Cell Host & Microbe
Letters

Improving Development of the Molecular Signature for Diagnosis of Acute Respiratory Viral Infections

Alexander Statnikov,[1,2,*] Lauren McVoy,[3] Nikita Lytkin,[1] and Constantin F. Aliferis[1,3,4]
[1]Center for Health Informatics and Bioinformatics
[2]Department of Medicine
[3]Department of Pathology
New York University School of Medicine, NY 10016, USA
[4]Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232, USA
*Correspondence: alexander.statnikov@med.nyu.edu
DOI 10.1016/j.chom.2010.01.003

Acute respiratory viral infections cause significant morbidity and mortality in the United States and worldwide. Unfortunately, clinicians do not currently have practical means to make a timely and accurate diagnosis of acute viral respiratory infections, leading to unnecessary antibiotic treatment that increases healthcare costs and facilitates development of antibiotic resistance. In a recent breakthrough paper in *Cell Host & Microbe*, Zaas et al. (2009) provided a novel approach for diagnosis of acute respiratory infections based on microarray gene-expression profiles of blood from infected and uninfected subjects. They developed an ''acute respiratory viral response,'' a 30-gene panviral signature that can accurately diagnose symptomatic subjects (with influenza A, HRV, and RSV) from uninfected individuals and validated this signature using data from an independent study that contained influenza A patients and healthy controls (Ramilo et al., 2007). Overall, the study of Zaas et al. made a significant contribution toward improved diagnosis of infectious diseases. In this brief communication, we first propose several ways to improve the analysis that led to development of the 30-gene panviral signature and then provide an example of how the new analysis protocol can lead to an improved molecular signature.

We suggest several approaches to improve the analysis protocol that led to discovery of the acute respiratory viral response signature. First, to obtain an unbiased estimate of predictive accuracy, genes should be selected using the training set of subjects as opposed to selecting genes from the entire data set as was done in the study of Zaas et al. (2009). The latter gene selection procedure is known to typically lead to over-optimistic predictive accuracy estimates. Second, the cross-validation procedure employed by Zaas et al. should be modified to prohibit the use of samples from the same subjects both for developing signature and estimating its predictive accuracy, as this is another potential source of over-optimism. Third, the employed factor analysis-based gene selection method does not control for false discovery rate and may output redundant genes that are not located in the pathway causing the phenotype (i.e., they may not allow proper mechanistic interpretation and, e.g., may be ''passenger genes''). In addition, the choice of 30 genes used by Zaas et al. is arbitrary. There exist methods that circumvent all these issues (Aliferis et al., 2010). Fourth, the independent data set of Ramilo et al. (2007) used for validation of the signature does not contain data for two out of three viruses (RSV and HRV) and originates mostly from pediatric subjects, whereas the data used for development of the 30-gene panviral signature spanned all three viruses and is based on adult subjects. Therefore, more similar data sets should be sought for ''apples-to-apples'' validation. Finally, fifth, it would be useful to not only show the existence of a single signature to discriminate symptomatic subjects from uninfected individuals, but also to seek all possible maximally predictive signatures of the phenotype that do not contain redundant genes. Such analysis allows improvement in the discovery of the underlying biological mechanisms by not missing genes that are implicated mechanistically in the disease processes, and computationally efficient methods have been recently introduced to solve this problem (Statnikov, 2008). In summary, we suspect that the procedures employed in Zaas et al. to discover genes and signatures likely provide redundant genes, over-optimistic estimates of predictive accuracy, and biologically ''false positive'' (i.e., noncausative) and ''false negative'' (i.e., biologically significant but overlooked) genes.

Below we provide an example of how a causal graph-based analysis can lead to a more parsimonious signature that can predict phenotype with high accuracy and eliminates known sources of over-optimistic estimation of predictive accuracy. We undertook an additional analysis of the gene-expression data of Zaas et al. To select genes, we used HITON-PC, a supervised multivariate biomarker discovery method (Aliferis et al., 2010). This method is designed to discover local pathway members around the response variable of interest. In addition, the genes selected under certain broad assumptions exhibit maximal predictive accuracy for the data set at hand combined with maximum parsimony, beyond which predictive accuracy is compromised. Once the genes were selected, we applied support vector machine (SVM) classifiers to develop molecular signatures. These classifiers are robust even with the high variable-to-sample ratio, they can learn efficiently complex classification functions, they employ powerful regularization principles to avoid overfitting, and they are fairly insensitive to the large number of irrelevant variables. In order to obtain an unbiased estimate of predictive accuracy that will hold in future applications of signatures to unseen patients, gene selection, and development of molecular signatures was performed by repeated 10-fold cross-validation. Finally, to ensure signature reproducibility, we applied the procedure for assessing statistical significance of multivariate signatures. This procedure involves creating 1000

permutations of the sample classification labels and building molecular signatures and estimating their predictive accuracy for the permuted sample labels (Aliferis et al., 2009b).

Using the original data set of Zaas et al., we applied the HITON-PC method for biomarker discovery and fitted SVM models to diagnose symptomatic subjects from uninfected individuals using only training sets of samples within the repeated 10-fold cross-validation protocol. Since there are two samples for each subject who remained asymptomatic (one from baseline and another one from peak time), we randomly assigned these samples together either to the training or to the testing set, thus avoiding situations where we will train and test on the same subjects. The above procedure yields an unbiased cross-validated estimate of predictive accuracy, 0.94 AUC (95% confidence interval [0.89; 0.99] AUC). On average HITON-PC selected ten genes depending on the training set of cross-validation. Genes that were selected by HITON-PC in more than 20% of the training sets are listed in Figure S1A. Next, HITON-PC and SVM were applied to the data for all samples, resulting in a 12-gene panviral signature (see Figure S1B). Notice that all these 12 genes except for DEGS1 were also among the most frequently selected by HITON-PC during cross-validation. This 12-gene signature yields 0.99 AUC (95% confidence interval [0.98; 1.00] AUC) in the data of (Ramilo et al., 2007), which is statistically indistinguishable from the predictive accuracy of the 30-gene signature of Zaas et al.

Genes that participate in the 12-gene signature discovered by HITON-PC are the following: *GRAMD1C*, *OSBPL10*, *ID3*, *IGHD*, *C13orf18*, *MS4A1*, *RAPGEF6*, *GTF2I*, *DEGS1*, *FCGR1B*, *IFI44L*, and *RSAD2*. Only two of these genes (*IFI44L* and *RSAD2*) are among the original group of 30 genes that were included in the panviral signature of Zaas et al. *FCGR1B* was not reported by Zaas et al. in the panviral signature, but was found in the RSV-specific signature. Among ten unique genes in the panviral signature reported here, the following genes are directly involved in immune responses: *GTF2I*, *DEGS1*, *ID3*, *IGHD*, *FCGR1B*, and *MS4A1*. *GTF2I* or general transcription factor II-i is involved in T cell activation and proliferation as well as regulation of the immunoglobulin promoter (Sacristán et al., 2009; Tantin et al., 2004). *DEGS1*, the degenerative spermatocyte homolog 1, is upregulated in natural killer cells, which are an important component of the innate immune response (Dybkaer et al., 2007). *ID3* or inhibitor of DNA binding 3 promotes development of γΔ T cells, enabling them to become competent to produce interferon-γ, and also plays an integral role in a mouse model of the autoimmune disease, Sjogren's syndrome (Lauritsen et al., 2009; Li et al., 2004). *IGHD* or immunoglobulin heavy chain D is produced by B cells, although the role of IgD is not clearly defined in comparison to the other immunoglobulins. *FCGR1B* encodes a receptor for the constant region of IgG. It is expressed on myeloid cells and upregulated by interferon-γ (Eichbaum et al., 1994). *MS4A1* encodes CD20, which is expressed on the plasma membrane of mature B cells. To summarize, 8 of the 12 genes that differentiate symptomatic from asymptomatic individuals are directly involved in immune responses. Further, four of these genes (*IFI44L, RSAD2, ID3,* and *FCGR1B*), are either upstream or downstream of interferon production, which is widely involved in the immune response to viral infection.

As mentioned above, HITON-PC discovers genes in the local pathway of the response variable of interest under the assumptions stated (Aliferis et al., 2010). Below we provide an interpretation of the HITON-PC results in light of possible violations of its key assumptions in the data set of Zaas et al. First, because of small sample size, some statistical tests of conditional independence may be underpowered, which leads to false negatives in the output of the method. Notice that because the discovered genes provide a very high value of predictive accuracy (0.94 AUC), any such false negatives are fairly insignificant because they can uniquely account for only 0.06 AUC (= 1.0–0.94). Second, because of the possible presence of hidden variables in the local pathway of the response variable (i.e., not anywhere in the network), some of the genes discovered by HITON-PC may be false positives (i.e., confounded by local unmeasured variables). Given that the output of the method contains only 12 genes, further validation of their mechanistic role is much easier than working with substantially larger gene lists returned by other methods.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure and can be found with this article online at doi:10.1016/j.chom.2010.01.003.

ACKNOWLEDGMENTS

We thank the authors of Zaas et al. for promptly and generously sharing with us their data, codes, and details about their analyses. This research was supported in part by grants R56 LM007948-04A1 and U54 RR024386-01A2.

REFERENCES

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X.D. (2010). J. Mach. Learn. Res., in press.

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Schildcrout, J.S., Shepherd, B.E., and Harrell, F.E., Jr. (2009b). PLoS ONE *4*, e4922.

Dybkaer, K., Iqbal, J., Zhou, G., Geng, H., Xiao, L., Schmitz, A., d'Amore, F., and Chan, W.C. (2007). BMC Genomics *8*, 230.

Eichbaum, Q.G., Iyer, R., Raveh, D.P., Mathieu, C., and Ezekowitz, R.A. (1994). J. Exp. Med. *179*, 1985–1996.

Lauritsen, J.P., Wong, G.W., Lee, S.Y., Lefebvre, J.M., Ciofani, M., Rhodes, M., Kappes, D.J., Zúñiga-Pflücker, J.C., and Wiest, D.L. (2009). Immunity *31*, 565–575.

Li, H., Dai, M., and Zhuang, Y. (2004). Immunity *21*, 551–560.

Ramilo, O., Allman, W., Chung, W., Mejias, A., Ardura, M., Glaser, C., Wittkowski, K.M., Piqueras, B., Banchereau, J., Palucka, A.K., and Chaussabel, D. (2007). Blood *109*, 2066–2077.

Sacristán, C., Schattgen, S.A., Berg, L.J., Bunnell, S.C., Roy, A.L., and Rosenstein, Y. (2009). Eur. J. Immunol. *39*, 2584–2595.

Statnikov, A. (2008). Ph.D. thesis, Vanderbilt University, Nashville, Tennessee.

Tantin, D., Tussie-Luna, M.I., Roy, A.L., and Sharp, P.A. (2004). J. Biol. Chem. *279*, 5460–5469.

Zaas, A.K., Chen, M., Varkey, J., Veldman, T., Hero, A.O., 3rd, Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., et al. (2009). Cell Host Microbe *6*, 207–217.