



Published in final edited form as:

Cell Host Microbe. 2010 February 18; 7(2): 100–101. doi:10.1016/j.chom.2010.01.003.

Improving development of the molecular signature for diagnosis of acute respiratory viral infections

Alexander Statnikov^{1,2}, Lauren McVoy³, Nikita Lytkin¹, and Constantin F. Aliferis^{1,3,4}

¹Center for Health Informatics and Bioinformatics, New York University School of Medicine, NY 10016, USA

²Department of Medicine, New York University School of Medicine, NY 10016, USA

³Department of Pathology, New York University School of Medicine, NY 10016, USA

⁴Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232, USA

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 and they often resort 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. provided a novel approach for diagnosis of acute respiratory infections based on microarray gene expression profiles of the blood (Zaas et al., 2009). They developed an “acute respiratory viral response” 30-gene panviral signature that can accurately diagnose symptomatic subjects (with influenza A, HRV, and RSV) from uninfected individuals and validated this signature in 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 from peripheral blood. In the present 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.

From the data-analytic perspective there are 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 cross-validation as opposed to selecting genes on the entire dataset as was done in the study of Zaas et al. 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 avoid another potential source of over-optimism by prohibiting the use of samples from the same subjects both for developing signature and estimating its predictive accuracy. *Third*, the employed factor analysis-based gene selection method does not control for false discovery rate and may output redundant genes that are furthermore not located on the causal pathway of the phenotype (i.e., they do not have 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 problems (Aliferis et al., 2009).

© 2009 Elsevier Inc. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Fourth, the independent dataset 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 from mostly children, whereas the data used for development of the 30-gene panviral signature spanned over adults with all three viruses. Therefore, more similar datasets should be sought for “apples-to-apples” validation. Finally *fifth*, it is informative not only to 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 to improve 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 and Aliferis, 2009). In summary, we suspect that the procedures employed in Zaas et al. to discover genes and signatures likely provide predictively redundant genes, over-optimistic estimates of predictive accuracy and biologically “false positive” (i.e., non-causative) and “false negative” (i.e., biologically significant but overlooked) genes.

We examine the above issues in a forthcoming extended technical paper. Below we provide an example of how a causal graph-based analysis can lead to a novel and more parsimonious signature that can predict phenotype with high accuracy and does not suffer from known sources of over-optimistic estimation of predictive accuracy. To this end, 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., 2009). This cutting-edge method provably discovers the local pathway membership around the response variable of interest. In addition, the selected genes under broad assumptions exhibit maximal predictive accuracy for the dataset 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 to 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 (i.e., neither over-optimistic nor under-optimistic) 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 that involves permutations of the sample classification labels.

Using the original dataset 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 10 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 Supplementary Figure 1(a). Next, HITON-PC and SVM were applied to the data for all samples, resulting in a 12-gene panviral signature, see Supplementary Figure 1(b). Thus, 2.5 fold reduction of genes was accomplished in comparison to the 30-gene signature of Zaas et al. 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 novel 12-gene signature discovered by HITON-PC are: GRAMD1C, OSBPL10, ID3, IGHD, C13orf18, MS4A1, RAPGEF6, GTF2I, DEGS1, FCGR1B, IFI44L, 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 10 novel genes that we identified in the panviral signature, several are involved in immune responses to infection while others are involved in more general cellular processes such as signal transduction and regulation of transcription. The following genes are directly involved in immune responses: *GTF2I*, *DEGS1*, *ID3*, *IGHD*, *FCGR1B*, *MS4A1*. *GTF2I* or general transcription factor II-i is involved in T-cell activation and proliferation as well as regulation of the immunoglobulin promoter (Sacristan et al., 2009; Tantin et al., 2004). *DEGS1*, the degenerative spermatocyte homolog 1, is up-regulated 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 gamma-delta T-cells, enabling them to become competent to produce interferon gamma 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 up-regulated by interferon gamma (Eichbaum et al., 1994). *MS4A1* encodes CD20, which is expressed on the plasma membrane of mature B-cells. To summarize, eight of the twelve 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 it was mentioned above, HITON-PC provably discovers genes in the local pathway of the response variable of interest under assumptions stated in (Aliferis et al., 2009). Below we provide an interpretation of the HITON-PC results in light of possible violations of its key assumptions in the dataset 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the authors of (Zaas et al., 2009) 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 CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*. 2009 In press.

Cell Host Microbe. Author manuscript; available in PMC 2011 February 18.

- Dybkaer K, Iqbal J, Zhou G, Geng H, Xiao L, Schmitz A, d'Amore F, Chan WC. Genome wide transcriptional analysis of resting and IL2 activated human natural killer cells: gene expression signatures indicative of novel molecular signaling pathways. *BMC. Genomics* 2007;8:230. [PubMed: 17623099]
- Eichbaum QG, Iyer R, Raveh DP, Mathieu C, Ezekowitz RA. Restriction of interferon gamma responsiveness and basal expression of the myeloid human Fc gamma R1b gene is mediated by a functional PU.1 site and a transcription initiator consensus. *J. Exp. Med* 1994;179:1985–1996. [PubMed: 8195721]
- Lauritsen JP, Wong GW, Lee SY, Lefebvre JM, Ciofani M, Rhodes M, Kappes DJ, Zuniga-Pflucker JC, Wiest DL. Marked induction of the helix-loop-helix protein Id3 promotes the gammadelta T cell fate and renders their functional maturation Notch independent. *Immunity* 2009;31:565–575. [PubMed: 19833086]
- Li H, Dai M, Zhuang Y. A T cell intrinsic role of Id3 in a mouse model for primary Sjogren's syndrome. *Immunity* 2004;21:551–560. [PubMed: 15485632]
- Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM, Piqueras B, Banchereau J, Palucka AK, Chaussabel D. Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 2007;109:2066–2077. [PubMed: 17105821]
- Sacristan C, Schattgen SA, Berg LJ, Bunnell SC, Roy AL, Rosenstein Y. Characterization of a novel interaction between transcription factor TFII-I and the inducible tyrosine kinase in T cells. *Eur. J. Immunol* 2009;39:2584–2595. [PubMed: 19701889]
- Statnikov, A.; Aliferis, CF. Submitted. 2009. Analysis and computational dissection of molecular signature multiplicity. Preprint is available from the authors
- Tantin D, Tussie-Luna MI, Roy AL, Sharp PA. Regulation of immunoglobulin promoter activity by TFII-I class transcription factors. *J. Biol. Chem* 2004;279:5460–5469. [PubMed: 14645227]
- Zaas AK, Chen M, Varkey J, Veldman T, Hero AO III, Lucas J, Huang Y, Turner R, Gilbert A, Lambkin-Williams R, Oien NC, Nicholson B, Kingsmore S, Carin L, Woods CW, Ginsburg GS. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host. Microbe* 2009;6:207–217. [PubMed: 19664979]