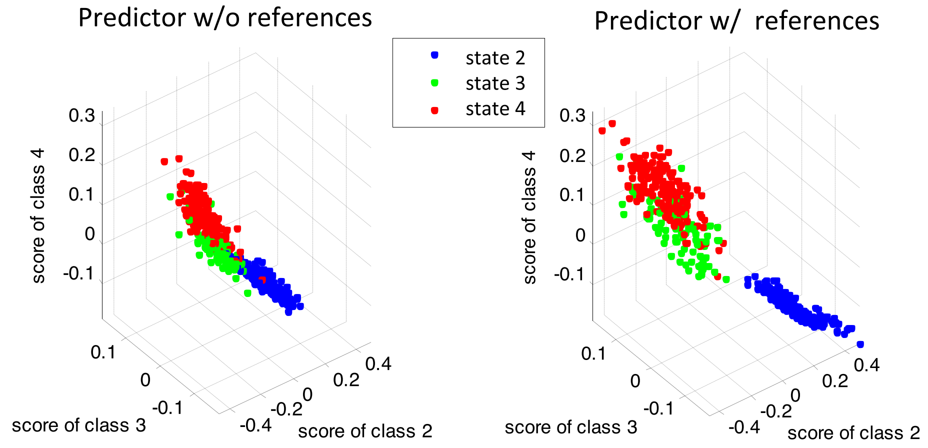

Fig. 7 Scatter plots of the vector of scores over the H3N2 samples. Shown are scatterplots of the score functions in the multi-class classifier evaluated for the standard predictor (w/o reference) at left and for the reference-aided predictor (w reference) at right. Points with a particular denote the vector of scores for states 2, 3, and 4 for samples labeled with a particular state. The scores are better separated when the references are taken into account. Especially note that the class 2 (UnInf) scores (labeled in blue) are much better separated

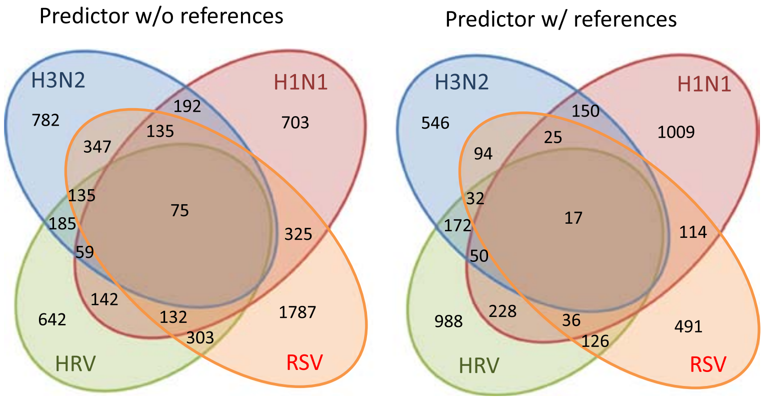

Fig. 8 Venn diagram of the genes selected in HRV, RSV, H3N2 and H1N1 datasets. Indicated are the intersections of the genes that were selected at least once by the standard predictor (left) and reference-aided predictor (right) in each of the virus-specific datasets. The list of the genes in the intersection among all 4 datasets is listed in Table 3. These are called pan-viral predictive genes

Many of the genes that were selected by both the predictors (with and without baseline reference) are well known transcription factors in host immune response. Specifically, interferon-stimulated genes, such as IFI44L, produce cellular factors that protect cells against invading viral pathogens [32]. OAS1 has been identified to be relevant to apoptosis, which eliminates the cells that have damaged DNA or have experienced uncontrolled proliferation [33]. Therefore, it may prevent viral replication by eliminating virus-infected cells. Indeed we observe the steady up-regulation of OAS1 during acute-infection.

Table 3 Biomarkers that have been selected in every virus study. Notice that these genes are in the intersection among the 4 virus studies in Fig. 8

Method	Genes
w/o baseline reference	ACP2 ADM AFFX-r2-Bs-dap-3_at AMFR ASGR2 ATP9A BAIAP3 BTN2A2 C4BPA CDC20 CHI3L1 CYP26B1 DAP DCXR DSP EIF2AK2 ERC1 FBXL8 FOLR3 GDF9 H1F0 HMBOX1 HPCAL4 HRASLS3 IFI27 IFI44 IFI44L IFIT1 IFIT3 IGFBP6 IGFBP7 IGHV1-69 IGLV4-60 IL1F9 IRF5 IRF9 ISG15 LOC643224 LY6E MAP7 MFAP3 MICB MMP1 MX1 MYOM1 NARFL NGFRAP1 NKX3-1 NQO2 NUDT4 OAS1 OLFM1 PAPSS2 PGGT1B PODXL PRRG4 PRSS21 RGS20 RIMBP2 RSAD2 SCO2 SERGEF SERPINE2 SLC30A4 SP100 SPATA20 STAT1 TCL1B TMEM140 TNNT1 TSPAN15 TTLL4 TUBB6 UAP1L1 ZNF701
w/ baseline reference	C7orf58 CCR1 H1F0 IFI27 IFI44L IFITM3 IL1F9 IRF9 ISG15 MTMR12 NUDT13 OAS1 ORM2 RSAD2 STAT1 TSPAN8 TSTA3

The role of ISG15 in innate immunity to viral infection has been studied in [34], and has been found to be highly expressed upon viral infection. IL1F9 is reported in [35] to be be up-regulated in cells involved in immune responses induced by HRV. IRF9 is one of the transcriptional activators, along with STAT1 in the ISGF3 transcriptional complex, which stimulates the expression of the interferon-inducible genes, e.g., IRF7 for antiviral responses [36]. IRF7 is one of the interferon regulatory factors, which regulates transcriptional activities to induce cellular response to the invasion of viruses. It has been reported to induce the interferon inducible genes like IFI27 in infected cells [37, 38]. Further studies suggest that IRF7 controls both the innate immunity and adaptive immunity [39, 40]. Several of the pan-viral predictive genes, e.g, IFI44L, IFI27, and OAS1 are related to type-I interferon antiviral response, and have been reported to constitute pathways regulating inflammatory response [5, 41–43].

In this paper, several viral challenge study datasets were used to demonstrate the intrinsic value of the proposed reference-aided method for biomarker selection and improved performance in predicting symptomatic infection. These findings are, of course, specific to the setting of our challenge studies and the value for clinical applications needs to be further explored. Two issues stand in the way of direct generalization of our findings to clinical medicine.

The first issue is that each enrolled subject's healthy reference sample was collected within 24 h of exposure to the viral pathogen. An open question is whether the demonstrated performance advantages of the proposed method would generalize to the clinically relevant case where the reference sample collected in the more distant past. Such a generalization will require testing our method on observational data collected over a longer baseline

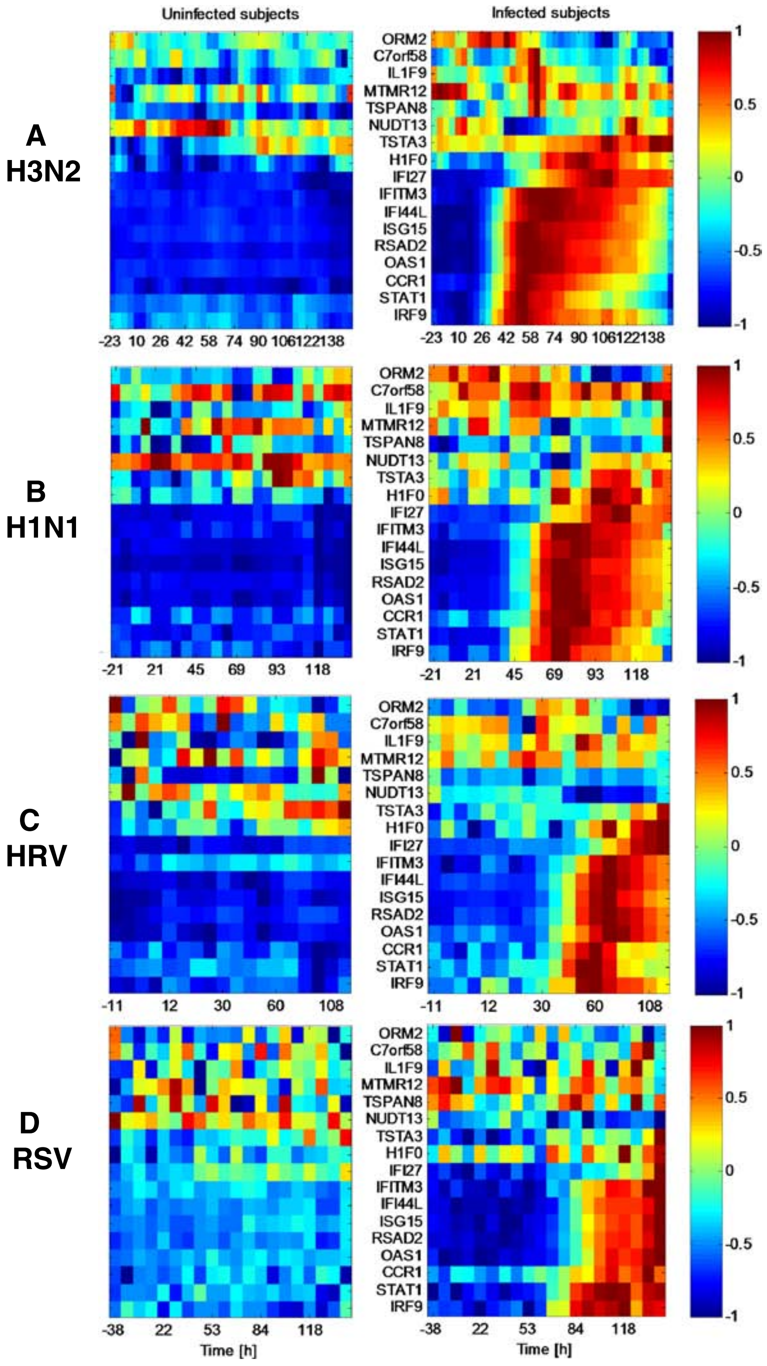

Fig. 9 Expression profiles of reference-aided pan-viral predictive genes. Average expression profiles of the pan-viral predictive genes discovered by the reference-aided predictor (genes listed at bottom of Table 3) averaged over the uninfected subjects (*left*) and infected subjects (*right*) in each virus-specific dataset (**a** H3N2, **b** H1N1, **c** HRV, and **d** RSV). The expression levels are normalized such that the maximum and minimum of each gene achieve 1 and −1 respectively

period than 24 h prior to exposure. Given the expanding interest in discovery of temporal pathways for processes such as biochronicity, immunity, and aging, we can anticipate that such data will become available in the not so distant future.

The second issue is that the findings reported here are restricted to the subset of enrolled individuals who unambiguously reported health or illness, as measured by concordance between viral shedding and self-reported symptoms. All predictors have low average accuracy when ambiguous reports are included in the training data, even though the proposed reference-aided predictor maintains significant performance advantages over the other predictors to classify the state of infection (see Additional file 1: Sec. 5 for details). This poor performance on ambiguous subjects may signal the need for more complex non-linear modeling of gene expression for these subjects. On the other hand, such ambiguities may simply reflect the inadequacy of viral shedding and self reported symptom as reliable proxies for symptomatic illness.

In spite of these caveats, the framework presented here may be relevant to personalized medicine, where preventative and diagnostic medical testing could possibly benefit from availability of a recent personalized baseline reference. The reported results establish that, when used with a carefully designed classifier, inclusion of such a reference can improve the accuracy of classifiers of early onset infection based on gene expression assays. Furthermore, the variables selected by the predictor can give insight into the molecular discriminants that provide high contrast between healthy baseline and infection states. The referenced-based classifier framework we have developed can likely be extended to other diseases and diagnostic tasks, e.g., classifiying yellow fever [44, 45] or personal health monitoring [5]. For example, in a recently published paper by Chen et al. [46], the authors have demonstrated the ability of a personal ‘omics’ profile to reveal dynamic molecular and medical phenotypes by monitoring a single individual over 14 months. This might be modeled by our multi-block multi-class classifier framework, where the blocks partition the periods of health and sickness and the classes indicate different stages of infection and/or types of infection. The accuracy of such a multi-block classifier might also benefit from integration of other biomarkers, e.g., proteomic, metabolomic, antibody, into the predictor.

Conclusions

This paper developed reference-aided prediction as way to design personalized test panels and associated predictors when one can pair a target sample with a baseline reference sample from the same subject. The framework is applicable when a population of serial samples from multiple subjects are available. The proposed referenced-aided predictor uses the framework of learning sparse linear score functions in a multi-block multi-class support vector machine (SVM). However, other types of reference-aided predictors may also be worth investigating, e.g., using multi-block non-linear kernelized multi-class classifiers or multinomial logistic classifiers.

We used a large-scale respiratory virus challenge study to illustrate the advantages of reference-aided prediction. In this predictive health problem, pre-inoculation reference (baseline) samples of each subject are incorporated into the classifier along with post-inoculation target samples. Application of the reference-aided predictor demonstrated significant improvement in the accuracy of prediction of different stages of host immune response for infected and uninfected subjects. Furthermore, it achieved this improved accuracy using fewer biomarkers than a standard predictor that does not use a reference sample. Some of the biomarkers discovered by the reference-aided predictor are genes that exhibit high contrast in expression between the target and reference samples. Other biomarkers were discovered to be low contrast genes that use the reference sample to normalize the target sample.

With minor modification, the prediction algorithm used in this paper applies to more general serially sampled diagnostic tests than the single-reference/single-target predictor. For example, consider a predictor that labels the state of a particular subject at the time a target sample is acquired using both the target sample and at least one previously acquired sample. This framework is applicable to a wide range of applications in diagnostic medicine, drug discovery and biology. For example, when using a panel of biomarkers to test for health or disease of the subject, the anterior samples might correspond to the same panel taken when the subject was at a baseline of health. When testing for the specific stage of an advancing disease the anterior samples may be panels of previously acquired target samples. It is likely that in these situations, reference-aided predictors will similarly show accuracy benefits.

Availability of supporting data

The data has been deposited to the GEO database (accession number GSE73072).

Additional file

Additional file 1: Supplemental materials for the paper: An individualized predictor of health and disease using paired reference and target samples. (PDF 14,201 kb)

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

CWW, AKZ and GSG conceived and designed the experiments. TB and LPP participated in the subject designation. TL and AOH contributed analysis tools and analyzed the data. TL and AOH wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA), under the Predicting Health and Disease (PHD) and Biochronicity programs.

Author details

[1]Electrical Engineering and Computer Science Department, University of California, Berkeley CA, USA. [2]Center for Applied Genomics and Precision Medicine, Department of Medicine, Duke University, Durham NC, USA. [3]Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor MI, USA. [4]Center for Computational Biology and Bioinformatics, University of Michigan, Ann Arbor MI, USA.

Received: 2 February 2015 Accepted: 6 January 2016

[redacted]

References

1. Meldrum C, Doyle MA, Tothill RW. Next-generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. 2011;32(4):177.
2. Bortz E, García-Sastre A. Predicting the pathogenesis of influenza from genomic response: a step toward early diagnosis. Genome Med. 2011;3(10):67.
3. Lecuit M, Eloit M. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. Front Cell Infect Microbiol. 2014;4. doi:10.3389/fcimb.2014.00025.
4. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, Lucas J, et al. Gene expression signatures diagnose influenza and other symptomatic w respiratory viral infections in humans. Cell Host Microbe. 2009;6(3):207⊠17.
5. Huang Y, Zaas AK, Rao A, Dobigeon N, Woolf PJ, Veldman T, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. PLoS Genet. 2011;7(8):1002234. doi:10.1371/journal.pgen.1002234.
6. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. The Lancet. 2010;375(9725):1525⊠35.
7. Liu TY, Wiesel A, Hero AO. A Sparse Multiclass Classifier for Biomarker Screening. In: IEEE Global Conference on Signal and Information Processing (GloabalSIP). Piscataway, New Jersey, USA: IEEE; 2013. p. 77⊠83.
8. Woods CW, McClain MT, Chen M, Zaas AK, Nicholson BP, Varkey J, et al. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza h1n1 or h3n2. PloS One. 2013;8(1):52198.
9. Zaas AK, Burke T, Chen M, McClain M, Nicholson B, Veldman T, et al. A host-based rt-pcr gene expression signature to identify acute respiratory viral infection. Sci Transl Med. 2013;5(203)203ra126⊠203ra126. doi:10.1126/scitranslmed.3006280.
10. Jackson GG, Dowling HF, Spiesman IG, Boand AV. Transmission of the common cold to volunteers under controlled conditions: I. the common cold as a clinical entity. AMA Arch Intern Med. 1958;101(2):267⊠78.
11. Ronald BT. Ineffectiveness of intranasal zinc gluconate for prevention of experimental rhinovirus colds. Clin Infect Dis. 2001;33(11):1865⊠70.
12. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
13. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. Series B (Methodological). 1996;58(1):267⊠88.
14. Kreßel UHG. Pairwise classification and support vector machines. In: Advances in Kernel Methods. Brussels, Belgium: MIT Press; 1999. p. 255⊠68.
15. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. Neural Netw IEEE Trans. 2002;13(2):415⊠25.
16. Weston J, Watkins C. Support vector machines for multi-class pattern recognition. In: Proceedings of the Seventh European Symposium on Artificial Neural Networks; 1999. p. 219⊠24.
17. Bredensteiner EJ, Bennett KP. Multicategory classification by support vector machines. Comput Optim Appl. 1999;12(1):53⊠79.
18. Guermeur Y. Combining discriminant models with new multi-class svms. Pattern Anal Appl. 2002;5(2):168⊠79.
19. Crammer K, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. J Mach Learn Res. 2002;2:265⊠92.
20. Liu Y, Shen X. Multicategory ψ-learning. J Am Stat Assoc. 2006;101(474): 500⊠9.
21. Wang L, Shen X. On l1-norm multiclass support vector machines. J Am Stat Assoc. 2007;102(478):583⊠94.
22. Bach F, Jenatton R, Mairal J, Obozinski G. Optimization with sparsity-inducing penalties. Foundations Trends® Mach Learn. 2012;4(1): 1⊠106.
23. Liu TY. Statistical learning for sample-limited high-dimensional problems with application to biomedical data. PhD thesis. 2013.
24. Afonso MV, Bioucas-Dias JM, Figueiredo MAT. Fast image recovery using variable splitting and constrained optimization. Image Process IEEE Trans. 2010;19(9):2345⊠56.
25. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418⊠29.
26. Bühlmann P, Van De Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. Berlin Heidelberg: Springer; 2011, pp. 25⊠33.
27. Keerthi SS, Sundararajan S, Chang KW, Hsieh CJ, Lin CJ. A sequential dual method for large scale multi-class linear svms. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2008. p. 408⊠16.
28. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. Liblinear: A library for large linear classification. J Mach Learn Res. 2008;9:1871⊠4.
29. Combettes PL, Pesquet JC. Proximal splitting methods in signal processing. In: Fixed-point Algorithms for Inverse Problems in Science and Engineering. New York: Springer; 2011. p. 185⊠212.
30. Everitt AR, Clare S, Pertel T, John SP, Wash RS, Smith SE, et al. Ifitm3 restricts the morbidity and mortality associated with influenza. Nature. 2012;484(7395):519⊠23.
31. Miller AL, Gerard C, Schaller M, Gruber AD, Humbles AA, Lukacs NW. Deletion of ccr1 attenuates pathophysiologic responses during respiratory syncytial virus infection. J Immunol. 2006;176(4):2562⊠7.
32. Schoggins JW, Wilson SJ, Panis M, Murphy MY, Jones CT, Bieniasz P, et al. A diverse range of gene products are effectors of the type i interferon antiviral response. Nature. 2011;472(7344):481⊠5.
33. Chawla-Sarkar M, Lindner D, Liu YF, Williams B, Sen G, Silverman R, et al. Apoptosis and interferons: role of interferon-stimulated genes as mediators of apoptosis. Apoptosis. 2003;8(3):237⊠49.
34. Ritchie KJ, Hahn CS, Kim KI, Yan M, Rosario D, Li L, et al. Role of isg15 protease ubp43 (usp18) in innate immunity to viral infection. Nat Med. 2004;10(12):1374⊠8.
35. Bochkov Y, Hanson K, Keles S, Brockman-Schneider R, Jarjour N, Gern J. Rhinovirus-induced modulation of gene expression in bronchial epithelial cells from subjects with asthma. Mucosal Immunol. 2010;3(1):69⊠80.
36. Kawai T, Akira S. Innate immune recognition of viral infection. Nat Immunol. 2006;7(2):131⊠7.
37. Au WC, Yeow WS, Pitha PM. Analysis of functional domains of interferon regulatory factor 7 and its association with irf-3. Virology. 2001;280(2): 273⊠82.
38. Barnes BJ, Richards J, Mancl M, Hanash S, Beretta L, Pitha PM. Global and distinct targets of irf-5 and irf-7 during innate response to viral infection. J Biol Chem. 2004;279(43):45194⊠207.
39. Honda K, Yanai H, Negishi H, Asagiri M, Sato M, Mizutani T, et al. Irf-7 is the master regulator of type-i interferon-dependent immune responses. Nature. 2005;434(7034):772⊠7.
40. Kawai T, Akira S. Innate immune recognition of viral infection. Nat Immunol. 2006;7(2):131⊠7.
41. Stetson DB, Medzhitov R. Type i interferons in host defense. Immunity. 2006;25(3):373⊠81.
42. Samuel CE. Antiviral actions of interferons. Clin Microbiol Rev. 2001;14(4): 778⊠809.
43. Manderson AP, Botto M, Walport MJ. The role of complement in the development of systemic lupus erythematosus. Annu Rev Immunol. 2004;22:431⊠56.
44. Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol. 2008;10(1):116⊠25.
45. Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, et al. Systems biology of vaccination for seasonal influenza in humans. Nat Immunol. 2011;12(8):786⊠95.
46. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293⊠307.