Lung proteome and metabolome endotype in HIV-associated obstructive lung disease

Sarah Samorodnitsky[1], Eric F. Lock[1], Monica Kruk[1], Alison Morris[2], Janice M. Leung[3], Ken M. Kunisaki[4,1], Timothy J. Griffin[1] and Chris H. Wendt[1,4]

[1]University of Minnesota, Minneapolis, MN, USA. [2]University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [3]University of British Columbia, Vancouver, Canada. [4]Minneapolis VA Health Care System, Minneapolis, MN, USA.

Corresponding author: Chris Wendt (wendt005@umn.edu)

Shareable abstract (@ERSpublications)
BALF protein profile distinguishes HIV-associated obstructive lung disease. A unique endotype in black men with more severe obstructive lung disease was found that associated with signal transduction pathways, cell cycle and apoptosis. https://bit.ly/3u7j3Np

Cite this article as: Samorodnitsky S, Lock EF, Kruk M, *et al*. Lung proteome and metabolome endotype in HIV-associated obstructive lung disease. *ERJ Open Res* 2023; 9: 00332-2022 [DOI: 10.1183/23120541.00332-2022].

Received: 12 July 2022
Accepted: 24 Nov 2022

Abstract

Purpose Obstructive lung disease is increasingly common among persons with HIV, both smokers and nonsmokers. We used aptamer proteomics to identify proteins and associated pathways in HIV-associated obstructive lung disease.

Methods Bronchoalveolar lavage fluid (BALF) samples from 26 persons living with HIV with obstructive lung disease were matched to persons living with HIV without obstructive lung disease based on age, smoking status and antiretroviral treatment. 6414 proteins were measured using SomaScan® aptamer-based assay. We used sparse distance-weighted discrimination (sDWD) to test for a difference in protein expression and permutation tests to identify univariate associations between proteins and forced expiratory volume in 1 s % predicted (FEV_1 % pred). Significant proteins were entered into a pathway over-representation analysis. We also constructed protein-driven endotypes using K-means clustering and performed over-representation analysis on the proteins that were significantly different between clusters. We compared protein-associated clusters to those obtained from BALF and plasma metabolomics data on the same patient cohort.

Results After filtering, we retained 3872 proteins for further analysis. Based on sDWD, protein expression was able to separate cases and controls. We found 575 proteins that were significantly correlated with FEV_1 % pred after multiple comparisons adjustment. We identified two protein-driven endotypes, one of which was associated with poor lung function, and found that insulin and apoptosis pathways were differentially represented. We found similar clusters driven by metabolomics in BALF but not plasma.

Conclusion Protein expression differs in persons living with HIV with and without obstructive lung disease. We were not able to identify specific pathways differentially expressed among patients based on FEV_1 % pred; however, we identified a unique protein endotype associated with insulin and apoptotic pathways.

Introduction

Improved survival in persons with HIV has led to a higher prevalence of several chronic illnesses including obstructive lung disease, which affects an estimated 3–23% [1–8]. A major gap in knowledge is the inability to identify those at risk and a limited understanding of why HIV increases obstructive lung disease risk independent of smoking status. In the era prior to the common use of combination antiretroviral therapy pulmonary obstruction was mostly associated with advanced HIV/AIDS and frequent pulmonary infections [9]. In the antiretroviral treatment era, obstructive lung disease persists as a frequent comorbidity even in the absence of AIDS or frequent pulmonary infections [1–5, 8, 10–14]. While pulmonary infections may still have some role in obstructive lung disease development in persons living with HIV [15], it is highly likely that other factors are involved in HIV-associated obstructive lung disease,

including the HIV virus itself. This is particularly pertinent as the lung is a reservoir for HIV and site for HIV replication [16]. Currently no biomarker identifies risk or lends insight into the mechanisms that lead to a rapid decline of lung function and subsequent obstructive lung disease in persons living with HIV.

The goal of this study was to identify lung-specific biomarkers and corresponding biological pathways associated with HIV-associated obstructive lung disease using aptamer-based assays to profile the proteome in bronchoalveolar lavage fluid (BALF) in persons living with HIV comparing those with and without obstructive lung disease. Included in our analysis is targeted, mass spectrometry-based metabolite analysis. We employed statistical and computational methods to identify endotypes and proteomic pathways to lend insight into putative mechanisms of disease.

Methods

We performed a cross-sectional, matched case–control study using BALF and plasma samples previously collected from two cohorts.

Study population

Cases and controls were persons living with HIV selected from the Pittsburgh and Vancouver Lung HIV Cohorts [17, 18]. We identified 26 cases with HIV and obstructive lung disease with available BALF; obstructive lung disease was defined as the ratio of forced expiratory volume in 1 s/forced vital capacity (FEV_1/FVC) <lower limit of normal. Pulmonary function tests were obtained within 3 months of acquiring the BALF. Participants fasted prior to BALF collection, while the majority, but not all, fasted prior to plasma collection. Sample size was based on our previously reported study [19]. Controls consisted of 26 individuals with HIV and normal lung function (defined as FEV_1/FVC >lower limit of normal and FEV_1 >80% of predicted normal) matched on age (±5 years), antiretroviral treatment use and smoking status (current *versus* nonsmoker). Participants in the parent cohort studies provided informed consent for BALF collection and storage; the current study was approved by the University of Minnesota Institutional Review Board. Parent studies had Institutional Review Board Approval from Pittsburgh and Vancouver.

Sample processing

At study enrolment, BALF was collected as previously described [17, 18]. Samples were stored at −80°C prior to processing and underwent one freeze–thaw cycle. BALF samples were vortexed and centrifuged at 5000 ×g for 5 min at 4°C followed by separation of the pellet and supernatant for the removal of additional debris. Samples were prepared for proteomics following SomaScan® Assay specification, specifically 75 µL of sample concentrated to 200 $\mu L \cdot mL^{-1}$. To identify metabolites, 10 µL of plasma or 200 µL of BALF were loaded onto a Biocrates Life Sciences Absolute IDQ p400 HR (Biocrates Life Sciences catalog number 21018; Biocrates Life Sciences, Innsbruck, Austria) following the manufacturer's instructions and as previously reported [20]. Metabolite analysis was performed on a Thermo Scientific Q Exactive TM, Hybrid Quadrupole-Orbitrap TM (Thermo Fisher Scientific, Waltham, MA, USA), mass spectrometer equipped with a Thermo Scientific Ultimate 3000 UHPLC equipped with an autosampler, following manufacturer parameters.

Data cleaning

The SomaScan® proteomics assay data contained 7335 aptamers targeting human proteins. For each aptamer, we calculated an empirical lower limit of detection (LOD) to filter out those with over 50% of samples below the LOD. We calculated the LOD as $median(aptamer) + 4.5MAD$ where $MAD(\cdot)$ is the median absolute deviation between each sample and the median aptamer level [21]. LODs were calculated using the detected aptamer levels from buffer samples. We removed 3048 aptamers with over 50% of samples below the corresponding empirical LOD, retaining 4253 aptamers. We used Fisher's exact test to assess if the aptamers removed during this procedure are over-represented in cases *versus* controls and found no significant difference in samples above or below the LOD between cases and controls after adjusting for multiple comparisons. The remaining aptamers mapped to 3872 unique protein targets by the UniProt ID [22]. For the BALF metabolomic data we removed any metabolites that were either missing data or below the LOD for >50% of the cohort leaving 252 metabolites. There were 258 metabolites for plasma. We applied a log(1+x) transformation to both datasets, scaled and centred each protein and metabolite to have mean 0 and standard deviation 1.

Statistical analysis

We assessed the ability of the proteomic results to separate cases from controls using sparse distance-weighted discrimination (sDWD) [23], a method to classify subjects based on a large number of features (*i.e.*, proteins) in the setting when the data are high-dimensional, or when the number of features exceeds the number of samples. Cases were labelled as 1 while controls were labelled as −1. sDWD

calculates a score, a linear combination of protein expression levels, for each observation that captures how confidently cases and controls can be classified based on protein expression: highly positive scores for cases and highly negative scores for controls reflects good separation between the two classes. An L1 penalty on the coefficients of the proteins induces variable selection. We used cross-validation where we iteratively held out each case–control pair as a test set and trained the sDWD model on the remaining pairs. We compared the average, cross-validated sDWD scores for cases and controls using a paired t-test and the area under the curve (AUC) to assess the overall performance of the classifier. We used the sdwd R package version 1.0.5 by the authors [23] to implement this method.

We used the Global Lung Initiative standards to calculate FEV_1 per cent predicted (FEV_1 % pred) [24]. To investigate the relationship between each individual protein and FEV_1 % pred we used the Pearson correlation in a permutation testing framework to ensure our results were robust to deviations from normality, and to accommodate scenarios in which multiple aptamers map to the same protein. We first used a correlation test for each aptamer and FEV_1 % pred and combined the p-values for aptamers that mapped to the same protein using Fisher's method [25]. This yielded a chi-squared test statistic for each protein. Then, for 10 000 permutation iterations, we permuted the FEV_1 % pred values across the patient cohort, applied the t-test for the correlation between each aptamer and FEV_1 % pred under the permuted labelling scheme, and combined resulting p-values for aptamers mapping to the same protein using Fisher's method to obtain a chi-squared test statistic for each protein. This resulted in 10 000 combined chi-squared test statistics for each unique protein target. We computed a permutation p-value for each protein by taking the proportion of chi-squared test statistics that exceeded the chi-squared test statistic based on the original FEV_1 % pred observations. We applied a false discovery rate (FDR) correction [26] to the permutation p-values to correct for multiple comparisons. The proteins that were significant at the 0.05 level after permutation testing prior to multiple comparisons adjustment were then used in a pathway over-representation analysis using IMPaLA software [27]. We considered an analogous permutation testing framework to study the correlation between each protein and FEV_1/FVC and diffusing capacity of the lung for carbon monoxide (D_{LCO}) % predicted.

We identified protein-driven endotypes using K-means clustering. We ran the K-means algorithm 100 times with 10 random start points. For each of the 100 K-means replications, we saved the clustering scheme that achieved the minimum within-cluster variability across the 10 random start points to ensure an optimal solution [28]. We tested for individual protein differences between the clusters using the permutation testing framework described previously. At each permutation iteration, we permuted the cluster labels across the patient cohort and compared the permuted clusters using a t-test. Proteins that were significantly different between the clusters at an FDR level of 0.05 were used in pathway over-representation analysis. To quantify our uncertainty of the K-means clusters, we considered K-means clustering with 100 bootstrapping replications using the bootcluster package [29] in R.

For comparison, we also applied the same K-means clustering approach on metabolomics data collected from BALF and plasma in the same patient cohort. We tested for individual metabolite differences between the resulting clusters and quantified our uncertainty surrounding these clusters using bootstrapping.

Results

Study participant characteristics

Table 1 summarises the demographics of the individuals in our study. This cohort consisted largely of males (73.1%) with a mean age of 56.7 years. Over half (53.8%) of individuals identified as black and non-Hispanic and the same percentage identified as a current smoker (mean pack-years 23.1). Lung function (FEV_1) ranged from 21 to 90% of predicted normal in the obstructive lung disease cases (all of whom had FEV_1/FVC <lower limit of normal), compared to from 80 to 128% in the controls. D_{LCO} did not differ amongst the ∼80% of cases and controls that had it measured. Most individuals (92.3%) were treated with antiretroviral treatment at the time of sample collection. Two individuals among the cases and three among the controls exhibited viral loads >50 copies·mL^{-1}, among those for whom we had available viral load data.

BALF proteome differences in obstructive lung disease

We used sDWD to assess the collective power of BALF proteins to distinguish between cases and controls. This analysis shows a significant difference in the measured proteome between participants with obstructive lung disease and those without. Figure 1 displays the distribution of sDWD scores for cases and controls with cross-validation. Large differences in the average sDWD scores for cases and controls suggests better prediction accuracy based on protein expression. We compared the average scores for cases and controls under cross-validation using a paired t-test, which yielded a p-value of 0.0027 and an AUC of 0.6538.

TABLE 1 Demographics of the patient cohort considered in this study

	Case	Control	Total
Patients n	26	26	52
Sex n (%)			
Male	20 (76.9)	18 (69.2)	38 (73.1)
Female	6 (23.1)	8 (30.8)	14 (26.9)
Age years*			
Mean±SD	59.6 (8.58)	53.8 (7.30)	56.7 (8.41)
Median (range)	58.0 (44.0–80.0)	54.0 (42.0–76.0)	56.0 (42.0–80.0)
Ethnicity n (%)			
Black, non-Hispanic	16 (61.5)	12 (46.2)	28 (53.8)
White, Hispanic/Latino	10 (38.5)	13 (50.0)	23 (44.2)
Asian/Pacific Islander	0 (0)	1 (3.8)	1 (1.9)
Smoker n (%)			
Yes	14 (53.8)	14 (53.8)	28 (53.8)
Former	9 (34.6)	7 (26.9)	16 (30.8)
Never	3 (11.5)	5 (19.2)	8 (15.4)
Pack-years*			
Mean±SD	31.1 (28.3)	15.2 (13.7)	23.1 (23.4)
Median (range)	29.6 (0–120)	13.6 (0–38.0)	17.2 (0–120)
Receiving ART n (%)			
Yes	24 (92.3)	24 (92.3)	48 (92.3)
No	2 (7.7)	2 (7.7)	4 (7.7)
Viral load n (%)			
<50 copies	12 (46.2)	18 (69.2)	30 (57.7)
>50 copies	2 (7.7)	3 (11.5)	5 (9.6)
Missing	12 (46.2)	5 (19.2)	17 (32.7)
FEV_1 % pred***			
Mean±SD	68.0 (15.9)	104 (11.2)	85.8 (22.6)
Median (range)	68.3 (21.0–90.4)	102 (81.3–128)	86.2 (21.0–128)
FEV_1 L			
Mean±SD	2.07 (0.596)	3.25 (0.746)	2.66 (0.896)
Median (range)	2.09 (0.650–3.29)	3.08 (1.95–4.77)	2.59 (0.650–4.77)
FEV_1/FVC			
Mean±SD	0.556 (0.113)	0.795 (0.0567)	0.676 (0.150)
Median (range)	0.590 (0.293–0.679)	0.789 (0.689–0.905)	0.684 (0.293–0.905)
D_{LCO} % pred***			
Mean±SD	71.6 (26.3)	76.6 (23.1)	74.2 (24.6)
Median (range)	67.6 (36.3–139)	74.5 (14.4–117)	74.5 (14.4–139)
Missing n (%)	6 (23.1)	5 (19.2)	11 (21.2)

Cases and controls were matched based on age, smoking status and ART status. ART: antiretroviral treatment; FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity; D_{LCO}: diffusing capacity of the lung for carbon monoxide. Asterisks denote variables that were significantly different between cases and controls: *: significance at the 0.05 level; ***: significance at the 0.001 level.

We then considered the correlation between FEV_1 % pred and each BALF protein within a permutation testing framework. Table 2 summarises the top proteins most significantly correlated with FEV_1 % pred. We found that 1305 proteins were significantly correlated with FEV_1 % pred at the 0.05 level, prior to FDR adjustment and 575 proteins were significant at the 0.05 level after FDR adjustment. Proteins significant at the 0.05 level, prior to adjustment, were filtered into pathway over-representation analysis using IMPaLA software. Although we found many significant proteins, no pathways met significance after controlling for multiple comparisons (supplementary table 1S).

We found 1467 proteins were significantly correlated with FEV_1/FVC at the 0.05 FDR level (supplementary table 2S). No proteins were significantly correlated with D_{LCO} % pred after multiple comparisons adjustment.

Endotypes identified by cluster analysis

Heatmaps of the protein expression revealed a visually distinct subgroup of patients who tended to have lower FEV_1 % pred values (figure 2a). We thus constructed protein-driven endotypes using K-means clustering with 100 replications for K=2 clusters. K-means clustering results were stable across the 100

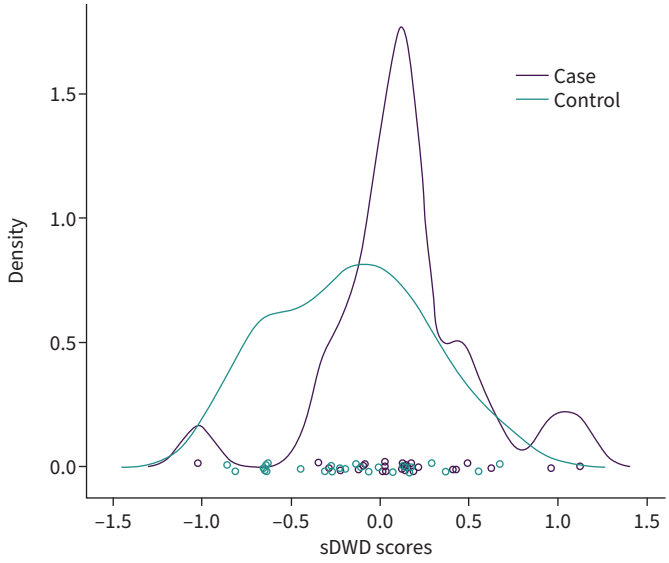

FIGURE 1 Densities of sparse distance-weighted discrimination (sDWD) scores for cases and controls based on bronchoalveolar lavage fluid protein expression. sDWD scores are a linear combination of protein expression levels. Distinct separation in the scores between cases and controls reflects more power to distinguish the classes based on protein expression.

replications, with each replication yielding the same clustering scheme of 10 individuals in one cluster and 42 in the other, referred to as Cluster 1 and Cluster 2, respectively. Figure 2b shows the protein expression across the patient cohort, with observations grouped by their assigned cluster. Table 3 demonstrates that Cluster 1 was largely male (80%) and on average older than Cluster 2 (62.5 *versus* 55.3 years). Cluster 1 also largely comprised individuals who identified as black (70%) compared to Cluster 2 where 50% of individuals identified as black. Cluster 1 also comprised individuals with lower average FEV_1 % pred (63.9 *versus* 91.0) and lower average FEV_1/FVC (0.514 *versus* 0.714), and 90% of individuals were diagnosed with obstructive lung disease compared to 40.5% in Cluster 2. D_{LCO} was available for six out of 10 individuals in Cluster 1 and 35 out of 42 in Cluster 2. The average D_{LCO} in Cluster 1 was 0.611 and 0.764 in Cluster 2 (p=0.15). The overall stability of this clustering scheme, as determined under bootstrap replications, was 89%.

We used an analogous permutation testing scheme to compare proteins across the clusters. The top 10 proteins that were significant at an FDR level of 0.05 are shown in table 4, and all 1279 significant

TABLE 2 Top 11 proteins most significantly correlated with FEV_1 % pred

Protein	UniProt ID	Average correlation	Permutation p-value	Q-value
NFL	P07196	−0.5329880	0.0001000	0.0107545
TM190	Q8WZ59	−0.5138893	0.0001000	0.0107545
RASN	P01111	−0.5341348	0.0001000	0.0107545
Ephrin-A2	O43921	−0.5215288	0.0001000	0.0107545
EPN4	Q14677	−0.5653480	0.0001000	0.0107545
CACO2	Q13137	0.5132561	0.0001000	0.0107545
NAL10	Q86W26	−0.5643206	0.0001000	0.0107545
FLRT3:ECD	Q9NZU0	0.5840321	0.0001000	0.0107545
kallikrein 8	O60259	−0.5329102	0.0001000	0.0107545
IL-18 Ra	Q13478	−0.6274007	0.0001000	0.0107545
RHG05	Q13017	0.5339263	0.0001000	0.0107545

Proteins are ordered based on FDR-adjusted p-values from permutation testing with the Pearson correlation test. Correlation was calculated by averaging the correlation across all aptamers that map to each protein. FEV_1: forced expiratory volume in 1 s.

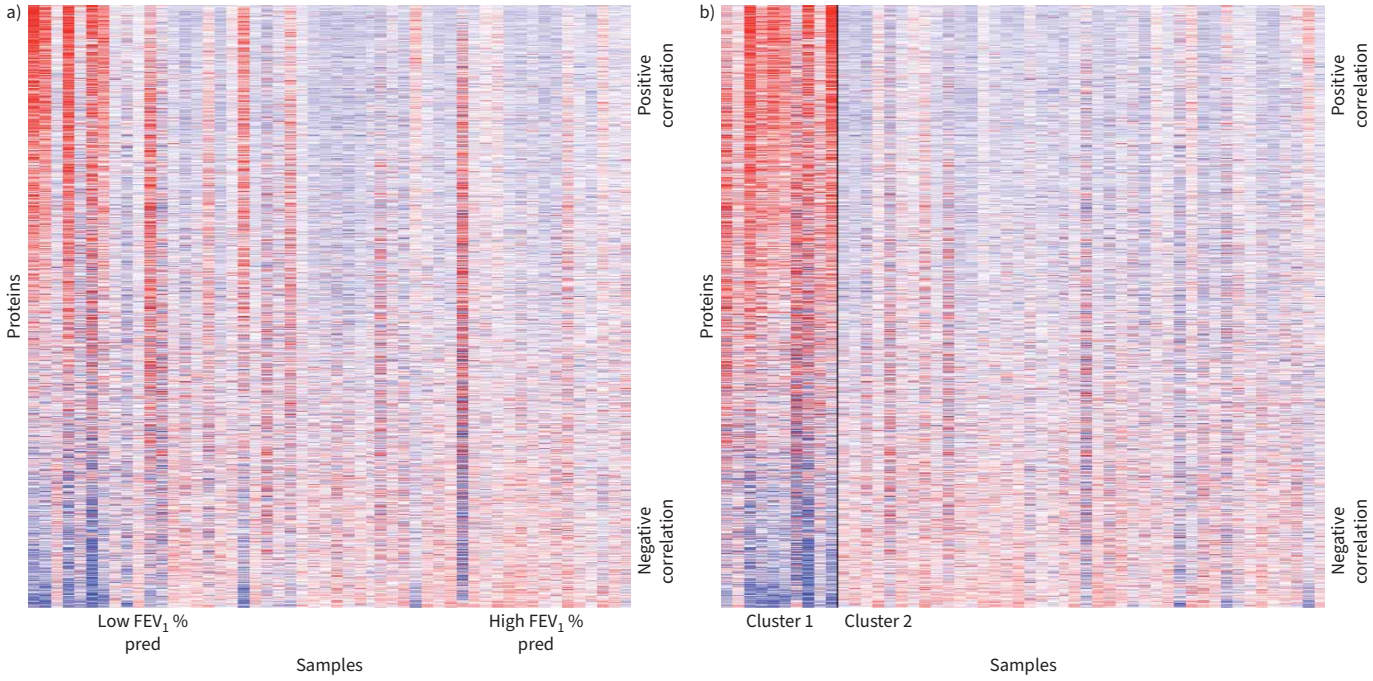

FIGURE 2 Heatmaps show protein expression across individuals in study cohort. Columns reflect samples while rows reflect proteins. In both heatmaps, proteins are ordered based on the direction and magnitude of their correlation with forced expiratory volume in 1 s % predicted (FEV_1 % pred). a) Heatmap samples ordered by FEV_1 % pred. b) Samples are grouped within their respective K-means clusters and the solid black line separates the two groups. Within clusters, samples are ordered by FEV_1 % pred.

proteins after FDR adjustment were used for pathway analysis. The top pathway distinguishing these two clusters involved insulin, with an FDR-adjusted p-value of 0.0312 (table 5). Other top pathways involved FOXO transcription factors, apoptosis, RNA metabolism and retinol metabolism.

To determine if the proteomic findings extended to metabolomic expression we performed a K-means clustering on the BALF metabolomics data. We found that the metabolomic expression yielded similar results to the BALF proteomics data with similar heatmaps of metabolite expression (figure 3). We obtained a consistent clustering scheme across 100 replications: one cluster with 10 individuals and another cluster with 42 (supplementary table 4S). However, no cluster was identified using plasma metabolites (supplementary figure 1S). Asparagine was the most significant BALF metabolite in cluster 1, while overall acylcarnitines were the predominant metabolites in this cluster (supplementary table 5S). While the cluster sizes were identical to those obtained using the BALF protein expression data, the composition of the metabolite clusters differed slightly. There was an overlap of six individuals in the smaller cluster of 10 between the protein-driven and metabolite-driven clusters (supplementary table 6S). These six individuals were all male with an average age of 65.7 years and all identified as black.

Discussion

We found that the BALF proteome in persons living with HIV distinguishes those with obstructive lung disease from those without obstructive lung disease. However, we did not identify any pathways composed of proteins that were differentially expressed among individuals with high FEV_1 % pred and those with low FEV_1 % pred. We did identify an endotype driven by both proteomic and metabolomic BALF molecular contributors, but not plasma metabolites. Based on the differentially expressed BALF proteins, this endotype exhibited over-representation of insulin- and apoptosis-related pathways, suggesting that signal transduction pathways with FOXO and cell cycle are important regulators.

Previous studies identified plasma proteins associated with pulmonary dysfunction in HIV. These include elevated plasma interleukin (IL)-6, C-reactive protein, endothelin-1 [30] and activation of inflammatory pathways [31, 32]. We found similar findings in the START cohort where higher plasma levels of IL-6, high-sensitivity C-reactive protein and serum amyloid A associated with lower FEV_1 and FVC [33].

TABLE 3 Demographics of two K-means clusters determined using protein expression

	Cluster 1	Cluster 2	Total
Patients n	10	42	52
Sex n (%)			
Male	8 (80.0)	30 (71.4)	38 (73.1)
Female	2 (20.0)	12 (28.6)	14 (26.9)
Age years*			
Mean±SD	62.5 (8.61)	55.3 (7.84)	56.7 (8.41)
Median (range)	62.0 (49.0–76.0)	54.5 (42.0–80.0)	56.0 (42.0–80.0)
Ethnicity n (%)			
Black, non-Hispanic	7 (70.0)	21 (50.0)	28 (53.8)
White, Hispanic/Latino	3 (30.0)	20 (47.6)	23 (44.2)
Asian/Pacific Islander	0 (0)	1 (2.4)	1 (1.9)
Smoker n (%)			
Yes	5 (50.0)	23 (54.8)	28 (53.8)
Former	4 (40.0)	12 (28.6)	16 (30.8)
Never	1 (10.0)	7 (16.7)	8 (15.4)
Pack-years			
Mean±SD	31.0 (25.2)	21.3 (22.9)	23.1 (23.4)
Median (range)	32.5 (0–80.0)	16.2 (0–120)	17.2 (0–120)
Receiving ART n (%)			
Yes	9 (90.0)	39 (92.9)	48 (92.3)
No	1 (10.0)	3 (7.1)	4 (7.7)
Viral load n (%)			
<50 copies	3 (30.0)	27 (64.3)	30 (57.7)
>50 copies	0 (0)	5 (11.9)	5 (9.6)
Missing	7 (70.0)	10 (23.8)	17 (32.7)
FEV_1 % pred#			
Mean±SD	63.9 (22.6)	91.1 (19.4)	85.8 (22.6)
Median (range)	62.8 (21.0–100)	90.5 (46.7–128)	86.5 (21.0–128)
FEV_1 L			
Mean±SD	1.87 (0.676)	2.85 (0.843)	2.66 (0.896)
Median (range)	1.79 (0.650–3.06)	2.83 (1.32–4.77)	2.59 (0.650–4.77)
FEV_1/FVC#			
Mean±SD	0.514 (0.146)	0.714 (0.124)	0.676 (0.150)
Median (range)	0.549 (0.293–0.694)	0.750 (0.413–0.905)	0.684 (0.293–0.905)
D_{LCO} % pred			
Mean±SD	61.1 (21.1)	76.5 (24.7)	74.2 (24.6)
Median (range)	54.7 (43.0–103)	75.0 (14.4–139)	74.5 (14.4–139)
Missing n (%)	4 (40.0)	7 (16.7)	11 (21.2)
Case–control status n (%)*			
Case	9 (90.0)	17 (40.5)	26 (50.0)
Control	1 (10.0)	25 (59.5)	26 (50.0)

ART: antiretroviral treatment; FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity; D_{LCO}: diffusing capacity of the lung for carbon monoxide. *: significant differences between clusters at 0.05 level; #: significant differences between clusters at the 0.005 level.

In addition, unique phenotypes associated with inflammatory pathways have been identified by cluster analysis in HIV-associated obstructive lung disease [34]. Unlike our current study, these studies were obtained from blood samples, not direct sampling of the lung. Since the BALF proteome in healthy persons living with HIV differs significantly from non-HIV controls [35], we sought to identify lung-specific biomarkers of HIV-associated obstructive lung disease.

This study was designed as a case–control study (obstructive lung disease present/absent) where pairs were matched based on age, antiretroviral treatment status and smoking status. Though we found the BALF proteome moderately differentiates cases and controls, we primarily considered FEV_1 % pred as an outcome rather than case–control status due to the heterogeneity in lung function within each group, which also improved study power. We found 1305 proteins that were significantly correlated with FEV_1 % pred at the 0.05 level and 575 proteins that were significant after FDR adjustment. Our finding that BALF IL-18 Ra is highly correlated with FEV_1 % pred is consistent with Imaoka and colleagues [36] who reported

TABLE 4 Top 10 proteins significantly different between two K-means clusters determined based on protein expression

Protein	UniProt ID	Average test statistic	p-value	Q-value
Beclin-1	Q14457	−8.960709	0.0001000	0.0035519
Hepatocyte nuclear factor 4-α	P41235	6.592431	0.0001000	0.0035519
Mothers against decapentaplegic homolog 3	P84022	6.590634	0.0001000	0.0035519
Ankyrin repeat domain-containing protein 1	Q15327	−4.938720	0.0001000	0.0035519
Neurotrimin	Q9P121	−5.468840	0.0001000	0.0035519
Triple functional domain protein	O75962	−4.330947	0.0001000	0.0035519
RING finger protein 122	Q9H9V4	7.671931	0.0001000	0.0035519
Adhesion G-protein-coupled receptor F1	Q5T601	10.490794	0.0001000	0.0035519
Calsequestrin-1	P31415	7.972240	0.0001000	0.0035519
Transaldolase	P37837	9.138911	0.0001000	0.0035519

p-values determined by permutation testing. Average test statistic calculated by averaging the t-test statistic across aptamers that correspond to each protein.

IL-18 as highly expressed in lung tissue among COPD patients without HIV and that increased expression associated with a decrease in FEV_1 % pred. This has not been previously reported in obstructive lung disease associated with HIV. We also found the protein ephrin-A2 to be highly associated with FEV_1 % pred and its gene, *EFNA2*, is associated with weight loss among patients with non-HIV COPD [37]. Despite identifying many proteins significantly correlated with FEV_1 % pred, we were not able to detect pathways reflected by these proteins. This may be due to lack of statistical power in our relatively small sample size.

We identified two endotypes driven by proteomic expression in BALF that were also identified in the BALF metabolome. These BALF endotypes were consistently detected across 100 replications with 10 random start values of the K-means clustering algorithm, suggesting these clusters are robust to many different initialisations of the algorithm. We detected these endotypes in both the BALF proteome and the metabolome but not the plasma metabolome, though the compositions of the clusters differed slightly between the BALF proteomic and metabolomic platforms. Both the metabolome and the proteome had a smaller endotype consisting of 10 individuals, which showed clear differential expression in heatmaps. There were six individuals who were consistently grouped into this smaller cluster between both sources, suggesting they possess a unique BALF expression profile apparent in both the proteome and metabolome (supplementary table 6S). These six individuals all identified as black non-Hispanic males who met the criteria for obstructive lung disease and exhibited lower FEV_1 % pred compared to Cluster 2.

Pathway analysis using the significant proteins identified in cluster 1 revealed several pathways. Among those were insulin, regulation of FOXO transcription factors and apoptosis pathways. Many of the proteins

TABLE 5 Top nine pathways based on proteins that were differentially expressed between K-means clusters and were significant at an FDR level of 0.05

Pathway	Source	Overlapping genes	All pathway genes	p-value	Q-value
Insulin mammalian	INOH	22	31 (82)	1.33e-05	0.0312
Regulation of localisation of FOXO transcription factors	Reactome	9	9 (12)	4.19e-05	0.0384
Intrinsic pathway for apoptosis	Reactome	16	21 (52)	5.28e-05	0.0384
Metabolism of RNA	Reactome	57	113 (584)	5.29e-05	0.0384
Retinol metabolism	SMPDB	12	14 (37)	6.42e-05	0.0384
Vitamin A deficiency	SMPDB	12	14 (37)	6.42e-05	0.0384
Programmed cell death	Reactome	29	48 (142)	6.57e-05	0.0384
Apoptosis	Reactome	27	44 (127)	8.03e-05	0.0417
Fas	INOH	13	16 (24)	8.94e-05	0.0419

FDR: false discovery rate; INOH: Integrating Network Objects with Hierarchies database; SMPDB: The Small Molecule Pathway Database.

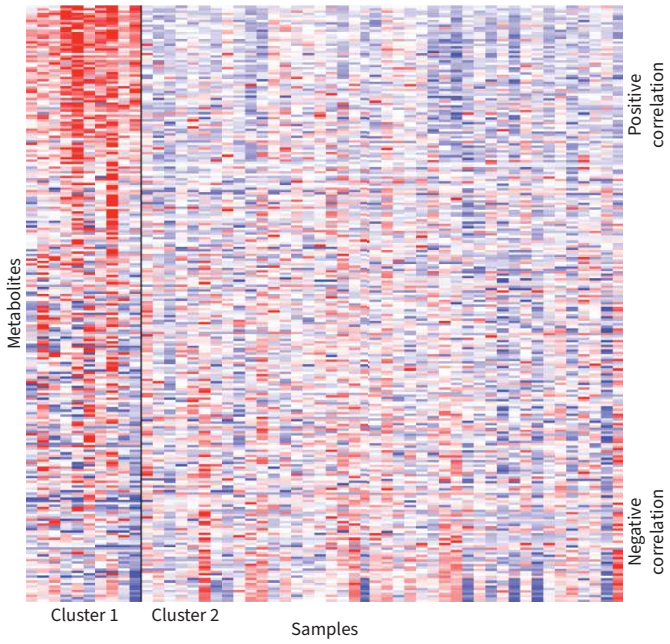

FIGURE 3 Heatmap shows metabolite expression for the study cohort. Columns reflect samples and rows reflect metabolites. The metabolites are ordered by the direction and magnitude of their correlation with forced expiratory volume in 1 s % predicted. The samples are grouped within their respective K-means clusters and the solid black line separates the two groups.

expressed in these pathways are involved in signal transduction and cell cycle regulation. Two promoter forkhead box (FOX) proteins were among the proteins in the insulin and FOXO pathways. FOX promoter expression is increased in epithelial cells in non-HIV COPD patients; however, in patients with mucus hyperexcretion phenotype it is depleted. This transcription factor is involved in goblet cell differentiation in the airway epithelium, and aberrant methylation patterns have been described in non-HIV COPD lung epithelium [38–41].

Pathway analysis of our endotype also revealed an intrinsic pathway for apoptosis as being significant. Enhanced apoptosis in lung endothelial and epithelial cells is found in non-HIV COPD and is felt to be a critical step in COPD pathogenesis [42, 43]. In non-HIV COPD, enhanced apoptosis has been associated with the emphysema phenotype. Computed tomography imaging was not available to quantify emphysema in our study and the D_{LCO}, which correlates with the emphysema phenotype, did not vary between endotypes, although the relatively small sample size of Cluster 1 likely limited statistical power. Accelerated apoptosis is one mechanism proposed for the loss of $CD4^+$ T-lymphocytes in HIV infection [44]. It is also postulated that HIV-infected persons have increased susceptibility to apoptosis because the HIV proteins Tat and Nef induce endothelial cell apoptosis [45, 46]. Further investigations are necessary to determine whether these apoptotic pathways are associated with lymphocyte or lung parenchymal cells.

Our study has a few limitations. Owing to our small sample size and large number of proteins, we were unable to detect any significantly over-represented pathways reflected by proteins associated with FEV_1 % pred. In a future study with more participants, we may have more power to identify pathways associated with lung function decline. It would be interesting to recruit HIV-negative controls to assess if differentially expressed proteins and protein-driven pathways are unique to HIV-modulated obstructive lung disease or are exhibited across the population of patients with obstructive lung disease. A larger sample size and validation cohort would also be beneficial to corroborate the endotypes we detected in our study and their clinical outcomes. Though our study was a matched case–control study, individuals were not matched based on race or sex, and additional studies would be necessary to validate if these findings are race- or sex-specific. Though pairs were matched based on smoking status, we did not account for this in our correlation analysis, which is a limitation. Although not a limitation of the study, we did not find differences reflected in the plasma metabolome, thus limiting the feasibility in applying these BALF findings as global biomarkers of lung disease. In addition, a longer, longitudinal study in which protein

expression is measured over time may highlight proteins that are relevant to lung function change, as previous research has shown that proteins associate differently with COPD outcomes at a single time point *versus* longitudinally [47].

In conclusion, the proteomic BALF profile distinguishes HIV-associated obstructive lung disease. Furthermore, a unique endotype was identified in both BALF proteomic and metabolomic profiles that predominantly were in black men with more severe obstructive lung disease, and this cluster was not found in the plasma metabolome. The proteomic pathways that were differentially expressed within this endotype were linked to signal transduction pathways, cell cycle and apoptosis.

Provenance: Submitted article, peer reviewed.

Conflict of interest: No conflict of interest related to the work in this manuscript for any of the authors. K.M. Kunisaki reports consulting (Allergan/AbbVie), and independent data safety and monitoring boards (Nuvaira and Organicell), outside of this work.

Support statement: Supported by National Institutes of Health grant R01 HL140971-01A1 (all authors). This material is also the result of work supported with resources and the use of facilities at the Minneapolis Veterans Affairs Medical Center, Minneapolis, MN, USA. The views expressed in this article are those of the authors and do not reflect the views of the US Government, the Department of Veterans Affairs, the funders, the sponsors or any of the authors' affiliated academic institutions. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Drummond MB, Merlo CA, Astemborski J, *et al.* The effect of HIV infection on longitudinal lung function decline among IDUs: a prospective cohort. *AIDS* 2013; 27: 1303–1311.
2 George MP, Kannass M, Huang L, *et al.* Respiratory symptoms and airway obstruction in HIV-infected subjects in the HAART era. *PLoS One* 2009; 4: e6328.
3 Gingo MR, George MP, Kessinger CJ, *et al.* Pulmonary function abnormalities in HIV-infected patients during the current antiretroviral therapy era. *Am J Respir Crit Care Med* 2010; 182: 790–796.
4 Hirani A, Cavallazzi R, Vasu T, *et al.* Prevalence of obstructive lung disease in HIV population: a cross sectional study. *Respir Med* 2011; 105: 1655–1661.
5 Crothers K, Huang L, Goulet JL, *et al.* HIV infection and risk for incident pulmonary diseases in the combination antiretroviral therapy era. *Am J Respir Crit Care Med* 2011; 183: 388–395.
6 Cui Q, Carruthers S, McIvor A, *et al.* Effect of smoking on lung function, respiratory symptoms and respiratory diseases amongst HIV-positive subjects: a cross-sectional study. *AIDS Res Ther* 2010; 7: 6.
7 Madeddu G, Fois AG, Calia GM, *et al.* Chronic obstructive pulmonary disease: an emerging comorbidity in HIV-infected patients in the HAART era? *Infection* 2013; 41: 347–353.
8 Kristoffersen US, Lebech AM, Mortensen J, *et al.* Changes in lung function of HIV-infected patients: a 4.5-year follow-up study. *Clin Physiol Funct Imaging* 2012; 32: 288–295.
9 Raynaud C, Roche N, Chouaid C. Interactions between HIV infection and chronic obstructive pulmonary disease: clinical and epidemiological aspects. *Respir Res* 2011; 12: 117.
10 Crothers K, Butt AA, Gibert CL, *et al.* Increased COPD among HIV-positive compared to HIV-negative veterans. *Chest* 2006; 130: 1326–1333.
11 Diaz PT, King MA, Pacht ER, *et al.* Increased susceptibility to pulmonary emphysema among HIV-seropositive smokers. *Ann Intern Med* 2000; 132: 369–372.
12 Drummond MB, Kirk GD, McCormack MC, *et al.* HIV and COPD: impact of risk behaviors and diseases on quality of life. *Qual Life Res* 2010; 19: 1295–1302.
13 Drummond MB, Kirk GD, Ricketts EP, *et al.* Cross sectional analysis of respiratory symptoms in an injection drug user cohort: the impact of obstructive lung disease and HIV. *BMC Pulm Med* 2010; 10: 27.
14 Fitzpatrick ME, Singh V, Bertolet M, *et al.* Relationships of pulmonary function, inflammation, and T-cell activation and senescence in an HIV-infected cohort. *AIDS* 2014; 28: 2505–2515.
15 Attia EF, McGinnis KA, Feemster LC, *et al.* Association of COPD with risk for pulmonary infections requiring hospitalization in HIV-infected veterans. *J Acquir Immune Defic Syndr* 2015; 70: 280–288.
16 White NC, Agostini C, Israel-Biet D, *et al.* The growth and the control of human immunodeficiency virus in the lung: implications for highly active antiretroviral therapy. *Eur J Clin Invest* 1999; 29: 964–972.
17 Cribbs SK, Uppal K, Li S, *et al.* Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome* 2016; 4: 3.
18 Akata K, Leung JM, Yamasaki K, *et al.* Altered polarization and impaired phagocytic activity of lung macrophages in people with HIV and COPD. *J Infect Dis* 2022; 225: 862–867.

19 Hodgson S, Griffin TJ, Reilly C, *et al.* Plasma sphingolipids in HIV-associated chronic obstructive pulmonary disease. *BMJ Open Respir Res* 2017; 4: e000180.
20 Wendt CH, Castro-Pearson S, Proper J, *et al.* Metabolite profiles associated with disease progression in influenza infection. *PLoS One* 2021; 16: e0247493.
21 Simats A, Ramiro L, Montaner J, *et al.* Application of an aptamer-based proteomics assay (SOMAscan) in rat cerebrospinal fluid. *Methods Mol Biol* 2019: 2044: 221–231.
22 UniProt Consortium. UniProt: the universal protein knowledge base in 2021. *Nucleic Acids Res* 2021: 49: D480–D489.
23 Wang BZ H. Sparse distance weighted discrimination. *J Comput Graph Stat* 2016; 25: 826–838.
24 Quanjer PH, Stanojevic S, Cole TJ, *et al.* Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012; 40: 1324–1343.
25 Elson RC. On Fisher's method of combining p-values. *Biom J* 1991; 33: 339–345.
26 Benjamini Y, Drai D, Elmer G, *et al.* Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* 2001; 125: 279–284.
27 Kamburov A, Cavill R, Ebbels TM, *et al.* Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 2011; 27: 2917–2918.
28 James G, Witten D, Hastie T, *et al.*, eds. An Introduction to Statistical Learning. New York, Springer, 2013.
29 Yu H, Chapman B, Di Florio A, *et al.* Bootstrapping estimates of stability for clusters, observations and model selection. *Comput Stat* 2019; 34: 349–372.
30 Fitzpatrick ME, Nouraie M, Gingo MR, *et al.* Novel relationships of markers of monocyte activation and endothelial dysfunction with pulmonary dysfunction in HIV-infected persons. *AIDS* 2016; 30: 1327–1339.
31 Jan AK, Moore JV, Wang RJ, *et al.* Markers of inflammation and immune activation are associated with lung function in a multi-center cohort of persons with HIV. *AIDS* 2021; 35: 1031–1040.
32 North CM, Muyanja D, Kakuhikire B, *et al.* Brief report: systemic inflammation, immune activation, and impaired lung function among people living with HIV in rural Uganda. *J Acquir Immune Defic Syndr* 2018; 78: 543–548.
33 MacDonald DM, Zanotto AD, Collins G, *et al.* Associations between baseline biomarkers and lung function in HIV-positive individuals. *AIDS* 2019; 33: 655–664.
34 Qin S, Vodovotz L, Zamora R, *et al.* Association between inflammatory pathways and phenotypes of pulmonary dysfunction using cluster analysis in persons living with HIV and HIV-uninfected individuals. *J Acquir Immune Defic Syndr* 2020; 83: 189–196.
35 Nguyen EV, Gharib SA, Crothers K, *et al.* Proteomic landscape of bronchoalveolar lavage fluid in human immunodeficiency virus infection. *Am J Physiol Lung Cell Mol Physiol* 2014; 306: L35–L42.
36 Imaoka H, Hoshino T, Takei S, *et al.* Interleukin-18 production and pulmonary function in COPD. *Eur Respir J* 2008; 31: 287–297.
37 Kumar PL, Wilson AC, Rocco A, *et al.* Genetic variation in genes regulating skeletal muscle regeneration and tissue remodelling associated with weight loss in chronic obstructive pulmonary disease. *J Cachexia Sarcopenia Muscle* 2021; 12: 1803–1817.
38 Lange P, Ahmed E, Lahmar ZM, *et al.* Natural history and mechanisms of COPD. *Respirology* 2021; 26: 298–321.
39 Song J, Heijink IH, Kistemaker LEM, *et al.* Aberrant DNA methylation and expression of SPDEF and FOXA2 in airway epithelium of patients with COPD. *Clin Epigenetics* 2017; 9: 42.
40 Choi W, Choe S, Lin J, *et al.* Exendin-4 restores airway mucus homeostasis through the GLP1R-PKA-PPARgamma-FOXA2-phosphatase signaling. *Mucosal Immunol* 2020; 13: 637–651.
41 Chen G, Korfhagen TR, Xu Y, *et al.* SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. *J Clin Invest* 2009; 119: 2914–2924.
42 Demedts IK, Demoor T, Bracke KR, *et al.* Role of apoptosis in the pathogenesis of COPD and pulmonary emphysema. *Respir Res* 2006; 7: 53.
43 Kasahara Y, Tuder RM, Cool CD, *et al.* Endothelial cell death and decreased expression of vascular endothelial growth factor and vascular endothelial growth factor receptor 2 in emphysema. *Am J Respir Crit Care Med* 2001; 163: 737–744.
44 Cummins NW, Badley AD. Mechanisms of HIV-associated lymphocyte apoptosis: 2010. *Cell Death Dis* 2010; 1: e99.
45 Tuder RM, McGrath S, Neptune E. The pathobiological mechanisms of emphysema models: what do they have in common? *Pulm Pharmacol Ther* 2003; 16: 67–78.
46 Micoli KJ, Pan G, Wu Y, *et al.* Requirement of calmodulin binding by HIV-1 gp160 for enhanced FAS-mediated apoptosis. *J Biol Chem* 2000; 275: 1233–1240.
47 Serban KA, Pratte KA, Bowler RP. Protein biomarkers for COPD outcomes. *Chest* 2021; 159: 2244–2253.