# AJRCCM Instructions for Contributors

Manuscripts should:

* 12-point type
* 2.5 centimeters (1 inch) margins
* Maximum word count of 3500 words
* Double spacing should be used throughout
* All papers should be organized to include:
  + title page
  + abstract
  + text
  + acknowledgments
  + references
  + figure legends
  + footnotes
  + tables
  + figures
* Each of the preceding elements should begin on a separate page. While the Introduction should begin on a separate page, it is not necessary for the Methods, Results, and Discussion section to begin on a separate page; instead, they should follow after the Introduction, with two spaces separating each section.
* Pages should be numbered consecutively, beginning with the abstract.

# Title Page

Gene Expression Signatures of COPD Disease Activity

**AUTHORS:** Virginia Y. Chen1, \*, Casey P. Shannon1, \*, Yuree Jang1, Zsuzsanna Hollander1, Robert Balshaw1,2, Scott J. Tebbutt1,3,6,7, Bruce M. McManus1,3,4,7, Raymond T. Ng1,3,5,7, †Don D. Sin3,6,7

**AFFILIATIONS:** 1PROOF Centre of Excellence, Vancouver, BC, Canada; 2BC Centre for Disease Control, Vancouver, BC, Canada; 3 Institute for Heart and Lung Health, Vancouver, BC, Canada; 4Department of Pathology and Laboratory Medicine, UBC, Vancouver, BC, Canada; 5Department of Computer Science, UBC, Vancouver, BC, Canada; 6Department of Medicine, Division of Respiratory Medicine, UBC, Vancouver, BC, Canada; 7Centre for Heart Lung Innovation, UBC, Vancouver, BC, Canada.

\* These authors contributed equally to the work

† Corresponding author –include address in full

Virginia Y. Chen: [virginia.chen@hli.ubc.ca](mailto:virginia.chen@hli.ubc.ca)

Casey P. Shannon: [casey.shannon@hli.ubc.ca](mailto:casey.shannon@hli.ubc.ca)

Yuree Jang: [yj.yuree.jang@gmail.com](mailto:yj.yuree.jang@gmail.com)

Zsuzsanna Hollander: [zsuzsanna.hollander@hi.ubc.ca](mailto:zsuzsanna.hollander@hi.ubc.ca)

Robert Balshaw: [robert.balshaw@bccdc.ca](mailto:robert.balshaw@bccdc.ca)

Scott J. Tebbutt: [scott.tebbutt@hli.ubc.ca](mailto:scott.tebbutt@hli.ubc.ca)

Bruce M. McManus: [bruce.mcmanus@hli.ubc.ca](mailto:bruce.mcmanus@hli.ubc.ca)

Raymond T. Ng: [rng@cs.ubc.ca](mailto:rng@cs.ubc.ca)

Don D. Sin: [don.sin@hli.ubc.ca](mailto:don.sin@hli.ubc.ca)

**Author’s Contributions:**

VYC:

CPS:

YJ:

ZH:

RB:

SJT:

BMM:

RTN:

DDS:

**Funding:** Funding was provided by Genome Canada, Genome British Columbia, Genome Quebec, the Canadian Institutes of Health Research, PROOF Centre of Excellence, St. Paul’s Hospital Foundation, Providence Health Care, and the Canadian Respiratory Research Network.

**Running Title:**

**Subject Category Descriptor:**

**Word Count (abstract – max 250): 250**

**Word Count (body – max 3500): 4123**

# Abstract

**Rationale** Chronic obstructive pulmonary disease (COPD) is the 4th leading cause of death worldwide, accounting for 5% of global mortality. COPD is characterized by loss of lung function and sporadic worsening of symptoms termed acute exacerbations of COPD (AECOPD). These events are associated with significant morbidity and mortality, are a major cause of COPD-related hospitalizations, and have been associated with accelerated disease progression and permanent lung function decline. No biomarker exists capable of predicting imminent AECOPD.

**Objective:** To identify blood-based biomarkers of disease activity in COPD.

**Methods:** Whole blood gene expression profiling was carried out in two large clinical cohorts. Weighted Gene Co-expression Network Analysis (WGCNA) was first applied to 226 subjects with stable COPD to identify co-regulated gene modules. These modules were used for biomarker discovery in 20 imminent exacerbators (IE; AECOPD occurring within 60 days of the blood draw) versus 122 non-exacerbators (NE), and 64 active AECOPD vs. 45 stable COPD. Their discriminative performance was confirmed in independent samples (157 active AECOPD vs. 57 stable COPD).

**Measurements and Main Results** Twenty-three gene modules were identified. Three modules yielded panels capable of discriminating IE from NE with cross-validation AUC greater than 0.65 (ranging from 0.67 – 0.71), and able to discriminate between AECOPD and stable COPD when tested off-the-shelf with AUC 0.76 – 0.82. We confirmed the latter in a large independent cohort (AUC 0.74 – 0.84).

**Conclusions:** These biosignatures were highest during AECOPD and showed decreasing levels during convalescence, tracking the clinical manifestation of symptoms, and may be useful monitoring tools.

**Word Count (max – 250): 250**

**Key Words (MeSH terms preferred):**

# Introduction

Chronic obstructive pulmonary disease (COPD) affects an estimated 328 million individuals worldwide(1). In 2013, COPD was the 4th leading cause of death worldwide, accounting for 2.9 million deaths or approximately 5% of global mortality (2). COPD is a progressive disease that is characterized by loss of lung function, and sporadic worsening of symptoms (primarily dyspnea, cough, and sputum production) termed acute exacerbation of COPD (AECOPD). These events are associated with significant morbidity and mortality(3, 4), are a major cause of COPD-related hospitalizations, and have been associated with accelerated disease progression(5, 6). AECOPDs are also important contributors to the economic burden of COPD(7) and, consequently, there has been much interest in developing tools to better diagnose them or predict their onset in order to improve patient management and treatment.

Currently, diagnosis of AECOPD is based on the worsening of day-to-day COPD symptoms to levels deemed clinically significant enough to require acute treatment. These diagnoses are thus subjective assessments. While knowledge that a patient has an underlying COPD disease can assist in diagnosis of AECOPD, symptoms tend to be non-specific and overlap with those of myocardial infarction, pulmonary embolism, congestive heart failure, and others [00], presenting a challenge to physicians. Furthermore, appropriate treatment varies with underlying AECOPD etiology, which may consist of bacterial infection, viral infection, or other triggers [00]. Timely diagnosis of AECOPD and prompt initiation of appropriate treatment is important in preventing full-blown lung attacks. As no simple clinical tools currently exist to risk-stratify patients with COPD or identify the underlying etiology of AECOPD, patients are often over- or under-treated, leading to suboptimal health outcomes, cost inefficiencies, and significant morbidity and mortality [00]. Thus, there is a pressing need for biomarkers to improve prevention and management of AECOPD.

This study improves on the state of predictive and diagnostic biomarkers of AECOPD, by using an untargeted approach to identify modules of co-expressed genes in COPD patients, then assessing these modules for their ability to differentiate between different levels of AECOPD activity: non-exacerbating, upcoming exacerbation, and current exacerbation. Specifically, we discover transcriptomic signatures that are predictive of what we term imminent exacerbation (occurring within 60 days post-blood draw), and we show that these signatures also track the resolution of AECOPD to convalescence. These signatures could be useful in anticipating clinical manifestation of AECOPD symptoms, and informing COPD patient management in a way that reduces the burden of disease on both patients and healthcare providers.

# Methods

## Study Populations

Single time point blood samples from stable COPD subjects were obtained from the Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) cohort(16). The ECLIPSE study was conducted in accordance with the Declaration of Helsinki, and was approved by the ethics and review boards at participating sites. The present study using ECLIPSE samples was approved by The University of British Columbia Research Ethics Board number (H11-00786). A total of 367 samples were analyzed for this study, of which 226 were obtained from subjects’ 3 month time point and were used for gene regulatory module identification. The remaining samples were obtained from subjects’ 1 year time point, and were classified into 20 “imminent exacerbators” (IE; exacerbation occurring within 60 days post-blood draw) versus 122 non-exacerbators (NE; no exacerbation for ≥365 days post-blood draw) for prognostic biomarker discovery. These were subjects with fewer than 2 exacerbations requiring antibiotics or corticosteroids in the prior year, and who were also exacerbation-free for at least 30 days pre-blood draw.

The Rapid Transition cohort included patients admitted for AECOPD to Vancouver General Hospital or St. Paul’s Hospital in Vancouver, Canada. The diagnosis of AECOPD had to be verified by general internists or pulmonologists, and had to occur without evidence of pneumonia or heart failure on a chest x-ray. Blood samples were collected at the time of admission to the hospital (representing their AECOPD state), day 3 post-admission, day 7 post-admission, discharge, and at either day 30 or 90 post- admission if they were clinically stable (representing their convalescent state). Stable COPD patients (no current exacerbation symptoms, and no exacerbation in the past 2 weeks) were also enrolled and their blood collected during routine visits to the COPD clinic. The Rapid Transition study was approved by the University of British Columbia Clinical Research Ethics Board (certificate numbers H11-00786 and H13-00790). Written informed consent was provided by each participant in accordance with the Ethics Board. A total of 165 subject-timepoints, corresponding to 112 unique subjects enrolled between July 2012 and October 2014, were analyzed as part of the biomarker selection phase of this study. A total of 336 subject-timepoints, corresponding to 293 unique subjects enrolled between July 2012 and January 2017 (non-overlapping with the previous subjects), were analyzed as part of the replication phase of this study.

An overview of the study cohorts with respect to subjects’ time to/from exacerbation is shown in Figure 2. An overview of the study cohorts with respect to biomarker discovery and replication is shown in Figure 3.

## Sample processing

Blood was collected in PAXgene (PreAnalytix, Switzerland) and EDTA tubes, and stored at -80oC. A complete blood count, including leukocyte differential (CBC/diff) was obtained from the EDTA blood. Total RNA was extracted from PAXgene blood on the QIAcube (Qiagen, Germany), using the PAXgene Blood miRNA kit from PreAnalytix, according to manufacturer’s instructions. Human Gene 1.1 ST 96-well array plates (Affymetrix, United States) were used to measure mRNA abundance, and this was carried out at The Scripps Research Institute DNA Array Core Facility (TSRI; La Jolla, CA). Samples were pseudo-randomly assigned to plates to prevent confounding of phenotype with plate effects.

## Statistical analysis

Statistical analyses were performed in R(17), using packages sourced from CRAN and Bioconductor(18). Patient demographics were assessed using ANOVA and Kruskal-Wallis for continuous variables, and Fisher’s exact test for categorical variables.

### Normalization

Gene expression data from ECLIPSE and from Rapid Transition were normalized together to enable comparisons. Raw CEL files were processed using the *oligo* package(19) to assess quality and perform the Robust Multi-array Average normalization method(20), with probes summarized at the transcript cluster level. Batch correction was performed using the *ComBat* algorithm(21) to correct for plate-to-plate artifacts, without incorporating any phenotypic information.

### Gene-level Summarization

To ensure biologically meaningful results, transcript cluster-level data was summarized at the gene level using Affymetrix Human Gene 1.1 ST transcript cluster annotations (version 36). Un-annotated transcripts and transcripts that mapped to multiple genes were removed from the data. Genes with multiple transcripts assigned to them were assigned a value equal to the average of their corresponding transcripts. The result of this summarization will be referred to as the gene expression data.

### Gene regulatory module identification

We carried out an untargeted discovery of co-expressed gene modules using the gene expression data of 226 ECLIPSE subjects with month 3 samples. Many systemic inflammatory markers are elevated during AECOPD and it is reasonable to assume this is associated with changes in gene expression in various component cells of the blood. This untargeted approach may identify novel pathways and mechanisms specific to AECOPD. Weighted Gene Co-expression Network Analysis (WGCNA)(22) was used to identify modules of co-expressed genes, with the idea that co-expression may serve as a surrogate for regulatory relationships. WGCNA performs hierarchical clustering on a weighted gene-gene correlation matrix that has approximately scale-free topology. Unsigned correlations were used and the minimum module size was set at 50 genes.

### Pathway Enrichment analysis

The modules of genes identified by WGCNA were compared to approximately 13,000 annotated gene sets from the Broad Institute’s MSigDB collections(23, 24), 350 Blood Transcription Modules(25), and 73 tissue-specific gene sets(26). A hypergeometric test was used to identify which of these annotated genes sets were statistically over-represented in the modules identified by WGCNA, with correction for multiple testing performed using the Benjamini-Hochberg procedure(27).

### Module-based Biomarker discovery

We identified 20 imminent exacerbators (IE; patients who exacerbate without 60 days post-blood draw) and 122 non-exacerbators (NE; patients who are exacerbation-free for ≥365 days post-blood draw) from the ECLIPSE cohort. These were subjects with year 1 samples, whose data was not previously used to derive the co-expression modules. Differential gene expression between IE and NE patients was assessed on a per-module basis using the moderated *t*-test provided by the Linear Models for Microarray Data (LIMMA)(28) package. Genes with p-value <0.05 were selected as candidates for the biomarker panels. We then applied elastic net regularized logistic regression(29) to each list of candidate genes, separately for each module, to build binary classifiers for IE from NE. As elastic net performs regularization in addition to regression, only a subset of the candidate genes is given non-zero weights. We call the set of non-zero weighted genes in a given elastic net model a “biomarker panel”.

Area under the receiver operating characteristics curve (AUC) was selected as the primary performance metric of interest, since it summarizes discriminative power independently of decision thresholds. To obtain out-of-sample AUC estimates for each of the elastic net models, 10-fold stratified cross-validation was applied to the entire biomarker discovery process described above, for each gene module. Briefly, samples were divided into 10 partitions that were approximately balanced for the phenotype of interest. Data from all but one partition was used to identify differentially expressed genes with p-value <0.05 and train an elastic net classifier. This classifier was then tested using the samples in the left-out partition to obtain out-of-sample probabilities and AUC. This process was repeated for each of the 10 partitions, and AUCs across the 10 partitions were averaged. Cross-validation was repeated a total of 100 times, and average AUC across all runs was reported for each biomarker panel.

### Biomarker Panel Selection

Biomarker panels with more than 50 genes were eliminated from consideration, in the interest of potential future implementation in a clinical setting. We then selected the best biomarker panels using a two-step process. First, we identified the biomarker panels with cross-validation AUC >0.65 for discriminating between IE and NE. We tested these biomarker panels in the first Rapid Transition subcohort, which consisted of 62 AECOPD, 15 day 30 convalescent, 54 day 90 convalescent and 34 stable COPD subjects. We then identified biomarker panels with the best AUCs and most promising visual patterns. That is, we looked for patterns consistent with our hypothesis that a true marker of COPD disease activity should rise with upcoming AECOPD, peak during AECOPD onset, then fall during convalescence and remain low during periods of stability.

### Biomarker Panel Replication

The top biomarker panels were replicated in the second Rapid Transition subcohort, which consisted of 177 AECOPD, 20 day 30 convalescent, 82 day 90 convalescent and 57 stable COPD subjects. There was no overlap between these subjects and the subjects previously used to select the best biomarker panels. We evaluated the success of the replication using AUCs and visual assessment.

# Results

This study used 226 stable COPD subjects from the ECLIPSE cohort to derive gene co-expression modules, 20 IE versus 122 NE subjects from the ECLIPSE cohort (non-overlapping with the 226) to identify prognostic biomarkers of AECOPD, 112 subjects from the Rapid Transition cohort to select the most promising biomarker panels, and 293 (non-overlapping with the 112) additional subjects from the Rapid Transition cohort to replicate the top biomarker panels. The demographics of these study populations are shown in Table 1. Note that in the Rapid Transition cohort, 16 of 112 subjects and 59 of 293 subjects did not have a baseline AECOPD or stable COPD sample; we did not use these to calculate performance metrics, and they are not included in Table 1.

There was a significant difference (p<0.05) in FEV1, FEV1 %predicted, and lymphocytes when comparing IE and NE, with IE having worse lung function and lower lymphocytes as a percentage of total white blood cells. Within the subjects used to select biomarker panels, there were significant differences between AECOPD and stable COPD in bronchodilator use, inhaled corticosteroid use, white blood cells (overall and composition), and NT-proBNP. The AECOPD group saw lower inhaled corticosteroid use, higher white blood cells and neutrophils, and elevated NT-proBNP. The replication subjects also showed significant differences between AECOPD and stable COPD in bronchodilator use, inhaled corticosteroid use, white blood cells (overall and composition), and NT-proBNP, but also in ethnicity and smoking status. Again, the AECOPD subjects seemed to have lower inhaled corticosteroid use, higher white blood cells, and elevated NT-proBNP, and also had a higher proportion of current smokers.

Normalizing the Human Gene 1.1 ST microarray data and summarizing it at the transcript cluster level produced 33,397 probe sets. After we removed un-annotated probe sets (11,041) and transcripts assigned to multiple genes (2121), then summarized genes with multiple probe set mappings (676), the resulting expression data consisted of 19,245 unique genes. WGCNA identified 23 distinct modules of co-expressed genes in the 226 ECLIPSE subjects. The sizes of these modules and their biological annotation are included in Table 2.

For each of the gene modules identified, we carried out differential gene expression analysis in 20 IE versus 122 NE from the ECLIPSE cohort, and built classifier panels using elastic net regression. We obtained out-of-sample estimates for the AUC of each of the resulting classifiers via 10-fold cross-validation, repeated 100 times. The results of these biomarker discovery analyses, including which module the genes came from, number of differentially expressed genes, biomarker panel size, cross-validation AUC estimate, and 95% confidence interval (CI) for the AUC, are shown in Table 3. A total of six biomarker panels had cross-validation AUC >0.65 while consisting of fewer than 50 genes.

When these six biomarker panels were applied off-the-shelf to the first Rapid Transition subcohort, four of them showed a peak in predicted disease activity at the time of AECOPD followed by gradually decreasing levels at day 3, day 30, day 90, and in stable COPD subjects. Of these, three (derived from the salmon, green, and light-cyan WGCNA modules) demonstrated AUC >0.75 for discriminating AECOPD from stable COPD. The AUCs and patterns described here can be seen in Table 3 and Figure 4. Finally, the performance of these three selected biomarker panels was confirmed in the second Rapid Transition subcohort in AECOPD, day 30, day 90, and stable COPD samples, with AUCs ranging 0.74 – 0.84 between AECOPD and stable COPD (Table 3, Figure 5).

Biological characterization of these 3 modules revealed that, while entirely distinct at the gene level, there was significant functional overlap between the salmon and light-cyan modules, with both modules reflecting T-cell activation and differentiation, as well as recruitment (e.g. BTM modules: T cell activation (I) (M7.1) and T cell activation (III) (M7.4), enriched in T cells (I) (M7.0) and enriched in T cells (II) (M223), T cell differentiation (M14), T cell differentiation (Th2) (M19), T cell surface signature (S0), cell adhesion (GO) (M117), receptors, cell migration (M109), and IL2, IL7, TCR network (M65)). The salmon module was additionally enriched in NK cell-specific BTMs (e.g. BTM modules: enriched in NK cells (I) (M7.2), enriched in NK cells (III) (M157), and enriched in NK cells (receptor activation) (M61.2)). Comparing module membership to cell-specific gene signatures derived from peripheral whole blood(26), we again found the salmon and light-cyan modules to have very similar profiles: significantly enriched for genes characteristic of CD4+ and CD8+ T-cells, as well as Treg cells. Once again, the salmon module was significantly enriched in NK cell-specific genes, while the light-cyan module was not. While the green module did not correspond well to any BTM, it was significantly enriched in genes involved in endoplasmic reticulum stress and unfolded protein responses (MSigDB Hallmark Collection: Unfolded Protein Response).

# Discussion

Diagnosis of AECOPD is largely based on subjective assessments. AECOPD symptoms tend to be non-specific and overlap with those of many other co-morbidities, making correct diagnosis in an ED setting challenging. Even when correctly diagnosed, treatment is often not informed by the underlying etiology. Patients are often over- or under-treated, leading to significant morbidity and mortality [00]. Predicting and preventing AECOPD episodes is an important primary care goal(32), but, currently, no clinically useful score exists capable of predicting short term AECOPD risk (00). Thus, there is a pressing need for biomarkers to improve prevention and management of AECOPD.

The purpose of the current work was to identify COPD-specific co-expressed gene modules and derive modifiable biomarkers. Modifiable biomarkers track disease activity, and may be used to anticipate worsening disease and guide interventions when providing patient care. Hospitalization due to AECOPD is are major cost to the healthcare system and the ability to anticipate episodes of increased disease activity would result in significant savings to the healthcare system. We leveraged two large COPD patient cohorts and gene expression profiling in peripheral whole blood to attempt to identify such modifiable biomarkers. Starting from 23 COPD-specific co-expressed gene modules, each representing distinct functional biological units coordinately regulated in stable COPD, we constructed biomarker panels prognostic of imminent AECOPD, identified which of these signatures tracked with disease activity, and confirmed this behavior in an independent test set of >300 COPD patient.

We identified three suitable panels, derived from the salmon, light-cyan, and green WGCNA modules. These three panels had reasonable performance in cross-validation when predicting AECOPD ≤ 60 days (AUC = 0.71, 0.68, and 0.67 for the salmon, light-cyan, and green module derived panels, respectively), and good off-the-shelf classification performance in distinguishing AECOPD from stable COPD controls (AUC = 0.84, 0.77, and 0.74 for the salmon, light-cyan, and green module derived panels, respectively). Moreover, the temporal patterns of the biomarker scores obtained from these panels suggest that they may be used to monitor disease activity in COPD patients. These results are particularly interesting given that no additional tuning was carried out by us. The panels were trained on different patient cohorts using conceptually related, but different phenotypes (imminent, but non-symptomatic vs. active exacerbation), with gene expression profiles generated at different times (though batch-correction was applied) and achieved reasonable classification performance in this unseen dataset.

## Comparison with existing AECOPD prognostics

Several sociodemographic, physiological, psychological, and clinical factors are associated with a higher risk of hospitalizations in COPD. Multi-dimensional indices that capture these risk factors, such as St George´s Respiratory Questionnaire (SGRQ), the Clinical COPD Questionnaire (CCQ) (00), BODE (body mass index [BMI], airflow obstruction, dyspnea, and exercise capacity) (00), BODEX (BMI, airflow obstruction, dyspnea, and previous severe exacerbations) (00), ADO (age, dyspnea, and airflow obstruction) (00), DOSE (dyspnea, airflow obstruction, smoking status, and exacerbation frequency) (00), CODEX (comorbidity, obstruction, dyspnea, and previous severe exacerbations) (00), and SCOPEX (short-term [6-month] risk of COPD exacerbations) (00) scores have all been assessed for their ability to predict exacerbation in the medium-long term (1 year) (AUC = 0.65 – 0.69). Risk of exacerbation naturally varies throughout the course of the disease and disease GOLD stage itself has been shown to be moderately predictive of AECOPD (AUC = 0.69) during 1 year follow-up. Importantly, these multi-dimensional scores cannot be used to identify imminent AECOPD and guide intervention/treatment.

Modifiable biomarkers have also been investigated. Thomsen and others (00) found that elevated CRP, fibrinogen, and leukocyte counts (relative to baseline levels) were associated with frequent exacerbation (1.2-, 1.7-, and 3.7-fold for 1, 2, and 3 elevated markers, respectively). Addition of these modifiable inflammatory markers to a model including basic demographic and clinical factors significantly improved classification performance (C statistics = 0.73) when trying to identify frequent exacerbators. Eosinophilia has also been associated with an increased risk of severe exacerbations among individuals with COPD in the general population (1.76-fold) (00). Urokinase-type plasminogen activator receptor is elevated during AECOPD relative to baseline and can discriminate day 1 of AECOPD from day 7 (00). The ability of these modifiable biomarkers to identify imminent AECOPD is yet to be evaluated.

Short-term re-exacerbation risk may be predictable using various demographic and clinical factors (00), but in stable COPD, the task is more challenging. Here there have been a few promising studies leveraging telehealth monitoring. By monitoring breathing rate, oxygen saturation, or other respiratory signals, various groups were able to identify imminent AECOPD with relatively high accuracy (76.0% – 84.7%) (00). We note, however, that performance metrics were estimated from the training data, and are likely to be optimistic. Ultimately, because these measure respiratory symptoms, they may detect AECOPD too late for effective intervention to prevent hospitalization (1-5 days prior).

## Comparison with existing AECOPD diagnostics

Many studies have attempted to identify molecular biomarkers diagnostic of AECOPD. Refer to a recent review by our group for a good overview (13).

Bafadel et al. report on a number of potential single-molecule biomarkers for discriminating between AECOPD and stable COPD, and between bacterial\*, viral†, and eosinophilic‡ AECOPD(30). Their independent validation showed the following performance: serum CRP (AUC = 0.70\*), sputum IL-1β (AUC = 0.73\*), serum CXCL-10 (AUC = 0.65†), and serum eosinophil percent (AUC = 0.95‡). However, they found no single biomarker capable of discriminating between AECOPD and stable COPD with AUC greater than 0.70, whereas our biomarker panels achieve replication AUCs ranging 0.74 – 0.84 for the same comparison.

Hurst et al. found that CRP had the best performance (AUC = 0.73) out of 36 plasma biomarkers, for discriminating between AECOPD and stable COPD(15). Combining CRP with other biomarkers did not significantly increase performance, but combining CRP with an observation of at least one major clinical symptom improved performance greatly (AUC = 0.88, 95% CI 0.82 – 0.93). Given that their definition of exacerbation required at least one major symptom, this is not entirely unexpected. One of our panels (derived from the salmon module, AUC = 0.84, 95% CI 0.78 – 0.90) has comparable performance, while our other two panels perform significantly below this level. However, we believe that a diagnostic which does not require assessment of symptoms by a clinician, has different utility and potential applications. Moreover, CRP alone does not track with AECOPD severity or outcome(31).

* Lacoma et al. report on CRP (AUC = 0.53), procalcitonin (AUC = 0.52), and neopterin (AUC = 0.61) for discriminating between bacterial AECOPD and stable COPD in a large clinical cohort (217 AECOPD; 46 stable COPD). In all cases, including a clinical symptom improved performance of these single-molecule biomarkers: CRP (AUC = 0.68), procalcitonin (AUC = 0.66), and neopterin (AUC = 0.70).

Copeptin on the other hand has been reported as associated with disease severity and outcomes in COPD and may be more specific, at least relative to heart failure.

## Limitations

Given the emergency setting, patients were often treated before consent could be obtained and blood drawn. In the Rapid Transition cohort, which was used for biomarker selection and replication, most of AECOPD time points had received prednisone prior to blood draw. Prednisone has an important effect on whole blood gene expression (00). This is a significant confounding effect and it may be that many or all panels derived from this dataset are discriminating between prednisone status (prednisone +ve vs. prednisone -ve), instead of disease activity (AECOPD vs. convalescent). This is unlikely, however, given COPD gene module identification and biomarker discovery were both carried out in a well-controlled, stable COPD population (ECLIPSE). Nevertheless, work is ongoing to address this possibility.

More importantly, the performance of all 3 signatures needs to be assessed prospectively, preferably in a longitudinal cohort to determine clinical utility. The ability of the signatures to identify imminent AECOPD in particular needs to be confirmed in independent samples. Specificity of the signatures also needs to be assessed. Gene expression signatures derived from peripheral blood gene expression are likely to reflect immune processes and may not be very specific. In an ED setting for example, AECOPD symptoms tend to be non-specific and overlap with those of myocardial infarction, pulmonary embolism, and congestive heart failure. We would like to be able to correctly identify them as non-AECOPD, but we have not yet assessed whether the signatures are able to do so.

## Conclusion

We identified 23 distinct COPD-specific co-expressed gene modules. The genes of each module were used to carry out biomarker discovery to identify biomarker panels of disease activity, and panels with good performance in cross-validation were back-tested in an off-the-shelf manner, to the diagnostic or prognostic question, respectively. Three biomarker panels derived from the salmon, green, and light-cyan co-expressed COPD gene modules, exhibited the desired characteristics and may represent modifiable biomarkers of disease activity. Annotation of these modules provides biological plausibility. The green module was enriched in genes related to the unfolded protein response and endoplasmic reticulum stress, a hallmark of viral infection(33), while the salmon and light-cyan modules reflected T-cell recruitment, activation, and differentiation. The salmon module was additionally enriched in NK cell-specific genes. Taken together, modules that yielded biomarker panels of disease activity appear consistent with response to viral infection, which are present in 22-64% of AECOPD(34) and have been causally linked to triggering AECOPD(35). These three panels should now be evaluated prospectively in a longitudinal cohort to confirm their ability to track (and anticipate) disease activity and meaningfully impact the care of COPD patients.

# Acknowledgments

Research coordinators, study nurses, study participants, etc.

# References

1. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, Shibuya K, Salomon JA, Abdalla S, Aboyans V, Abraham J, Ackerman I, Aggarwal R, Ahn SY, Ali MK, AlMazroa MA, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Bahalim AN, Barker-Collo S, Barrero LH, Bartels DH, Basáñez M-G, Baxter A, Bell ML, *et al.* Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 2012;380:2163–2196.

2. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Lond Engl* 2015;385:117–171.

3. Anzueto A, Sethi S, Martinez FJ. Exacerbations of Chronic Obstructive Pulmonary Disease. *Proc Am Thorac Soc* 2007;4:554–564.

4. Anzueto A. Impact of exacerbations on COPD. *Eur Respir Rev* 2010;19:113–118.

5. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van Weel C, Zielinski J. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2007;176:532–555.

6. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *The Lancet* 2007;370:765–773.

7. López-Campos JL, Tan W, Soriano JB. Global burden of COPD: Global burden of COPD. *Respirology* 2016;21:14–23.

8. Koutsokera A, Stolz D, Loukides S, Kostikas K. Systemic biomarkers in exacerbations of copd: The evolving clinical challenge. *Chest* 2012;141:396–405.

9. Gompertz S, Bayley DL, Hill SL, Stockley RA. Relationship between airway inflammation and the frequency of exacerbations in patients with smoking related COPD. *Thorax* 2001;56:36–41.

10. Eagan TML, Ueland T, Wagner PD, Hardie JA, Mollnes TE, Damås JK, Aukrust P, Bakke PS. Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. *Eur Respir J* 2010;35:540–548.

11. Eagan TM, Damås JK, Ueland T, Voll-Aanerud M, Mollnes TE, Hardie JA, Bakke PS, Aukrust P. Neutrophil gelatinase-associated lipocalin: a biomarker in COPD. *Chest* 2010;138:888–895.

12. Donaldson GC, Seemungal TAR, Patel IS, Bhowmik A, Wilkinson TMA, Hurst JR, Maccallum PK, Wedzicha JA. Airway and systemic inflammation and decline in lung function in patients with COPD. *Chest* 2005;128:1995–2004.

13. Chen Y-WR, Leung JM, Sin DD. A Systematic Review of Diagnostic Biomarkers of COPD Exacerbation. *PLOS ONE* 2016;11:e0158843.

14. Chen H, Wang Y, Bai C, Wang X. Alterations of plasma inflammatory biomarkers in the healthy and chronic obstructive pulmonary disease patients with or without acute exacerbation. *J Proteomics* 2012;75:2835–2843.

15. Hurst JR, Donaldson GC, Perera WR, Wilkinson TMA, Bilello JA, Hagan GW, Vessey RS, Wedzicha JA. Use of Plasma Biomarkers at Exacerbation of Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2006;174:867–874.

16. Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L. Evaluation of COPD longitudinally to identify predictive surrogate end-points (eclipse). *Eur Respir J* 2008;31:.

17. R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. at <http://www.R-project.org>.

18. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 2015;12:115–121.

19. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinforma Oxf Engl* 2010;26:2363–2367.

20. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat Oxf Engl* 2003;4:249–264.

21. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE* 2011;6:.

22. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;4:.

23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;102:15545–15550.

24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;27:1739–1740.

25. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol* 2013;15:.

26. Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT r... - PubMed - NCBI. at <https://www.ncbi.nlm.nih.gov/pubmed/20410506>.

27. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol* 1995;57:289–300.

28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47.

29. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* 2005;67:301–320.

30. Bafadhel M, McKenna S, Terry S, Mistry V, Reid C, Haldar P, McCormick M, Haldar K, Kebadze T, Duvoix A, Lindblad K, Patel H, Rugman P, Dodson P, Jenkins M, Saunders M, Newbold P, Green RH, Venge P, Lomas DA, Barer MR, Johnston SL, Pavord ID, Brightling CE. Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers. *Am J Respir Crit Care Med* 2011;184:662–671.

31. Antonescu-Turcu AL, Tomic R. C-reactive protein and copeptin: prognostic predictors in chronic obstructive pulmonary disease exacerbations. *Curr Opin Pulm Med* 2009;15:120–125.

32. Curtis JL, Martinez CH. Preventing COPD exacerbations: new options for a crucial and growing problem. *Fed Pract Health Care Prof VA DoD PHS* 2014;31:18S–24S.

33. He B. Viruses, endoplasmic reticulum stress, and interferon responses. *Cell Death Differ* 2006;13:393–403.

34. Hewitt R, Farne H, Ritchie A, Luke E, Johnston SL, Mallia P. The role of viral infections in exacerbations of chronic obstructive pulmonary disease and asthma. *Ther Adv Respir Dis* 2016;10:158–174.

35. Hutchinson AF, Ghimire AK, Thompson MA, Black JF, Brand CA, Lowe AJ, Smallwood DM, Vlahos R, Bozinovski S, Brown GV, Anderson GP, Irving LB. A community-based, time-matched, case-control study of respiratory viruses and exacerbations of COPD. *Respir Med* 2007;101:2472–2481.

# Figure Legends

**Figure 1: Descriptive title statement.** (A) Schematic representation of cohorts and subjects used for this study, in relation to COPD disease activity. (B) Flowchart of analyses performed for this study.

**Figure 2: Descriptive title statement.** Performance of the top 6 prognostic biomarker signatures, when applied to independent samples from AECOPD, convalescing, and stable COPD subjects.

**Figure 3: Descriptive title statement.** Replication performance of the top 3 signatures of AECOPD activity when applied to independent samples from AECOPD, convalescing, and stable COPD subjects.

**Figure 4: Descriptive title statement.**

(A) Pathway over-representation of BTM gene sets in the selected COPD modules. We compared the module membership of selected WGCNA modules to the Blood Transcriptome Modules (BTM), a collection of annotated gene sets derived from a large-scale co-expression clustering analysis of >50,000 publicly available blood transcriptome profiling datasets, to better understand their biological function. Our COPD modules are completely distinct at the gene level (by design), but there is significant overlap in function between the light-cyan and salmon modules. (B) Cell-specificity of the selected WGCNA modules. We compared the module membership of selected WGCNA modules to cell-specific gene signatures derived from peripheral whole blood(26). Their relative enrichment across these cell-specific gene signatures is visualized in a radar plot. The salmon and light-cyan modules have very similar profiles, being significantly enriched for genes characteristic of CD4+ and CD8+ T-cells, as well as Treg cells. The salmon module was significantly enriched in NK cell-specific genes, while the light-cyan module was not. The green module did not exhibit any cell-specificity.

# Tables

**Table 1.** Study population demographics.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **ECLIPSE** | | | | **Rapid Transition** | | | | | | |
| **Co-expression** | **Biomarker discovery** | | | **Biomarker selection** | | | **Biomarker replication** | | | **P** |
| **WGCNA**  **(N=226)** | **IE**  **(N=20)** | **NE**  **(N=122)** | **P** | **AECOPD**  **(N=62)** | **Stable**  **(N=34)** | **P** | **AECOPD**  **(N=177)** | **Stable**  **(N=57)** | **P** |
| Age | 64.2 ± 6.2 | 63.7 ± 6.8 | 64.1 ± 6.1 | 0.782 | 68.3 ± 11.5 | 63.8 ± 11.6 | 0.070 | 66.5 ± 12.0 | 65.6 ± 13.2 | 0.626 | 0.335 |
| Male (%) | 63.7% | 75.0% | 70.5% | 0.794 | 56.5% | 76.5% | 0.075 | 63.3% | 69.6% | 0.425 | 0.207 |
| Caucasian (%) |  |  |  |  | 85.5% | 85.3% | 1.000 | 84.5% | 96.4% | 0.019 | 0.094 |
| BMI | 28.2 ± 6.0 | 24.5 ± 5.3 | 26.8 ± 6.1 | 0.103 | 22.8 ± 12.5 | 25.5 ± 5.7 | 0.250 | 24.9 ± 8.6 | 26.7 ± 7.3 | 0.208 | 0.214 |
| FEV1 (L) | 1.38 ± 0.54 | 1.30 ± 0.49 | 1.55 ± 0.51 | 0.045 | 1.54 ± 0.74 | 1.76 ± 0.88 | 0.334 | 1.47 ± 0.69 | 1.54 ± 0.84 | 0.589 | 0.350 |
| FEV1 %predicted | 50.0 ± 16.1 | 45.4 ± 17.2 | 53.9 ± 15.0 | 0.024 | 55.6 ± 20.9 | 55.7 ± 24.2 | 0.993 | 51.1 ± 23.1 | 53.8 ± 24.0 | 0.538 | 0.743 |
| FVC (L) | 3.00 ± 0.89 | 3.11 ± 1.07 | 3.34 ± 0.91 | 0.316 | 2.92 ± 1.10 | 3.45 ± 1.28 | 0.133 | 2.79 ± 1.02 | 2.88 ± 0.98 | 0.620 | 0.039 |
| FVC %predicted |  |  |  |  | 82.3 ± 20.6 | 85.1 ± 22.3 | 0.644 | 75.3 ± 23.9 | 79.6 ± 21.1 | 0.324 | 0.192 |
| FEV1/FVC ratio | 46.4 ± 12.4 | 42.7 ± 12.0 | 46.8 ± 10.9 | 0.130 | 52.5 ± 14.3 | 50.6 ± 14.5 | 0.650 | 53.3 ± 16.0 | 52.7 ± 19.7 | 0.847 | 0.903 |
| GOLD code |  |  |  | 0.600 |  |  | 0.414 |  |  | 0.435 | 0.634 |
| 1 |  |  |  |  | 23.8% | 22.6% |  | 13.0% | 20.0% |  |  |
| 2 | 44.7% | 45.0% | 56.6% |  | 23.8% | 35.5% |  | 37.7% | 24.4% |  |  |
| 3 | 44.2% | 45.0% | 36.1% |  | 42.9% | 22.6% |  | 29.9% | 35.6% |  |  |
| 4 | 11.1% | 10.0% | 7.4% |  | 9.5% | 19.4% |  | 19.5% | 20.0% |  |  |
| Unknown |  |  |  |  | 66.1% | 8.8% |  | 56.5% | 21.1% |  |  |
| Bronchodilator (%) | 71.2% | 60.0% | 60.7% | 1.000 | 98.4% | 84.8% | 0.018 | 98.9% | 92.5% | 0.027 | <0.001 |
| Inhaled steroids (%) | 77.9% | 70.0% | 59.0% | 0.461 | 41.9% | 72.7% | 0.005 | 39.2% | 83.0% | <0.001 | <0.001 |
| Smoking status |  |  |  | 0.470 |  |  | 0.754 |  |  | 0.004 | 0.037 |
| Never |  |  |  |  | 6.5% | 8.8% |  | 7.3% | 5.4% |  |  |
| Former | 96.5% | 40.0% | 51.6% |  | 41.9% | 47.1% |  | 30.5% | 55.4% |  |  |
| Current | 3.5% | 60.0% | 48.4% |  | 51.6% | 44.1% |  | 62.1% | 39.3% |  |  |
| Pack years | 45.8 ± 26.3 | 38.2 ± 19.3 | 49.6 ± 27.0 | 0.073 | 40.4 ± 33.5 | 31.1 ± 34.8 | 0.278 | 35.4 ± 33.9 | 23.8 ± 28.6 | 0.182 | 0.316 |
| White blood cells | 7.68 ± 2.04 | 8.01 ± 3.72 | 7.52 ± 1.87 | 0.370 | 8.65 ± 2.98 | 7.12 ± 2.68 | 0.023 | 10.3 ± 4.7 | 7.16 ± 2.21 | <0.001 | <0.001 |
| Neutrophil (%) | 64.3 ± 8.6 | 67.7 ± 10.6 | 62.7 ± 9.8 | 0.107 | 78.0 ± 13.4 | 63.8 ± 9.1 | <0.001 | 80.5 ± 13.3 | 64.7 ± 12.8 | <0.001 | <0.001 |
| Monocyte (%) | 6.58 ± 2.19 | 6.42 ± 2.49 | 6.08 ± 2.19 | 0.348 | 6.59 ± 3.75 | 7.93 ± 1.95 | 0.038 | 5.09 ± 3.18 | 5.59 ± 2.46 | 0.092 | <0.001 |
| Eosinophil (%) | 3.13 ± 2.05 | 2.87 ± 2.00 | 2.67 ± 1.68 | 0.708 | 1.06 ± 1.48 | 3.08 ± 2.44 | <0.001 | 0.67 ± 1.34 | 2.84 ± 1.97 | <0.001 | <0.001 |
| Basophil (%) | 0.34 ± 0.23 | 0.33 ± 0.26 | 0.34 ± 0.17 | 0.276 | 0.59 ± 0.52 | 0.89 ± 0.42 | 0.001 | 2.23 ± 8.59 | 1.50 ± 6.77 | 0.004 | <0.001 |
| Lymphocyte (%) | 25.7 ± 7.4 | 22.7 ± 8.6 | 28.2 ± 9.2 | 0.035 | 13.8 ± 9.9 | 24.3 ± 7.9 | <0.001 | 11.1 ± 7.8 | 24.1 ± 10.3 | <0.001 | <0.001 |
| NT-proBNP (IQR) | 109 [37, 225] | 107 [61, 300] | 133 [59, 244] | 0.903 | 273 [115, 786] | 132 [80, 332] | 0.017 | 468 [216, 1340] | 63 [14, 155] | <0.001 | <0.001 |

Patient population characteristics are reported as mean ± standard deviation or percentages. For continuous variables, analysis of variance was used to calculate p-values for group differences. For categorical variables, Fisher’s exact test was used to calculate p-values for group differences.

**Table 2.** WGCNA modules and their biological significance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **# of genes** | **Blood Transcriptome Module** | **FDR** |
| black | 324 | enriched in cell cycle (M167) | 6.95E-05 |
| blue | 1416 | translation initiation (M227) | 3.20E-06 |
| brown | 578 | enriched in monocytes (II) (M11.0) | 3.95E-20 |
| cyan | 169 | enriched in NK cells (I) (M7.2) | 6.97E-36 |
| darkgreen | 97 | platelet activation - actin binding (M196) | 3.30E-25 |
| darkred | 109 | enriched in B cells (I) (M47.0) | 4.01E-90 |
| darkturquoise | 56 | N/A | N/A |
| **green** | **421** | **mitochondrial cluster (M235)** | **1.07E-03** |
| greenyellow | 254 | regulation of transcription, transcription factors (M213) | 1.09E-03 |
| grey60 | 148 | immune activation - generic cluster (M37.0) | 1.57E-08 |
| **light-cyan** | **149** | **T cell differentiation (M14)** | **3.45E-17** |
| light-green | 145 | enriched in neutrophils (I) (M37.1) | 1.01E-36 |
| light-yellow | 142 | enriched in monocytes (IV) (M118.0) | 4.71E-05 |
| magenta | 264 | intracellular transport (M147) | 1.02E-02 |
| midnightblue | 157 | translation initiation factor 3 complex (M245) | 2.40E-13 |
| pink | 297 | enriched in monocytes (II) (M11.0) | 9.79E-16 |
| purple | 260 | N/A | N/A |
| red | 397 | heme biosynthesis (I) (M171) | 7.03E-17 |
| royalblue | 109 | type I interferon response (M127) | 2.32E-21 |
| **salmon** | **192** | **T cell activation (I) (M7.1)** | **3.58E-30** |
| tan | 218 | cell cycle (I) (M4.1) | 1.73E-10 |
| turquoise | 1958 | mismatch repair (I) (M22.0) | 1.96E-03 |
| yellow | 434 | nuclear pore, transport; mRNA splicing, processing (M143) | 5.03E-05 |

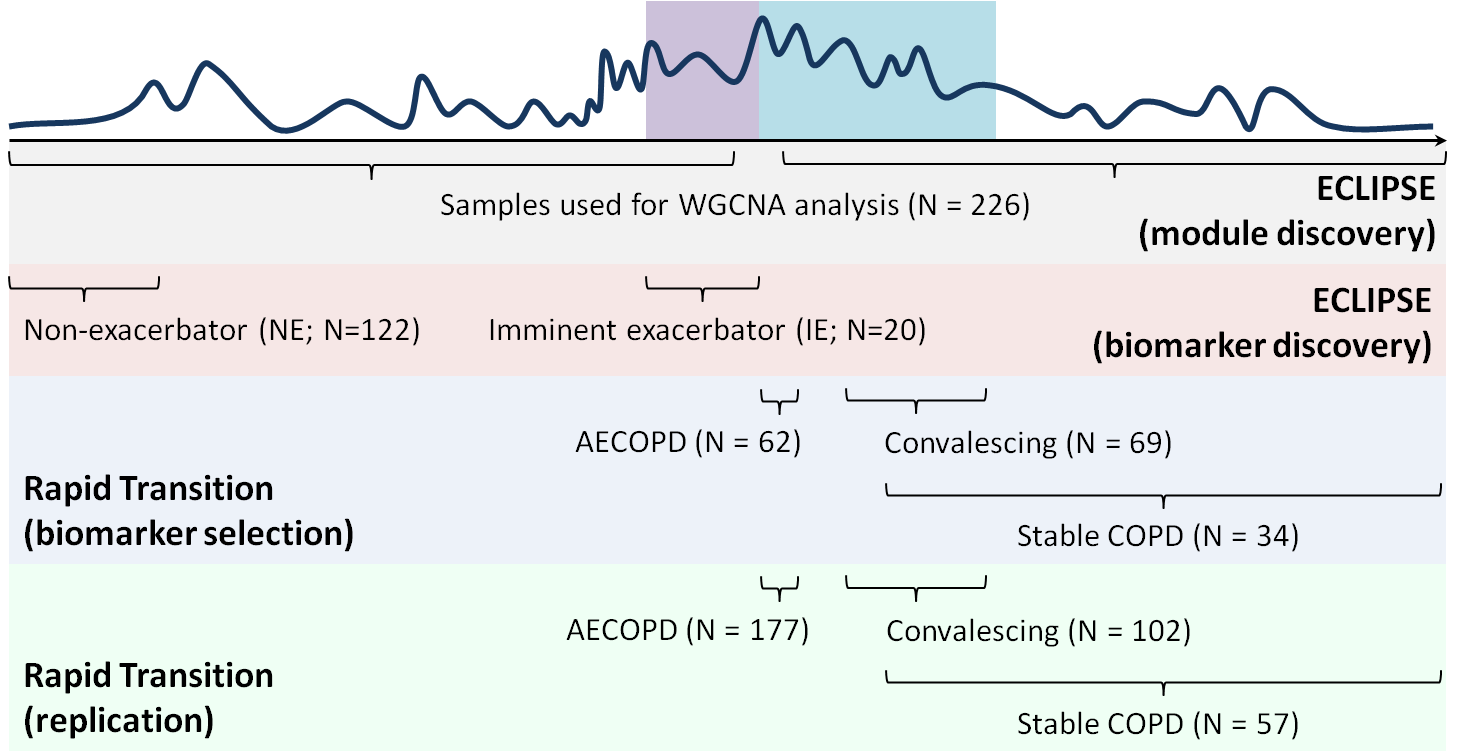
**Table 3.** Module-based biomarker discovery, biomarker selection, and biomarker replication results, with 95% confidence intervals for the AUC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Module** | **# genes with p<0.05** | **Panel size** | **Biomarker discovery AUC\*** | **Biomarker selection AUC\*** | **Biomarker replication AUC\*** |
| All transcripts | 709 | 84 | 0.63 [0.57, 0.69] | 0.70 [0.59, 0.81] | 0.74 [0.68, 0.81] |
| black | 16 | 12 | 0.65 [0.57, 0.71] |  |  |
| blue | 304 | 96 | 0.58 [0.52, 0.63] |  |  |
| brown | 180 | 72 | 0.68 [0.62, 0.75] |  |  |
| cyan | 3 | 3 | 0.58 [0.51, 0.65] |  |  |
| darkgreen | 28 | 20 | 0.63 [0.57, 0.68] |  |  |
| darkred | 3 | 3 | 0.48 [0.38, 0.58] |  |  |
| darkturquoise | 2 | 2 | 0.57 [0.49, 0.64] |  |  |
| **green** | **97** | **50** | **0.67 [0.60, 0.74]** | **0.77 [0.67, 0.86]** | **0.74 [0.67, 0.81]** |
| greenyellow | 11 | 9 | 0.64 [0.56, 0.71] | 0.31 [0.20, 0.43] |  |
| grey60 | 7 | 5 | 0.62 [0.53, 0.69] |  |  |
| **light-cyan** | **22** | **16** | **0.68 [0.63, 0.74]** | **0.76 [0.67, 0.86]** | **0.77 [0.71, 0.83]** |
| light-green | 15 | 10 | 0.64 [0.58, 0.68] |  |  |
| light-yellow | 5 | 5 | 0.43 [0.35, 0.51] |  |  |
| magenta | 38 | 28 | 0.64 [0.57, 0.70] |  |  |
| midnightblue | 14 | 7 | 0.57 [0.51, 0.63] |  |  |
| pink | 49 | 28 | 0.67 [0.61, 0.73] | 0.39 [0.27, 0.51] |  |
| purple | 12 | 10 | 0.72 [0.63, 0.79] | 0.74 [0.63, 0.83] |  |
| red | 2 | 2 | 0.44 [0.33, 0.56] |  |  |
| royalblue | 11 | 5 | 0.52 [0.43, 0.59] |  |  |
| **salmon** | **53** | **35** | **0.71 [0.65, 0.76]** | **0.82 [0.73, 0.90]** | **0.84 [0.78, 0.90]** |
| tan | 32 | 24 | 0.58 [0.53, 0.65] |  |  |
| turquoise | 526 | 99 | 0.71 [0.66, 0.76] |  |  |
| yellow | 155 | 70 | 0.59 [0.53, 0.67] |  |  |

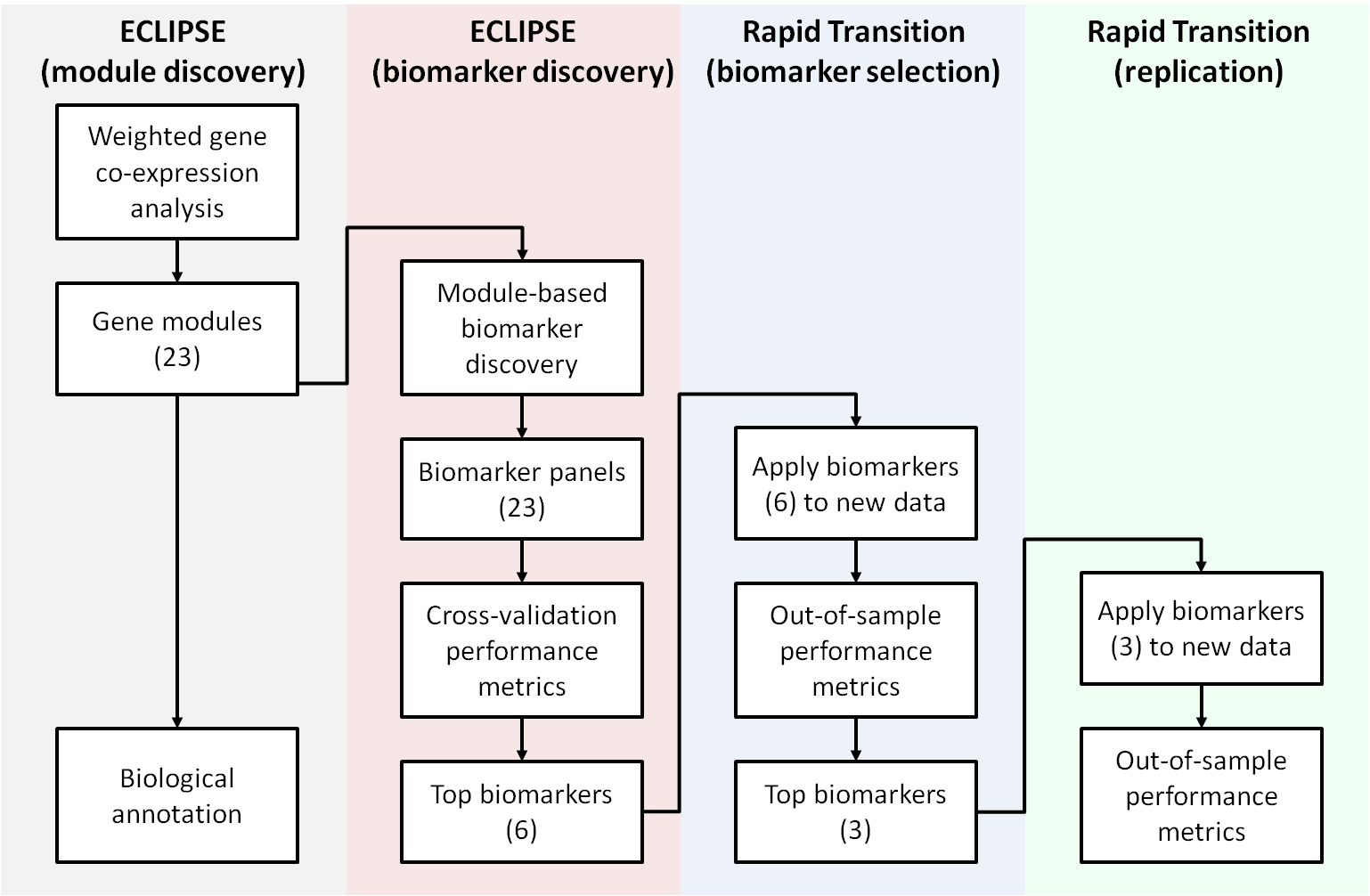
\*The 95% confidence interval for the AUC was calculated empirically for the biomarker discovery, using cross-validation performance. The 95% confidence intervals for biomarker selection and replication were calculated by bootstrapping the out-of-sample probabilities with 1000 iterations per panel.

# Figures

**Figure 1A.** Schematic representation of cohorts and subjects used for this study, in relation to COPD disease activity.



**Figure 1B.** Flowchart of analyses performed for this study.

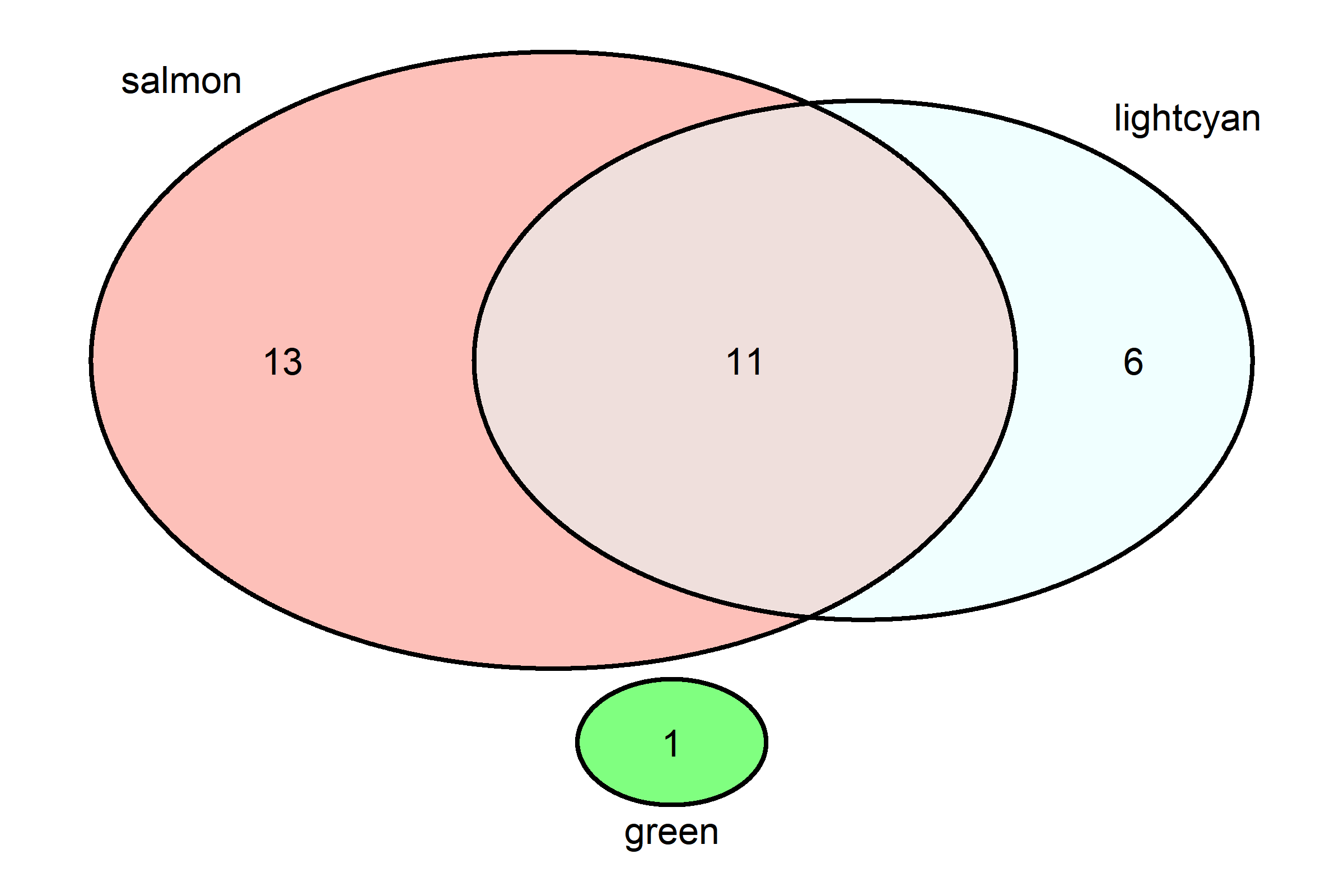


**Figure 2.** Performance of the top 6 prognostic biomarker signatures, when applied to independent samples from AECOPD, convalescing, and stable COPD subjects.

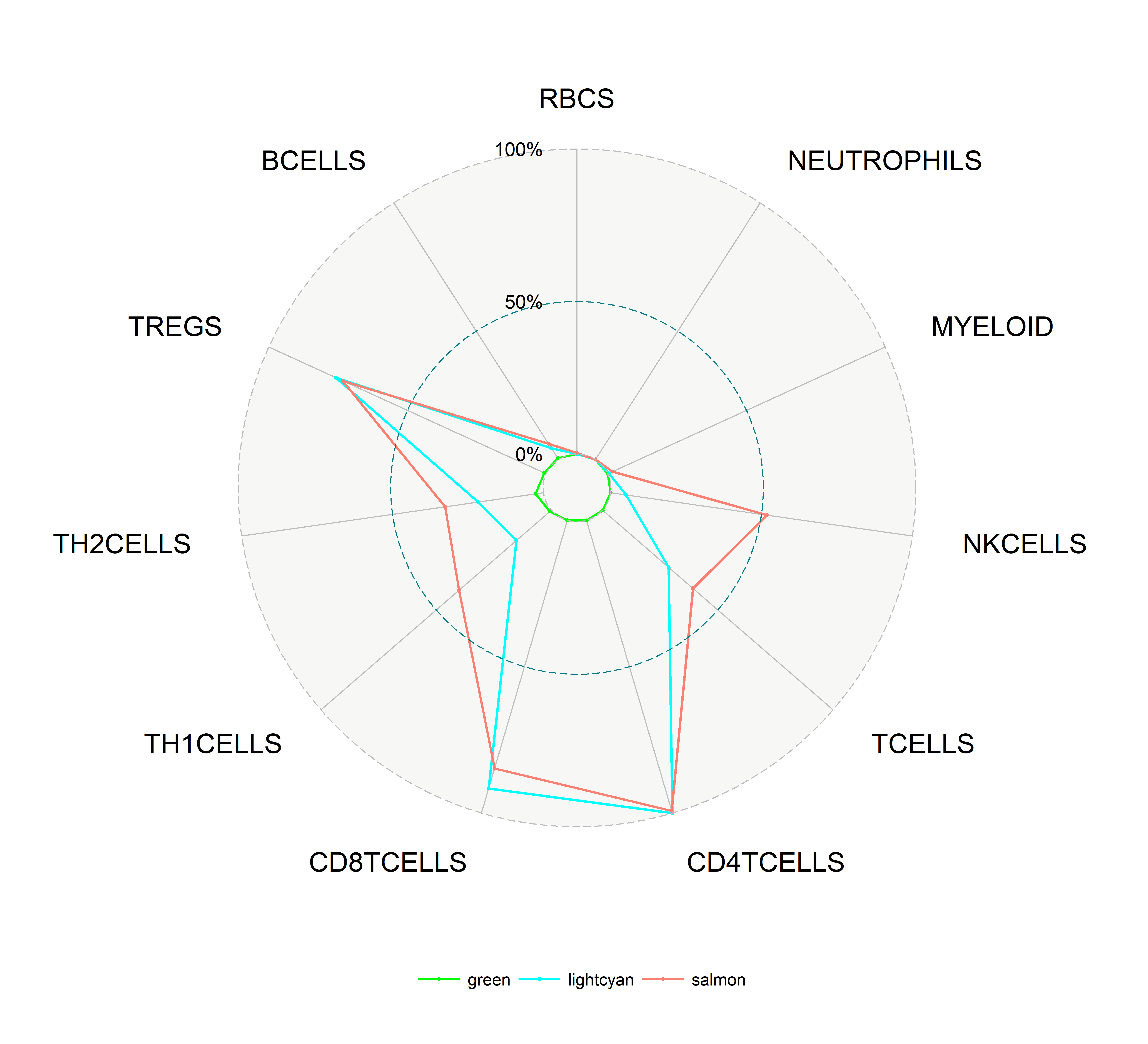


**Figure 3.** Replication performance of the top 3 signatures of AECOPD activity when applied to independent samples from AECOPD, convalescing, and stable COPD subjects.

**Figure 4A.** **Pathway over-representation of BTM gene sets in the selected COPD modules.** We compared the module membership of selected WGCNA modules to the Blood Transcriptome Modules (BTM), a collection of annotated gene sets derived from a large-scale co-expression clustering analysis of >50,000 publicly available blood transcriptome profiling datasets, to better understand their biological function. Our COPD modules are completely distinct at the gene level (by design), but there is significant overlap in function between the light-cyan and salmon modules.

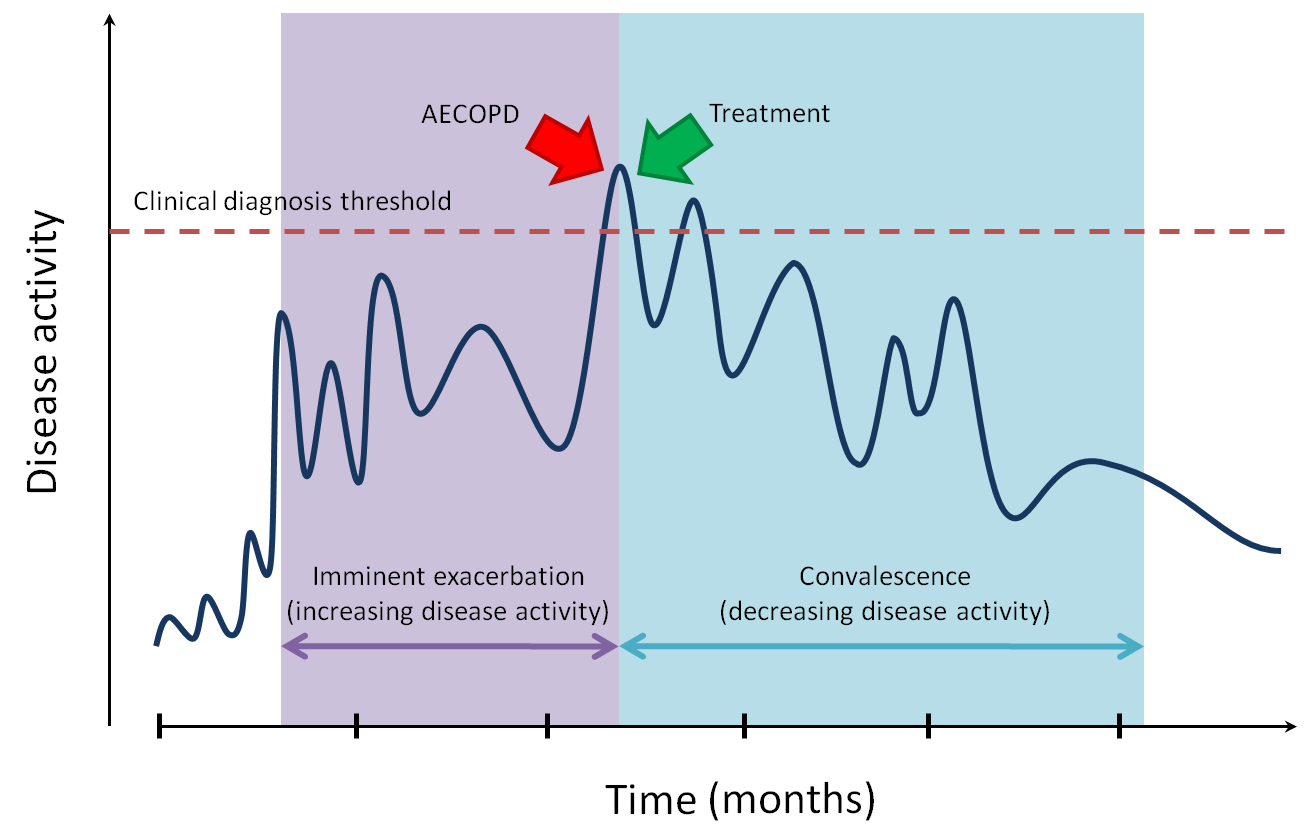


**Figure 4B.** **Cell-specificity of the selected WGCNA modules.** We compared the module membership of selected WGCNA modules to cell-specific gene signatures derived from peripheral whole blood(26). Their relative enrichment across these cell-specific gene signatures is visualized in a radar plot. The salmon and light-cyan modules have very similar profiles, being significantly enriched for genes characteristic of CD4+ and CD8+ T-cells, as well as Treg cells. The salmon module was significantly enriched in NK cell-specific genes, while the light-cyan module was not. The green module did not exhibit any cell-specificity.

****

# Supplementary Figures

**Figure S1.** Disease activity model of COPD exacerbation.



**Figure S2.** **Pathway over-representation of BTM gene sets in the COPD modules.** We compared our COPD-specific modules to the Blood Transcriptome Modules (BTM), a collection of annotated gene sets derived from a large-scale co-expression clustering analysis of >50,000 publicly available blood transcriptome profiling datasets, to better understand their biological function. Our COPD modules are completely distinct at the gene level (by design), and there is generally little overlap in significantly over-represented BTMs across modules, suggesting that they are also largely distinct at a functional level. Notable exceptions include the black and brown (shown in blue), and light-cyan and salmon modules (shown in red).

