META-ANALYSIS OF CONTINUOUS PHENOTYPES IDENTIFIES A GENE SIGNATURE THAT CORRELATES WITH COPD DISEASE STATUS

MADELEINE SCOTT*

Stanford University School of Medicine, Stanford University
Stanford, CA 94305, USA

FRANCESCO VALLANIA*

Stanford Institute for Immunity, Transplantation, and Infection, Stanford University
Stanford, CA 94305, USA

PURVESH KHATRI

Stanford Institute for Immunity, Transplantation, and Infection, Stanford University
Dvision of Biomedical Informatics Research, Department of Medicine, Stanford University
Stanford, CA 94305, USA
Email: pkhatri@stanford.edu

* *authors contributed equally to this work*

The utility of multi-cohort two-class meta-analysis to identify robust differentially expressed gene signatures has been well established. However, many biomedical applications, such as gene signatures of disease progression, require one-class analysis. Here we describe an R package, MetaCorrelator, that can identify a reproducible transcriptional signature that is correlated with a continuous disease phenotype across multiple datasets. We successfully applied this framework to extract a pattern of gene expression that can predict lung function in patients with chronic obstructive pulmonary disease (COPD) in both peripheral blood mononuclear cells (PBMCs) and tissue. Our results point to a disregulation in the oxidation state of the lungs of patients with COPD, as well as underscore the classically recognized inflammatory state that underlies this disease.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive, debilitating lung disease that affects one in 20 people across the globe.[1] It is characterized by declining lung function, as measured by Forced Expiratory Volume (FEV) or Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage.[2,3] FEV is the amount of air that a COPD patient can expel in one second, and decreases as the disease progresses. GOLD scoring is the result of a global effort to reach an agreement on spirometric thresholds for COPD diagnosis and is considered the gold standard of COPD severity.[2,3] An increasing GOLD stage reflects declining lung function, where GOLD stage of 0 represents at-risk patients, while a stage of 4 identifies patients with predicted FEV <30%.[3] The rate of COPD progression varies widely from patient to patient, and there are no current treatment options that effectively halt the disease.[4] There is an urgent, critical unmet need to identify pathways that are robustly and reproducibly associated with COPD severity in order to identify novel targets for therapy.

We have previously described a multi-cohort analysis framework for integrated analysis of heterogeneous datasets, and repeatedly demonstrated its successful application across diverse set of diseases including organ transplant, cancer, and infectious diseases for identifying diagnostic, prognostic, and therapeutic signatures.[5–10] At its core, our multi-cohort analysis framework uses random effects inverse variance meta-analysis to identify differentially

expressed genes between two groups of samples (e.g., cases vs controls). However, despite its demonstrated utility, its application is limited to two-class comparisons. One of the drawbacks of this framework is that it does not take into account the stage of disease of the patients.[11,12] Further, many biomedical applications, such as those looking to identify signatures of disease progression, require one-class analysis. Such analyses are indispensible for identifying higher risk patients for more personalized care, and to discover pathways associated with disease progression,[12] which in turn could improve our understanding of the disease.

We have implemented an R package, MetaCorrelator, that addresses this challenge and extends the utility of our multi-cohort analysis framework to analyze continuous phenotypes across multiple datasets (Figure 1). MetaCorrelator follows principles of our framework to identify robust signatures for continuous phenotypes. It provides flexibility to use with different continuous phenotypes and widely heterogenous data.

2. Methods

2.1. *Integration of correlation coefficients across independent datasets*

MetaCorrelator starts by computing a correlation coefficient between a designated continuous phenotype and every gene measured in a given discovery dataset. The correlation coefficients can be computed as Pearson's r, Spearman's ρ, or Kendall's τ. Because Spearman's ρ is defined as the Pearson's r calculated on the ranks,[13] it can used directly as r for the rest of the analysis. Kendall's τ need to be converted into r[14] according to

$$r = \sin(\pi * 0.5 * \tau) \tag{1}$$

Then, each correlation coefficient r is converted into a Fisher's Z effect size, defined as:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{2}$$

with variance, V_z, and standard error, SE_z, defined as

$$V_z = \frac{1}{n-3} \tag{3}$$

and

$$SE_z = \sqrt{V_z} \tag{4}$$

where n is the total number of samples used for correlation. Next, we combine Fisher's z for every gene across all discovery datasets into a summary effect size using a random-effects inverse variance model,[15] which assumes that the true effect sizes across each study are not identical but rather sampled from a distribution of true effects. The summary effect size is calculated as

$$Z_s = \frac{\sum_i^n W_i Z_i}{\sum_i^n W_i} \tag{5}$$

and the corresponding summary standard error was computed as

$$SE_s = \sqrt{\frac{1}{\sum_i^n W_i}} \tag{6}$$

where z_i is the Fisher's Z for a given dataset i and W_i is a weight defined as

$$W_i = \frac{1}{V_i + T^2} \tag{7}$$

where V_i is the variance of the Fisher's Z effect size for a given gene within dataset i and T^2 indicates the in-between-dataset variation. Finally, every gene is assigned a p-value calculated using a two-tailed test defined as

$$p = 2[1 - 2(\phi(|\frac{Z_s}{SE_s}|)] \tag{8}$$

The p-value is then corrected for multiple hypothesis testing using Benjamini-Hochberg.

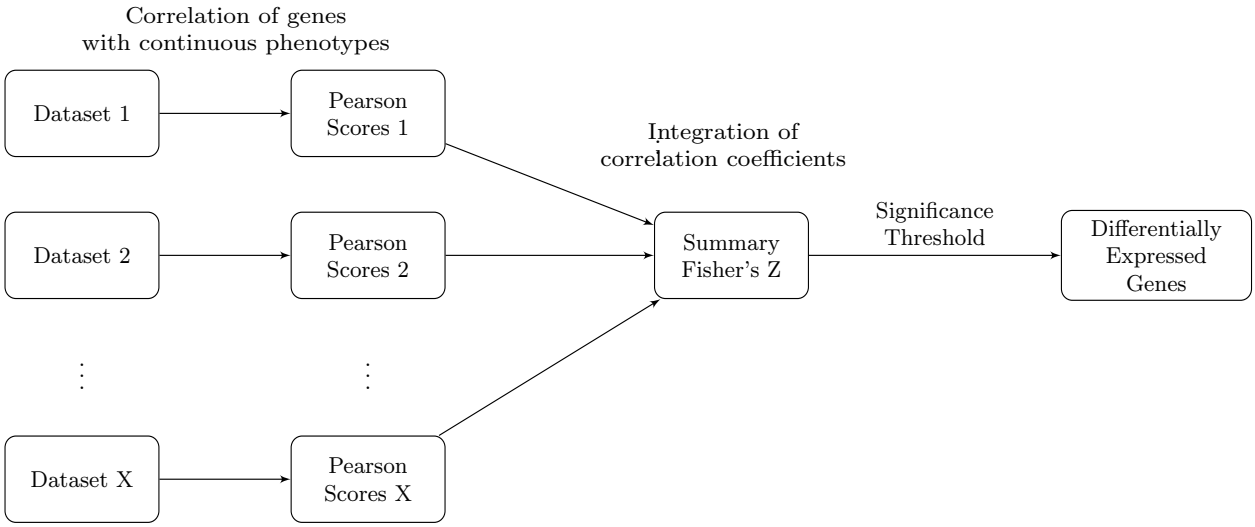

Fig. 1. Schematic Overview of MetaCorrelator. Each dataset is correlated with a continuous phenotype to compute correlation coefficients (example: Pearson's correlation coefficients). These coefficents are then combined into a summary Fisher's Z effect size. A significance threshold is determined to produce the final list of differentially expressed genes.

2.2. *Datasets*

We identified publicly available gene expression datasets from the NCBI GEO that provided lung function in COPD patients using GOLD stage or FEV. In total, we identified six datasets with 642 samples. All six had expression data that was pre-normalized. We used three datasets for discovery and three as validation (Tables 1 and 2). The datasets were highly heterogeneous; five were from lung biopsies and two from PBMCs, spanning collection over seven years and three countries. All probes were matched to Gene IDs based on the platform information available on GEO. Three of the datasets did not have a control group as all of the samples originated from patients with COPD.

2.3. *Selection and validation of COPD signature*

We used FDR < 5% to identify genes significantly correlated with COPD severity as defined by GOLD stage or FEV. We performed Gene Ontology enrichment analysis using iPathwayGuide (http://www.advaitabio.com). All other statistical analyses were performed using the statistical programming language R.

2.4. *Availability*

Source code is available at http://khatrilab.stanford.edu/metacorrelator. The Khatri lab will provide full results upon request.

Table 1. Datasets Used in Discovery of Lung Function Signature

GEO ID	Tissue	Phenotype	Cases
GSE47460	Lung Biopsy	GOLD Stage and FEV	75
GSE69818	Lung Biopsy	GOLD Stage	70
GSE76705	PBMCs	FEV	229
3 Datasets	2 Tissues		324

Table 2. Datasets Used to Validate Lung Function Signature

GEO ID	Tissue	Phenotype	Cases
GSE42057	PBMCs	FEV	136
GSE38974	Lung Biopsy	GOLD Stage	32
GSE11906	Lung Biopsy	GOLD Stage	150
3 Datasets	2 Tissues		318

3. Results

3.1. *Functional Analysis of Differentially Expressed Genes Identified by MetaCorrelator*

We identified six independent datasets of 692 lung biopsies or PBMC samples from COPD patients that also provided either GOLD stage or FEV for each patient. The samples included in these datasets came from patients across all stages of COPD and covered all the lobes in the lung. We selected three datasets composed of 374 PBMC samples or lung biopsies as discovery datasets (Table 1), and the rest as validation datasets (Table 2). We choose three discovery datasets such that they increased heterogeneity in the discovery. Two datasets were from lung tissue, and had annotation describing the GOLD stage of the samples. Of the two lung tissue datasets, GSE698181 had COPD patients with and without emphysema. Although GSE47460 had both GOLD stage and FEV annotation, only GOLD stage was used for discovery of the gene signature.

MetaCorrelator identified 108 genes (FDR < 25%) that are consistently correlated with COPD severity as measured by GOLD stage or FEV in the three discovery datasets. We performed Gene Ontology enrichment analysis (Figure 2) to explore the functions of the

identified genes. Our enrichment analysis highlighted the role of oxidative stress in COPD progression. We identified the disulfide oxidoreductase activity pathway as a highly significant in COPD progression. This is consistent with previous literature that has identified oxidative stress as a sign of COPD progression.[16]

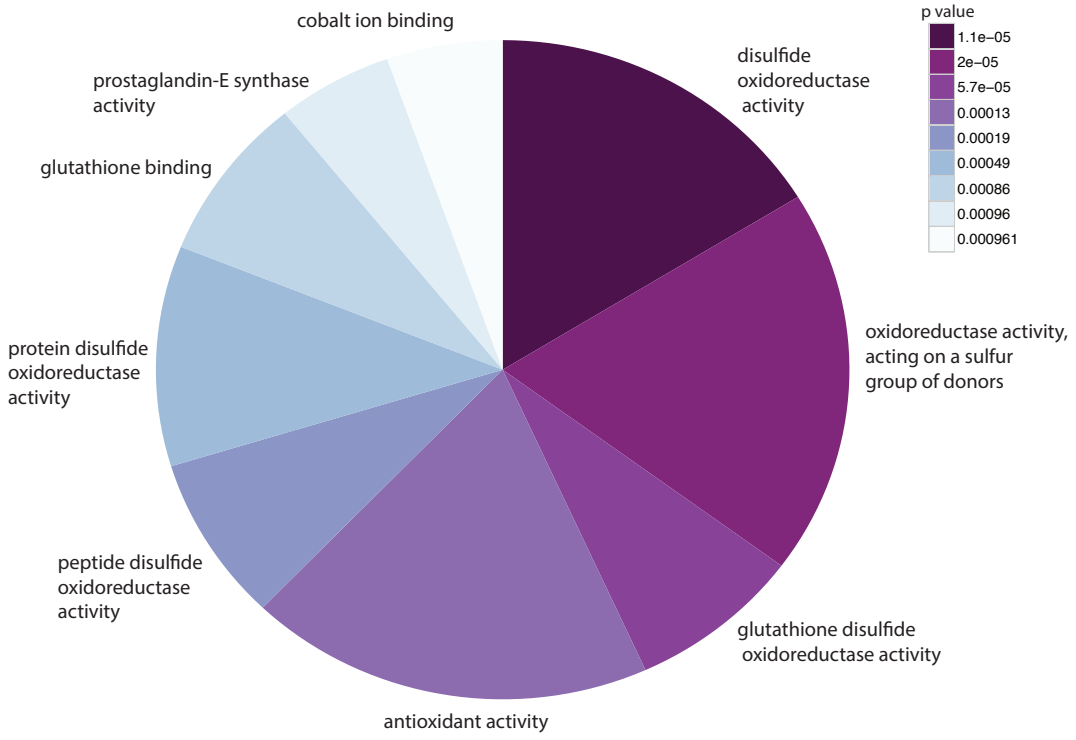

Fig. 2. Functional categories enriched in genes identified by MetaCorrelator

3.2. *Identification and Analysis of a 25-Gene Signature Correlates with COPD Progression*

It is difficult to translate 108-gene signature into a clinical practice. Therefore, to reduce the number of genes, we increased the stringency of our selection criteria by reducing the FDR to 5e-5. We identified 25 genes (7 over-expressed, 18 under-expressed) that are significantly correlated with the COPD severity in the discovery datasets. Enrichment analysis using Gene Ontology of the 25 genes identified matrix remodeling and inflammation as pathways associated with the progression of COPD. Specifically, a subset of the 25 genes, including UDP-Glucuronate Decarboxylase 1 (*UXS1*) and Tetraspanin 13 (*TSPAN13*) are underexpressed genes known to be involved in ECM production and cell adhesion. Importantly, a double tetraspanin knockout mouse (Tetraspanin 28 and 29) has been shown to develop a COPD-like phenotype, underscoring the importance of this protein family to healthy lung function.[18] Inflammatory mediators such as *TLR2* and *FKBP5* are over-expressed in the gene signature, reflecting the inflammatory state of COPD. Previous literature has shown that *TLR2* is over-

expressed on Lung CD8+ and CD4+ T-cells as well as CD8+NK T-cells, demonstrating that our results reflect validated biology.[19] Two differentially expressed genes, *TNIK* and *PTPRK* are involved in both ECM and inflammation. Taken together, our results align with previously published literature, which describes COPD as a disregulation of the immune system and subsequent breakdown of the ECM.[20–22]

Next, we defined Lung Function Score (LFS) for a sample as the geometric mean of the expression value of the 25 genes as previously described.[5–7] We observed strong significant correlation between our signature score and FEV in GSE76705 (r = -0.50; p-value = 5.81e-14) (Figure 3). In datasets where GOLD staging was available, we observed a significant score increase in concordance with increasing GOLD stage (JT test; GSE47460: p-value = 9.4e-10; GSE69818: p-value = 5e-4). Interestingly, although only the GOLD stages in GSE47460 were used in the discovery, the LFS strongly correlated with FEV score in GSE47460 (r = -0.57; p-value = 6.23e-4).

3.3. *Validation of the Gene Signature in Three Independent Cohorts*

We validated the 25-gene LFS in three independent cohorts of 318 lung biopsy and PBMC samples from COPD patients (Table 2, Figure 4). Across all three validation cohorts, the LFS was significantly correlated with lung function in the COPD patients (summary effect size = 0.46, p = 3.98e-3). In individual datasets, we observed a significant negative correlation between FEV and LFS in GSE42057 (r = -0.41; p-value = 5.03e-7) and a positive significant correlation with GOLD stages in our remaining two independent cohorts (GSE38974; p-value = 5.539e-6; GSE11906; p-value = 0.02514).

3.4. *Differential Expression in Current vs Never Smokers*

To explore the broader implications of our results, we examined whether any of our 25 identified genes were also significantly expressed in smokers compared to healthy controls. We downloaded seven publicly available datasets from NCBI GEO for a total of 200 samples from smokers and 158 from never smokers (GSE11952, GSE17913, GSE19667, GSE3320, GSE5056, GSE5057, GSE5059). Using the 25-gene LFS derived from MetaCorrelator, two genes, *TSPAN13* and *NR3C2*, were found to be differentially expressed in smokers compared to non-smokers with p value < 0.01. The tetraspanin family has been shown to be critical to normal lung function, and *NR3C2* has been implicated in lung morphogenesis.[23] These results demonstrate the flexibilty of MetaCorrelator to highlight patterns of biological relevance in conjunction with two-class analysis.

4. Discussion

Availability of large amounts of heterogeneous molecular data has necessitated the development of new frameworks to identify patterns and extract new information from these data. We have repeatedly shown the effectiveness of our multi-cohort analysis framework for diagnostic and therapeutic applications across a broad spectrum of human diseases.[5–10] However, this framework is limited to analysis of case-control experiments, and is not suitable for analysis of one-class quantitative phenotypes. Here, we extend our previously established framework to include analysis of gene expression with quantitative quantitative.

Correlation analysis has been a powerful tool for decades, but at this time there does not exist a single framework that can take a collection of datasets and different quantitative

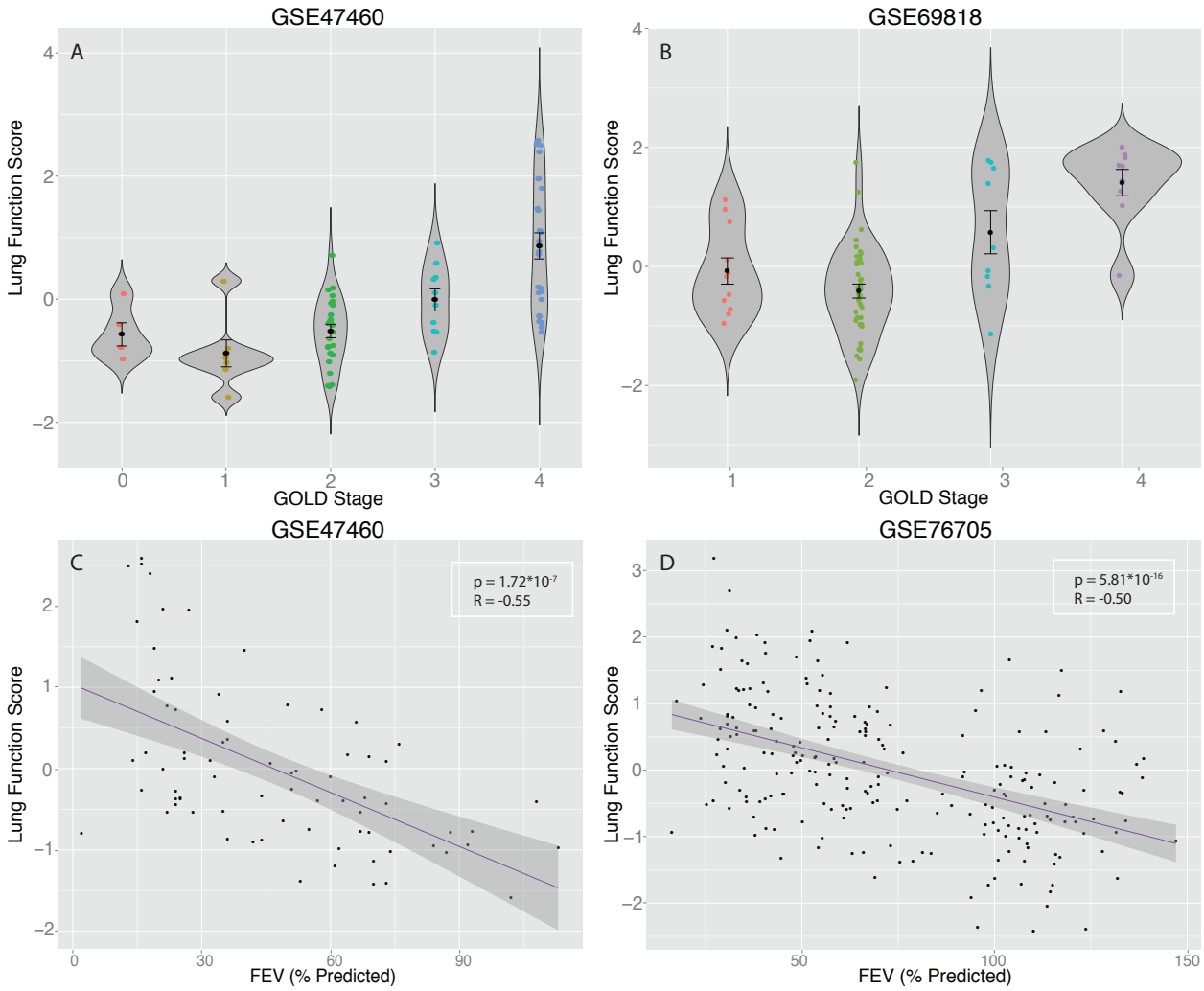

Fig. 3. Lung Function Scores in training cohorts: (A) Violin plots of Lung Function Scores (LFS) in patients distinguished by progressively increasing GOLD stage from GSE47460. (B) Same as (A) but for dataset GSE69818. (C) Correlation plot between LFS and FEV scores in individual patients from GSE47460. (D) Same as (C) but for dataset GSE76705.

phenotypes as input and produce a correlated gene expression signature. There are currently available packages in R, such as metacor, that can compute Fisher's Z values from correlation coefficients; however, MetaCorrelator is uniquely positioned to take multiple datasets as input and correlate gene expression with heterogeneous phenotypes. This is especially relevant in the realm of human disease; methods that are able to integrate different but related organ function phenotypes, such as FEV and GOLD stage, would allow for more powerful analysis that could identify new markers for disease progression.

Our method enables the identification of a gene signature across tissues, thus highlighting the globally relevant differentially expressed genes. By integrating PBMC and lung tissue data, we were able to distill out a gene signature that represents the global differential gene expression of COPD progression. These results emphasize the advantage of integrating multiple tissues. The genes in our signature suggest the importance of inflammation (*TLR2*, *FKBP5*)

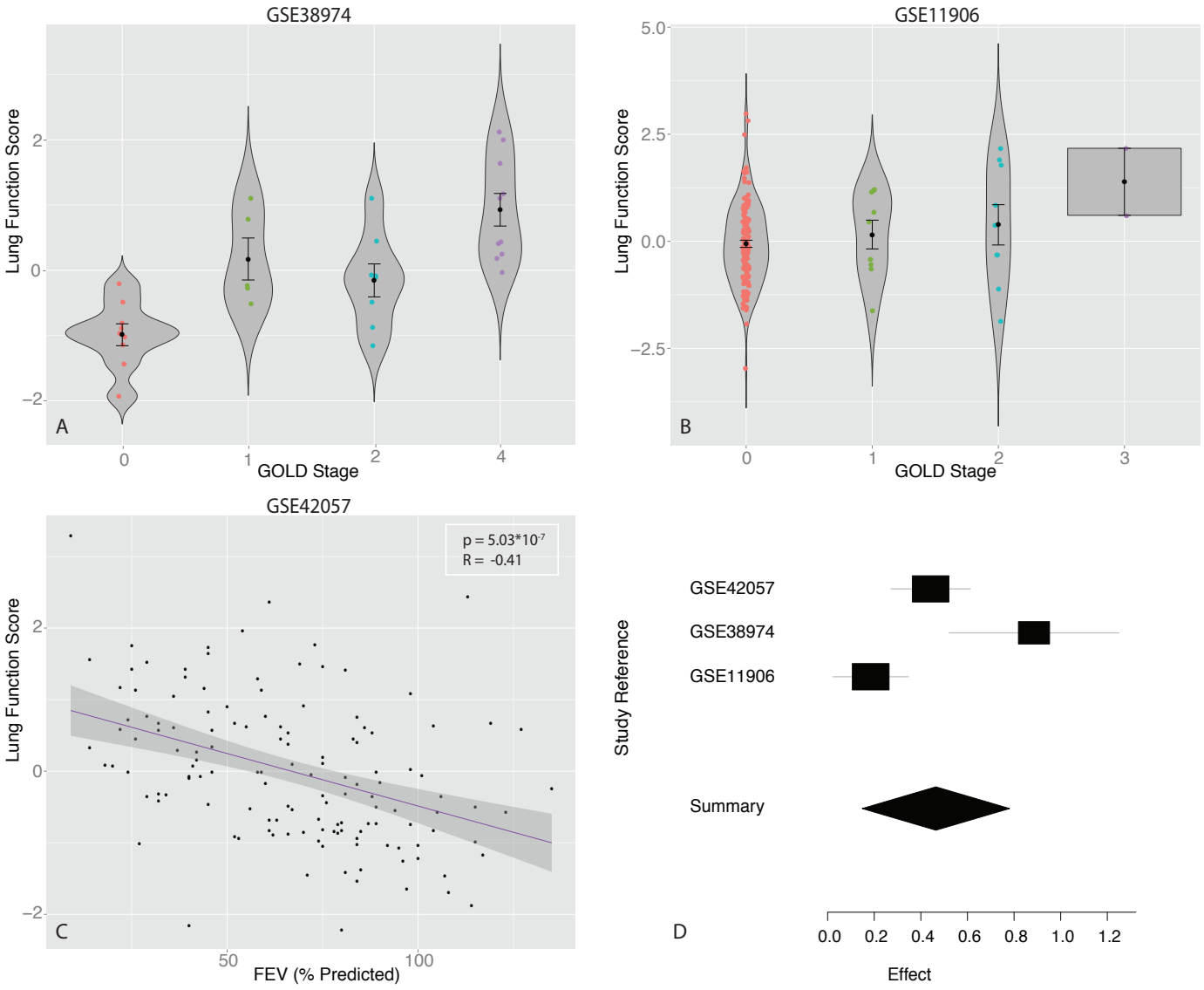

Fig. 4. Lung Function Scores in validation cohorts: (A) Violin plots of Lung Function Scores (LFS) in patients distinguished by progressively increasing GOLD stage from GSE38974. (B) Same as (A) but for dataset GSE11906. (C) Correlation plot between LFS and FEV scores in individual patients from GSE42057. (D) Forest plot representing Fisher's Z values for each of the validation datasets. Squares indicate individual dataset Fisher's Z, with square-size proportional to sample size and horizontal lines indicating individual standard errors (GSE42507 was reverted in sign because of the inverse relationship between GOLD score and FEV). Rhombus indicates summary Fisher's Z with width corresponding to summary standard error.

as well as cell adhesion (*TSPAN13*, *UXS1*), which demonstrates that our framework is able to recapitulate known biology. By integrating the MetaCorrelator framework with established two-class analysis, we can select genes of particular interest. For instance, after identifying differentially expressed genes between smokers and non-smokers, we could further focus on two genes that MetaCorrelator had identified as correlating with COPD progression. MetaCorrelator can be used to correlate any continuous disease phenotype with disease progression. For example, one could identify a gene signature that correlates with prostate specific antigen, a marker of prostate cancer progression. Alternatively, one could correlate a gene signature

with ejection fraction of the heart. In summary, MetaCorrelator provides a framework that can correlate whole genome transcriptome across multiple independent datasets with a quantitative phenotype, which in turn can be further explored in case-control studies using the multi-cohort analysis framework.

5. Conclusion

In this study we developed a meta-analysis framework that can integrate multiple gene expression datasets to identify gene signatures that correlate with quantitative phenotypes. Importantly, this method uses the inherent heterogeneity present in multiple cohorts to identify consistently correlated genes and is applicable to datasets that have a single class of sample. Our method can be used in conjunction with other methods that separate samples by class, for example, in order to further differentiate a single group of patients. We applied our method to COPD patients and extracted a 25-gene signature that correlated with lung function in three datasets (two in tissue, one in PBMCs). We then successfully validated our gene signature on three independent datasets. We demonstrated the ability to identify a robust signature with heterogeneous data and phenotypes by correlating the tissue datasets with increasing GOLD stage, and the PBMC dataset with decreasing FEV. Our results suggest an increasing immune response in later stage COPD patients, which has been noted by others, as well as point to a under-appreciated role in sulfur-related oxidative stress. In summary, MetaCorrelator provides a powerful framework to extract a gene signature that is linked to disease progression.

References

1. Halbert, R. J., et al. *European Respiratory Journal* **28**, 523 (2006)
2. Miravitlles, Marc, et al. *Thorax* **64**, 863 (2009).
3. Pauwels, Romain A., et al. *American journal of respiratory and critical care medicine* bf 163, 1256 (2012).
4. Rabe, Klaus F., et al. *American journal of respiratory and critical care medicine* **176**, 532 (2007)
5. Khatri, Purvesh and Roedder, Silke and Kimura, Naoyuki and De Vusser, Katrien and Morgan, Alexander A and Gong, Yongquan and Fischbein, Michael P and Robbins, Robert C and Naesens, Maarten and Butte, Atul J and Sarwal MM. *J. Exp. Med.* **210**, 2205 (2013)
6. M. Andres-Terre, H. M. McGuire, Y. Pouliot, E. Bongen, T. E. Sweeney, C. M. Tato and P. Kha- tri, Immunity **43**, 1199 (December 2015).
7. Timothy E. Sweeney, Aaditya Shidham, Hector R. Wong, and Purvesh Khatri. *Science Translational Medicine* **7**, 287 (2015)
8. Sweeney, Timothy E., Hector R. Wong, and Purvesh Khatri. *Science Translational Medicine* **8** , 346 (2016)
9. R. Chen, P. Khatri, P. K. Mazur, M. Polin, Y. Zheng, D. Vaka, C. D. Hoang, J. Shrager, Y. Xu, S. Vicent, A. J. Butte and E. A. Sweet-Cordero, *Cancer Research* **74**, 2892 (May 2014).
10. T. E. Sweeney, L. Braviak, C. M. Tato and P. Khatri, *The Lancet Respiratory Medicine* **4**, 213 (2016).
11. Walker, Esteban, Adrian V. Hernandez, and Michael W. Kattan. *Cleveland Clinic Journal of Medicine* **75**, 431 (2008)
12. Nordmanna, Alain J., Benjamin Kasendaa, and M. Briel. *Swiss Med Wkly* **142**, w13518 (2012)
13. Myers, Jerome L.; Well, Arnold D. *Research Design and Statistical Analysis* (Routledge, New York, 2003)
14. David A. Walker *JMASM* **2** 525 (2003)
15. M. Borenstein, L.V. Hedges, J.P.T Higgins, and H.R. Rothstein. *Introduction to Meta-Analysis* (Wiley, UK, 2011)
16. Rahman I, Kinnula VL. *Expert Review of Clinical Pharmacology* **5**, 293 (2012).
17. Avrum Spira, Jennifer Beane, Victor Pinto-Plata, Aran Kadar, Gang Liu, Vishal Shah, Bartolome Celli, and Jerome S. Brody. *American Journal of Respiratory Cell and Molecular Biology*, **31**, 601 (2004)
18. Jin, Yingji, et al. *American Thoracic Society* **42**, 633 (2014)
19. Freeman, Christine M., et al. *Respiratory research* **14** (2013)
20. Chung, K. F., and I. M. Adcock. *European Respiratory Journal* **31**, 1334 (2008)
21. Oudijk, EJ D., et al. *Thorax* **60**, 538 (2005)
22. Zandvoort, Andre, et al. *Respiratory research* **9**, 10.1186/1465-9921-9-83 (2008)

23. Duga, Balazs, et al. *Molecular cytogenetics* **7**, 10.1186/1755-8166-7-36 (2014)