Supplementary Information

Chronic marijuana use is associated with gene expression in bronchoalveolar lavage

Cuining Liu[1,2], Sunita Sharma[2], Jeanette Gaydos[2], Rebecca Johnson-Paben[2], Katerina Kechris[1], Ellen L. Burnham[2]

[1] Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO
[2] Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado School of Medicine, Aurora, CO

Key Links

- **Analysis Code, Processed Data, Results**
  https://github.com/chooliu/arjccm_marijuana_rnaseq

- **Raw RNA-seq data** (NCBI GEO)
  https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155213

Contents

Supplementary Text 1: Demographic Features (“Table 1”)

Table A. Participant characteristics by smoking group. Continuous variables are represented as median (interquantile range) and categorical variables are represented as percent in each group. Missing data is indicated in brackets as the number of subjects without a record. Low p-values indicate that that a feature differs in at least one of three groups, as assessed by a univariate F-test for continuous variables and Fisher’s exact test for categorical variables.

	Marijuana (M)	Non-Smoker (C)	Tobacco (T)	*P* (three-group)
N	15	10	16	
Demographic Information				
… Sex (Female)	26.70%	20.00%	37.50%	0.69
… Age (years)	32.0 (30.0, 38.5)	31.0 (26.2, 36.8)	39.0 (37.2, 47.8)	0.0026 **
… BMI	25.6 (23.1, 26.5)	31.2 (27.7, 33.7)	27.9 (25.6, 33.0)	0.042 *
… Race (White)	73.30%	90.00%	75.00%	0.71
% Differential Cell Types				
... Monocyte/Macrophage	95.3 (94.2, 97.9) [1]	94.7 (91.0, 96.7) [2]	92.9 (91.2, 95.9) [3]	0.73
... Lymphocyte	1.9 (0.9, 3.5) [1]	4.3 (1.7, 7.4) [2]	4.3 (3.4, 7.7) [3]	0.51
... PMNs	0.9 (0.6, 1.0) [1]	0.9 (0.5, 1.0) [1]	0.8 (0.3, 2.2)	0.73
... Airway	0.4 (0.0, 1.3) [1]	0.2 (0.0, 1.0) [2]	0.0 (0.0, 0.4) [3]	0.33
... Eosinophils	0.0 (0.0, 0.0) [1]	0.0 (0.0, 0.0) [1]	0.0 (0.0, 0.5)	0.063
Tobacco Consumption				
... Age Began Smoking Tobacco	-	-	18.0 (16.5, 23.0) [1]	
... Number of Years Smoking Tobacco	-	-	21.5 (17.8, 23.5)	
… Packs Per Day	-	-	0.5 (0.5, 1.0) [1]	
... Estimated Pack-Years	-	-	11.5 (10.0, 23.0) [1]	
Marijuana Consumption				
... Age Began Smoking Marijuana	16.0 (14.0, 17.5)	-	-	
... Number of Years Smoking Marijuana	15.0 (12.0, 21.0)	-	-	
… Times Smoked Per Day	5.0 (3.8, 8.0) [3]			
... Estimated Joint-Years	45.0 (30.0, 52.0)	-	-	

Supplementary Text 2: Detailed Methods

Participant Recruitment.

The study was approved by the Colorado Multiple Institutional Review Board. All participants provided written informed consent prior to participation. All participants were recruited between 2015-2017 using approved print and electronic advertisements from the Denver, Colorado metropolitan area. Of note is that both medical and recreational marijuana use was legal, and commercial marijuana products were available locally during this time period.

Screening, informed consent, and protocol completion were performed at the University of Colorado in the Clinical and Translational Research Center. Screening blood and urine, a chest radiograph, and spirometry were obtained in all participants(E1, E2). Current marijuana using participants reported <1 pack-year cigarette smoking history and daily or near-daily marijuana smoking equivalent to at least 20 joint-years, where a joint-year is defined as the number of joints consumed per day multiplied by the number of years smoking marijuana. Tobacco smoking participants reported daily or near-daily cigarette smoking, no active marijuana consumption, and <1 joint-year of marijuana use. Non-smoking control participants reported no active use of either substance, <1 joint-year of marijuana use, and <1 pack-year of tobacco use and no active use of either substance. Urine toxicology screenings for THC and the tobacco metabolite cotinine were used to verify recent use patterns reported by participants. An approximately age and sex-matched subset of the parent cohort (41 of the 146 participants available) were selected for RNA sequencing due to budget limitations; tobacco-only smokers were generally older than the other groups.

In an effort to minimize potential confounding related to comorbidities, exclusion criteria for all participants (regardless of category) included: age <21 or >55 years; an alcohol use disorders identification test (AUDIT)-C of ≥4 in men or ≥3 in women(E3); prior medical history of liver disease or cirrhosis, total bilirubin > 2.0 mg/dL or albumin <3 g/dL; prior medical history of myocardial infarction or congestive heart failure; prior medical history of end-stage renal disease or serum creatinine >3 mg/dL; positive toxicology screen for opiates, cocaine, or methamphetamines; history of diabetes mellitus; history of chronic obstructive pulmonary disease (COPD) or asthma; history of HIV; peripheral white blood cell count of less than 3000/µL; abnormal chest radiograph; spirometry of < 60% predicted for either FEV_1 and FVC; use of systemic antibiotics for any reason in the 4 weeks; and current pregnancy.

Sample Collection and Cell Profiling.

Bronchoscopy with bronchoalveolar lavage (BAL) was performed under conscious sedation using previously reported methods(E1, E2, E4). Briefly, the bronchoscope was wedged into a subsegment of either the right middle lobe or lingula, followed by instillation of 50-mL aliquots of sterile, room temperature 0.9% saline. Each 50-mL aliquot (up to four total) was immediately hand aspirated back into the syringe, then transferred to sterile 50-mL conical tubes, whereupon they were transported to the laboratory. The first aliquot of BAL was not used in these experiments. Second and subsequent BAL aliquots were combined, then centrifuged to separate cellular and acellular components. Cell counts were performed, and ≥ 95% BAL cell viability was determined for all participants via trypan blue exclusion. Differentials were assessed after Diff-Quik staining by a single blinded observer (EB). A proportion of BAL cells were snap frozen for RNA-sequencing and RT-PCR.

RNA-Sequencing.

After RNA extraction from BAL cell pellets using manufacturer protocols (RNeasy Mini Kit with QiaShredder columns; Qiagen, Venlo, Netherlands), library creation and sequencing was performed in two batches at the University of Colorado Genomics and Microarray Core: Batch 1 (TruSeq Stranded mRNA Library Prep Kit, Illumina HiSeq4000; Illumina, San Diego, CA, USA) and Batch 2 (Illumina TruSeq Stranded Total RNA kit, HiSeq2500).

Read Quality Control.

We obtained a median read count of 25.3 and 23.4 million reads in Run 1 and 2, respectively. Using skewer(E5) v0.2.2, we removed Illumina library adaptors, trimmed the end of the read until bases had Phred Q-Score ≥ 10, and removed reads with length ≤ 31bp after trimming. Default skewer settings were used otherwise.

Transcript Quantification.

We used kallisto(E6) v0.46.0 to convert the quality-controlled reads into transcript counts, with reference to the Ensembl(E7) v79 *Homo sapiens* genome assembly (pre-indexed by the kallisto developers; homo_sapiens.tar.gz, github.com/pachterlab/kallisto-transcriptome-indices/releases). Due to the library preparation kits used, we ran kallisto with "first read reverse" stranded-ness and with an assumed mean fragment length of 200 bases with standard deviation of 20 bases. The output was a count table of transcripts (155,188 transcripts detected) for each individual's BAL sample (41 participants).

Transcript-to-Gene Summarization.

All subsequent analyses were performed in the R statistical package, v3.6.0. The kallisto transcript-level counts were then summarized into 28,775 gene-level counts (i.e., unique ENSG identifiers) using tximport(E8) v1.4.0. Annotations converting between gene symbols and Ensembl v79 gene identifiers were obtained using the biomaRt(E9) R package, v2.40.4.

Gene-Level Filtering.

In order to mitigate across-batch technical effects and low statistical power for detecting differential expression in lowly expressed genes, we carried the 17,602 genes with non-zero counts in at least 2/3rd of Batch 1 samples as well as 2/3rd of Batch 2 samples forward to subsequent analyses.

Differential Expression.

We used edgeR(E10) v3.26.7 to model the counts of each gene using a quasi-likelihood negative binomial generalized linear model (glmQLFit defaults). edgeR normalization factors were calculated based on tximport developer recommendations. We regressed the count of each gene on smoking group (non-smoking control, marijuana smoker, or tobacco smoker), adjusting for age (years), sex (male or female), obesity (BMI < 30 or ≥ 30 kg/m^2), technical batch (Batch 1 or Batch 2), and three RUVSeq(E11) components measuring unmeasured expression heterogeneity (details about batch effect and selection of RUVSeq components in **Supplementary Text 3**). We corrected for this limited number of potential confounders due to the pilot sample size of our study. We defined statistically significant differences between groups using a quasi-likelihood F-test (glmQLFTest), after correcting for multiple testing using a Benjamini-Hochberg(E12) false discovery rate of 5% applied to each pairwise comparison (FDR < 0.05; p.adjust in R).

Residual Expression.

For visualization purposes, it was useful to obtain expression measures omitting the effect of age, sex, obesity, batch effects, and RUVSeq factors, but keeping differences between smoking groups intact. We thus fit the edgeR model with all of the covariates used in the primary analysis except for smoking group. Deviance residuals were extracted from the resulting model, yielding approximately Gaussian "residual expression" measures that were used in principal components analysis (letter **Figure 1B**).

Over-Representation Analysis.

We used over-representation analysis to evaluate if the sets of differential expressed genes between groups significantly overlapped with "biological process" gene lists defined by the Gene Ontology project(E13). We used the hypergeometric test implementation and annotations in WebGestalt(E14) (GO database from 01/14/2019, as curated by WebGestalt team). The list of 17,602 genes tested for differential expression were used as the background gene set for the analysis. We used a less stringent FDR < 0.10 as statistically significant enrichment in a GO term due to the correlation between statistical tests (e.g., due to highly similar lists of genes) as well as the more exploratory nature of gene set analysis.

RT-PCR Technical Validation.

RNA leftover from RNA-sequencing was used to create cDNA using the iScript cDNA Synthesis Kit (Bio-Rad). Because of the limited amount of RNA available and the number of RT-PCR targets, 20µl reactions in a 384 well format was chosen. Assays (including a no-template control) were performed in triplicate under standard real-time PCR conditions (50°C for 2 min, 95°C for 10 min, and 40 cycles of 95°C for 15 s, followed by 60°C for 1 min) using a real-time PCR system (Quant Studio 6; Thermo Fisher Scientific) with sequence-detection software. Inventoried TaqMan Gene Expression Assays (Thermo Fisher Scientific) were used to measure mRNA expression levels for *ADRB2* (assay ID Hs_00240532_s1), *ARNT* (assay ID Hs_00121918_m1), *CASP1* (assay ID Hs_00354836_m1), *HIF1α* (assay ID Hs_00153153_m1), *SIRT* (assay ID Hs_01009006_m1), and *TLR6* (assay ID Hs_01039989_s1). TATA box-binding protein (assay ID Hs00427620_m1) was used as the endogenous control based on previous work by our laboratory. We used linear regression to test for mean differences in mRNA levels by modeling each gene's $2^{-\Delta CT}$ value(E15) by smoking group, accounting for age, sex, and obesity.

Supplementary Text 3: Batch Effects and RUVSeq

Evaluating Batch Effects in Gene Expression Data.

Due to funding availability for this pilot study, RNA library preparation and sequencing were performed in two technical batches marked by distinct library preparation and sequencing methods. We observed a batch effect in our dataset, as reflected by greater variance in relative log-expression(E16) values in Batch 2 as well as clustering by batch in principal components analysis (PCA; **Figure S1**). The PCA plot below was constructed from model residuals from the edgeR quasi-likelihood model fit on smoking group, age, obesity, and sex (but excluding any batch correction), thus illustrating that a batch effect exists even after accounting for known participant features in the data.

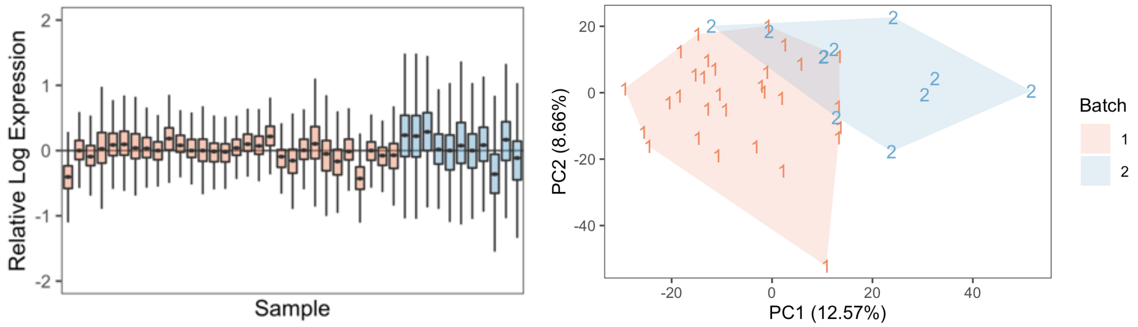


Figure S1. Visualization of batch effects using RLE and PCA plots. RLE (left): Each sample's relative log-expression (gene count log$_2$-transformed, minus the sample median) was plotted as a boxplot and colored by batch. There is increased variance in Batch 2 samples. **PCA (right)**: The resulting model deviance residuals for the top 2,500 genes by coefficient of variation were plotted by principal components 1 and 2. Batch 1 and Batch 2 clustered separately, suggesting the presence of a batch effect.

Assessing Batch Corrections and Use of RUVSeq.

Between-batch differences visually decreased in PCA and RLE plots after we included a fixed effect for batch (Batch 1 or 2) in edgeR modeling (full diagnostics not shown). However, there appeared to be other sources of unmeasured heterogeneity, as suggested by differences in residual variance between participants within a given batch. These may include residual confounding by batch, unmeasured technical artifacts (e.g., differences between RNA extraction dates), or unmeasured biological/clinical covariates (e.g., BAL cell differentials, smoking intensity). We thus used the RUVr procedure in RUVSeq(E17) v1.18.0 procedure) to infer covariates representing unmeasured sources of variation that are orthogonal to variables already in the model.

Number of RUVSeq Factors.

By iterating the number of RUVr components (*k*), we observed that using *k* = 3 components yielded an "elbow" or plateau in a plot of the variance explained in gene counts by covariates as a function of *k*. The percent variance was calculated using PERMANOVA using the adonis function in vegan(E18) v2.5-6 (1000 permutations, Euclidean distance matrix on log-transformed counts). The selection of *k* = 3 also minimized between-sample variability in different metrics (e.g., between-participant differences in the interquartile range of gene counts). Although increasing *k* continued to make samples more uniform and therefore may have captured more undesired variability, we erred on including less components / chose the "elbow" value to mitigate overfitting. Patterns in the resulting edgeR results following RUVSeq with *k* = 3 are presented below (**Figure S2**).

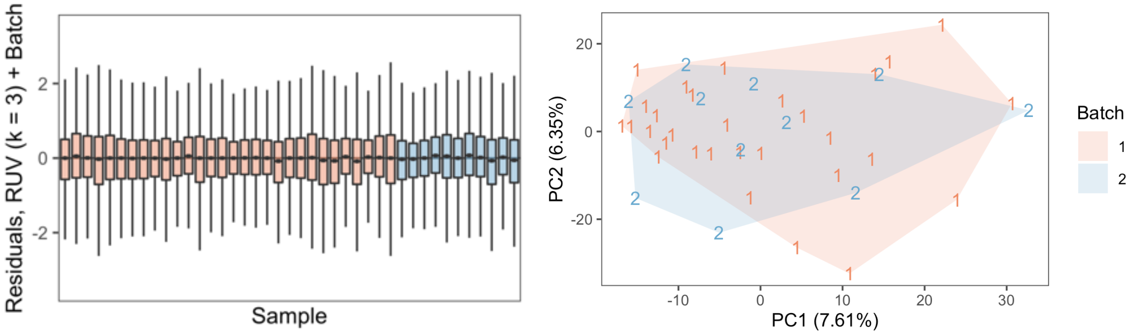


Figure S2. Between-batch patterns after adjusting for batch and RUVr components. Residuals (left): The distribution of edgeR residuals between participants and between batches appear more similar after including an adjustment for batch and adjustment for three RUV components. **PCA (right):** In principal components analysis, the batches no longer appear to cluster separately.

Correlations with RUV Components and Participant Features

As a standard procedure, our group examines correlations between RUVSeq components and measured clinical covariates (including features not included in primary regression models) to understand the sources of variation within our datasets. Interestingly, the RUVSeq components were correlated with BAL cell types, lung function, and metrics of marijuana/tobacco exposure (including THC metabolite concentration measured on a small subset of participants; **Table A**).

High correlations imply that a feature is associated with BAL gene expression, and that RUVSeq may implicitly correct for these features in our statistical modeling. However, as these inferred covariates are derived from factor analysis on the RNA-seq data, RUVSeq features may also represent unmeasured, unknown confounders that are not recorded in our clinical database.

Table B. Pearson correlations between participant features and RUVSeq components. The Pearson correlation between each feature and each of the three RUV components are presented below. For example, BMI had a weak negative correlation of -0.08 with the second RUV inferred covariate. Correlations with magnitude ≥ 0.3 are starred. Note that some variables were not measured in all participants (e.g., joint-years only applies to marijuana smoking participants; urine THC-Gluc was not measured in all participants), so the number of measurements used to calculate the correlation is displayed in the final column.

	RUV1	RUV2	RUV3	N
BMI	-0.10	-0.08	0.09	41
FEV1 (%p)	0.06	-0.18	-0.18	37
FEV1/FVC	-0.01	-0.41*	0.50*	22
FVC (%p)	0.06	0.06	-0.25	37
Lymphocytes (BAL Cell %)	-0.32*	-0.05	0.38*	39
Macrophage/Monocyte (BAL Cell %)	0.28	0.03	-0.39*	39
Marijuana Joint-Years	0.50*	-0.17	-0.03	15
Marijuana Smoking Frequency (Times/Day)	0.53*	0.06	-0.70*	12
Marijuana Vaping Frequency (Times/Day)	0.19	-0.15	-0.26	14
THC Concentration: BAL Cell	0.66*	0.05	-0.38*	12
THC Concentration: Urine THC-Gluc	0.26	0.05	-0.78*	6
THC Concentration: Urine THC-COOH	0.51*	0.15	0.05	13
Tobacco Pack-Years	0.03	0.06	0.24	15
Tobacco Smoking Frequency (Packs/Day)	0.09	-0.01	0.17	15

Supplementary Text 4: Sensitivity Analysis for Cell Composition

Rationale for Sensitivity Analysis.

Cell composition slightly differed between smoke exposure groups in our dataset (**Supplementary Text 1**). We observed a mean decrease in lymphocytes among marijuana smokers. While this decrease was not statistically significant, it is consistent with animal models(E19, E20). In addition, we expect different cell types likely have different gene expression profiles. Thus, because cell composition is associated with both our primary explanatory variable of interest (smoking group) as well as our outcome (gene expression), cell type is a potential confounder for the effect of smoke exposure on gene expression.

However, we worried that cell composition could be a mediator on the causal path between smoking group and BAL gene expression. That is, different smoke exposures could induce differences in cell composition, which then subsequently alters BAL gene expression. In general, adjusting for confounders is desirable but adjusting for putative mediators may mask smoke-expression associations(E21). However, if cell composition is a confounder, RUVSeq may implicitly control for it in the analysis (**Supplementary Text 2**).

While we hypothesized that cell composition was more likely to be a mediator, little is understood about the biological impact of chronic marijuana smoking at this time. We therefore performed the following sensitivity analysis to guard for cell composition being a confounder.

Approach.

Differential expression analysis was repeated after adding $\log_{10}$(lymphocyte % in BAL) to the edgeR model. Because cell differentials were not available on two participants, they were excluded from the analysis, resulting in a total sample size of N = 39. We elected to use lymphocyte %s (the second most prevalent cell type) in lieu of monocyte/macrophage %s because differentials for the former were approximately normally distributed after log-transformation whereas the latter was not. As most participants had >90% monocyte/macrophage cell differentials, this predictor would have had less variability and thus may have had less power to control for cell composition.

To account for unmeasured confounders beyond age, sex, obesity, batch, and $\log_{10}$(lymphocyte %), we derived k = 3 new RUVSeq components (same strategy to determine *k* as for the primary analysis; see **Supplementary Text 2**). We again used the edgeR quasi-likelihood model and F-test to assess genes with different average expression levels between groups after accounting for other covariates, and for comparison to the primary results. We also used the F-test to examine which genes varied on average with $\log_{10}$(lymphocyte percentage).

Results.

We detected 596 genes that significantly varied with $\log_{10}$(lymphocyte % in BAL) after accounting for other covariates ($FDR < 0.05$).

With $\log_{10}$(lymphocyte %) in the model, we detected 964, 2737, 3244 differentially expressed genes in respective MvC, MvT, and TvC pairwise comparisons ($FDR < 0.05$). Of the genes deemed significant in the primary analysis, 62.6%, 73.5%, and 95.4% were also statistically significant at a 5% FDR in this secondary analysis. However, there was high correlation between p-values and log-fold change values between analyses (**Figure S3**).

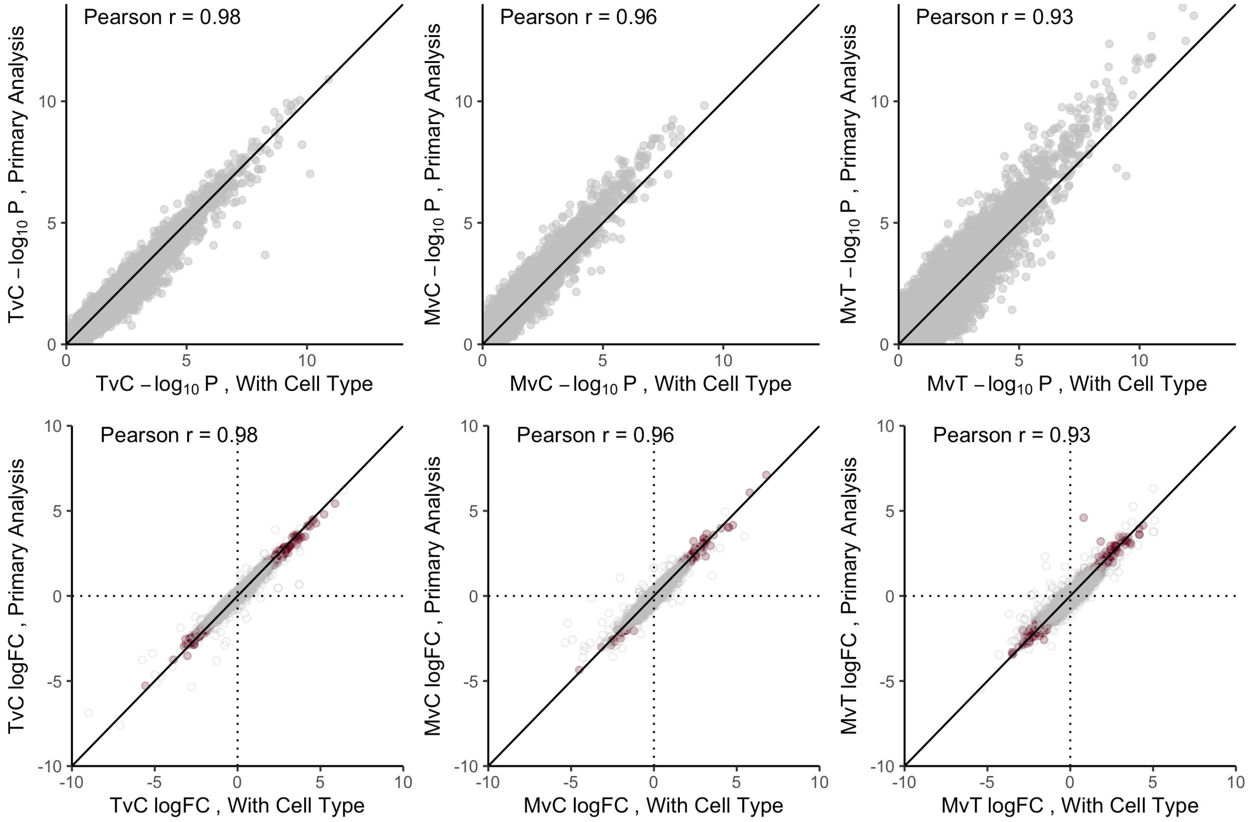


Figure S3. Pairwise comparisons with and without inclusion of $\log_{10}$(lymphocyte %). Top panels: concordance in p-values. Each point represents a gene's -$\log_{10}$(P-value) in the sensitivity analysis (x-axis) and primary analysis (y-axis) and is plotted against an identity line with intercept 0 and slope 1. **Bottom panels:** concordance in log fold change values. The log fold change in each analysis is displayed. Genes considered differentially expressed ($FDR < 0.05$) in the primary analysis have red, solid circles in contrast to grey open circles. Points in the lower-left or upper-right quadrants of these plots indicate that a gene had the same directionality in both analyses.

P-values were highly correlated between the primary analysis and the sensitivity analysis, although genes in the latter tended to have larger p-values than in the primary analysis (i.e., fewer genes significant). Log_2-fold change values were also highly correlated between the two analyses, with the effect direction of all significant differentially expressed genes in the primary analysis remaining the same in the sensitivity analysis.

The decreased number of results in the latter thus may be attributed in part to the decreased sample size due to participants excluded for missing cell differentials, as well as fewer degrees of freedom due to the added model parameter. Moreover, the effect of some differentially expressed genes may be decreased by masking of the smoke-expression association, if cell composition is a mediator. Regardless, due to high correlation between the two analyses, we conclude that our analysis appears to be robust to participant-level variability in cell composition.

Supplementary Text 5: Reproducing Tobacco vs. Control Alveolar Macrophage DEGs

Rationale

In an attempt to evaluate the external validity of our dataset, we compared our list of differentially expressed genes in tobacco smokers versus non-smoking controls to that reported in three other differential expression studies in alveolar macrophage. (At time of writing, we are not aware of other marijuana differential expression studies, outside of experimental models in other tissues/cell-types.)

Concordance between our BAL differential results and alveolar macrophage studies would help evaluate whether our methods, analysis procedure, and tobacco/non-smoker populations are potentially generalizable beyond our cohort. Moreover, we elected to focus on BAL cells without isolation of alveolar macrophages to minimize changes in gene expression due to *ex vivo* handling; while cell differentials indicated that we found majority alveolar macrophages and AM progenitors sin our BAL, and adjusted for cell type both implicitly and explicitly (RUVSeq, sensitivity analysis on lymphocyte count), overlap with existing AM literature would support that we are primarily measuring an AM signature in our BAL samples.

Three studies were found using literature searches in Google Scholar and the National Institutes of Health PMC database “differential expression” AND “alveolar macrophage” AND ("tobacco" OR "cigarette"). Studies required either the availability of raw data plus participant-level metadata (at minimum, smoker or non-smoker classification), or a table listing differentially expressed genes with smoking. All three studies evaluated below sequenced alveolar macrophages isolated from BAL using adherence-based methods. We compared the direction of tobacco effect and identity of differentially expressed genes between our results and these existing studies.

Woodruff, et al. (2005)

Woodruff and colleagues(E22) detected 124 microarray probes to be differentially expressed between smokers (n = 15) and non-smokers (n = 15). Participant ages were approximately a decade older on average than our study. The smokers had substantially higher pack-years than our cohort (47 average pack-years, standard deviation of 27) and were not necessarily "healthy smokers": the authors report that at least seven participants have clinical measurements indicative of mild to moderate COPD/emphysema.

Of the differentially expressed probes reported by the authors (their supplemental material Table E1; Bonferroni significance, obtained electronically June 2019), 97 probes mapped to gene symbols also tested in our dataset, across 73 unique genes. As our analysis was conducted on the gene level, for comparability we took the probe with the lowest p-value to represent the gene.

We sorted the 73 genes into the 2x2 table on the next page. Each gene was classified based on whether it was significant in our analysis (yes/no) and if the direction of effect reported was the same in our cohort (same/different).

		Significant in Our Cohort		
		yes	no	
Direction	same	54	17	71
	different	0	2	2
		54	19	73

An example interpretation of this 2x2 table is as follows. Of the genes significant in the Woodruff analysis and also present in our analysis, 54/73 = 74% were also significant in our dataset. Of the genes deemed significant in both datasets, all 54/54 had the same direction of effect. Of the 19 that were not significant in our dataset, 17/19 = 89% had the same effect.

Shaykhiev, et al. (2009)

We also compared tobacco smokers versus non-smoker genes to Shaykhiev, et al.(E23). The age of non-smokers was approximately a decade older on average. Smokers had approximately 17 more pack-years on average than our cohort (27 average pack-years, standard deviation of 18). In addition, most participants in our cohort identified as non-Hispanic white race/ethnicity whereas the Shaykhiev cohort was majority black.

As these authors did not publish a full list of differentially expressed genes between healthy smokers and healthy non-smokers (focusing instead on M1/M2 macrophage polarization genes), we downloaded the metadata and raw Affymetrix Human Genome U133 Plus 2.0 microarray data publicly shared by the authors in GEO (accession GSE13896; obtained June 2019) and re-analyzed the data. Participants with COPD and one smoker missing age metadata were excluded from this analysis, resulting in n = 24 controls and n = 33 smokers included. We used the oligo package(E24) v1.48.0 to read .cel files into the R statistical computing environment, apply Robust Multiarray Average(E25) normalization, and summarize probes into probesets. Genes with high expression (in the top 75th percentile of expression, quantified by median across-participant $\log_2$-intensity) were tested for differences between smokers and non-smokers. To compare features at the gene-level, when multiple probesets were assigned to the same gene symbol, we excluded the probeset with lower higher coefficient of variation from statistical testing. Differential expression was assessed using limma(E26) v3.40.6, adjusting for age, sex, ancestry, and k = 5 RUV components (where *k* was determined by same method described in **Supplementary Text 3**). We ultimately detected 3,446 differentially expressed genes between smokers and non-smokers at FDR < 0.05 in the Shaykhiev dataset, 2,579 of which were also tested in our dataset.

The 2,579 genes that were (i) statistically significant in the Shaykhiev dataset and (ii) tested in our cohort were then again classified into a 2x2 table. Most tobacco-associated genes had the same effect direction in our dataset (73% among genes significant in both cohorts). There was limited overlap in the genes deemed significant in both analyses (26%).

		Significant in Our Cohort		
		yes	no	
Direction	same	486	1132	1618
	different	179	782	961
		665	1914	2579

Morrow, et al. (2019)

Morrow, et al.(E27) discusses RNA-sequencing from COPD-diagnosed participants and controls from the COPDgene study. The manuscript focuses on comparing COPD-relevant traits across multiple tissues, including tests of differential expression in alveolar macrophage with smoking status, accounting for age, sex, race, and surrogate variable analysis factors. Smoking status was coded as an ordinal variable: never smoker (n = 4), former smoker (n = 10), and current smoker (n = 7). Participants were older than our cohort, with a mean of >54 years old at enrollment.

The 1,056 smoking-associated genes that were (i) statistically significant (q-value < 0.05; their electronic supplement Text Table S5, "smoking" tab; obtained June 2019) and that (ii) were tested in our cohort are classified below. There was high concordance with effect direction (99% of genes significant in both cohorts had same direction).

Significant in Our Cohort

		yes	no	
Direction	same	279	613	892
	different	2	162	164
		281	775	1056

Summary of Tobacco versus Non-Smoker Comparisons

There was high concordance in tobacco smokers and non-smokers between our BAL expression study and existing studies in alveolar macrophage. For all three studies examined, the majority of previously identified genes had the same direction in our study. However, previously identified genes may not have been statistically significant in our study, potentially in part due to limited sample sizes (small in all studies considered; largest sample size was Shaykhiev, et al. at n = 57 total) as well as true differences between the populations studied. These differences in populations are a key caveat, as these existing macrophage studies did not use similar inclusion/exclusion criteria. Namely, we emphasized obtaining BAL gene expression from “healthy” participants with no underlying lung disease.

Supplementary Text 6: RT-PCR Results

Summary

Thirty-nine of the 41 participants analyzed had sufficient RNA remaining for technical validation of six clinically relevant genes using RT-PCR, four of which are discussed in the research letter. This supplementary text shows the RNA-sequencing and RT-PCR results side-by-side, including visual examination of each gene (**Figure S4**) and a summary of effect sizes after adjusting for age, sex, and BMI (**Table B**). The RT-PCR data is presented in the form $2^{\Delta Ct}$ so both RNA-seq and RT-PCR data on a linear scale roughly proportional to the number of mRNA copies, rather than a logarithmic scale. Here, ΔC_T is defined as the C_T in the gene of interest minus the C_T for TATA-Box Binding Protein (TBP), our RT-PCR endogenous control for normalization.

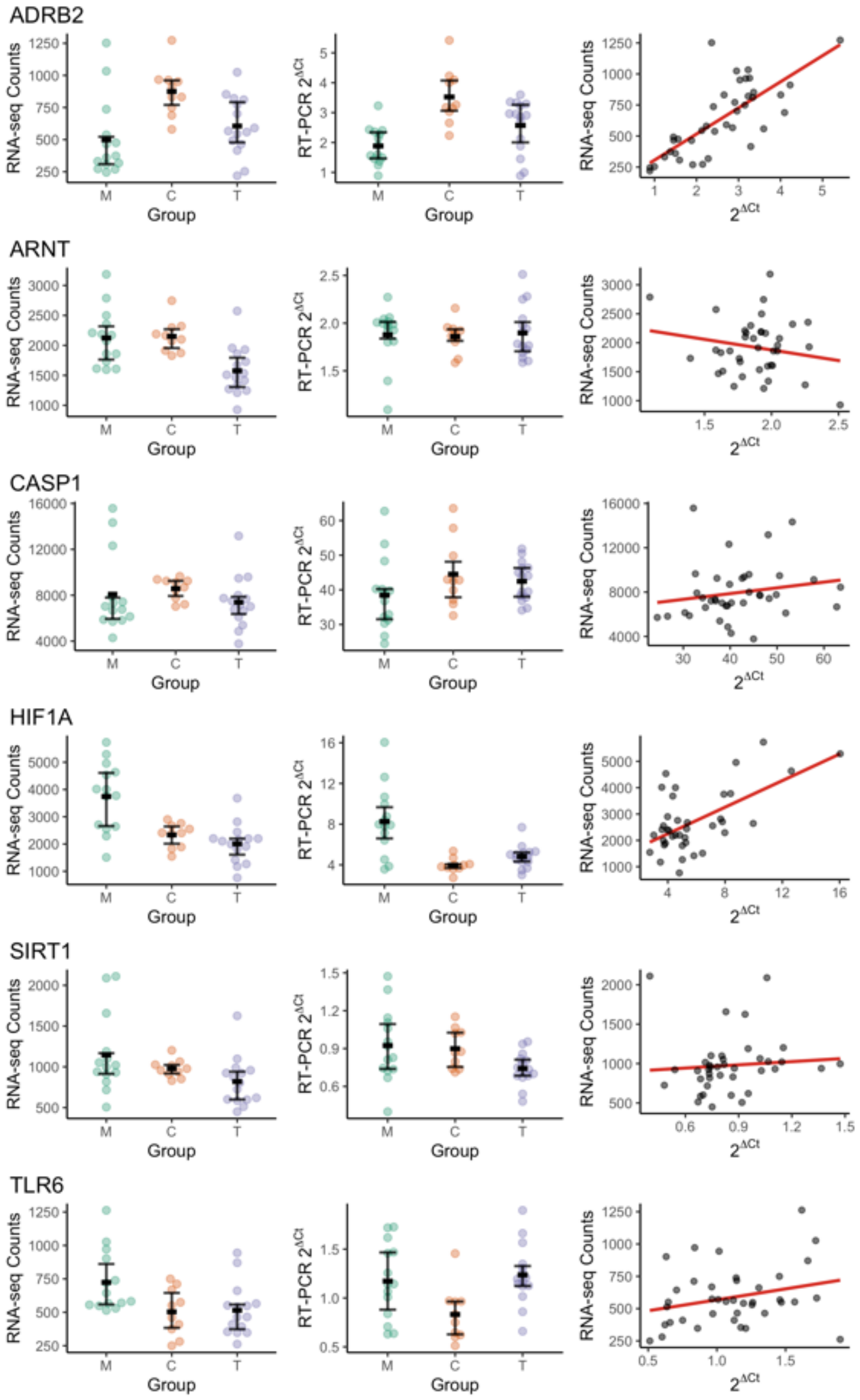
In the subsequent figures and tables, we use the abbreviations for marijuana smokers (M), tobacco smokers (T) and non-smoker controls (C).

Table C. **Log$_2$-fold changes in gene levels, adjusting for covariates**. The log$_2$-fold changes measured by RNA-sequencing counts (edgeR effect size) and by RT-PCR copies ($2^{\Delta Ct}$ linear model), after adjusting for age, sex, and obesity. The log-fold change is negative if the first group has lower expression relative to the second; for example, ADRB2 is lower in marijuana smokers than controls. The edgeR model also adjusts for batch and RUVr components. An asterisk is shown if the fold-change is statistically significant. The Pearson correlation between counts and copies is also reported.

Gene	**RNA-Seq**			**RT-PCR**			**Pearson *r***
	MvC	MvT	TvC	MvC	MvT	TvC	
ADRB2	-1.08*	-0.69 *	-0.39	-0.91 *	-0.36	-0.55 *	0.75
ARNT	0.02	0.28 *	-0.26 *	0.05	0.00	0.05	-0.20
CASP1	-0.28 *	-0.40 *	0.12	-0.30	-0.16	-0.14	0.18
HIF1A	0.72 *	0.88 *	-0.16	0.86 *	0.79 *	0.07	0.60
SIRT1	0.00	0.19 *	-0.19 *	0.04	0.20	-0.16	0.081
TLR6	0.67 *	0.47 *	0.20	0.40 *	0.11	0.29	0.27

Results

Figure S4. Visual examination of genes measured by both RNA-seq and RT-PCR. From left-to-right, the six figures depict: **(i)** RNA-seq count times the edgeR normalization factor to account for sequencing depth), overlaying the median and interquantile range. **(ii)** $2^{\Delta Ct}$ values for the gene. **(iii)** A scatter plot of RNA-seq and RT-PCR copies, with a fitted linear trend line.



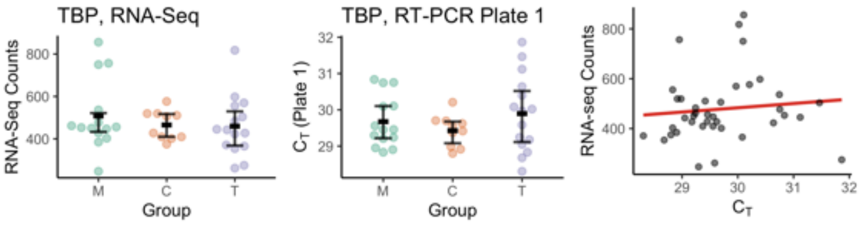
Summary

The direction of effect between groups was highly consistent for the majority of genes, and the majority of comparisons. In particular, we find that the marijuana versus control (MvC) comparisons are very similar between RNA-seq and RT-PCR. However, we did not consistently reproduce the direction of effect between groups for some smaller effect sizes (e.g., tobacco versus control comparison for *CASP1*) and some lowly expressed genes (small *ARNT, SIRT,* and *TLR6* ΔC_T values across all participants).

A Note on RT-PCR Reference Gene. As standard practice, we examined whether levels of the endogenous reference gene (TBP) appeared to vary in its distribution by smoking group. TBP C_T-values had a higher mean and variance among the marijuana smokers and tobacco smokers. However, there was no statistically significant between-group difference (ANOVA F-test P = 0.38, or P = 0.86 after correcting for covariates). This pattern was also present in the RNA-seq counts for TBP, albeit not as conspicuously (**Figure S5**). TBP was not statistically significant in the RNA-seq analysis (all pairwise comparisons P > 0.11, FDR > 0.275).

We have previously used this marker in the context of alcohol use disorders(E28), but after performing the RT-PCR assay found a reference suggesting that TBP is highly variable among alveolar macrophage derived from COPD patients(E29). It is unclear if TBP is thus truly constitutively expressed among our participants and between BAL in smokers in general, but this technical artifact may be one cause of the limited replication of low C_T genes and lower effect sizes. The need for reference genes is also a limitation of reference-based approaches like RT-PCR. Unfortunately, no further RNA is available to repeat these assays with another reference gene.

Figure S5. Levels of RT-PCR endogenous control/housekeeping gene TBP. Figure S4, but with TBP. In lieu of a $2^{\Delta Ct}$ (not relevant since TBP is the reference gene), C_T is presented in the second figure; note that this understates the differences between smoking groups because C_T is on a logarithmic scale rather than a linear copy scale.



References

E1. Bailey KL, Wyatt TA, Katafiasz DM, Taylor KW, Heires AJ, Sisson JH, Romberger DJ, Burnham EL. Alcohol and Cannabis Use Alter Pulmonary Innate Immunity. *Alcohol* 2018;doi:10.1016/j.alcohol.2018.11.002.

E2. Fini MA, Gaydos J, McNally A, Karoor V, Burnham EL. Alcohol abuse is associated with enhanced pulmonary and systemic xanthine oxidoreductase activity. *Am J Physiol - Lung Cell Mol Physiol* 2017;313:L1047–L1057.

E3. Afshar M, Burnham EL, Joyce C, Clark BJ, Yong M, Gaydos J, Cooper RS, Smith GS, Kovacs EJ, Lowery EM. Cut-Point Levels of Phosphatidylethanol to Identify Alcohol Misuse in a Mixed Cohort Including Critically Ill Patients. *Alcohol Clin Exp Res* 2017;41:1745–1753.

E4. Gaydos J, McNally A, Guo R, William Vandivier R, Simonian PL, Burnham EL. Alcohol abuse and smoking alter inflammatory mediator production by pulmonary and systemic immune cells. *Am J Physiol - Lung Cell Mol Physiol* 2016;310:L507–L518.

E5. Jiang H, Lei R, Ding SW, Zhu S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 2014;15:1–12.

E6. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–527.

E7. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, *et al.* Ensembl 2018. *Nucleic Acids Res* 2017;46:754–761.

E8. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 2016;4:1521.

E9. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009;4:1184–1191.

E10. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;26:139–140.

E11. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;32:896–902.

E12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995;

E13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: Tool for the unification of biology. *Nat Genet* 2000;25:25–29.

E14. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 2019;47:W199–W205.

E15. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods* 2001;25:402–408.

E16. Gandolfo LC, Speed TP. RLE Plots: Visualising Unwanted Variation in High Dimensional Data. 2017;1–9.doi:10.1371/journal.pone.0191629.

E17. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;32:896–902.

E18. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. vegan: Community Ecology Package. 2016;at <http://cran.r-project.org/package=vegan>.

E19. Ignatowska-Jankowska B, Jankowski M, Glac W, Swiergel AH. Cannabidiol-induced lymphopenia does not involve NKT and NK cells. *J Physiol Pharmacol* 2009;60:99–103.

E20. Khuja I, Yekhtin Z, Or R, Almogi-Hazan O. Cannabinoids reduce inflammation but inhibit lymphocyte recovery in murine models of bone marrow transplantation. *Int J Mol Sci* 2019;20:1–16.

E21. Schisterman EF, Cole SR, Platt RW. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiology* 2009;20:488–495.

E22. Woodruff PG, Koth LL, Yang YH, Rodriguez MW, Favoreto S, Dolganov GM, Paquet AC, Erle DJ. A distinctive alveolar macrophage activation state induced by cigarette smoking. *Am J Respir Crit Care Med* 2005;172:1383–1392.

E23. Shaykhiev R, Krause A, Salit J, Strulovici-Barel Y, Harvey B-G, O'Connor TP, Crystal RG. Smoking-Dependent Reprogramming of Alveolar Macrophage Polarization: Implication for Pathogenesis of Chronic Obstructive Pulmonary Disease. *J Immunol* 2009;183:2867–2883.

E24. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 2010;26:2363–2367.

E25. 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. *Sel Work Terry Speed* 2012;601–616.doi:10.1007/978-1-4614-1347-9_15.

E26. 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.

E27. Morrow JD, Chase RP, Parker MM, Glass K, Seo M, Divo M, Owen CA, Castaldi P, Demeo DL, Silverman EK, Hersh CP. RNA-sequencing across three matched tissues reveals shared and tissue-specific gene expression and pathway signatures of COPD. *Respir Res* 2019;20:1–12.

E28. Burnham EL, Phang TL, House R, Vandivier RW, Moss M, Gaydos J. Alveolar Macrophage Gene Expression Is Altered in the Setting of Alcohol Use Disorders. *Alcohol Clin Exp Res* 2011;35:284–294.

E29. Ishii T, Wallace AM, Zhang X, Gosselink J, Abboud RT, English JC, Paré PD, Sandford AJ. Stability of housekeeping genes in alveolar macrophages from COPD patients. *Eur Respir J* 2006;27:300–6.