ORIGINAL ARTICLE

Novel Blood-based Transcriptional Biomarker Panels Predict the Late-Phase Asthmatic Response

Amrit Singh[1,2,3], Casey P. Shannon[2], Young Woong Kim[1,2], Chen Xi Yang[1,2], Robert Balshaw[4], Gabriela V. Cohen Freue[5], Gail M. Gauvreau[6], J. Mark FitzGerald[7,8], Louis-Philippe Boulet[9], Paul M. O'Byrne[6], and Scott J. Tebbutt[1,2,8]

[1]Centre for Heart Lung Innovation, St. Paul's Hospital, [3]Department of Pathology and Laboratory Medicine, [5]Department of Statistics, and [8]Division of Respiratory Medicine, Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada; [2]Prevention of Organ Failure Centre of Excellence, Vancouver, British Columbia, Canada; [4]Centre for Healthcare Innovation, University of Manitoba, Winnipeg, Manitoba, Canada; [6]Department of Medicine, McMaster University, Hamilton, Ontario, Canada; [7]Vancouver Coastal Health Research Institute, Vancouver General Hospital, Vancouver, British Columbia, Canada; and [9]Quebec Heart and Lung Institute, Laval University, Quebec City, Quebec, Canada

ORCID IDs: 0000-0002-7475-1646 (A.S.); 0000-0002-5687-3156 (C.P.S.); 0000-0002-3470-6728 (Y.W.K.); 0000-0002-8033-4769 (C.X.Y.); 0000-0002-2455-8792 (R.B.); 0000-0003-4526-0175 (G.V.C.F.); 0000-0002-6187-2385 (G.M.G.); 0000-0002-5367-5226 (J.M.F.); 0000-0003-3485-9393 (L.-P.B.); 0000-0003-0979-281X (P.M.O'B.); 0000-0002-7908-1581 (S.J.T.).

Abstract

Rationale: The allergen inhalation challenge is used in clinical trials to test the efficacy of new treatments in attenuating the late-phase asthmatic response (LAR) and associated airway inflammation in subjects with allergic asthma. However, not all subjects with allergic asthma develop the LAR after allergen inhalation. Blood-based transcriptional biomarkers that can identify such individuals may help in subject recruitment for clinical trials as well as provide novel molecular insights.

Objectives: To identify blood-based transcriptional biomarker panels that can predict an individual's response to allergen inhalation challenge.

Methods: We applied RNA sequencing to total RNA from whole blood ($n = 36$) collected before and after allergen challenge and generated both *genome-guided* and *de novo* datasets: genes, gene-isoforms (University of California, Santa Cruz, UCSC Genome Browser), Ensembl, and Trinity. Candidate biomarker panels were validated using the NanoString platform in an independent cohort of 33 subjects.

Measurements and Main Results: The Trinity biomarker panel consisting of known and novel biomarker transcripts had an area under the receiver operating characteristic curve of greater than 0.70 in both the discovery and validation cohorts. The Trinity biomarker panel was useful in predicting the response of subjects that elicited different responses (accuracy between 0.65 and 0.71) and subjects that elicit a dual response (accuracy between 0.70 and 0.75) upon repeated allergen inhalation challenges.

Conclusions: Interestingly, the biomarker panel containing novel transcripts successfully validated compared with panels with known, well-characterized genes. These biomarker–blood tests may be used to identify subjects with asthma who develop the LAR, and may also represent members of novel molecular mechanisms that can be targeted for therapy.

Keywords: gene expression profiling; biomarkers; asthma

(Received in original form January 13, 2017; accepted in final form October 30, 2017)

Supported by funding from AllerGen NCE Inc. (Allergy, Genes, and Environment Network), Prevention of Organ Failure Centre of Excellence, the British Columbia Lung Association, and Mitacs (Mitacs-Accelerate Program); A.S. is the recipient of the Canadian Institutes of Health Research Doctoral Award–Frederick Banting and Charles Best Canada Graduate Scholarship; Y.W.K. and C.X.Y. are recipients of Mitacs Accelerate Studentships.

Author Contributions: A.S., G.M.G., and S.J.T. designed the study; G.M.G., P.M.O'B., J.M.F., L.-P.B., and S.J.T. participated in provision of samples; C.P.S. and C.X.Y. helped generate the STAR RNA-sequencing dataset; C.P.S. helped annotate the Trinity RNA-sequencing contigs; sample preparation and NanoString nCounter assays were performed by A.S., Y.W.K., and S.J.T.; R.B. and G.V.C.F. provided statistical support; A.S. performed the statistical and computational analyses and wrote the manuscript; all authors contributed to the editing of the final manuscript.

Correspondence and requests for reprints should be addressed to Scott J. Tebbutt, Ph.D., Department of Medicine, University of British Columbia Centre for Heart Lung Innovation, Room 166-1081 Burrard Street, Vancouver, BC, V6Z1Y6 Canada. E-mail: scott.tebbutt@hli.ubc.ca.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Am J Respir Crit Care Med Vol 197, Iss 4, pp 450–462, Feb 15, 2018

Originally Published in Press as DOI: 10.1164/rccm.201701-0110OC on October 31, 2017
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Attenuation of the allergen-induced late-phase asthmatic response (LAR) is used to determine the efficacy of novel asthma therapies. Recruitment, for clinical trials, of subjects who develop the late response is performed using an allergen inhalation challenge. However, only a subgroup of individuals with mild asthma develop the LAR.

What This Study Adds to the Field: This study identified and validated novel RNA transcripts that are predictive of the LAR. These biomarkers may be used to enrich clinical trials with subjects who are susceptible to developing the LAR.

The allergen-induced late-phase asthmatic response (LAR), occurs reproducibly in a subgroup of individuals with mild allergic asthma after allergen inhalation challenge (1), and shares many characteristics with chronic asthma, such as prolonged airway contraction, persistent airway inflammation, mucus hypersecretion, and airway remodeling (2, 3). Although inhaled corticosteroids are effective in abolishing the LAR (4), they do not alter the natural course of the disease (5). Airway inflammation (e.g., presence of epithelial shedding, T cells, and eosinophils in mucosal biopsies) is already present in subjects with mild disease, and worsens in subjects with persistent asthma (6). Given the heterogeneity of the disease within an individual and over time, reproducible biomarkers that can help risk stratify individuals may help improve diagnosis, selection of therapy, and prevention of overtreatment (7–11).

The purpose of this study was to identify transcriptional biomarker panels (combinations of single molecules) using whole blood that could predict a subject's response to allergen inhalation challenge, before and 2 hours after challenge (Figure 1). Following the Institute of Medicine guidelines for the development of clinical biomarker tests (12), biomarker panels were identified in a discovery cohort with rigorous clinical phenotypic characterization. The biomarker panels were developed using RNA-sequencing (RNA-Seq) data, and the performance of the computational model was assessed using cross-validation. The identified biomarker candidates were then transferred to the more clinically relevant NanoString platform and the biomarker panel formulas were locked down. Finally, the performance of these biomarker panels was assessed in an independent external validation cohort. Some of the results of these studies have been previously reported in the form of abstracts (13–15).

Methods

The institutional review boards of the participating institutions—University of British Columbia, McMaster University, and Université Laval—approved this study.

Subjects

Written informed consents were obtained from participants undergoing allergen inhalation challenges. Subjects were nonsmokers with mild atopic asthma, free of other lung diseases and any cardiovascular disease. All subjects had physician diagnosed, clinically stable asthma, with a baseline FEV_1 of 70% or greater of predicted value, and baseline provocative concentration of methacholine that caused a 20% drop in FEV_1 (PC_{20}) less than 16 mg/ml. All subjects developed the early asthmatic response (at least 20% fall in FEV_1 within 2 hours after allergen inhalation). Exclusion criteria included the use of inhaled corticosteroids, and use of other asthma medication, with the exception of infrequently inhaled β_2-agonist, which was withheld for 8 hours before spirometry measurements (additional details can be found in References 16–22).

Methacholine and Allergen Inhalation Challenge

Methacholine and allergen challenge were performed on triad visits. Methacholine inhalation tests were performed on Days 1 (pre-methacholine) and 3 (post-methacholine) in order to determine airway hyperresponsiveness, defined as the allergen-induced shift ($[PC_{20}]_{pre}/[PC_{20}]_{post}$). Allergen inhalation challenge was performed on Day 2 using allergen extracts in doubling doses until a drop in FEV_1 of at least 20% was achieved; subsequently, FEV_1 was measured at regular intervals up to 7 hours after challenge. All subjects demonstrated an FEV_1 drop of 20% between 0 and 2 hours (early asthmatic response) after allergen inhalation challenge. Participants who demonstrated a maximum drop in FEV_1 greater than or equal to 15% between 3 and 7 hours after allergen challenge (LAR) were classified as dual responders (DRs). Subjects who did not meet the 15% drop in FEV_1 during the LAR were classified as early responders (ERs).

Cohorts

The discovery cohort consisted of 36 subjects. For all subjects of the discovery cohort, blood samples were collected for molecular analysis immediately before and 2 hours after allergen inhalation challenge. For the validation cohort, pre-challenge blood samples from 33 subjects were used for molecular analyses. Furthermore, 17 pre-challenge samples were collected from 7 subjects who elicited a different response during repeat allergen inhalation challenges, whereas 20 pre-challenge samples were collected from 9 subjects who elicited only the dual response during repeat allergen inhalation challenges. Standard operating protocols were used at each participating center for all blood collection procedures (*see* the online supplement for blood collection and processing details).

RNA-Seq

Total RNA from blood samples of the discovery cohort was purified and sent to Génome Québec (Centre d'Innovation Genome Quebec et Université McGill) for whole-transcriptome sequencing (eight samples per lane). Ribosomal RNA and globin-depleted stranded cDNA libraries were sequenced using an Illumina HiSeq 2000 as 100-bp paired end reads. All RNA-Seq files passed quality-control metrics based on FastQC standards. After trimming (Seqtk, version 1.0), the left and right concatenated files were used by the Trinity software (version r20131110) (23) to construct a *de novo* assembly of the blood transcriptome (reverse forward–stranded library type was specified). The abundance estimates of Trinity contigs were estimated using RSEM (RNA-Seq by Expectation Maximization, version 1.2.11) (24) using the Bowtie aligner. The abundance

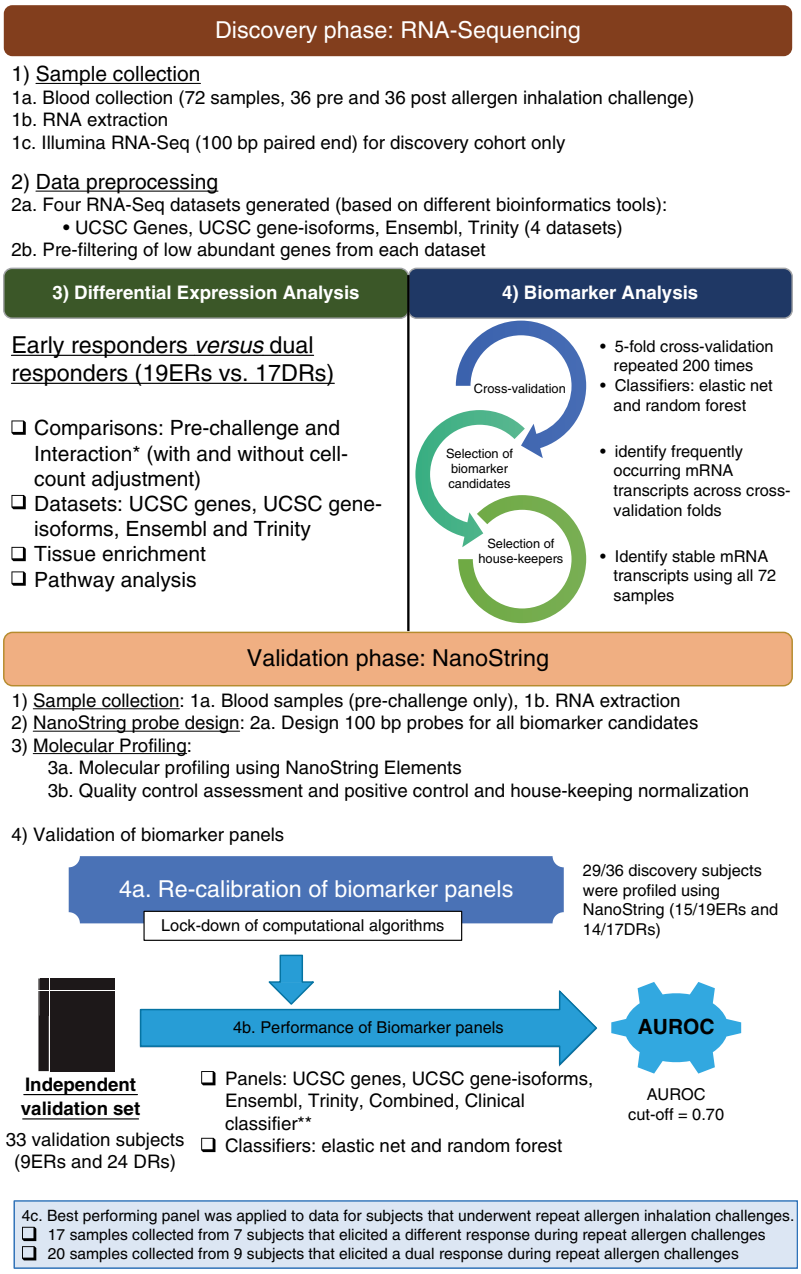

Figure 1. Overview of analysis. The study was divided into a discovery and validation phase. Unbiased whole-genome expression profiling was performed using RNA sequencing, and predictive biomarker panels of the late asthmatic response were identified before challenge and at the interaction (after challenge normalized to before challenge). Selected transcripts were profiled using the NanoString platform on all samples. The biomarker panels were recalibrated using the discovery samples and applied to the additional samples to determine the classification performance. *Interaction: for each mRNA transcript, the expression of the pre-challenge samples was subtracted from the corresponding post-challenge sample before comparing ERs and DRs. **Clinical classifier consisted of baseline FEV_1, PC_{20} before challenge, total leukocyte counts, and white blood cell–count frequencies (neutrophils, lymphocytes, monocytes, eosinophils, and basophils). AUROC = area under the receiver operating characteristic curve; DR = dual responder; ER = early responder; PC_{20} = provocative concentration of methacholine that caused a 20% drop in FEV_1; UCSC = University of California, Santa Cruz.

estimates of University of California, Santa Cruz (UCSC) genes and gene-isoforms were estimated using RSEM (version 1.2.19) with the Bowtie2 (version 2.2.4) aligner. STAR (Spliced Transcripts Alignment to a Reference, version 2.5.0a), a fast RNA-Seq alignment software (25), was also used to align the paired end reads to the human genome using annotations from the GENCODE project (GENCODE release v21) (26). Feature counts (27) in the Subread/Rsubread package (version 1.5.0) was used to estimate the abundance of Ensembl transcripts. *See* the online supplement for complete details on RNA-Seq data preprocessing and normalization.

nCounter Elements TagSets

Oligonucleotide probes (100 bp) were designed for all biomarker candidates using the nDesign portal (http://www.nanostring.com/) and with the help of the bioinformatics team at NanoString Technologies. For uncharacterized candidates, the sequences were selected based on coverage plots of RNA-Seq data and provided to NanoString Technologies (*see* the online supplement for probe sequences). RNA samples from both the discovery and validation cohorts as well as repeated challenges were profiled for biomarker candidates using custom nCounter Elements TagSets (NanoString Technologies). All NanoString data were assessed for various quality-control metrics (28) and normalized according to their data analysis guide (*see* the online supplement for complete details).

Statistical Analyses

Differential expression analysis was performed using the limma (linear models for microarray and RNA-Seq data) R-library (version 3.32.4; 29). Two comparisons were considered: ERs versus DRs before challenge and ERs versus DRs in the interaction (post-minus pre-challenge levels). A *t* test was used to compare ERs and DRs for all quantitative clinical and demographics variables (e.g., cell counts, age) when the data were deemed to be normally distributed (assessed using the Shapiro-Wilk normality test and normal Q-Q plots). The Wilcoxon rank-sum test was used when the normality assumption was not met. Tissue enrichment analysis was performed using the sear R-library (version 0.1), whereas gene set analysis was

performed using Enrichr (30). When applicable, the Benjamini-Hochberg correction procedure for multiple testing (31) was used, and a false discovery rate (FDR) of 0.15 was used to determine statistical significance.

Elastic net (32) and random forest (33) were used to identify biomarker panels that could distinguish ERs from DRs for each RNA-Seq dataset: UCSC genes, UCSC gene-isoforms, Ensembl, and Trinity. Classification performance of biomarker panels for each dataset was estimated using a 5-fold cross-validation repeated 200 times (200 × 5-fold cross-validation), using an elastic net and random forest classifier. For each cross-validation run, the area under the receiver operating characteristic curve (AUROC) was computed and averaged over 200 repeats for each dataset. For both the discovery and validation phase an AUROC cut-off of 0.70 was used as per recommended guidelines for clinical biomarker implementation (34). Biomarker selection was based on the frequently occurring (stable) transcripts in biomarker panels identified within cross-validation folds. To select stable biomarker candidates, the most frequently occurring transcripts among the 1,000 panels (5-fold × 200 repeats) for both elastic net and random forest applied to each dataset were determined. The overlap between the top 100 ranked transcripts based on elastic net and random forest were identified for each dataset and annotated, from which a subset of candidates was selected to be transferred to the NanoString platform. Housekeeping transcripts per dataset were identified using the geNorm algorithm (35) in the NormqPCR R-library (version 1.16.0). Complete details on the statistical methodology used in this study can be found in the online supplement.

Results

Discovery and Validation Cohorts

The discovery cohort consisted of 36 subjects (Table 1): 19 were classified as ERs and 17 were classified as DRs. The post-challenge PC_{20} was significantly lower in DRs compared with ERs, resulting in a significantly increased allergen-induced shift in DRs compared with ERs. No significant changes between ERs and DRs were identified for total leukocyte counts and specific white blood cell frequencies before challenge (Table 1). In the discovery cohort, white blood cell count frequencies were also not significantly different after

Table 1. Subject Demographics of the Discovery Cohort

Clinical Variable	Isolated ERs (*n* = 19)	DRs (*n* = 17)	*P* Value
Female, %	68	65	
Height, cm*	168.19 ± 8.65	168.53 ± 9.26	0.91
Weight, kg†	62.70 (58.55–78.25)	71.00 (63.35–87.15)	0.23
Age, yr†	28.00 (21.00–34.50)	23.00 (21.00–37.00)	0.65
Baseline FEV_1, L/s*	3.29 ± 0.79	3.27 ± 0.81	0.95
Predicted FEV_1, L/s†	3.43 (3.15–4.23)	3.53 (3.19–4.13)	0.91
% drop in FEV_1 during the EAR†	30.40 (25.45–40.15)	34.80 (30.70–44.10)	0.12
% drop in FEV_1 during LAR*	6.51 ± 5.86	34.75 ± 7.66	2.86×10^{-14}
Pre-methacholine PC_{20}, mg/ml†	5.25 (1.88–11.15)	2.24 (0.59–4.56)	0.11
Post-methacholine PC_{20}, mg/ml†	5.71 (1.00–9.50)	0.54 (0.23–2.26)	0.004
AIS*‡	1.89 ± 1.53	3.71 ± 1.91	0.01
Allergen			
Cat	9	4	
Fungus	1	1	
Grass	3	2	
HDM	5	7	
Horse	0	1	
Ragweed	1	2	
Site			
Laval	11	7	
McMaster	8	9	
UBC	0	1	
Blood cell counts and frequencies before challenge			
Leukocytes, $\times 10^9$ cells/L*	6.38 ± 1.44	5.87 ± 1.35	0.31
Neutrophils, %*	0.57 ± 0.08	0.53 ± 0.09	0.19
Lymphocytes, %*	0.31 ± 0.08	0.32 ± 0.09	0.70
Monocytes, %*	0.08 ± 0.01	0.09 ± 0.02	0.23
Eosinophils, %†	0.03 (0.02–0.05)	0.04 (0.03–0.07)	0.27
Basophils, %†	0.01 (0.004–0.01)	0.0 (0.0–0.01)	0.36

Definition of abbreviations: AIS = allergen-induced shift; DRs = dual responders; EAR = early asthmatic response; ERs = early responders; HDM = house dust mite; LAR = late-phase asthmatic response; PC_{20} = provocative concentration of methacholine that caused a 20% drop in FEV_1; UBC = University of British Columbia.

Grass refers to grass mix, orchard grass, or timothy grass. HDM refers to *Dermatophagoides farinae* or *D. pteronysinnus*.

*Variable is assumed to be normally distributed. Descriptive statistics are presented as mean ± SD. A *t* test was used to compare ERs and DRs.

†Variable is assumed to not be normally distributed. Descriptive statistics are presented as median (25–75th percentiles). A Wilcoxon rank-sum test was used to compare ERs and DRs.

‡$[PC_{20}]_{pre}$ divided by $[PC_{20}]_{post}$.

Table 2. Subject Demographics of the Validation Cohort

Clinical Variable	Isolated ERs ($n = 9$)	DRs ($n = 24$)	*P* Value
Female, %	56	46	
Height, cm*	168.11 ± 10.21	171.15 ± 8.91	0.43
Weight, kg*	69.00 ± 11.87	72.21 ± 11.55	0.51
Age, yr†	31.00 (27.00–36.00)	25.50 (21.75–41.00)	0.35
Baseline FEV_1, L/s*	3.05 ± 0.51	3.50 ± 0.81	0.14
Predicted FEV_1, L/s*	3.67 ± 0.84	3.68 ± 0.62	0.98
% drop in FEV_1 during the EAR*	31.92 ± 9.32	36.39 ± 8.83	0.21
% drop in FEV_1 during LAR†	7.20 (4.90–11.10)	23.75 (17.50–30.88)	1.39×10^{-5}
Pre-methacholine PC_{20}, mg/ml†	5.13 (0.64–5.45)	1.58 (0.69–4.08)	0.49
Post-methacholine PC_{20}, mg/ml†	5.11 (2.18–7.98)	0.76 (0.34–1.55)	0.04
AIS*‡	1.35 ± 1.13	2.92 ± 1.69	0.04
Allergen			
Cat	7	9	
Fungus	0	0	
Grass	2	4	
HDM	0	8	
Horse	0	1	
Ragweed	0	2	
Site			
Laval	6	9	
McMaster	1	12	
UBC	2	3	
Blood cell counts and frequencies before challenge			
Leukocytes, $\times 10^9$ cells/L*	5.13 ± 1.25	6.12 ± 1.52	0.10
Neutrophils, %*	0.58 ± 0.10	0.51 ± 0.08	0.04
Lymphocytes, %*	0.29 ± 0.08	0.37 ± 0.07	0.02
Monocytes, %*	0.07 ± 0.02	0.08 ± 0.01	0.62
Eosinophils, %†	0.03 (0.02–0.06)	0.04 (0.02–0.04)	0.70
Basophils, %†	0.00 (0.002–0.006)	0.01 (0.005–0.01)	0.17

For definition of abbreviations, *see* Table 1. Grass refers to grass mix, orchard grass, or timothy grass. HDM refers to *Dermatophagoides farinae* or *D. pteronyssinus*.

*Variable is assumed to be normally distributed. Descriptive statistics are presented as mean ± SD. A *t* test was used to compare ERs and DRs.

†Variable is assumed to not be normally distributed. Descriptive statistics are presented as median (25–75th percentiles). A Wilcoxon rank-sum test was used to compare ERs and DRs.

‡$[PC_{20}]_{pre}$ divided by $[PC_{20}]_{post}$.

challenge (normalized to before challenge, also called the interaction) between ERs and DRs. The validation cohort consisted of 33 subjects (Table 2): 9 were classified as ERs and 24 were classified as DRs. Significant changes between ERs and DRs with respect to pre-challenge neutrophil and lymphocyte frequencies were also identified (Table 2).

Differential Expression Analysis Reveals a Strong Discriminatory Signal between ERs and DRs before Challenge

At an FDR cut-off of 0.15, thousands of transcripts were identified to be differentially expressed between ERs and DRs before challenge with or without adjusting for total leukocyte counts (Figure 2A). Specifically, 1,177 UCSC genes (125 up in DRs, and 1,052 down in DRs), 683 UCSC gene-isoforms (69 up and 614 down), 1,032 Ensembl (305 up and 727 down), and 866 Trinity contigs (30 up and 836 down) were differentially expressed before challenge with or without adjustment for cell counts (Figure 2B). However, at the same FDR cut-off, no significant transcripts between ERs and DRs were identified in the interaction comparison (Figure 2A). Annotation of all differentially expressed transcripts resulted in 2,051 unique gene symbols, which were used for tissue and pathway enrichment analysis. Tissue enrichment analysis attributed the differentially expressed genes before challenge to T cell–specific genes, namely CD4 and CD8 T cells (Figure 2C). Enrichr identified 85 significant BioCarta pathways, 24 KEGG pathways, and 62 WikiPathways. Figure 2D depicts the overlap between the top 10 significant pathways with a pathway membership similarity of greater than 20%. Significantly enriched pathways included the T-cell receptor signaling pathway, IL-2 receptor β chain in T-cell activation, and IL-1, TNF, TGF-β, and B-cell receptor signaling pathways (Figure 2D). Additional details regarding dataset preprocessing can be found in the online supplement.

Biomarker Analysis of RNA-Seq Data Identifies RNA Transcripts That Are Predictive of the LAR

Given that the discriminatory signal between ERs and DRs was stronger before challenge, we focused our efforts on identifying risk biomarker panels that could discriminate ERs from DRs before allergen inhalation challenge. Using a 200×5-fold cross-validation, estimates of the test performance of each of the RNA-Seq dataset classifiers as well as the clinical classifier (based on baseline FEV_1, before methacholine PC_{20}, total leukocyte counts, and white blood cell count frequencies [neutrophils, lymphocytes, monocytes,

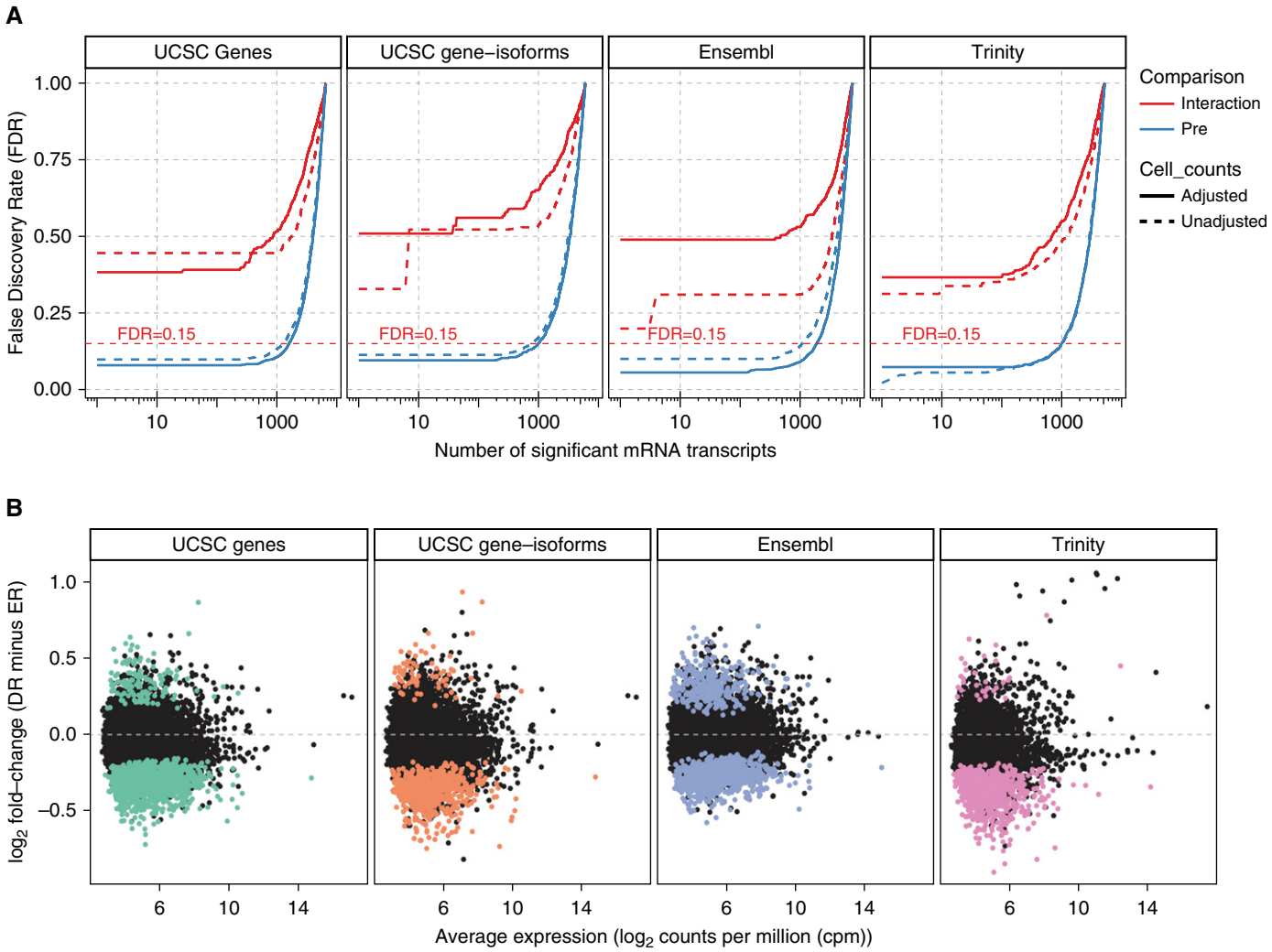

Figure 2. Differential expression analysis. (*A*) Benjamini-Hochberg false discovery rate (FDR) versus the number of significant mRNA transcripts, before challenge and at the interaction (before minus after), with and without adjusting for total leukocyte counts, for each dataset. (*B*) Minus–average plots depicting the fold change versus the average expression of mRNA transcripts across all pre-challenge samples, for each dataset. The colored points represent differentially expressed transcripts (comparing early responders [ERs] and dual responders [DRs] before challenge) common between the with- and without-adjustment for total leukocyte count comparisons, whereas the black points represent nondifferentially expressed transcripts. (*C*) Tissue enrichment of common (with and without adjusted for total leukocyte counts) differentially expressed transcripts between ERs and DRs before challenge from all datasets. (*D*) Top 10 significantly enriched pathways from the BioCarta, KEGG, and WikiPathways databases using the entire list of differentially expressed genes from all datasets. Connected pathways have a Jaccard similarity of greater than 20%. UCSC = University of California, Santa Cruz.

eosinophils, and basophils]) were determined (Figure 3A). The performance of the Trinity biomarker panel met an AUROC cut-off of 0.70 with average classification performances of AUROC of 0.71 (elastic net) and AUROC of 0.74 (random forest). The clinical panel had the lowest classification performance (AUROC of ∼50%).

Setting the tuning parameter ($\alpha = 0.9$) resulted in approximately 13–17 transcripts per elastic net biomarker panel (an equivalent-sized random forest classifier was chosen based on the importance score ranking) within the cross-validation folds. Biomarker panel transcripts across the 1,000 panels (200 × 5-fold = 1,000-fold) were tallied for both the elastic net and random forest panels for each dataset (*see* Methods for details). The overlaps between the top 100 ranked transcripts in the elastic net and random forest panels were determined for each dataset and annotated, from which a subset of candidates was selected to be transferred to the NanoString platform: 33 UCSC genes, 35 UCSC gene-isoforms, 35 Ensembl transcripts, and 14 Trinity contigs (*see* Figure 3B for an overlap between transcripts selected from each dataset). The RNA biomarker transcripts retained their discriminative ability in separating ERs from DRs 2 hours after allergen inhalation challenge, based on principal component analysis (Figure 2C).

Nine housekeeping transcripts were identified across three datasets: UCSC genes, Ensembl, and Trinity (Figure 2D). However, because the Trinity housekeeping contigs were well above the expression of all other selected transcripts, they were removed from the final NanoString assay design,

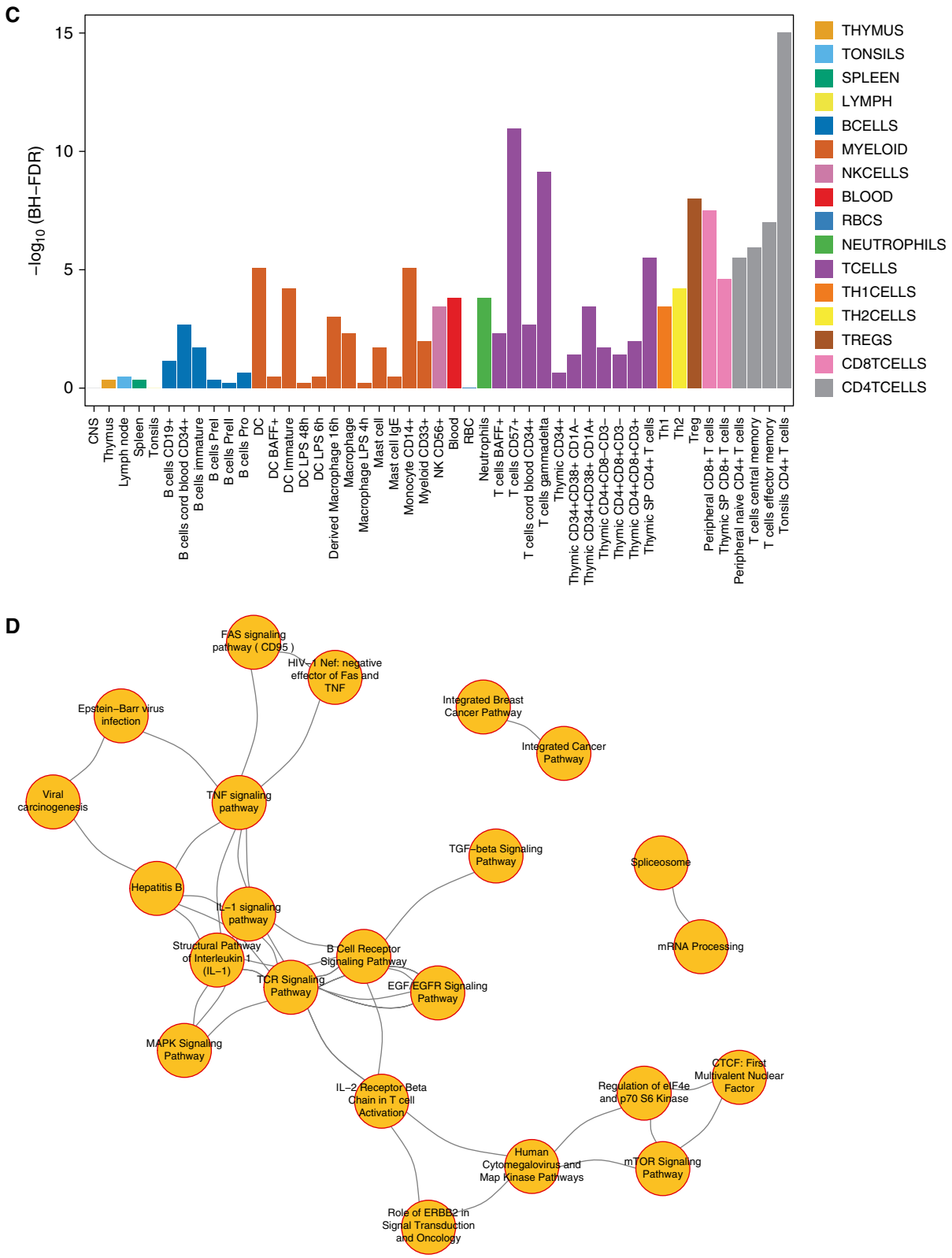

Figure 2. (Continued).

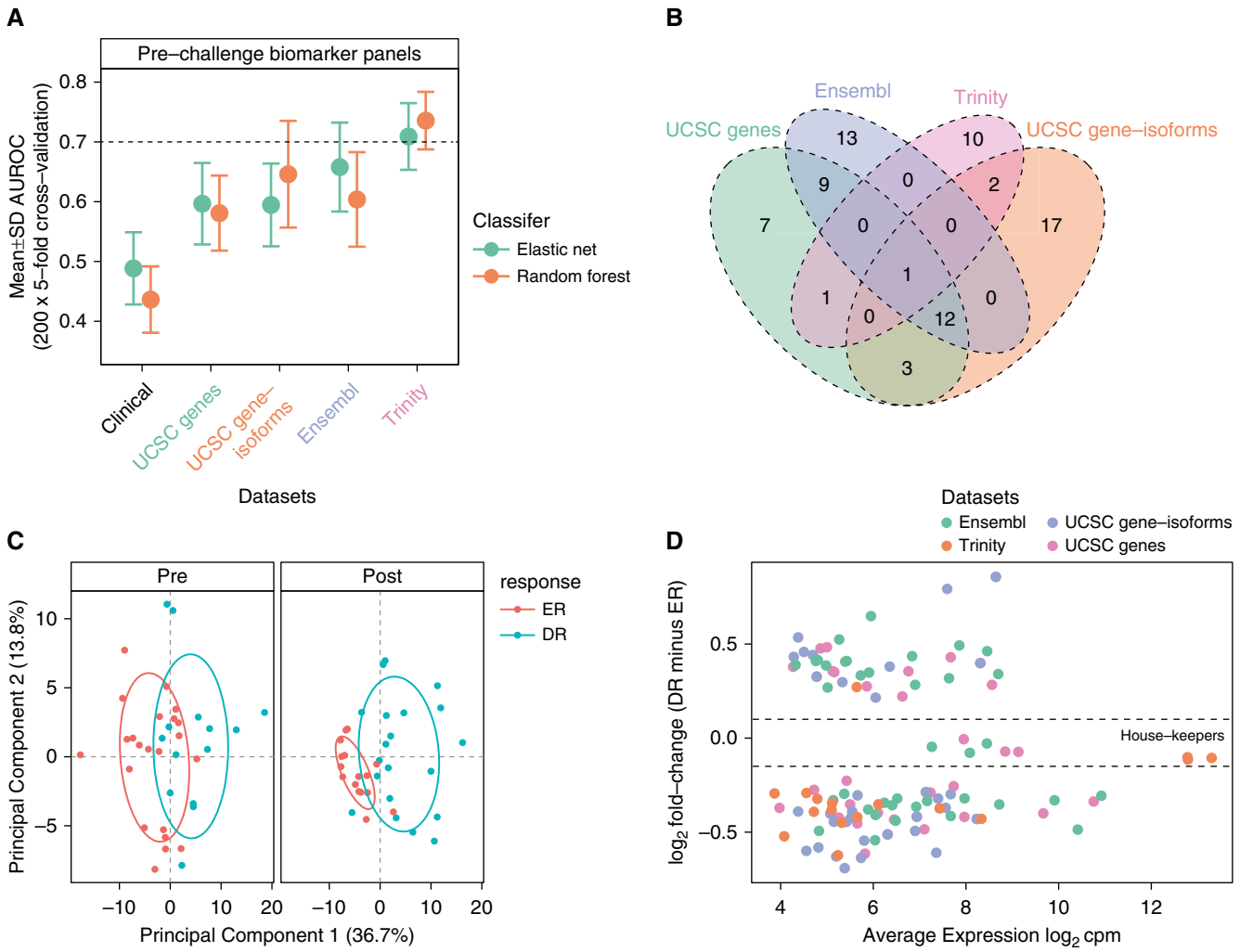

Figure 3. Discovery of pre-challenge biomarker panels. (*A*) Classification performance of pre-challenge biomarker panels based on a 5-fold cross-validation repeated 200 times (200 × 5-fold cross-validation) using an elastic net and random forest classifier. (*B*) Overlap between the selected mRNA biomarker transcripts from each dataset; each set of transcripts serves as the resulting biomarker panel for each dataset. (*C*) Principal component analysis plot of all samples (before and after) using all biomarker transcripts in *B*. The ellipses (with a 68% confidence interval) depict a separation between the early responders (ERs) and dual responders (DRs) before challenge, which is sustained at 2 hours after challenge. (*D*) Minus–average plot of all biomarker transcripts as well as housekeeping transcripts (within dashed lines). Note: complete clinical data were present for 17/19 ERs and 15/17 DRs. AUROC = area under the receiver operating characteristic curve; UCSC = University of California, Santa Cruz.

resulting in six housekeeping gene candidates: *MED13*, *TOR1AIP2*, and *WAC* (Ensembl dataset) and *ARPC4*, *TMBIM6*, and *RHOA* (UCSC gene dataset), all of which were transferred to the NanoString platform. A NanoString assay can reliably measure up to two million counts per sample. However, if some transcripts are highly abundant (e.g., globin in a whole-blood sample), they may saturate the assay such that most counts would be attributed to the highly expressed targets. Therefore, the highly abundant Trinity housekeeping transcripts were not selected for downstream NanoString analyses.

Panel Recalibration and Validation Using the NanoString Platform

Six housekeeping probes as well as 75 probes corresponding to the 117 biomarker candidates (due to gene redundancy across datasets) were designed. Transcripts were annotated to their corresponding gene symbol, and a probe was selected from the nDesign portal (www.nanostring.com). Custom probes were designed for Trinity contigs that mapped either to intronic regions of known genes or to uncharacterized loci of the human genome, based on visualizing the nucleotide coverage of each transcript, and sent to NanoString technologies, which selected 100 bp from this sequence as the final probe sequence (*see* Table E1 in the online supplement).

All samples of the discovery and validation cohorts (Figure 4A) were profiled using the NanoString platform for all biomarker and housekeeping probes. Quality-control assessment and normalization based on positive spike-in controls and housekeeping probes were performed (*see* the online supplement). Five biomarker panels (UCSC genes, UCSC gene-isoforms, Ensembl, Trinity, and Combined [all transcripts]) and a clinical panel were locked down (fixed weights of transcripts in the biomarker

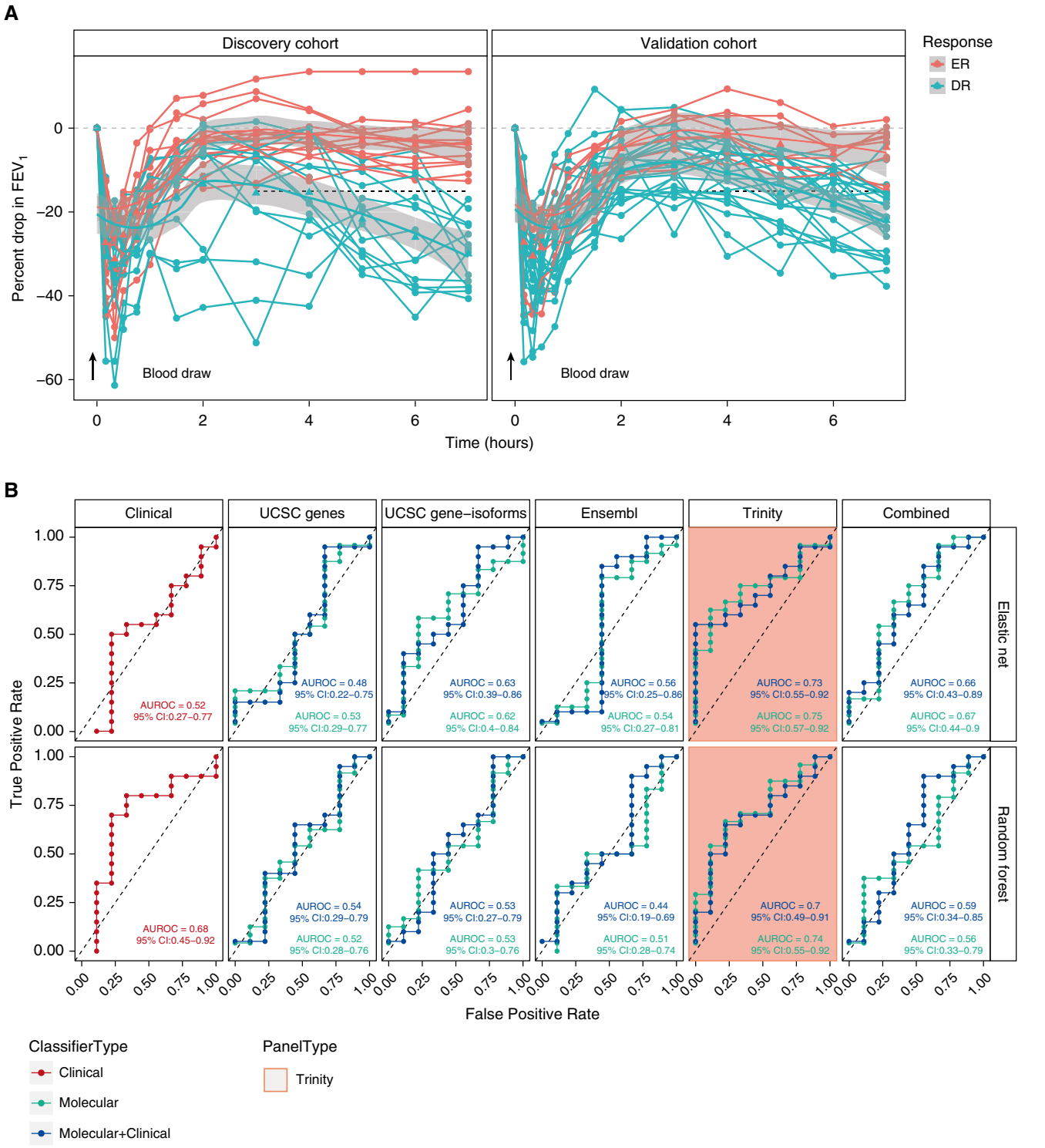

Figure 4. Validation of pre-challenge biomarker panels. (*A*) FEV_1 drop from baseline of subjects in the discovery and validation cohorts undergoing allergen inhalation. The LOESS (locally weighted scatterplot smoothing) curve was fitted to each group separately, with a 95% confidence interval. (*B*) Area under the receiver operating characteristic curve (AUROC) for all molecular panels (from each dataset) and in conjunction with clinical variables, such as baseline FEV_1, provocative concentration of methacholine that causes a 20% drop in FEV_1 before challenge, total leukocyte counts, and white blood cell count frequencies (neutrophils, lymphocytes, monocytes, eosinophils, and basophils). The Trinity panel (pink) achieved an AUROC of greater than or equal to 0.70 by itself and in conjunction with clinical variables. Notes: given the limited RNA quantity available for samples in the discovery cohort, molecular profiling was performed for 15/17 early responders (ERs) and 14/17 dual responders (DRs). Complete clinical data were present for 14/15 ERs and 12/14 DRs in the discovery cohort and 9/9 ERs and 20/24 DRs in the validation cohort. CI = confidence interval; UCSC = University of California, Santa Cruz.

panels) by fitting a model (elastic net and random forest) to NanoString data from discovery samples (15 ERs and 14 DRs; some samples were lost due to the lack of sufficient RNA material remaining for some samples). An "off the shelf" test was performed using an independent cohort of 9 ERs and 24 DRs for all biomarker panels, and the AUROC was determined (Figure 4B). The Trinity biomarker panels performed well using both elastic net (AUROC = 0.75; 95% confidence interval = 0.57–0.92) and random forest (AUROC = 0.74; 95% confidence interval = 0.55–0.92) biomarker panels (Figure 4B), with both panels having an accuracy of 0.72 (using a probability threshold of 0.5).

The Trinity Biomarker Panel and Its Application to Subjects Undergoing Repeated Allergen Inhalation Challenges

The Trinity biomarker panel consisted of 14 transcripts (both known and uncharacterized) that demonstrated the strongest discriminatory signal among all molecular panels, as well as the clinical panel. Figure 5A depicts the separation of the responder groups (using samples from both the discovery and validation cohort) along the first principal component. The first principal component is a linear combination of the 14 transcripts weighted by the contribution of each individual transcript, leading to responder group separation observed in Figure 5A. Figure 5B shows the contribution of each transcript that leads to the separation depicted in Figure 5A. Overall, the expression levels of the Trinity panel transcripts were higher in ERs compared with DRs (Figure 5C).

The Trinity panel–based elastic net and random forest classifiers were applied to 17 samples corresponding to 7 subjects who elicited a different response during repeated allergen inhalation challenges. The Trinity panel had an accuracy of 0.71 and 0.65 for the elastic net and random forest classifier, respectively (Figure 5D, upper panel). The Trinity panel–based elastic net and random forest classifiers were applied to subjects who elicited only the dual response upon repeated allergen challenges and achieved an accuracy of 0.70 and 0.75, respectively (Figure 5D, lower panel).

Discussion

We identified four predictive biomarker panels of the LAR, of which the Trinity biomarker panel, consisting of novel RNA transcripts, was validated using an independent cohort, achieving an AUROC of greater than 0.70. Interestingly, the Trinity panel demonstrated utility in predicting the response of subjects who underwent repeated allergen inhalation challenges. This may indicate that the Trinity biomarker panel was sensitive to molecular fluctuations within an individual that are predictive of the type of response a given subject will elicit on the day of the allergen inhalation challenge. These results may imply inherent differences in the underlying molecular mechanisms that can be detected in peripheral whole blood, which predispose individuals with asthma for the LAR.

All biomarker transcripts, with the exception of adenosine deamimase 2 (*CECR1*), were downregulated in DRs compared with ERs. CECR1 is secreted by antigen-presenting cells, such as monocytes, and leads to the proliferation of T cells and macrophages (36). Other Trinity panel transcripts included those with multiple cellular functions, such as splicing factor 3b subunit 1 (*SF3B1*), which is involved in DNA damage response (37), and quaking homolog (*QKI*), which is involved in pre-mRNA splicing, protein translation, blood vessel development, and apoptosis (38). Downregulation of these transcripts, as well as caspase 8 (*CASP8*), another promoter of apoptosis, may suggest a reduced capacity of cellular maintenance in DRs compared with ERs. Reduced senataxin (SETX) levels have been shown to lead to viral resistance through inhibition of antiviral gene expression (39). Lysosomal trafficking regulator (*LYST*) has been shown to control Toll-like receptor-3– and -4–induced endosomal TRIF (TIR domain-containing adapter-inducing IFN-β) signaling pathways (40). Collectively, these results suggest a dysregulation of complex immune signaling networks that predispose an individual with asthma to the late-phase response upon allergen inhalation. A dampened antiinflammatory response was observed through reduced gene expression of the formyl peptide receptor (FPR) family, *FPR1_intron* and *FPR2*, in DRs compared with ERs. FPR1 and FPR2 are G protein–coupled receptors expressed by phagocytic cells, such as monocytes, macrophages, and dendritic cells, which play roles in host defense against pathogens (41). Lower levels of receptor expression of *FPR1* and *FPR2* in DRs suggest a reduced capacity to bind antiinflammatory ligands, thus impairing the resolution of inflammation.

Apart from known genes, uncharacterized transcripts that were constructed *de novo* were also present in the Trinity biomarker panel. These included intronic sequences in genes, such as *FPR1*, *IFRD1*, *TNFRSF10C*, and noncoding RNAs (LOC254896, LOC101927568, and LOC105378945). A previous study indicated that "dark matter" (unassigned functional role or unannotated) RNA represents a significant proportion of total RNA (42) of sequenced reads in normal and neoplastic tissues. The authors found that 50–65% of all nonribosomal, nonmitochondrial RNA was dark matter RNA, suggesting a greater discovery space for identifying novel transcripts associated with disease. Noncoding RNAs are known to regulate gene expression through mechanisms, such as chromatin remodeling, post-transcriptional regulation, and mRNA degradation (43). Given the unexplored territory of noncoding RNAs in the context of asthma pathogenesis, the transcripts identified in the present study may just be the low-hanging fruit, given the depth of sequencing. The functional role of these transcripts at present remains unknown. Given their reproducibility across platforms and cohorts, these transcripts may not be technical artifacts, but rather surrogates of the underlying biology. However, additional work will be required to delineate their biological function.

A prospective study will be required to demonstrate the utility of the Trinity biomarker panel. Briefly, the same inclusion/exclusion clinical criteria will be retained for future subjects. A blood sample will be collected from all subjects and the Trinity biomarker panel will be applied, such that each subject is assigned a probability score that describes the likelihood that the subject will elicit a dual response upon allergen inhalation challenge. Subjects with a score greater than or equal to 0.5 will be retained as potential candidates for entry into clinical trials for new asthma drugs. These subjects will then undergo

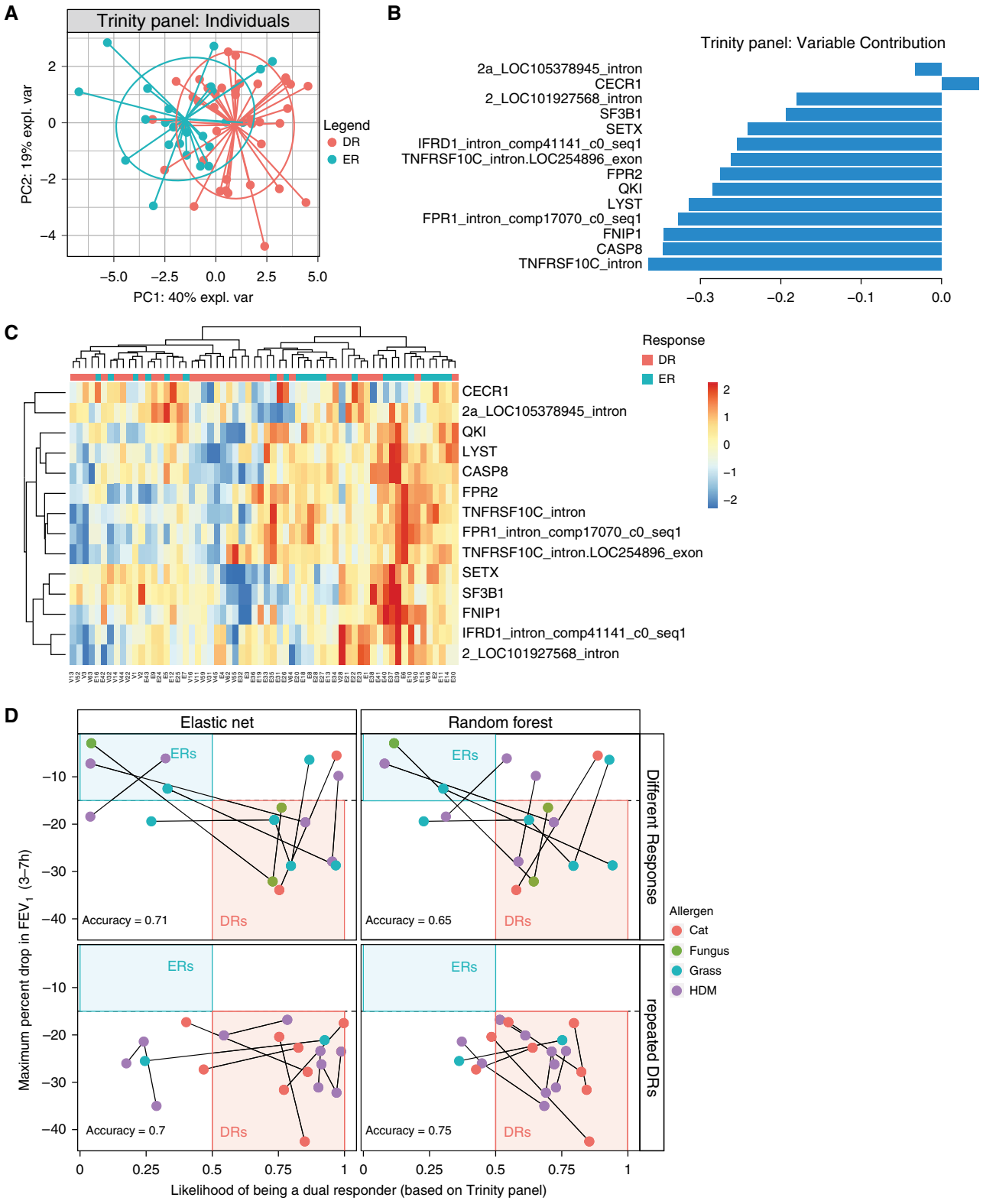

Figure 5. The Trinity biomarker panel. (*A*) Principal component analysis (PCA) plot of all samples (discovery + validation) using the transcripts of the Trinity biomarker panel. (*B*) Variable contributions that led to the separation between early responders (ERs) and dual responders (DRs) in the PCA. (*C*) Heatmap of the Trinity biomarker panel transcripts. All transcripts with the exception of CECR1 are upregulated in ERs compared with DRs. (*D*) Maximum drop in FEV_1 during the late asthmatic response and the Trinity biomarker panel–based predicted probabilities of repeatedly challenged subjects, using their baseline expression data before allergen inhalation challenge. HDM = house dust mite.

methacholine challenge tests as well as allergen inhalation challenges to determine their response to allergen challenge. This prospective study may demonstrate that fewer subjects are required for the screening process (allergen inhalation challenge) to obtain sufficient numbers of subjects for clinical trials. This study is not only a standard biomarker study, but also depicts the novelty of tapping into the world of noncoding RNA through the use of total RNA and *de novo* transcriptome assembly. Noncoding RNA may serve as novel predictive biomarkers, as well as provide additional insights into the biological mechanisms of asthmatic responses.

The present study uses a rigorous statistical approach in the identification and validation of whole-blood RNA biomarker panels that are predictive of the LAR before allergen inhalation challenge. It may be possible that the "state" of circulating immune cells may predispose individuals with asthma to the LAR. Additional studies that apply these panels to larger cohorts may provide appropriate estimates of positive and negative predictive values. These biomarker transcripts will be further investigated to elucidate the underlying molecular mechanisms of asthmatic responses, and may be used to develop novel asthma therapies. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank the research participants for taking part in these studies, as well as Johane Lepage, Philippe Prince, Joanne Milot, and Mylène Bertrand (Laval University, Quebec City, Quebec, Canada); Richard Watson, George Obminski, Heather Campbell, Abbey Torek, Tara Strinich, and Karen Howie (McMaster University, Hamilton, Ontario, Canada); and Linda Hui and Joyce Kum (University of British Columbia, Vancouver, British Columbia, Canada) for their expertise and assistance with participant recruitment, allergen challenge, and sample collection, as part of the AllerGen NCE Clinical Investigator Collaborative. They also thank the Genome Quebec Innovation Centre for performing the RNA sequencing, and the Prevention of Organ Failure Centre of Excellence for additional support.

References

1. Gauvreau GM, El-Gammal AI, O'Byrne PM. Allergen-induced airway responses. *Eur Respir J* 2015;46:819–831.
2. Samitas K, Delimpoura V, Zervas E, Gaga M. Anti-IgE treatment, airway inflammation and remodelling in severe allergic asthma: current knowledge and future perspectives. *Eur Respir Rev* 2015;24: 594–601.
3. Cockcroft DW, Davis BE. Mechanisms of airway hyperresponsiveness. *J Allergy Clin Immunol* 2006;118:551–559, quiz 560–561.
4. Pelikan Z. Effects of inhaled corticosteroids on the dual asthmatic response. *Allergy Asthma Proc* 2013;34:e47–e58.
5. Guilbert TW, Morgan WJ, Zeiger RS, Mauger DT, Boehmer SJ, Szefler SJ, *et al*. Long-term inhaled corticosteroids in preschool children at high risk for asthma. *N Engl J Med* 2006;354: 1985–1997.
6. Vignola AM, Chanez P, Campbell AM, Souques F, Lebel B, Enander I, *et al*. Airway inflammation in mild intermittent and in persistent asthma. *Am J Respir Crit Care Med* 1998;157:403–409.
7. Croteau-Chonka DC, Qiu W, Martinez FD, Strunk RC, Lemanske RF Jr, Liu AH, *et al*.; Asthma BioRepository for Integrative Genomic Exploration (Asthma BRIDGE) Consortium. Gene expression profiling in blood provides reproducible molecular insights into asthma control. *Am J Respir Crit Care Med* 2017;195:179–188.
8. Verrills NM, Irwin JA, He XY, Wood LG, Powell H, Simpson JL, *et al*. Identification of novel diagnostic biomarkers for asthma and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2011;183: 1633–1643.
9. Dweik RA, Sorkness RL, Wenzel S, Hammel J, Curran-Everett D, Comhair SAA, *et al*.; National Heart, Lung, and Blood Institute Severe Asthma Research Program. Use of exhaled nitric oxide measurement to identify a reactive, at-risk phenotype among patients with asthma. *Am J Respir Crit Care Med* 2010;181:1033–1041.
10. Yang CX, Singh A, Kim YW, Conway EM, Carlsten C, Tebbutt SJ. Diagnosis of western red cedar asthma using a blood-based gene expression biomarker panel. *Am J Respir Crit Care Med* 2017;196: 1615–1617.
11. Kicic A, Hallstrand TS, Sutanto EN, Stevens PT, Kobor MS, Taplin C, *et al*. Decreased fibronectin production significantly contributes to dysregulated repair of asthmatic epithelium. *Am J Respir Crit Care Med* 2010;181:889–898.
12. IOM (Institute of Medicine). Evolution of translational omics: lessons learned and the path forward. Washington, DC: National Academy Press; 2012.
13. Singh A, Shannon CP, Gauvreau GM, O'Byrne PM, FitzGerald J, Boulet L-P, *et al*. Blood biomarkers of the late phase asthmatic response using RNA-Seq [abstract]. *Allergy Asthma Clin Immunol* 2014;10:A61.
14. Singh A, Shannon CP, Kim YW, DeMarco ML, Gauvreau GM, FitzGerald JM, *et al*. Identifying molecular mechanisms of the late-phase asthmatic response by integrating cellular, gene, and metabolite levels in blood [abstract]. *Ann Am Thorac Soc* 2016;13:S98.
15. Singh A, Shannon CP, Kim YW, DeMarco ML, Le Cao K-A, Gauvreau G, *et al*. Multi-omic biomarker signatures are predictive of the allergen-induced late phase asthmatic response [abstract]. *Am J Respir Crit Care Med* 2017;195:A4968.
16. Diamant Z, Gauvreau GM, Cockcroft DW, Boulet L-P, Sterk PJ, de Jongh FHC, *et al*. Inhaled allergen bronchoprovocation tests. *J Allergy Clin Immunol* 2013;132:1045–1055.e6.
17. El-Gammal A, Oliveria J-P, Howie K, Watson R, Mitchell P, Chen R, *et al*. Allergen-induced changes in bone marrow and airway dendritic cells in asthmatic subjects. *Am J Respir Crit Care Med* 2016;194: 169–177.
18. Gauvreau GM, Boulet L-P, Leigh R, Cockcroft DW, Killian KJ, Davis BE, *et al*. A nonsteroidal glucocorticoid receptor agonist inhibits allergen-induced late asthmatic responses. *Am J Respir Crit Care Med* 2015; 191:161–167.
19. Gauvreau GM, Jordana M, Watson RM, Cockroft DW, O'Byrne PM. Effect of regular inhaled albuterol on allergen-induced late responses and sputum eosinophils in asthmatic subjects. *Am J Respir Crit Care Med* 1997;156:1738–1745.
20. Gauvreau GM, Boulet L-P, Cockcroft DW, Fitzgerald JM, Carlsten C, Davis BE, *et al*. Effects of interleukin-13 blockade on allergen-induced airway responses in mild atopic asthma. *Am J Respir Crit Care Med* 2011;183:1007–1014.
21. Tworek D, Smith SG, Salter BM, Baatjes AJ, Scime T, Watson R, *et al*. IL-25 receptor expression on airway dendritic cells after allergen challenge in subjects with asthma. *Am J Respir Crit Care Med* 2016; 193:957–964.
22. Gauvreau GM, Lee JM, Watson RM, Irani A-MA, Schwartz LB, O'Byrne PM. Increased numbers of both airway basophils and mast cells in sputum after allergen inhalation challenge of atopic asthmatics. *Am J Respir Crit Care Med* 2000;161:1473–1478.
23. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;8:1494–1512.
24. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;12:323.
25. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21.
26. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, *et al*. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res* 2012;22: 1760–1774.

27. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;30:923–930.
28. nCounter expression data analysis guide. Seattle, WA: NanoString Technologies, Inc.; 2017.
29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47.
30. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;44:W90–W97.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 1995;57:289–300.
32. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;67:301–320.
33. Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
34. Sin DD, Hollander Z, DeMarco ML, McManus BM, Ng RT. Biomarker development for chronic obstructive pulmonary disease: from discovery to clinical implementation. *Am J Respir Crit Care Med* 2015;192:1162–1170.
35. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, *et al*. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002;3:RESEARCH0034.
36. Zavialov AV, Gracia E, Glaichenhaus N, Franco R, Zavialov AV, Lauvau G. Human adenosine deaminase 2 induces differentiation of monocytes into macrophages and stimulates proliferation of T helper cells and macrophages. *J Leukoc Biol* 2010;88:279–290.
37. Wang L, Brooks AN, Fan J, Wan Y, Gambe R, Li S, *et al*. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. *Cancer Cell* 2016;30:750–763.
38. Chénard CA, Richard S. New implications for the QUAKING RNA binding protein in human disease. *J Neurosci Res* 2008;86:233–242.
39. Miller MS, Rialdi A, Ho JSY, Tilove M, Martinez-Gil L, Moshkina NP, *et al*. Senataxin suppresses the antiviral transcriptional response and controls viral biogenesis. *Nat Immunol* 2015;16:485–494.
40. Westphal A, Cheng W, Yu J, Grassl G, Krautkrämer M, Holst O, *et al*. Lysosomal trafficking regulator Lyst links membrane trafficking to toll-like receptor-mediated inflammatory responses. *J Exp Med* 2017;214:227–244.
41. Migeotte I, Communi D, Parmentier M. Formyl peptide receptors: a promiscuous subfamily of G protein–coupled receptors controlling immune responses. *Cytokine Growth Factor Rev* 2006;17:501–519.
42. Kapranov P, St. Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, *et al*. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol* 2010;8:149.
43. Whitehead J, Pandey GK, Kanduri C. Regulation of the mammalian epigenome by long noncoding RNAs. Biochim Biophys Acta 2009; 1790:936–947.