2019

Approaches for identifying lung cell type responses to perturbation

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

APPROACHES FOR IDENTIFYING LUNG CELL TYPE RESPONSES TO PERTURBATION

by

SEAN EDWARD CORBETT

B.A., Clark University, 2011
M.S., Boston University, 2016

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader _____

Avrum Spira, M.D., M.Sc.
Professor of Medicine

Second Reader _____

Joshua Campbell, PhD
Assistant Professor of Medicine

DEDICATION

This dissertation is dedicated to Claudette. Je t'aime.

ACKNOWLEDGMENTS

My Academic Family **To my advisors, Marc and Avi:** thank you for your unending patience, your willingness to accept my quirks and intellectual particulars, for knowing when to hold my hand and when to press me stand to on my own, and for sharing all the powerful pieces of wisdom I needed to get by, at just the right moments. I feel incredibly fortunate to have been able to grow as a scientist and thinker under your guidance.

To Josh: it seems like there are paltry few moments during one's doctoral work when fortune provides you with a project that is exciting enough to inspire you to push on, and the opportunity to work with you was precisely that. Thank you for your collaboration, good cheer, kindness, and unfathomably abundant patience. It was an honor to have the opportunity to work with you in the late phases of my doctorate.

To Sarah: your candor and advice were deeply important to me, and helped me survive in times of great inner turmoil. I'm thrilled to have watched you excel both as a postdoc and as faculty, and I can't wait to see what's next.

To Matt: your passion for your work as a physician is infectious. I hope someday to have fostered half of your ability to distill complex topics into something engaging and fun, and even a quarter of your diligence. I couldn't have asked for a better partner-in-crime.

To Iris: thank you for your humor, your boldness, and for lending me your

mathematical expertise on so many occasions. You served as a peer mentor to me at a point when I was increasingly afraid to ask for help out of embarassment; for this reminder that everyone around me is willing to help, I'm deeply grateful.

To Xu Ke, Jiarui, Beth, Xingyi, Zhe, Aaron, Reid and Yusuke: your collective talent, vivacity, passion for good science, and curiosity kept me on my toes through the phases of my doctorate where I felt like I had the most momentum. I'm excited to see the fruits of each of your respective academic voyages, though I'm already very proud of the growth you've all shown while we've worked together. Thank you for all of the tangential conversations, pushup contests, late-night pizza orders, and snark.

My Actual Family **To my siblings, Cat, Jimi and Bridget:** thank you for always being there as the only people on the planet capable of understanding my humor. I know our lives have become mutually unintelligible in many ways, but I'm glad you all are along for the ride regardless.

To my father Ed: thank you for reminding me not to worry about the stupid stuff.

My Queer Family **To Nick, my best friend:** I (in some ways literally) could not have survived my doctorate without your companionship. Our friendship has been a continual process of us enabling each other to become ourselves, and I hope that that never changes.

To Kevin, Allan, Leo, Brenden Michael, Paul, and Nick R: befoe meeting you, I did not know who I wanted to be, really. Now I do, and for that I am deeply grateful.

To the rest my queer family: Dusty, Shane, Alex, Chris, Kyle, Larwence, all three Mikes, Joey, Jason, Sam, Brenden, Mark, Tom S, Eric, Aditya, Phaedra, Evan, Aaron S, Brian, David, Flicky, Donald, Brett, Jake (like the Snake), Aaron C, Clark, Moon, Michael, George, Chad, Alpen, Victor, Nolan, Derek, Tom B, Tom O, Pete, Chad, Courtney, Tim, Jarrod, all of the innumerable faeries and freaks and forebears there wasn't room to mention, and Shannon Michael Cane. I had not realized what was most joyous about life before I met you, and I am immensely happy to have all of you in my life.

APPROACHES FOR IDENTIFYING LUNG CELL TYPE RESPONSES TO PERTURBATION

SEAN EDWARD CORBETT

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2019

Major Professor: Avrum Spira, Professor of Medicine, Professor of Pathology & Laboratory Medicine, Alexander Graham Bell Professor in Health Care Entrepreneurship

ABSTRACT

The use of genomic profiling can provide indications of underlying molecular responses to chemical perturbation, and the characterization of these responses can provide an increased understanding of the greater physiological effects of an exposure and inform clinical decision making. This approach has proven to be effective in understanding the effects of environmental exposures such as cigarette smoke on the airway epithelium, and how they may contribute to associated disease pathogenesis. Because of the existing body of work in genomic profiling towards understanding the effects of environmental exposures, it has relevant applications towards the study of the effects of emerging exposures such as electronic cigarettes, which remain poorly understood. Further, current approaches for genomic profiling could be improved through the development of data resources and computational methods which can identify not only tissue- or sample-level molecular responses to perturbation, but also responses specific to individual cells or cell types.

In light of these issues, I investigated the molecular response in airway

epithelium to a novel inhaled exposure, and developed methods to support more detailed characterization of such effects. In this dissertation, I describe a clinical observational study in which I examined the gene expression effects of electronic cigarettes on the airway epithelium, and compare these effects to those of conventional cigarettes (Aim 1). Next, I describe CELDA, a novel computational method for identifying cell subpopulations and the co-expressed modules of genes that identify them in single cell RNA-seq (scRNA-seq) data (Aim 2). Finally, I describe the Lung Connectivity Map (Lung CMap), a platform for interrogating lung cell type specific responses to a large set of chemical and molecular perturbations (Aim 3).

Collectively, this work encompasses both observational and computational approaches for detailed characterization of the molecular responses to perturbation, and the determination of the relative effects of these novel perturbations versus their more well-described counterparts.

CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ATCC	American Type Culture Collection
ANCOVA	Analysis of Covariance
CC	Replicate correlation; the correlation of the signatures produced by the biological replicates of a perturbagen
CELDA	Cellular Latent Dirichlet Allocation
CO	Carbon monoxide
CMap	Connectivity Map, a dataset of perturbational gene expression signatures generated by Subramanian et al.
CRISPR	Clustered regularly interspaced short palindromic repeats
CSNK1A1	Casein Kinase 1 Alpha 1
ECIG	Electronic cigarette, also known as a "vape" or vaporizer. Member of a class of devices known as Electronic Nicotine Delivery Systems, which are designed to deliver nicotine in an aerosolized mixture
FDR	False discovery rate
GEO	Gene Expression Omnibus
GSVA	Gene set variation analysis
HDAC	Histone deacetylase
L1000	Landmark 1000, a bead-based gene expression assay directly measuring the expression of 978 genes from which the expression of the rest of the transcriptome can be inferred
LCMap	Lung Connectivity Map
LDA	Latent Dirichlet Allocation

mRNA	Messenger RNA
NCS	Normalized Weighted Connectivity Score
PBMC	Peripheral blood mononuclear cell
PCA	Principal Components Analysis
RIN	RNA integrity number
RNA	Ribonucleic acid
scRNA-seq	Single-cell RNA sequencing
SS	Signature strength; the number of genes differentially expressed between a perturbation and its nearest DMSO control
STR	Short tandem repeat
t-SNE	tDistributed Stochastic Neighbor Embedding
TAS	Transcriptional Activity Score
Tau	A measure of similarity between a query gene set and a reference signature
TCIG	Tobacco cigarette
WTCS	Weighted Connectivity Score

CHAPTER 1
Introduction

1.0.1 The Transcriptome and Exposome in Human Health

The rapid advance in sophistication of methods in molecular biology has proven to be a continual boon towards enabling the study of the underlying mechanisms influencing human health and disease. Since the advent of DNA sequencing techniques more than 50 years ago, researchers have been able to investigate the specific roles of DNA and its associated molecules in both normal and pathogenic function of human tissues (Heather & Chain (2016)). RNA, the transcribed form of DNA, can perform many roles at a cellular level, such as serving as an intermediate molecule towards protein synthesis (mRNA), or by directly exerting a regulatory role (ncRNA). In a complementary fashion to genetic (DNA) profiling, the profiling of different species of RNA molecules in a cell (also known as transcriptomics) can provide crucial mechanistic insights with applications in basic biology, diagnostics, and personalized medicine.

One of the areas in which transcriptomics has proven to be particularly efficacious is in the field of exposomics. Exposomics encompasses research at the intersection of transcriptomics and the effects of environmental factors on human health, such as pollution or lifestyle choices (Vrijheid (2014)). For example, previous work has shown that there are discernable transcriptomic signatures in the airway epithelium associated with cigarette smoking (Spira et al. (2004)) and chronic obstructive pulmonary disease (Steiling et al. (2013)), and in buccal epithelium associated with inhalation of smoky coal (Wang et al. (2015)). This existing body of work serves not only as evidence that transcriptomics can provide important insights into the underpinnings of environmental exposures, but serves as a refer-

ence point for the study of yet-uncharacterized exposures which may have relative similarities and differences in terms of their transcriptomic effects.

1.0.2 Gene Expression Microarrays for Transcriptomic Profiling

Since the early 1990s, a rapid advance in the development of novel transcriptomic techniques have enabled the study of the human transcriptome with increasing speed, accuracy, and detail (Lowe et al. (2017)). These approaches include oligonucleotide ligation assays such as gene expression microarrays and bead-based platforms, as well as sequencing-based methods such as RNA-seq and single-cell RNA-seq (scRNA-seq).

High-density gene expression microarrays began to increase in popularity starting in 1995, and subsequently became the most common platform for transcriptomic assays (Lowe et al. (2017)). Microarrays presently represent a rapid, mature platform for transcriptomics experiments, with comparatively robust analytical conventions surrounding interpretation of their data. Microarray products, such as the Affymetrix Human Gene ST 1.0 arrays, function by hybridizing preparing RNA isolated from a biospecimen to pre-synthesized oligonucleotide probes arranged in a pre-specified pattern on a glass slide. These oligonucleotides are designed to complement the sequence of known transcripts from the human genome, and in the case of the Human Gene ST 1.0, cover upwards of 20,000 genes. A microarray scanner is then used, which is able to measure the amount of RNA from the provided sample hybridized to each transcript's corresponding probes probes via the amount of fluorescence emitted by hybridized probes. These fluorescence intensities can be compared to background intensities to derive a measure of how much of a given transcript was in the original sample, and allow a researcher to capture a gene expression profile for the condition in question (Gohlmann & Tal-

loen (2009)).

1.0.3 Bead-Based Transcriptomic Assays

Since the development of microarrays, bead-based oligonucleotide assays have emerged as a cheaper, higher throughput alternative. These platforms, most notably the Luminex®xMAP®protocol, leverage a similar oligonucleotide probe hybridization approach as microarrays, with the probes mounted on microbeads rather than a glass slide (Graham et al. (2019)). Using a mixture of beads, a similar fluorenscence based measure of gene expression can be derived. This approach can be performed at a much smaller experimental scale than microarrays, and often more cheaply, enabling highly multiplexed transcriptional profiling of many samples. One major disadvantage to this bead-based approach is that often only a subset of genes in the human genome are represented by the included oligonucleotide probes. This limitation can be circumnavigated by careful selection of probes, such as the L1000 approach (Subramanian et al. (2017)), in which the measurement of a core set of 978 genes can be used to linearly infer the expression of all unmeasured genes in the human genome.

1.0.3.1 The Connectivity Map

Subramanian *et al.* leveraged the speed and low cost of this assay to produce the Connectivity Map (CMap), a large database of gene expression profiles of many perturbagens such as cancer therapeutics, CRISPRs, and tool compounds exposed to a heterogeneous mix of cancer cell lines under a variety of experimental conditions (e.g. dose points, exposure length). Using a unique set of gene set enrichment methodologies and tooling, CMap enables researchers to investigate connections between exposomic perturbagens, as well as identify which of the assayed per-

turbagens may produce similar transcriptomic effects to a novel perturbagen of interest (Subramanian et al. (2017)).

An important limitation to the original CMap design is the lack of exclusion of non-cancer, non-immortalized cell lines. Because each perturbation in the dataset was performed in cancer cell lines, CMap's signatures may not be representative of peturbational responses in more phenotypically normal cell lines, and thus in more phenotypically normal tissues. It is possible that signatures of CMap perturbagens in non-cancer cell types may diverge from the signatures of the same perturbagens in cancer cell lines. A lack of perturbational signatures in more phenotypically normal cell lines thus could inhibit the discovery of cell-type responses to compound perturbation, which could reduce the utility of CMap for therapeutic discovery or development purposes.

1.0.4 Sequencing-Based Methods for Transcriptomic Profiling

RNA-seq, and more recently scRNA-seq, has begun to become increasingly popular in transcriptomic profiling. While DNA sequencing was prohibitively expensive at the inception of next-generation sequencing techniques, a rapid decrease in the cost of sequencing per genome has allowed researchers to leverage sequencing techniques in increasingly large-scale experiments (DNA). Recent advances in sequence library preparation have even enabled transcriptomic study of single cells, most often through single cell RNA-seq (scRNA-seq) protocols (Kalisky & Quake (2011)). These single-cell transcriptomic assays provide unparalled detail in the study of the gene expression mechanisms in human biospecimens, such as capturing the cellular heterogeneity and potential therapeutic implications in glioblastoma tumors in the brain (Patel et al. (2014)). Oligonucleotide probe based methods such as those previously described generally requires pooling of RNA

from all cells in a sample, and thus are often unable to capture cell-type-associated transcriptomic effects which could have crucial implications towards the interpretation of a system of interest.

While scRNA-seq protocols are powerful in the level of transcriptomic detail they can provide, the landscape of analytical methods for interpreting scRNA-seq data is still nascent. There is a lack of field-wide agreement on the best analytical methods for interpreting technical effects such as batch effects or dropouts, though there are a plethora of currently available options (Rostom et al. (2017)). An important question in analyzing scRNA-seq data is how to determine which subpopulations of cells are most similar in a scRNA-seq dataset. Understanding which subpopulations of cells exist in a sample and the relative proportions of these cell types has important implications towards the downstream biological interpretation of these datasets.

1.0.5 Dissertation Aims

The body of work in this dissertation aims to extend transcriptomic methods towards the characterization of novel inhaled exposures, as well as to develop analytical methods to capture cell-type-associated responses to perturbation via both high-throughput bead-based gene expression assays and scRNA-seq.

Aim 1: Utilize a whole-transcriptome profiling approach to study the effect of electronic cigarettes on the bronchial airway epithleium, and compare it to established signatures of smoking in order to contextualize its health effects relative to a more well understood exposure. The rapid increase in the use of electronic cigarettes (ECIGs) over the past decade despite a poor understanding of their effects remains an important open question for debate in the public health sphere. It is unclear what the short- and long-term effects of ECIG use in regards

to airway health may be, and it is further unclear how these effects may compare to those induced by conventional cigarettes. I will describe a clinical observational study profiling the transcriptomic effects of ECIGs in the bronchial airway epithelium. I will also describe how the transcriptomic profile of ECIGs compares to the profiles induced by both active cigarette use as well as cigarette cessation in the same tissue.

Aim 2: Develop a method for clustering cells in single cell RNA-seq data, and demonstrate its efficacy towards identifying subtle shifts in cell subpopulations which can provide crucial insights. I will describe a novel method for identifying cell subpopulations as well as the co-expressed modules of genes that define them, developed in collaboration with Dr. Joshua Campbell. I will further describe the application of this method to a publically available scRNA-seq dataset of peripheral blood mononuclear cells, and its efficacy in identifying small but biologically relevant cell subpopulations.

Aim 3: Develop a platform for wide-scale perturbational profiling of non-cancer lung cell lines, in order to determine the feasability of this approach in identifying lung cell-type-associated responses to compound perturbation. I will detail the development of the Lung Connectivity Map, a CMap-style dataset profiling the transcriptomic effects of compound exposures in primary, non-cancer lung cell types. I will show how this system demonstrates lung-cell-type-associated responses to compound perturbation, and detail how this finding argues for the inclusion of a wider variety of cell types in future expansions of the CMap dataset.

CHAPTER 2

Gene Expression Alterations in the Bronchial Epithelium of Electronic Cigarette Users

Adapted from the following manuscript:

Sean E. Corbett*, Matthew Nitzberg*, Elizabeth Moses, Eric Kleerup, Teresa Wang, Catalina Perdomo, Claudia Perdomo, Gang Liu, Sherry Zhang, Hanqiao Liu, David A Elashoff, Daniel R Brooks, George T O'Connor, Steven M Dubinett, Avrum Spira, Marc E Lenburg *(Manuscript under review)*

* Contributed equally

2.1 INTRODUCTION

Since their introduction in 2007, the use of electronic cigarettes (ECIGs) in the United States hs increased. Survey data indicates that 15.4% of adults 18 and older have tried ECIGs, with adults between the ages of 18-24 being the most likely to have used ECIGs at 23.5% (CDCMMWR (2017)). Current smokers and those who had recently quit were more likely to use ECIGs than never smokers. And, while ECIGs are not approved by the FDA for cessation, 55.4% of conventional tobacco cigarette (TCIG) smokers who tried to quit smoking in the previous year had tried an ECIG and 22% continued to actively use ECIGs (noa (2016)).

Despite the increased use of ECIGs (King et al. (2015)), their safety profile remains controversial. Popular perception and marketing of ECIGs assert that they are a safer alternative to TCIGs (King et al. (2013); Martínez-Sánchez et al. (2015)). In contrast, numerous studies have illuminated potential adverse health effects of ECIGs such as diminished cough reflex (Dicpinigaitis et al. (2016)), cytotoxicity from ECIG flavorings (American Thoracic Society (ATS) (2015)), endothe-

lial disruption (Schweitzer et al. (2015)), and DNA damage (Lee et al. (2018)). While a switch from TCIGs to ECIGs results in a reduction of users exposure to known toxicants and carcinogens (National Academies of Sciences, Engineering, and Medicine (2018)), it is difficult to reach a consensus on the safety of electronic cigarettes other than that their health risks likely lie somewhere between TCIG smoking and nonsmoking (Dinakar & O'Connor (2016)).

It is well-established that TCIG use leads to gene expression changes in airway epithelium. We have previously described these changes in bronchial and small airway epithelium of TCIG smokers (Hackett et al. (2012); Harvey et al. (2007); Hübner et al. (2009); Spira et al. (2004)), as well as how these patterns change after smoking cessation (Beane et al. (2007a); Chari et al. (2007)). TCIG-associated gene expression changes in the nasal epithelium overlap with changes seen in the bronchial epithelium demonstrating common TCIG-associated gene expression effects throughout the airway (Steiling et al. (2008)). Changes in airway gene expression have also been associated with lung diseases such as lung cancer (Silvestri et al. (2015); Spira et al. (2007)) and chronic obstructive pulmonary disease (Steiling et al. (2013)).

Prior work suggests that ECIG exposure can also alter airway gene-expression. We have described oxidative and xenobiotic stress associated gene-expression changes in an in vitro model of ECIG exposure, as well as ECIG dose-response changes in a marker of reactive oxygen species (Moses et al. (2017)). Work by Martin et al measured the expression of a target panel of immune and inflammation associated genes in cells collected from nasal brushings of TCIG smokers, ECIG users, and non-smokers (Martin et al. (2016)), finding that most of these genes were specifically downregulated by ECIG use.

To more comprehensively characterize the effects of ECIG use on airway gene-

expression, and to better understand their relationship with the effects of TCIGs, we sought to compare the transcriptome-wide impact of ECIGs and TCIGs in the bronchial airway epithelium. More specifically, we sought to investigate the gene expression effects of ECIG use in former smokers, as daily ECIG users are more likely to be former TCIG smokers than dual-users or de novo users (Coleman et al. (2017)). Toward this end, we recruited current TCIG smokers, former TCIG smokers using ECIGs, and former TCIG smokers to undergo voluntary bronchoscopy. Bronchial epithelial cells were collected in order to identify gene-expression changes associated with ECIG use and to compare the transcriptomic effects of ECIG and TCIG use.

2.2 MATERIALS AND METHODS

2.2.1 Study Population and Sample Collection.

We recruited volunteers at Boston University Medical center and the University of California Los Angeles Medical Center between December 2013 and March 2015. All participants were aged 18-55 years and were current or former TCIG smokers with a smoking history of at least 5 cigarettes per day for at least two years. Participants were excluded if they were using tobacco products such as chewing tobacco, snuff, hookah, or marijuana. Participants were also excluded if they were using intranasal or inhaled medications, or had a history of chronic lung disease or lung cancer. The remaining participants were enrolled and stratified into three study groups based on their cigarette use behavior at study recruitment: (1) TCIG group (n=9) defined as ongoing TCIG use with a minimum of 5 cigarettes daily and less than 2 lifetime ECIG uses; (2) ECIG group (n=15) defined as former TCIG smokers who have been tobacco abstinent for a minimum of 3 months prior to study re-

cruitment and have been using any brand or generation of ECIGs, with any brand of nicotine-containing liquid, with any type of vehicle or flavoring, at least 6 days per week for at least 1 month; (3) Former group (n=21) defined as former TCIG smokers who had been tobacco-abstinent for a minimum of 3 months prior to recruitment and were not using any form of nicotine replacement therapy. Urine cotinine levels were measured at baseline via the NicAlert ő assay to determine whether participants were using nicotine-containing products. Carbon monoxide (CO) levels were also measured in all ECIG and former smokers to rule out concurrent TCIG smoking (CoVita Bedfont Scientific piCO + Smokerlyzer Breath CO Monitor), and 2 self-reported ECIG users were excluded due to high exhaled CO levels. A detailed TCIG and ECIG use history was obtained for each participant. This study was approved (approval number H -32129) by the Boston Medical Center IRB #1 (Panel Green). All study participants provided written informed consent.

Bronchial airway epithelial cells were obtained from brushings of the right mainstem bronchus collected during fiberoptic bronchoscopy with an endoscopic cytobrush (Cellebrity Endoscopic Cytology Brush, Boston Scientific, Boston). The brushes were immediately placed in 1 mL of RNAprotect Cell Reagent (Qiagen, Valencia, CA) and kept at -80oC until RNA isolation was performed.

2.2.2 Microarray data acquisition and data preprocessing.

Total RNA was isolated using the miRNeasy Mini Kit (Qiagen, Valencia, CA). RNA integrity was assessed by Agilent BioAnalyzer, and RNA purity confirmed using a NanoDrop spectrophotometer. 100 ng of isolated RNA was processed and hybridized to Affymetrix Human Gene 1.0 ST Arrays (Affymetrix, Santa Clara, CA). Probeset normalization was performed using the Brainarray EntrezGene CDF

v17.0.0 and Robust Multiarray Average (Irizarry et al. (2003)). Statistical analysis was performed with R version 3.2.2. All microarray data and relevant clinical data is freely available through Gene Expression Omnibus (GEO) under accession GSE112073.

2.2.3 Microarray preprocessing and quality control.

Microarray quality was assessed using relative log expression, normalized unscaled standard error, and principal component analysis (PCA) metrics and all samples were suitable for subsequent analysis. Batch effects were corrected using ComBat (Johnson et al. (2007)) with a smoking status covariate.

2.2.4 Differential expression analysis.

We first identified genes differentially expressed between any of the study groups (ECIG, TCIG, and Former) via analysis of covariance (ANCOVA). Gene-expression was modeled as a linear function of smoking status while adjusting for age, RNA integrity (RIN), and months since last TCIG. A nested F-test was used to identify genes differentially expressed between ECIG users relative to former TCIG smokers, TCIG smokers relative to current TCIG smokers, or current TCIG smokers relative to ECIG users. Resulting p-values were adjusted to control the False Discovery Rate (FDR) using the Benjamini-Hochberg method (Benjamini & Hochberg (1995)). We used PCA on the log2-expression level of genes meeting these criteria to isolate the first two principal components (i.e. the two largest sources of variance) in the expression of these genes.

Using the linear model described above, we performed additional linear modelling with LIMMA (Ritchie et al. (2015); Smyth (2005a)) on the genes identified as being differentially expressed in TCIG smokers versus former smokers or ECIG

users versus former smokers as identified in the ANCOVA step. Genes demonstrating differential expression in ECIG users relative to former smokers were identified via the ECIG coefficients moderated t-test p-value < 0.05. These genes were divided into sub-clusters via Ward hierarchical clustering.

2.2.5 Functional enrichment.

We performed functional enrichment analysis on the gene clusters identified during differential expression via Enrichr (Chen et al. (2013a)). Only enrichment terms with an FDR-adjusted p-value < 0.05 were considered.

2.2.6 Gene Set Variation Analysis.

To analyze the similarity of the gene-expression effects of ECIGs and TCIGs, gene-expression pathways associated with TCIG use based on previously published data were projected into the study groups of the cross-sectional cohort using Gene Set Variation Analysis (GSVA) (Hänzelmann et al. (2013)). Microarray data from Beane et al (Beane et al. (2007b)), which compared current, former, and never cigarette smokers were scored for gene set activation for every gene set in the C2, C5, C7, and Hallmark collections of the Molecular Signatures Database (Liberzon et al. (2011)). Gene sets whose GSVA scores were significantly different between current smokers and non-current (former and never) smokers (Students t-test; FDR q < 0.05) in this dataset were categorized as differentially activated due to TCIG usage. These gene sets were then scored in the microarray data of the present study via GSVA. These GSVA scores were then compared between study groups via Students t-test to identify differences in gene set activity.

2.2.7 Quantitative Real-time Polymerase Chain Reaction (RT-PCR).

We profiled ADM, PGAM5, NCK2, and RSPH1 in 15 ECIG users and 15 former smokers with quantitative RT-PCR. These assays were performed with SYBR Green-based RT^2 qPCR Primer Assays (Qiagen, Valencia, CA). Primers for each candidate gene and the 18S ribosomal subunit as an endogenous control gene were designed and experimentally verified by Qiagen to ensure uniform and high PCR efficiencies under standardized amplification conditions. All real-time PCR experiments were carried out in triplicate on each sample, relative gene expression levels were calculated using the comparative CT method (Schmittgen & Livak (2008)), and differential expression was performed via Students t-test on the average expression across these replicates.

2.2.8 Comparison with in vitro and immune/inflammatory response dataset.

We compared our observations from the study cohort to an in vitro model of ECIG exposure, which profiled the gene expression effects of exposure to vapor from a single brand of disposable ECIGs on bronchial epithelial cells grown at an air-liquid interface (Moses et al. (2017)). Using data from an in vitro model of ECIG exposure, we generated a ranked list of genes differentially expressed between ECIG aerosol and air-exposed human bronchial epithelial cells via a Students t-test, and ranked genes by their corresponding t-statistic. The genes found to be associated with ECIG use in the current study were split into gene sets based on the clustering results, and Gene Set Enrichment Analysis (GSEA) (Subramanian et al. (2005)), was used to determine if these gene sets were significantly enriched toward the bottom or top of the in vitro ranked list.

2.3 RESULTS

2.3.1 Subject characteristics.

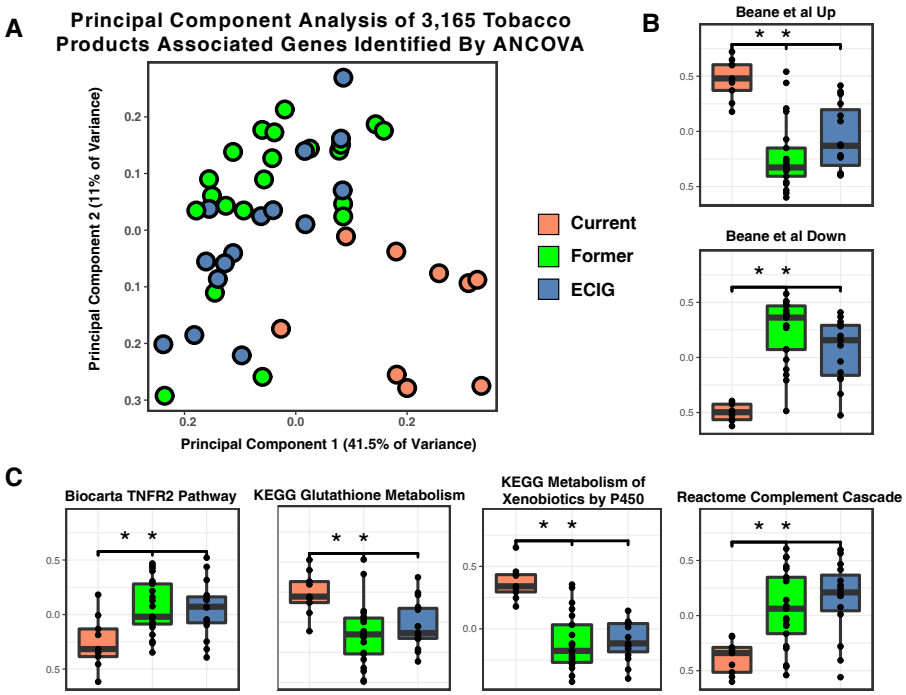
There were no significant differences in age, gender, race or pack-years between the three study groups (TCIG, ECIG, and former TCIG smokers) (Table 1). There was a statistically significant difference in time since TCIG cessation between the former smoker and ECIG groups (Students t-test $p < 0.05$). CO levels for all ECIG users and former smokers included in the final analysis were < 7 ppm. Urine cotinine levels in the TCIG and ECIG groups confirmed active nicotine use (>= 100 ng/ml) and were significantly higher than the former smoker group (Students t-test $p < 0.001$). Patterns of ECIG use, including frequency, nicotine dosage, generation of product, and product brand varied across ECIG users.

2.3.2 Airway gene-expression in former TCIG smokers who currently use ECIGs is more similar to former TCIG smokers than to active TCIG smokers.

Differential expression analysis via ANCOVA among the three study groups identified 3,165 genes whose expression is associated with either or both exposures (FDR-adjusted $p < 0.05$). Samples from TCIG smokers separate from the former- and ECIG-derived samples along the first principal component derived from the expression of these genes (Figure 2.1A), while there is little separation between the non-TCIG samples, suggesting that gene-expression differences associated with TCIG use are a strong driver of the differential expression detected by this analysis.

To identify the degree to which ECIGs induce TCIG-associated gene-expression changes, we used GSVA to determine the activation of genes up- or down-regulated between current TCIG smokers versus never smokers as identified by Beane et al.

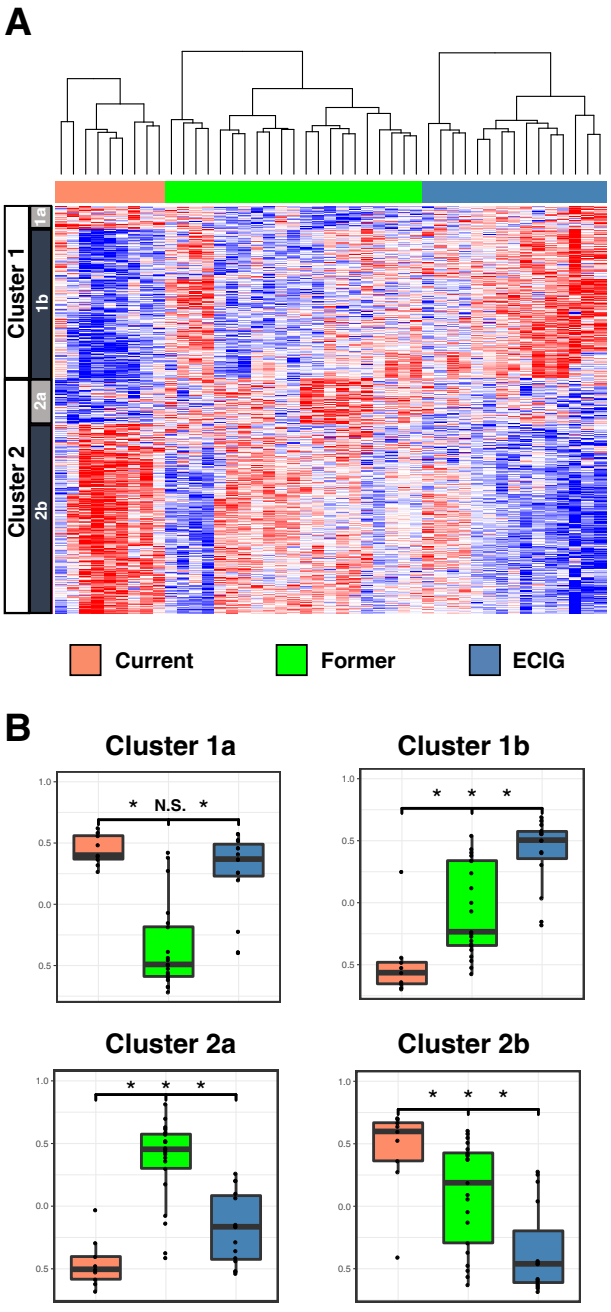
Figure 2.1



We found a significant difference in the expression of these TCIG-associated gene sets between TCIG smokers and ECIG users, and TCIG smokers and former smokers, but not between ECIG users and former smokers (Figure 2.1B). Additionally, we identified 280 pathway-related gene sets that have TCIG-associated expression differences in the Beane et al. cohort and used them in a similar metagene analysis of the current dataset via GSVA (Figure 2.1C). Expression of these select TCIG-associated pathways was also shown to be significantly altered in the TCIG group relative to the former and ECIG groups (Figure 2.1C; Students t-test, $p < 0.05$), but not significantly different between the Former and ECIG groups. Collectively, these data suggest that the gene-expression profiles of the ECIG users are more similar to the former group than the TCIG group both globally and for genes that are altered by TCIG use.

2.3.3 ECIG Associated Gene-expression Changes

Figure 2.2



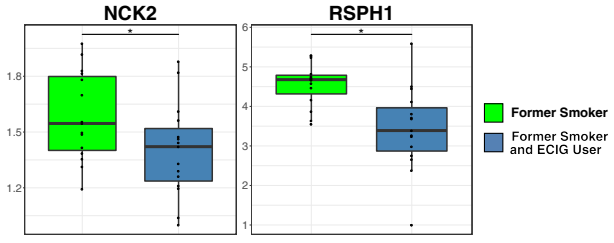
Post-hoc analysis of the 3,165 smoking-status-related genes identified by ANCOVA yielded 468 genes whose expression was associated with ECIG use status

(Figure 2.2A; moderated t-test $p < 0.05$). These genes were organized into four clusters via Ward hierarchical clustering (see Appendix A.1). We used the cluster gene sets in a metagene analysis via GSVA to identify the relationship between the expression levels of these genes in ECIG users relative to TCIG and former smokers (Figure 2.2B).

Genes in Cluster 1 (198 genes) were upregulated in the ECIG group compared to the former group. Cluster 1 was subdivided into genes whose expression were also upregulated in the TCIG group (Cluster 1a, 27 genes), or downregulated in the TCIG group (Cluster 1b, 171 genes). Genes in Cluster 2 (270 genes) were genes whose expression level were downregulated in the ECIG group compared to the former group. Cluster 2 was subdivided into genes whose expression level was also downregulated in the TCIG group (Cluster 2a, 52 genes), or upregulated in the TCIG group (Cluster 2b, 218 genes).

Cluster 1a, while small, contained several genes associated with interleukin receptor complexes. Cluster 1b contained ribosomal protein subunit associated genes, as well as genes associated with maturation of non-coding RNAs and translation. Cluster 2a contained genes associated with microtubule assembly and structure. Cluster 2b was similarly enriched for genes involved in microtubule assembly, as well as genes involved in regulation of RNA Polymerase II activity.

Figure 2.3

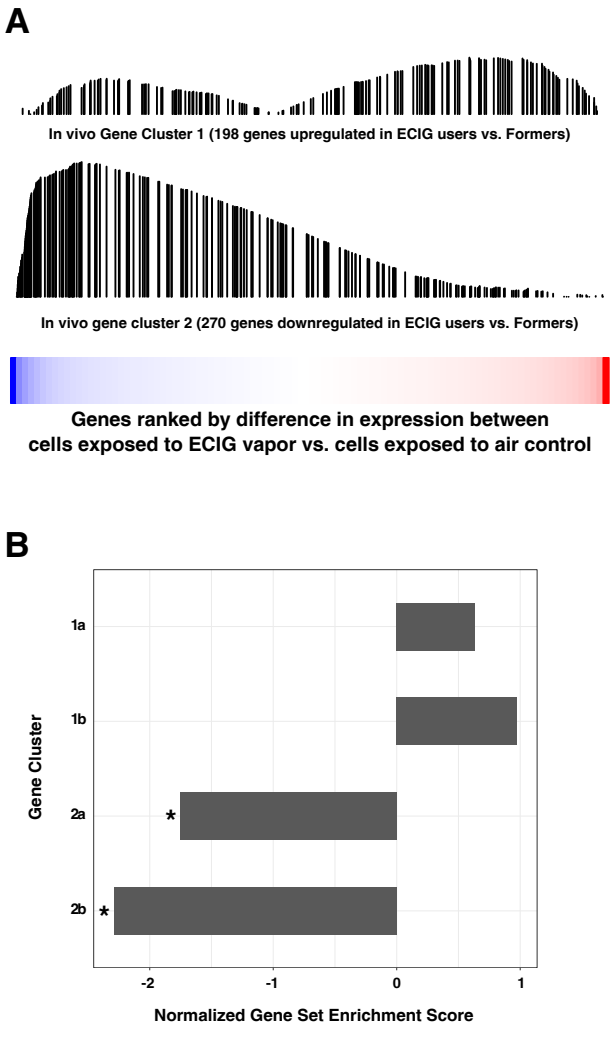


We performed qRT-PCR to validate the ECIG-associated differential expression

of ADM, PGAM5, NCK2, and RSPH1 as representative genes from Clusters 1a, 1b, 2a, and 2b, respectively. Concordant with the microarray data, NCK2 and RSPH1 demonstrated significantly lower expression in ECIG users compared to former smokers (Figure 2.3; Students t-test, $p < 0.05$). We were unable to confirm differential expression of ADM and PGAM5 by qRT-PCR.

2.3.4 Comparison with in vitro ECIG exposure.

Figure 2.4



To validate the gene-expression alterations we observed in ECIG users using

a variety of ECIG products, we determined if the genes we identified as differentially expressed in ECIG users were similarly altered in an in vitro dataset from a previously published experiment which examined the effects of ECIG aerosol on differentiated human bronchial epithelial cells (Moses et al. (2017)). Using GSEA, we found that genes downregulated in the bronchial epithelium of ECIG users are significantly enriched amongst the genes most downregulated with ECIG exposure in vitro (GSEA $p < 0.001$). In contrast, we did not detect significant enrichment of the genes upregulated in ECIG users amongst the genes most upregulated with ECIG exposure in vitro (Figure 2.4).

2.4 DISCUSSION

To our knowledge, this is the first study to comprehensively assess gene-expression changes in the bronchial airway epithelium of real-world ECIG users, and how these changes compare to those associated with TCIG smoking. Across all the genes we identified as differentially expressed between current TCIG smokers, former TCIG smokers now using ECIGs, and former TCIG smokers, we observed that the pattern of gene-expression in ECIG users was much more similar to former smokers than TCIG smokers. TCIG-associated gene sets derived in a previous study (Beane et al. (2007b)) were differentially expressed between the TCIG and ECIG as well as TCIG and former smoker groups, but not between the ECIG and former groups. We also found that the differential expression of genes in relevant TCIG-associated gene-expression pathways, specifically glutathione (Gould et al. (2011)) and xenobiotic metabolism (Pierrou et al. (2007)), Tumor necrosis factor receptor 2 signaling (Thum et al. (2006)), and complement cascade (Robbins et al. (1991)), were distinct in TCIG smokers but not significantly different between ECIG users and former smokers. These findings suggest that, overall, TCIG in-

duced transcriptional changes revert to baseline similarly in former TCIG smokers and former TCIG smokers who now use ECIGs.

While the ECIG and former TCIG smoker groups were not significantly different with regard to expression of genes changed in TCIG smokers, we were able to identify a set of genes whose expression changes in ECIG users relative to former smokers and that a small subset of these genes are similarly changed in TCIG smokers. This implies that ECIG use does impact the physiology of the bronchial epithelium in ways that are distinct from the effects associated with TCIG use, and that the degree of overlap between ECIG and TCIG associated effects is modest.

To validate that the gene-expression alterations in bronchial epithelium associated with ECIG use are likely due to the direct effects of exposure of this tissue to ECIG aerosols, we examined the expression of our ECIG signature in differentiated human bronchial airway epithelial cells that were exposed to ECIG aerosols in vitro (Moses et al. (2017)). We found that the genes we had identified as having decreased expression in ECIG users are enriched among the genes down-regulated in the in vitro ECIG aerosol exposed tissues. That we were able to identify common gene expression signatures between a heterogeneous in vivo exposure and a uniform in vitro exposure suggests the existence of a set of ECIG associated gene expression effects that are common and independent of the specific product used. Beyond suggesting that the gene-expression changes in ECIG users are likely the direct effects of ECIG aerosol exposure, these data also suggest that there is a common impact of ECIG use on bronchial epithelial gene-expression despite heterogeneity in ECIG product usage.

Amongst the genes in the ECIG signature we identified, those in Cluster 1a and Cluster 2a are similarly altered in both ECIGs and TCIGs, while those in Cluster 1b and 2b exhibit an ECIG-specific pattern. Pathway enrichment analysis of the genes

in each cluster provides potential insight into common and ECIG-specific effects (Table 2).

Genes in Cluster 1b were upregulated specifically in ECIG users and were significantly enriched for the Ribosome biogenesis GO Biological Process term. The over-expression of ribosomal genes in ECIG users might reflect increased oxidant stress (Gerashchenko et al. (2012)). Supporting this hypothesis, this cluster additionally contains NDUFB2 and NDUFA4L2, which code for subunits of the NADH-ubiquinone oxidoreductase complex and play a role in handling oxidant stress. ATP5H, a component of the mitochondrial electron transport chain, also appears in Cluster 1b. While these genes are only induced in ECIG users, both ribosomal structure and oxidative phosphorylation pathways have both been previously found to be induced by TCIG smoke (Spira et al. (2004)). Cluster 1b is also significantly enriched for targets of ATF2. Both of these transcription factors have been implicated as possible regulators of inflammation in the lung (Chishimba et al. (2010); Yu et al. (2014)). These findings suggest that ECIG use might modulate lung inflammation.

Cluster 2b contained genes that were downregulated specifically in ECIG users. This cluster is significantly enriched for targets of RFX3, a transcription factor involved in ciliary assembly and motility which has been specifically observed to cooperate with FOXJ1 in the process of ciliated cell differentiation in airway epithelium (Didon et al. (2013)). FOXJ1 is also significantly downregulated in airway epithelial cells following in vitro ECIG exposure (Moses et al. (2017)). We validated the downregulation of RSPH1 in ECIG users by qRT-PCR. RSPH1 is a gene in this cluster and mutations in RSPH1 are associated with ciliary defects and dyskinesia (Onoufriadis et al. (2014)). Taken together, these data suggest that ECIG use might impair ciliogenesis.

Genes in Cluster 2a were downregulated by both TCIG and ECIG use, was significantly enriched for genes associated with axon guidance. NCK2 is a gene from Cluster 2a that we validated by qRT-PCR is significantly downregulated by ECIG use. NCK2 is associated with EGFR signaling and cytoskeletal reorganization (Rivera et al. (2006)) and we hypothesize that this cluster reflects cytoskeletal and/or cilium-related effects of TCIG and ECIG use. Genes in Cluster 1a were are upregulated by both TCIG and ECIG use, and were associated with interleukin receptor complexes.

While the findings of the present study yield insights into the effects of ECIG use on bronchial epithelium, there were several limitations, most notably regarding sample sizes within each of the study groups. We observed that many fewer genes were affected by ECIG than TCIGs, which dominated differential expression when comparing all three groups (Figure 2.1). While we were able to identify a number of ECIG-specific gene-expression changes, it is likely that more subjects would be necessary to comprehensively identify the transcriptional impact of ECIGs. Future studies would also benefit from the inclusion of de novo users of ECIGs to be able to better isolate effects of ECIGs that might either be specific to former smokers or obscured in former smokers. The potential to perform experimental studies on volunteers with non-nicotine containing ECIGs would enable identification of the specific gene expression effects associated with vehicle and flavorings, which we are unable to discern in the present signature. This is potentially important as any specific ECIG additive might have an unusual but dramatic effect on airway biology.

Furthermore, we did not identify significant overlap between the genes we found to be differentially expressed in the bronchial epithelium of ECIG users and the genes found by Martin et al. in nasal epithelium as we expected based on

previous work establishing the commonalities of bronchial and nasal signatures of TCIG exposure (Sridhar et al. (2008); Zhang et al. (2010)).

Another limitation is the heterogeneity of ECIG products used by the ECIG users in our study. In order to recruit a sufficient study population, we did not limit the brand or type of ECIG device that participants could use. Additionally, though all participants met a minimum threshold of weekly ECIG usage, actual usage varied amongst participants. Most of the study participants were users of first-generation devices and it remains to be determined if new types of devices will elicit similar changes. However, our finding of common ECIG effects despite the heterogeneity of ECIG exposures within the study group, and our finding that many of the changes observed in users of heterogeneous ECIG products are similar to the effects observed with a single brand ECIG in vitro leads us to believe that the majority of these effects are independent of the vaporization device used, and that there is a common set of ECIG-related effects on airway epithelium.

Overall, our findings indicate that ECIG use does not lead to alterations in the expression of the vast majority of genes that are altered by TCIG use, but that there is a group of genes whose expression is specifically altered in ECIG users. When examining the expression of genes in key TCIG-associated pathways, we found that ECIG users had gene-expression profiles more similar to former TCIG smokers than current TCIG smokers. Our findings indicate that the use of ECIGs does impact the airway, which includes modulating the expression of a small set of genes that are altered in both ECIG users and TCIG smokers. Further study is required to identify the clinical significance of these findings and to fully evaluate the pulmonary impact of ECIG exposure.

2.5 ACKNOWLEDGEMENTS

Author Contributions:

- Sean E Corbett and Matthew Nitzberg performed all statistical analysis, created all figures, and prepared the original draft of this manuscript
- Teresa Wang performed preliminary data analysis
- Elizabeth Moses and Catalina Perdomo assisted with comparing this studys data to a related in vitro study
- George OConnor, Eric Kleerup, Claudia Perdomo and Daniel Brooks designed the study and enrollment questionnaire, and coordinated subject recruitment and sample collection
- Gang Liu, Xiaohui Xiao, and Hanqiao Liu prepared samples from the study participants and performed the microarray gene expression profiling experiments
- Steven M Dubinett, David Elashoff, Avrum Spira, and Marc E Lenburg designed the study and guided the analysis and interpretation of results
- All authors have revised this manuscript, provided final approval for its publication, and have agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Guarantor Statement: Avrum Spira takes responsibility for the content of the manuscript, including the data and analysis.

Financial disclosures: Dr. Avrum Spira reports the following conflicts of interest:

- Johnson and Johnson: Employee

CHAPTER 3

Bi-clustering of transcriptional states and cellular populations in discrete single-cell RNA-seq data

Adapted from the following manuscript:

Sean E. Corbett*, Yusuke Koga*, Shiyi Yang, Zhe Wang, Jianyuan Liu, Grant Duclos, Evan Johnson, Paola Sebastiani, Masanao Yajima, Joshua Campbell.

(Manuscript in preparation)

* Contributed equally

3.1 INTRODUCTION

Complex biological systems can be understood by dividing them into hierarchies. Each level of such a hierarchy is composed of different parts which perform distinct biological functions. For example, organisms can be subdivided into a collection of complex tissues: each complex tissue is composed of different cell types; each cell population is denoted by a unique combination of transcriptionally activated pathways (i.e. transcriptional states); and each transcriptional state is composed of groups of genes that are coordinately expressed to perform specific molecular functions. By identifying the basic building blocks at each level of the hierarchy as well as their composition, we can more easily conceptualize the inner workings of higher order biological systems.

Single-cell RNA-seq (scRNA-seq) has emerged as a powerful technique to quantify gene expression in individual cells, and is being used to elucidate the molecular and cellular building blocks of complex tissues. Rather than profiling RNA from a "bulk" sample, where only an average transcriptional signature across all the composite cells can be derived, scRNA-seq experiments can profile the tran-

scriptome of thousands of cells per sample. Thus, it represents an excellent opportunity to identify novel subpopulations of cells and to characterize transcriptional programs by examining co-varying patterns of gene expression across individual cells. However, analysis of scRNA-seq data has several challenges. For example, the data tends to be sparse due to the difficulty of amplifying the low amounts of RNA provided by individual cells. To combat noise from the amplification process, unique molecular identifiers (UMIs) are often incorporated, allowing researchers to measure counts of individual molecules. The use of these UMIs enables the measurement of discrete counts of mRNA transcripts within each cell, making models using discrete distributions an attractive approach for analyzing this type of data.

Discrete Bayesian hierarchical models have been widely used for unsupervised modeling of discrete data types. In the text mining field, a plethora of models have been developed which can identify hidden topics across documents and/or cluster documents into distinct groups. These models treat each document as a "bag-of-words" where each document is represented by a vector of counts for each word in a vocabulary. Each document cluster or hidden topic is represented by a Dirichlet distribution where words with higher probability are observed more frequently for the document cluster or topic (Blei et al. (2003)). Given the success of topic models with sparse text data and the discrete, sparse nature of count data generated by many scRNA-seq protocols, the application of such discrete Bayesian hierarchical models represents an appealing approach to characterize structure in scRNA-seq data. In this setting, cells are assumed to be a bag-of-transcripts and can be represented by a vector of gene counts.

Here, I present the details of three different models that can cluster cells into subpopulations (celda_C), cluster molecular features such as genes into modules (celda_CG), or simultaneously perform bi-clustering of cells into subpopulations

and genes into modules (celda_CG). While these models can perform clustering, they also offer probabilistic distributions which can be used to describe the contribution of each building block to each layer of the biological hierarchy. These distributions can also be viewed as reduced dimensional representations of the data that can be used for downstream exploratory analyses. I also detail the application of this modeling framework to a publicly available scRNA-seq dataset of peripheral blood mononuclear cells from a healthy donor.

3.2 MATERIALS & METHODS

3.2.1 Conventions

Below, I describe the derivation of the CELDA statistical models, the algorithms which leverage these models, and the software implementation of these algorithms and associated visualizations. Algorithms are specified in enumerated lists. Text appearing `in monospace font` refer to functions defined in the CELDA software package. I will also refer to CELDA meaning both the suite of statistical models and the software package interchangeably.

3.2.2 Statistical models

The CELDA models are an extension of the classic Latent Dirichlet Allocation model, previously derived by Blei *et al.* (Blei et al. (2003)). Each document is modeled as a multinomial distribution over a set of words in a vocabulary. Similarly, CELDA models each cell as a multinomial distribution of modules, and each module is a multinomial distribution over genes. However, CELDA additionally utilizes a technique first described in the sparseTM (Wang & Blei (2009)) that forces genes to belong to a single gene module, in order to produce a "hard-clustering"

behavior (Figure 3.1).

CELDA contains 3 sub-models:

1. **celda_C**, which clusters cells into cell subpopulations,
2. **celda_G**, which clusters genes into gene modules,
3. **celda_CG**, which performs bi-clustering of cells into cell subpopulations and genes into gene modules by leveraging celda_C and celda_G.

Below I provide an overview of the celda_CG model, which was used for the analysis of an example real-world dataset presented later in this chapter.

3.2.2.1 Generative Process

Topic models such as LDA and CELDA and be imagined as "generative processes," which explain how given the model's assumptions on the structure of the data it is designed for, a new dataset exhibiting these properties could be generated. The corresponding generative process for the CELDA model, as derived by Dr. Joshua Campbell, is as follows:

If S is the number of samples, M_i the number of cells in sample i, K be the number of cellular subpopulations, L the number of modules, G the number of genes, y_g the hidden state for gene g, $N_{i,j}$, the number of transcripts for cell j in sample i, $x_{i,j,t}$ the t^{th} transcript for cell j in sample i, and $w_{i,j,t}$ the hidden transcriptional state for transcript $x_{i,j,t}$, . The generative process for celda_CG can be described as:

3.2.2.2 Model Log-Likelihood and Perplexity

The likelihood function of a topic model such as CELDA is a measure of how likely a given clustering solution is, given the data being analyzed. The derivatiton of the

1. Draw $\eta \sim \text{Dir}_L(\gamma)$
2. For each gene $g \in \{1..G\}$, draw $y_g \sim \text{Mult}(\eta)$
3. For each transcriptional state distribution $l \in \{1..L\}$:
   (a) Define $Y_l = [y_g = l]_{g=1}^{G}$
   (b) Draw $\psi_l \sim \text{Dir}(\delta Y_l)$
4. For each sample $i \in \{1..S\}$, draw $\theta_i \sim \text{Dir}_K(\alpha)$
5. For each cell population $k \in \{1..K\}$, draw $\varphi_k \sim \text{Dir}_L(\beta)$
6. For each cell $j \in \{1..M_i\}$ in sample i:
   (a) Draw $z_{i,j} \sim \text{Mult}(\theta_i)$
   (b) For the t-th transcript in cell j in sample i, $t \in \{1..N_{i,j}\}$:
      i. Draw $w_{i,j,t} \sim \text{Mult}(\varphi_{z_{i,j}})$
      ii. Draw $x_{i,j,t} \sim \text{Mult}(\psi_{w_{i,j,t}})$

likelihood function of the celda_CG model appears in Appendix B. The likelihood of the celda_CG model is:

$$P(\eta, \psi, \theta, \varphi, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{X} | \alpha, \beta, \gamma, \delta) =$$
$$P(\eta|\gamma) \prod_{g=1}^{G} P(y_g|\eta) \prod_{l=1}^{L} P(\psi_l|\delta, \boldsymbol{Y}) \prod_{i=1}^{S} p(\theta_i|\alpha) \prod_{k=1}^{K} P(\varphi_k|\beta) \prod_{j=1}^{M_i} p(z_{i,j}|\theta_i) \prod_{t=1}^{N_{i,j}} P(w_{i,j,t}|\varphi_{z_{i,j}}) P(x_{i,j,t}|\psi_{w_{i,j,t}})$$

Perplexity is a function of the likelihood function, and explains how well a probability model predicts a data sample. In CELDA, perplexity is applied to re-samplings of the counts matrix being modeled in order to predict how well a clustering solution from the actual data predicts what should be a very similar (but not identical) count distribution. For the celda_CG model, the log-perplexity function ($\log(p(x))$) is defined as:

$$\log(p(x)) = \sum_{l=1}^{L} \eta_l^{|V_l|} + \sum_{i=1}^{S} \sum_{j=1}^{M_j} \log \left[\sum_{k=1}^{K} \theta_{i,k} \prod_{g=1}^{G} \left(\sum_{l=1}^{L} \varphi_{k,l} \psi_{l,g} \right)^{n_{i,j,g}} \right]$$

where $\theta_{i,k}$ is the probabiity of a cell belonging to a cell population k in sample i, $\varphi_{k,g}$ is the probability of transcriptional state l in cell population k, $\psi_{l,g}$ is the probability of gene g in transcriptional state l, $n_{i,j,g}$ is the count of gene g in cell j in sample i, η_l is the probability of a gene being assigned to transcriptional state l, and $|V_l|$ is the number of genes assigned to transcriptional state l. Note that the only non-zero $\psi_{l,g}$ will be where $y_g = l$ for gene g and thus the sum of the equation can be simplified to:

$$\log(p(x)) = \sum_{l=1}^{L} \eta_l^{|V_l|} + \sum_{i=1}^{S} \sum_{j=1}^{M_j} \log \left[\sum_{k=1}^{K} \theta_{i,k} \prod_{g=1}^{G} \left(\varphi_{k,y_g} \psi_{y_g,g} \right)^{n_{i,j,g}} \right]$$

3.2.2.3 Gibbs Sampling

The CELDA models are implemented using a Gibbs sampling approach (Geman & Geman (1984)), leveraging each model's log-likelihood function to evaluate step-wise permutation of cell subpopulation / gene module labels. Briefly, the Gibbs sampling procedure for the celda_CG model is as follows:

1. For each cell (column) in the counts matrix, draw a cell subpopulation label in the range [1,K] from a multinomial distribution, using a Dirichlet prior for the probability of each value in the range. For each gene (row) in the counts matrix, draw a gene module label in the range [1,L] from a seperate multinomial distribution, with its own Dirichlet prior for the probability of each value in the range.

2. For the first cell, draw a new cell subpopulation label from the multinomial distribution defined by the cell subpopulation labels of all of the other cells,

with consideration of the likelihood of that choice given the counts matrix being considered via the likelihood function.

3. Repeat the previous step for all subsequent cells, until each cell has been considered.

4. For the first gene, perform the analogus gene module label selection from the multinomial distribution defined by the gene module labels of all other genes.

5. Repeat the previous step for all subsequent genes, until each gene has been considered.

6. Repeat steps 2-5 for a prespecified (ideally, large) number of iterations

The above approach, over sufficiently many iterations, will converge on a set of cell subpopulation / gene module labels which are maximally likely, given the data being modeled, though there is no guarantee the solution returned is globally optimal. The software implementation of CELDA includes additional technical optimizations beyond the above specified algorithm in order to facilitate shorter model runtimes and to attempt to avoid local optima.

3.2.3 Software Implementation

In collaboration with Dr. Joshua Campbell, I implemented a software package in the R programming language to facilitate the application of the CELDA models to scRNA-seq datasets.

3.2.3.1 Package Architecture

The CELDA package is implemented in the R Programming Language version 3.5.2 R Core Team (2018). It leverages R's S4 object system, providing a consistent programming interface across each of the available CELDA models. Each of the CELDA models (celda_C, celda_G, celda_CG) is represented as its S4 own class, providing a data encapsulation scheme to contain the parameters used to generate a given model (such as K/L, α, β, etc) as well as the clustering solution(s) it derived (z/y). Each CELDA model object also contains a SHA1 hash of the counts matrix for which the corresponding clustering solution was generated, allowing users to compare an arbitrary counts matrix (by generating a new hash) with a CELDA model object. These sets of values are retrievable via a consistently named set of accessor functions, despite the heterogeneity of results that might be returned by the different models.

Aside from providing programmers a consistent means to access model parameters and clustering solutions, this object structure provides crucial benefits towards the software package's internals. The package's diagnostic plotting functions have model-type-specific implementations, but are able to identify the proper solution based off of the user-provided CELDA model object by dispatching based off of its class. This class-dispatch architecture allows for the future addition of arbitrary CELDA models, which would only require inclusion of the new class name in model-specific function implementations.

3.2.3.2 Model Selection

CELDA optionally allows users to tune model priors, such as $\alpha$and β, but requires that the user specify a K (number of cell subpopulations) and/or L (number of gene module) parameter. The choice of K/L corresponds to the more general open

problem regarding'choosing the right number of clusters' across clustering algorithms, such as choosing the parameter K in Kmeans clustering (Mirkin (2011)). A variety of approaches for both enabling user choice of this parameter or estimating the optimal choice exist.

The CELDA package provides both optimizations and visualizations to facilitate the choice of K/L for users. CELDA can be used to evaluate a range of L sequentially, using the celda_G model. This recursive process uses the previous L's final cluster assignments for each successive L as the initial cluster labels for Gibbs sampling, yielding faster convergence of clustering solutions for each L evaluated. After choosing a suitable L, the same process is applied to recursively model a range of K choices while initializing to L modules derived in the previous step. To identify which choice of L/K is suitable at each step of this process, CELDA provides a visualization of the perplexity of the cluster solutions at each L/K. .

3.2.3.3 Diagnostic Visualizations & Dimensionality Reduction

In combination with visualizing model fit performance via perplexity for a range of K/L parameters, CELDA provides several visualizations in order to inform choice of a clustering solution. CELDA provides utility functions to generate a t-SNE (Maaten & Hinton (2008)) representation of the user's count data, as well as to indicate on this t-SNE cell subpopulation clusters and the expression of user-provided genes on a per-cell basis.

3.2.4 Analysis of 10x 68k PBMC Dataset

To evaluate its ability to identify meaningful cell subpopulations in real-world scRNA-seq data, I applied CELDA to a scRNA-seq dataset prepared by 10x Genomics comprised of 68,000 peripheral blood mononuclear cells (PBMCs) from a

healthy volunteer (Zheng et al. (2017)). This dataset was created via 10x Genomics' droplet sequencing method utilizing Gel bead in EMulsion (GEM) microfluidics. Briefly, this system utilizes gel beads containing sequencing adapters and primers, as well as barcodes to uniquely identify the droplet and individual molecules that may be introduced. Cells are loaded into these droplets, where they are immediately lysed and their polyadenlyated mRNA molecules are reverse-transcribed and barcoded. cDNA generated within each droplet can be pooled, amplified and sequenced via conventional next generation sequencing methods. I retrieved the counts matrix generated by 10X Genomics via their Cell Ranger pipeline from their publically available website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/fresh_68k_pbmc_donor_a). In subsequent sections, this dataset is simply referred to as the "68k" dataset.

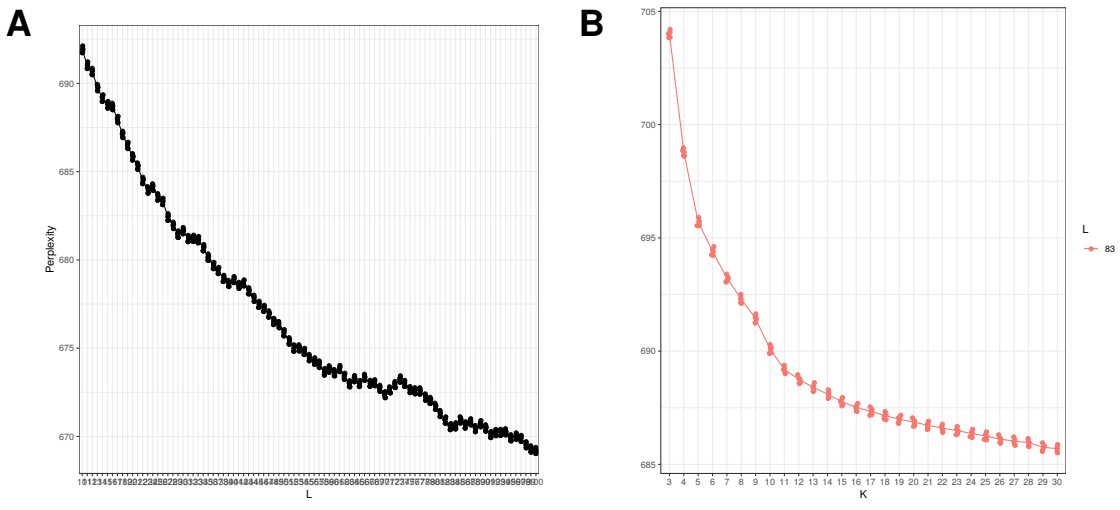
I filtered this downloaded counts matrix to transcripts which had at least 2 counts in at least 2 cells, and cells that had at least 2 counts for at least 2 transcripts. I then used CELDA's `recursiveSplitModule()` function for values of L from 10100 to determine a suitable choice of L. I chose L by referencing the perplexity performance at each value as visualized by `plotGridSearchPerplexity()`. After choosing L, I used CELDA's `recursiveSplitCell()` function to evaluate values of K from 330. I again used `plotGridSearchPerplexity()`, in combination with clustering t-SNEs generated by `plotDimReduceClusters()`, in order to choose a K with sufficiently well-performing perplexity while minimizing the amount of redundant clusters.

3.3 RESULTS

3.3.1 Cluster Size Determination

To determine a suitable selection of K and L for the 68k dataset, I used the recursive splitting procedure outlined above. I first ran celda_G models via the `recursiveSplitModule()` function, for values of L from 10100. Considering the resampled perplexity values of the model at each L (Figure 3.1A), I selected an L value of 83. I next ran celda_CG models via `recursiveSplitCell()`, and selected a K value of 16 by referencing the corresponding resampled perplexity values (Figure 3.1B), as well as by comparing clustering solutions for each value of K via `plotDimReduceClusters()`.

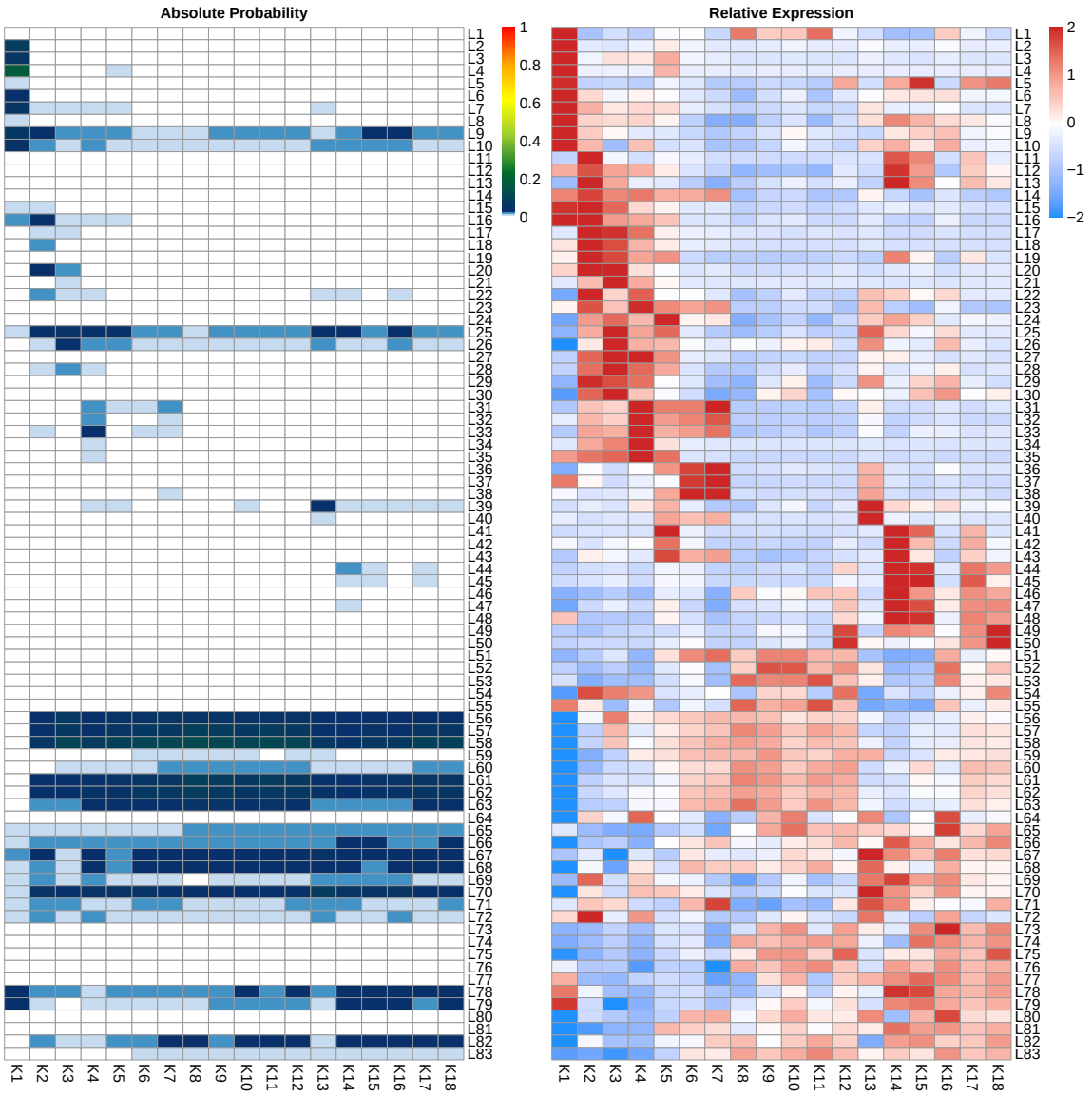
Figure 3.1



Using the clustering solution provided by the celda_CG model using K=16, L=83, CELDA is able to produced a reduced representation of the modeled counts matrix. This reduced representation can be used to elucidate the probability of a given gene module being expressed in a given cell subpopulation, and conversely to determine which gene modules may uniquely identify a given cell subpopula-

tion (Figure 3.2).

Figure 3.2

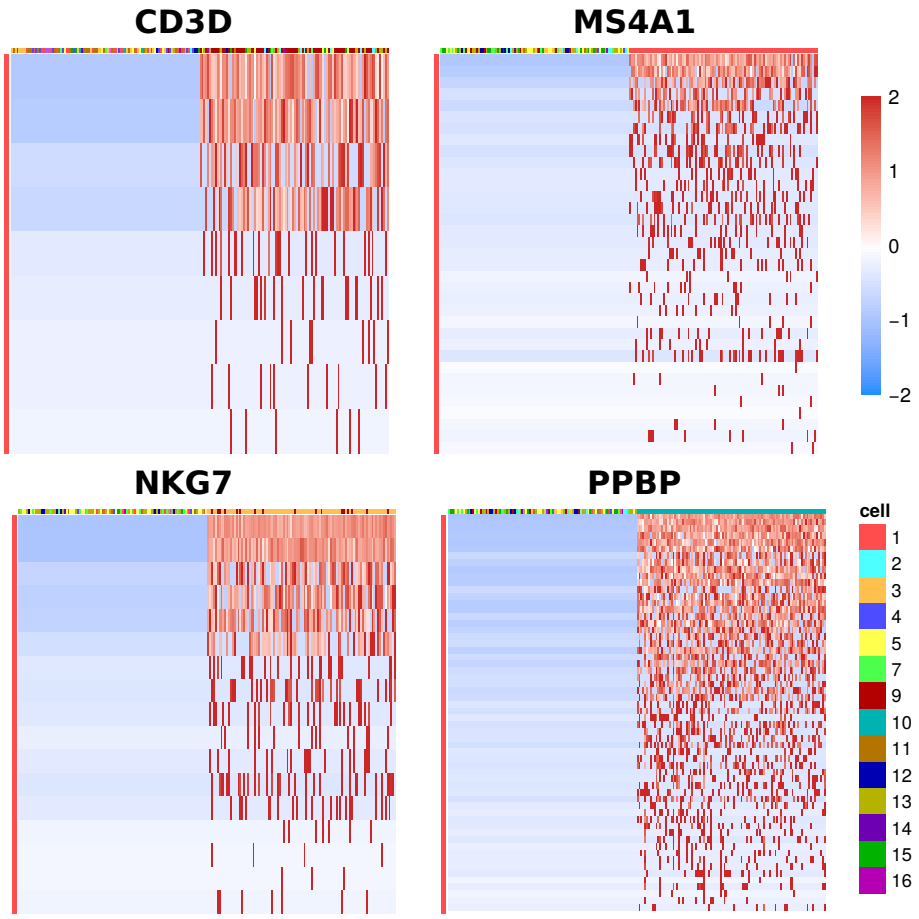


3.3.2 Identification of Broad Cell-Type Associated Clusters

The gene modules identified by CELDA for this dataset vary in terms of the amount of member genes, as well as how heterogeneously these genes are expressed be-

tween cell subpopluations. Modules containing broad cell-type associated marker genes, such as MS4A1 (B-cells) and PPBP (megakaryocytes) often exhibited strong association with a small subset of cell subpopulations in terms of expression (Figure 3.3).

Figure 3.3



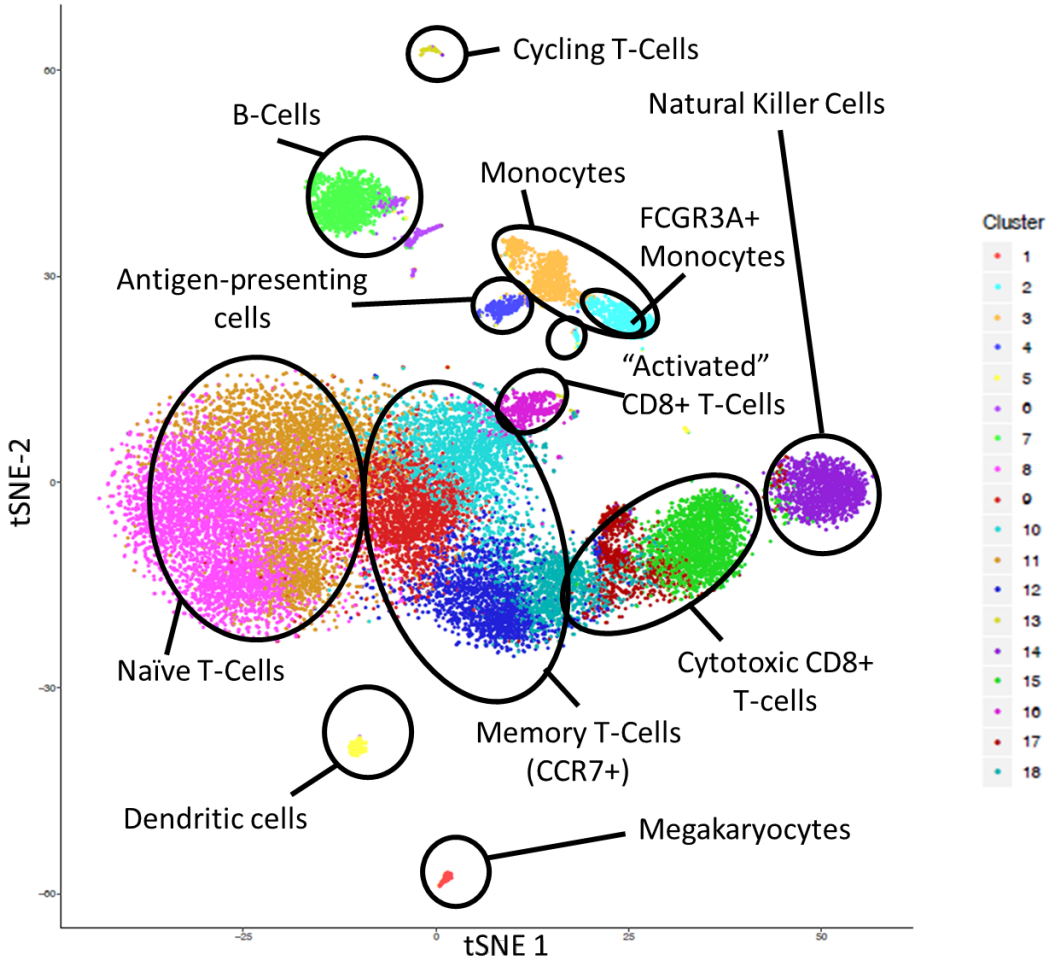
By considering these gene modules' association with cell-type-associated marker genes, as well as their exclusivity to certain subpopulations, I was able to identify broad classes of immune cell types associated with the cell subpopulation clustering solution determined by CELDA (Figure 3.4). CELDA's clustering solution aligns well with the structure determined by t-SNE dimensionality reduction, suggesting a relation in terms of the sub-classes of cells determined by both of these

disparate clustering algorithms.

3.3.3 Identification of a Population of "Activated" CD8+ T-Cells

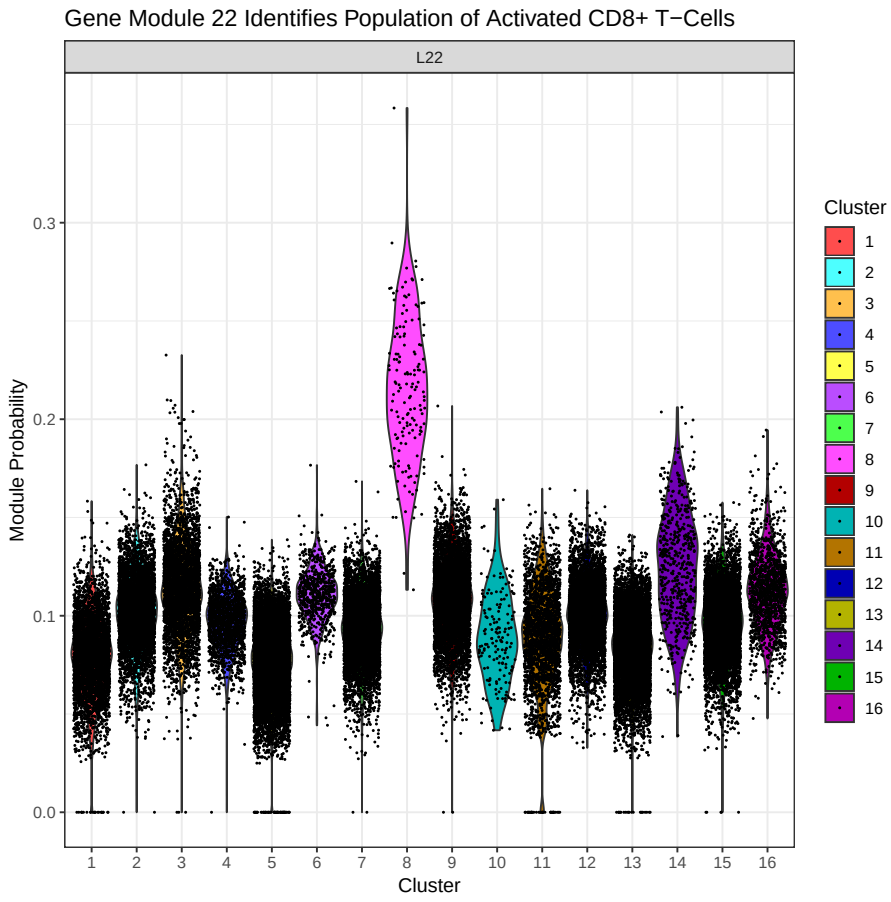
CELDA identified multiple sub-clusters within each broad immune cell type for the selected K and L parameters. For example, I observed several CD3D+ clusters, suggesting several subtypes of naive CD8+ T-cells (Figure 3.4). While most of these clusters demonstrated a consistent pattern of expression of multiple gene modules, I observed there was a gene module (L39) which appeared specifically

Figure 3.4



associated with only one of these cell subpoulations (K13; Figure 3.5). Further examination of this identifying module revealed an enrichment of genes established to be indicative of activated CD8+ T cells as well as cell-cycle associated genes (Su et al., 2014).

Figure 3.5



3.4 DISCUSSION

In this chapter, I describe a suite of hierarchical Bayesian models, CELDA, which can simultaneously assign cells and genes to biologically meaningful clusters in single-cell RNA-sequencing datasets. The models are able to discern subtle but

relevant transcriptional differences between cell subpopulations, including small but distinct subpopulations within broader cell types. I also described how these models were implemented in a software package.

When applying CELDA to publicly available, previously analyzed PBMC single-cell datasets (Zheng et al. (2017)), I was able to discern expected broad immune cell types, such as T-cells, B-cells, and megakaryocytes. I also observed that CELDA indicated heterogeneity within these cell types, such as several distinct clusters of CD3D+ CD8+ T-cells. Finally, I was able to identify small subpopulations of cells, a set of activated CD8+ T-cells, which were not identified in previously published analyses of this data. That I was able to identify these meaningful but subtle shifts in cell-subtype subpopulations suggests CELDAs sensitivity to relevant transcriptional changes that may be difficult to detect due to said subpopulations' correspondingly fewer amounts of cells and transcripts.

A relevant limitation in the use of CELDA and other clustering methods is the selection of the number of cell subpopulation / gene modules (K/L, respectively) to be modeled. While CELDA provides measures of model fit such as perplexity and models log-likelihood, users of CELDA will often be required to choose between several similarly performing clustering solutions. Using a combination of model fit metrics, differential expression between clusters to identify major transcriptional differences, and domain expertise, the process of choosing an appropriate K/L is feasible. While an automated solution for determining the optimal K/L would be ideal, doing so remains an open problem in computer science (Mirkin (2011)). In the future, incorporation of methods into the algorithm such as Dirichlet processes (Ferguson, 1973) may be able to facilitate automatic selection of good choices for these parameters, though ultimately the ideal choice of K/L will rely on knowledge of the experiment at hand.

Through a software implementation of Dr. Joshua Campbell's hierarchical Bayesian modeling approach for biclustering of scRNA-seq data, I was able to identify functionally distinct subpopulations of cells, and co-expressed genes which underlie these distinctions.

3.4.1 Package and Dataset Availability

The CELDA R package source code is freely available under an MIT license at the following URL: https://github.com/campbio/celda

The 68k cell PBMC dataset, generated by 10x Genomics, is freely available from their website at the following URL: https://support.10xgenomics.com/single-cell-gene-expression/datasets

3.4.2 Author Contributions

- Sean Corbett implemented much of the CELDA software, performed data analysis, prepared figures, and contributed to the authorship this manuscript
- Joshua Campbell derived all of the CELDA models, implemented the initial software, contributed to package development, supervised all data analysis, and contributed to the authorship of this manuscript
- Yusuke Koga contributed to package development, performed data analysis, prepared figures, and contributed to the authorship of this manuscript

CHAPTER 4

The Lung Connectivity Map: Identifying Lung Cell-Type-Associated Responses to perturbagen Perturbation

4.1 INTRODUCTION

The Connectivity Map (CMap) is an expansive compendium of gene expression signatures, capturing the gene expression responses of a large set of cancer cell lines exposures to a large set of cancer therapeutics, tool perturbagens, and molecular perturbations such as CRISPRs and gene knockdowns (Lamb et al., 2006; Subramanian et al. 2017). CMap also includes a suite of tools for the suitable normalization, summarization, and interrogation of these signatures. The utility of CMap is severalfold. For example, an input gene expression signature can be queried into the CMap tooling, which will provide a list of perturbations which induce a similar (or the same) gene expression signature in some or all of the assayed cell lines. By identifying a perturbagen with well-characterized molecular effects which induces a similar gene expression signature to a query signature corresponding to an uncharacterized biological phenomena, a researcher can gain additional information as to the differential biological effects used to derive their signature. The CMap querying functionality can also identify perturbagens which *reverse* a query gene expression signature. Such perturbagens may have some efficacy in reversing the underlying cause represented by the query gene expression signature. This signature querying approach has proven to be efficacious in recovering a set of known HDAC inhibitor perturbagens from a signature of HDAC inhibition , discovering a novel CSNK1A1 inhibitor, and identifying a molecule which reverses impaired collagen remodeling in lung fibroblasts *in vitro* (Lamb et al., 2006; Subramanian et al., 2017; Campbell et al., 2012).

Despite the effectiveness of the original CMap designs, the datasets have an inherent limitation in that they only assayed immortalized cancer cell lines, which were derived from a variety of tissues of origin. Because of this design, it is likely that these datasets omit relevant perturbational responses that would only be observed in normal, non-cancer cell lines. Further, the CMap datasets do not include multiple cell lines of the same type, decreasing the dataset's ability to portray, for example, epithelial cell associated responses to perturbagen perturbation. Given that there is also heterogeneity between cells of the same type from the same tissue of origin, it is imperative to investigate how much additional perturbational resolution could be achieved through a more redundant study design.

In this chapter, I will present the Lung Connectivity Map (LCMap), a novel CMap-style dataset intended to address these issues. I will describe observed patterns of gene expression effects of perturbagen perturbation both between a set of lung-derived, non-cancer cell lines, as well as between these non-cancer cell lines and CMap's cancer cell lines.

4.2 MATERIALS & METHODS

The below sections *Cell Culture, Cell Plating, Compound Exposure,* and *Cell Lysis* are included with permission by Dr. Elizabeth Moses (Moses (2017)). A signed letter of permission appears in Appendix B.

4.2.1 Cell Culture

Cell lines included BEAS-2B bronchial epithelial cells (ATCC #CRL-9609), HBEC bronchial epithelial cells (Cell Applications #502-05a), A549 bronchoalveolar carcinoma cells (ATCC #CCL-185), HUVEC human umbilical vein endothelial cells (ATCC #CRL-1730), NL20 bronchial epithelial cells (ATCC #CRL-2503), 1HAE bronchial

epithelial cells (gifted from Jim Hogg, UBC), IMR90 fetal lung fibroblasts (ATCC #CCL-186), HFL1 fetal lung fibroblasts (gifted from Jim Hogg, UBC) and WI38 fetal lung fibroblasts (ATCC #CCL-75). All cells were grown and maintained in Roswell Park Memorial Institute (RPMI) cell growth media supplemented with 10% fetal bovine serum (FBS) or ATCC recommended media, and subcultured with trypsin (0.25%)-EDTA and phosphate-buffered saline (PBS).

4.2.2 Cell Plating

Cells were plated into 384-well plates using an Apricot personal pipettor (Apricot) for the pilot experiments performed at BU or a Multidrop Combi (Thermo Fisher) at the Broad, at a density previously determined per cell line to ensure 100% confluency upon exposure (typically about 3500 cells per well). Total volume of cells was 45L per well. Plates were incubated for 20 minutes at room temperature after plating to promote adherence before being moved to a 37žC humidified atmosphere containing 5% CO2. After 24 hours, cell plates were removed from the incubator and prepared for treatment.

4.2.3 Compound Exposure

For the second study, 324 compounds from the drug management database at the Broad institute were selected for profiling (Table 4.4). Compounds were randomly selected from a list of 1000 compounds previously profiled in the LINCS database and readily available for plating, including small molecules and drugs of various mechanisms. Additionally, DMSO vehicle controls were included as well as positive control compounds, selected due to consistent performance in the original CMap (LINCS). Compound plating was performed at the Broad institute so that each well of a 384-well plate contained 10uL of compound at 10mM concentra-

tion. Cell treatment was performed using a 384-well CyBio liquid handling robot. In brief, a 384-well dilution plate was prepared containing 99L RPMI media (10% FBS) per well. The CyBio transferred 1uL of compound from the compound plate to 99L of RPMI on the dilution plate. The plate was then spun down for 1min at 1000rpm. The CyBio then transferred 5uL from the dilution plate to 45L of cells in the cell plates. Thus, the final concentration of compound exposure was 10M. Cell treatment was performed for 24 hours. Concentration and time point were selected to ensure comparability with the Connectivity Map, as the LINCS experiments used the same conditions.

4.2.4 Cell Lysis

Following treatment, cells were lysed using the BioTek microplate washer (BioTek) for the pilot study, or CyBio at the Broad. First, the machine removed 35L of media from each well of each cell plate. The machine then added 25L TCL Buffer (Qiagen) to each well. Following cell lysis, plates were sealed with foil and incubated at room temperature for 30 minutes. Afterwards plates were frozen at -80žC until ready for reverse transcription. For pilot experiments at BU, plates underwent an additional washing step prior to the addition of TCL buffer, to prevent transportation of potential carcinogenic compounds. For this step, the BioTek plate washer was used to wash the wells twice with 35L PBS.

4.2.5 Ligation Mediated Amplification and L1000 XMap Detection

Ligation mediated amplification and bead detection methods were followed as developed by the Broad, using the same protocols for both Lung CMap and LINCS CMap. 20L cell lysate from the previous steps was transferred to Turbocapture plates for mRNA capture, lysate was removed by unsealing the plate and cen-

trifuging face down on a super rag, and first strand cDNA was made using a 5L per well master mix containing dNTPs and MMLV reverse transcriptase. Upstream and downstream probes for each gene were annealed to the first strand cDNA and each probe contained a barcode for annealing to a Luminex bead in the detection process. cDNA was denatured at 95žC for 2 minutes and ramped down from 70žC to 40žC over 6 hours. The probe pairs were ligated together forming a template for PCR. A 5L master mix was prepared containing taq ligase in 1X ligase buffer, added to the plate and then incubated at 45žC for 60 minutes followed by a 65žC hold for 10 minutes. The ligated probe template was PCR amplified using a set of universal primers. A 15L master mix containing T3 and T7 primers, dNTPs and hot start taq in a 1X reaction buffer was prepared and added to the plate and then loaded into a Thermo Electron MBS 384 Satellite Thermal Cycler for PCR amplification. After an initial denature step at 95žC for 15 minutes, the plates cycled 29 times at 60žC. The primers annealed to the universal primer sites on the ligated probe pairs and the upstream primer contained biotin needed for staining. PCR amplicon was then hybridized to barcoded Luminex beads. The barcode incorporated in the amplicon was complementary to the barcode on each Luminex bead so that each gene annealed to a specific bead. A 5L aliquot of PCR amplicon was transferred to a well containing 30L bead mix (about 350 beads/gene/well) and the plate was sealed and incubated at 95žC for 2 minutes, and then at 45žC for 18 hours. Following incubation, the plate was spun at 3000rpm for 1 minute to pellet the beads. The plate was then washed and stained for biotin.

4.2.6 L1000 Data Processing And Normalization

Gene expression signature data from the above-described L1000 XMap gene expression assay was normalized per the standard pipeline as described by Subrama-

nian et al. Subramanian et al. (2017). Briefly, the normalization steps (also referred to as "Levels") from the raw fluorescence measurements to the final differential expression scores are as follows:

1. Raw fluorescence data,
2. Deconvolution, during which the fluorescence data is deconvoluted to identify the expression of each of the 978 landmark genes,
3. Normalization, during which the the data is scaled and normalized to a set of genes with invariant expression, and inference of unmeasured genes is performed
4. Differential expression, during which each perturbation signature's genes is z-score normalized against the expression of that gene in all other samples on its respective source plate,
5. Replicate-consensus, where the signatures derived for the previous step (which still contain individual replicates for each perturbation) are collapsed into a single consensus signature.

4.2.7 STR Validation of LCMap A549 Cells

Because A549 cell lines and served as an important point of comparison between LCMap and CMap, I collaborated with Dr. Sarah Mazzilli to validate that the A549 cell line profiled in LCMap was genetically identical to a reference signature. Dr. Mazzilli performed a short tandem repeat (STR) assay on LCMap A549 cell lines retained from the time of the original L1000 XMap experiment. This assay was performed via the Promega GenePrint 10 STR assay (Promega, Madison, WI). The GeneMapper v4 fragment analysis software (Applied Biosystems, Foster City, CA),

along with the Promega GenePrint10 allele panel bin files, were used to identify the alleles. We compared these STR results with the ATCC STR profile for A549, available at .

4.2.8 Statistical Analysis

All statistical analyses were performed using the R programming language (R Core Team (2018)), version 3.5.2, using utility functions from the cmapR package (Enache et al. (2017)).

4.2.8.1 Calculation of Transcriptional Activity Score (TAS)

The CMap normalization pipeline generates metadata indicating the quality of each perturbational signature. These measures include signature strength (SS; the number of genes differentially expressed in that perturbation relative to a DMSO perturbation on the same plate) and replicate correlation (CC; the correlation of the signatures between each of the 3 biological replicates for a given perturbation). These measures can be combined to produce a transcriptional activity score (TAS), which can be used as an indicator of overall strength and consistency of a perturbagen's effects. TAS is calculated as

$$TAS = \sqrt{(SS + max(CC, 0)/978}$$

A TAS of > 0.2 is defined as indicating a perturbation with a strong transcriptional effect (Subramanian et al., 2017).

4.2.8.2 *Calculation of Connectivity Scores*

In order to capture how similar two L1000 gene expression signatures are, Subramanian et al. define a score called a weighted connectivity score (WTCS). This non-parametric measure of similarity describes how a set of up- and down-regulated query gene sets are enriched in a gene expression signature. The formula for deriving WTCS to determine a given gene set's enrichment in a signature is as follows:

$$w_{q,r} = \begin{cases} (ES_{up} - ES_{down})/2, & \text{if } \mathrm{sgn}(ES_{up}) \neq \mathrm{sgn}(ES_{down}) \\ 0, & \text{otherwise} \end{cases}$$

Where ES_up is the enrichment of q_up in r and ES_down is the enrichment of q_down in r. WTCS ranges from -1 to 1, with 1 indicating exact similarity and -1 indicating exact dissimilarity, with a value of 0 indicating no relationship between the query gene sets and the signature being evaluated.

WTCS scores are subsequently normalized to allow for comparison across various perturbation and cell types, yielding a Normalized Connectiviy Score (NCS). NCS is calculated as:

$$NCS_{c,t} = \begin{cases} w_{c,t} \Big/ \mu^{+}_{c,t} & \text{if } \mathrm{sgn}(w_{c,t}) > 0 \\ w_{c,t} \Big/ \mu^{-}_{c,t} & \text{otherwise} \end{cases}$$

where $NCS_{c,t}$, $w_{c,t}$, $\mu^{+}_{c,t}$,and $\mu^{-}_{c,t}$ are the normalized connectivity scores, raw weighted connectivity scores, and signed means of the raw weighted connectivity scores (the mean of positive and negative values evaluated separately) within the subset of signatures corresponding to cell line c and perturbagen type t, respectively.

While meaningful comparisons can be made between the *NCS* values of reference signatures with respect to query q, it is also useful to assess if the connectivity between q and a particular signature r is significantly different from that observed between r and other queries. This is done by comparing each observed *NCS* value $ncs_{q,r}$ between the query q and a reference signature r to a distribution of NCS values representing the similarities between a reference compendium of queries (Q_ref) and r. This procedure results in a standardized measure Tau (τ) that ranges from 100 to +100 and represents the percentage of queries in Q ref with a lower $|NCS|$ than $|ncs_{q,r}|$, adjusted to retain the sign of $ncs_{q,r}$:

$$\tau_{q,r} = \mathrm{sgn}(ncs_{q,r}) \frac{100}{N} \sum_{i=1}^{N} \left[\left| ncs_{i,r} \right| < \left| ncs_{q,r} \right| \right]$$

To compare signatures in LCMap and CMap, by deriving the top- and bottom-50 most differentially expressed genes from each signature and using these as a query gene set for all signatures in either dataset, it is possible to calculate a Tau score to capture the pairwise similarity between all signatures in both datasets.

4.2.8.3 t-SNE Visualization of Gene Expression Signatures

In order to investigate broad patterns in terms of perturbagen similarity with respect to cellular context, I visualized all of the LCMap perturbation L1000 signatures using t-Distributed Stochastic Neighbor Embedding (Maaten & Hinton (2008)).

4.2.8.4 Network Comparison Method for Identifying Cell-Type-Associated Signatures

I used a network modeling approach to identify perturbagens with cell type specific L1000 gene expression signatures. I chose to investigate perturbagens which

had a statistically significant difference in the composition of their nearest neighbors in all of the cell lines in one lung cell type relative to all other cell lines in either Lung CMap or CMap. Such a perturbagen would only demonstrate such a pattern of connectivity if either all of these neighbors were not transcriptionally active in all of the other cell lines, or if the mechanism through which this perturbagen and all of the perturbagens which induced a very similar signature to it was only modified in the lung cell type in question. The procedure for generating a network representation of each cell line's perturbational profiles was as follows:

1. Calculate the pairwise Tau between every signature of every perturbagen in that cell line
2. Cast these scores into a symmetrical similarity matrix
3. To remove redundant entries, set the lower triangle of this similarity matrix to NA
4. For connectivity values >= 90, set the entry to 1, otherwise set to 0

We selected a connectivity score threshold of >=90 as a threshold for signature similarity . In this representation, each node in the network is a perturbation, and edges in the network are only drawn between nodes with a Tau signature similarity of at least 90.

4.2.8.5 Comparison of Cell Line Level Networks

Given these cell line networks, I performed all pairwise comparisons of them by adding their matrix representations, and counting the number of entries = 2, which represent edges shared between the same two perturbagens in both cell lines. In this context, cell lines with overall similar responses to perturbation are assumed to have more consistent edges than cell lines with fewer similar responses.

4.2.8.6 Identifying Compounds With Fibroblast Associated Activity

To identify compounds which may have cell-type associated effects, I identified compounds with the highest proportion of connected neighbor perturbations (Tau >= 90) across the 3 fibroblast cell lines relative to all other LCMap and CMap cell lines. I further selected compounds which had a *consistent* set of neighbor perturbations across each of the fibroblast cell lines.

In order to establish the significance of the consistency of these compound neighborhoods in fibroblast cell lines, I utilized a bootstrapping approach as follows:

1. Select any 3 cell lines from either Lung CMap or CMap
2. Generate the matrix-network representation for that cell line as described above
3. Randomly permute the edges in these networks, maintaining the frequency of compound connections while discarding their inherent biological information
4. Note whether any compound in these 3 random networks has a consistent neighborhood appearing in all 3 networks comprised of at least as many compounds as the observed compound neighborhood
5. Repeat 1-4 for 1000 iterations
6. The p-value for how often one should expect to see a consistent neighborhood of the observed size in a random network is calculated as

$$\frac{\#\,Times\ Similar\ Consistent\ Neighborhood\ Size\ Observed}{1000}$$

4.2.8.7 Identifying Compound Associated Genes

In order to identify genes associated with compounds demonstrating fibroblast-associated activity, I first used a LIMMA Smyth (2005b) linear modeling approach on signatures of every compound to identify genes at Level 5 (Z-score) which had significantly higher or lower Z-scores in fibroblasts relative to all other LCMap and CMap cell lines. The model utilized was as follows:

$$Zscore = Perturbation\_Involves\_Fibroblast + Batch$$

Where *Perturturbation_Involves_Batch* is a dummy variable indicating whether a perturbation involves a fibroblast cell line or not, and *Batch* is a dummy variable indicating whether the perturbation is in LCMap or CMap. I identified all genes for which this model has an FDR adjusted p-value < 0.05 (Benjamini & Hochberg (1995)) as being associated with a compound's activity in fibroblasts.

In order to understand how genes selected in this way may behave in unperturbed cell lines, I again used a LIMMA Smyth (2004) linear modeling approach. For the genes selected as being fibroblast-compound-associated in the previous step, I determined which were associated with differential expression between DMSO exposed fibroblasts and DMSO exposed non-fibroblast cell lines in the non-differential-expression level data (Level 3) via the following model:

$$Gene\ Expression = Perturbation\_Involves\_Fibroblast$$

where *Perturbation_Involves_Fibroblast* is a dummy variable indicating whether the DMSO perturbation involved one of the 3 fibroblast cell lines. This test was performed solely in the LCMap DMSO Level 3 data, excluding the need for the batch term used in the previous step. Genes with an FDR adjusted p-value < 0.05

(Benjamini & Hochberg (1995)) as being associated with unperturbed fibroblasts as well as fibroblasts perturbed by the compound in the previous step.

4.2.9 Functional Gene Set Enrichment

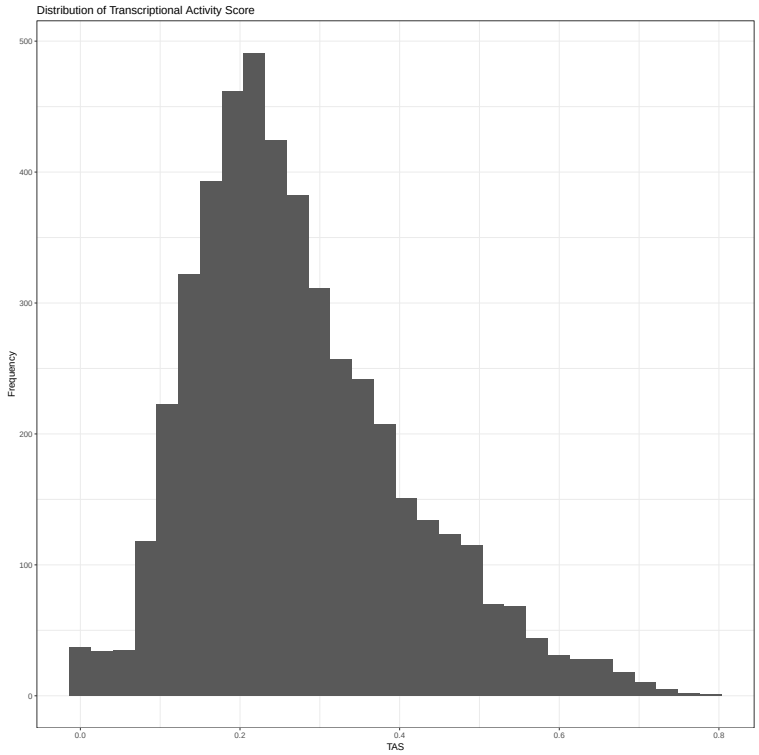
To functionally characterize compound associated genes, I used the Enrichr web tool (Chen et al. (2013b)) to determine gene sets significantly associated with these compound associated genes, identifying gene sets with an Enrichr-determined adjusted p-value < 0.05 as significant.

4.3 RESULTS

4.3.1 Characterization of Transcriptionally Active Perturbagens in LCMap Cell Lines.

To understand the proportion of perturbagens demonstrating a strong transcriptional effect in the LCMap cell lines, I calculated the TAS score of every perturbagen's signature in each cell line assayed (Figure 4.1). The Touchstone CMap dataset defined a TAS threshold of 0.2 as indicating a perturbagen with a strong transcriptional response, and indicated that while 92% of established drugs demonstrated TAS scores above this threshold in at least one cell line, only 15% of un-optimized perturbagens exceeded this threshold. 54% of the LCMap signatures yielded a TAS > 0.2, suggesting slightly less than half of the perturbations assayed did not meet the original dataset's standard for a strong L1000 transcriptional response (Subramanian et al., 2017). This could indicate that some of the perturbagens randomly selected for our dataset fell among these less transcriptionally active perturbagens, or that a batch-specific TAS threshold for L1000 datasets is more appropriate. Subramanian et al. also indicate that perturbagens with a TAS less than this threshold

Figure 4.1

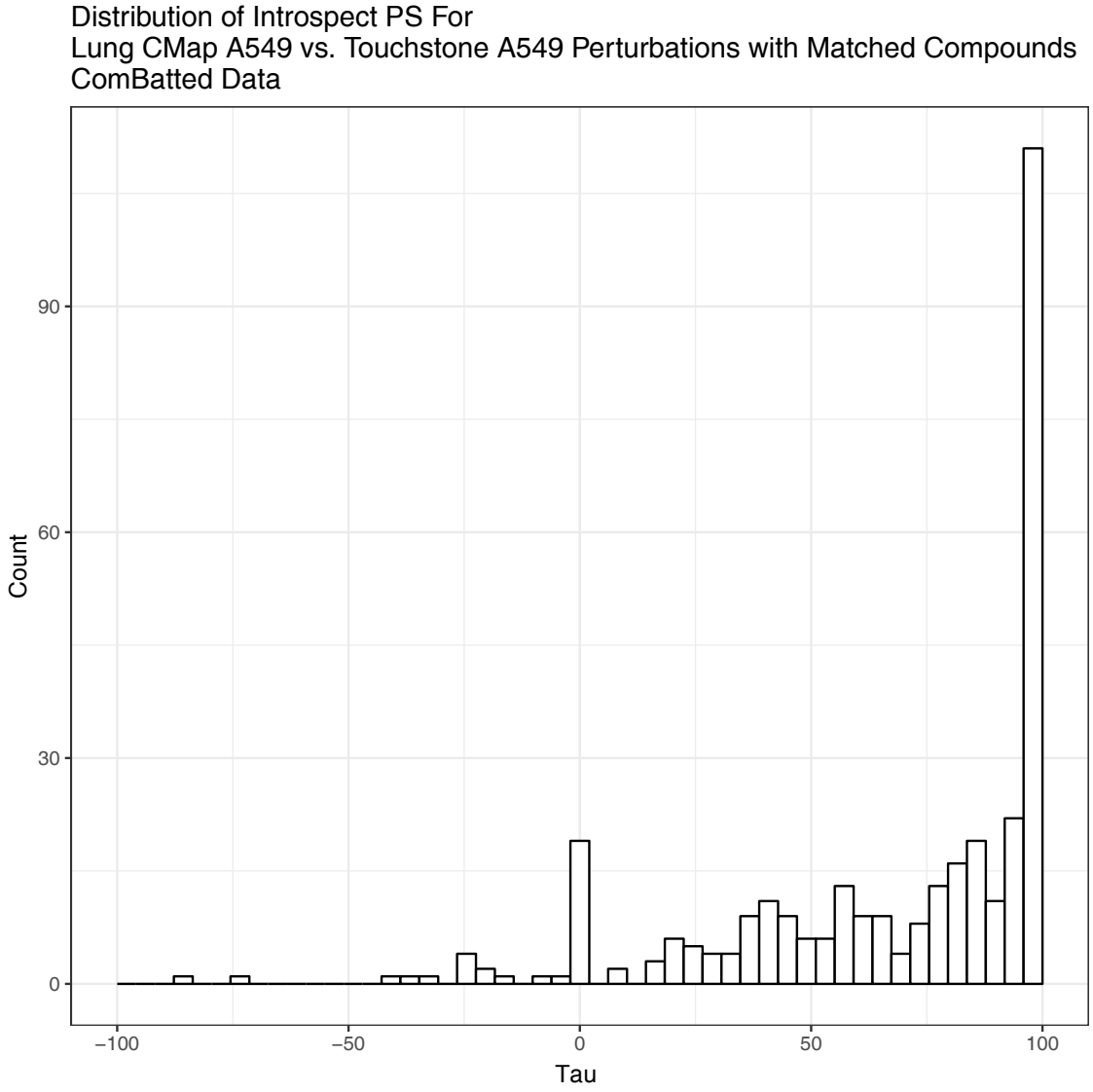


could possibly consist of perturbagens with cell-type associated effects. It is possible that the perturbagens we randomly selected are comprised of both inactive perturbagens and perturbagen with a lung cell type associated effects.

4.3.2 Similarity of Gene Expression Signatures in A549 Cells Between Datasets.

Because A549 cells were assayed in both the LCMap and CMap datasets, and because LCMap's perturbagens are a proper subset of those assayed in CMap, the LCMap A549 signatures should be expected to very closely replicate the corresponding CMap A549 signatures. This could also help indicate major batch effects between the LCMap dataset and CMap dataset. To verify this expected signature similarity, I calculated the Tau connectivity score between the signature of each perturbagen assayed in both the LCMap and CMap A549 cell lines. 31% of

Figure 4.2



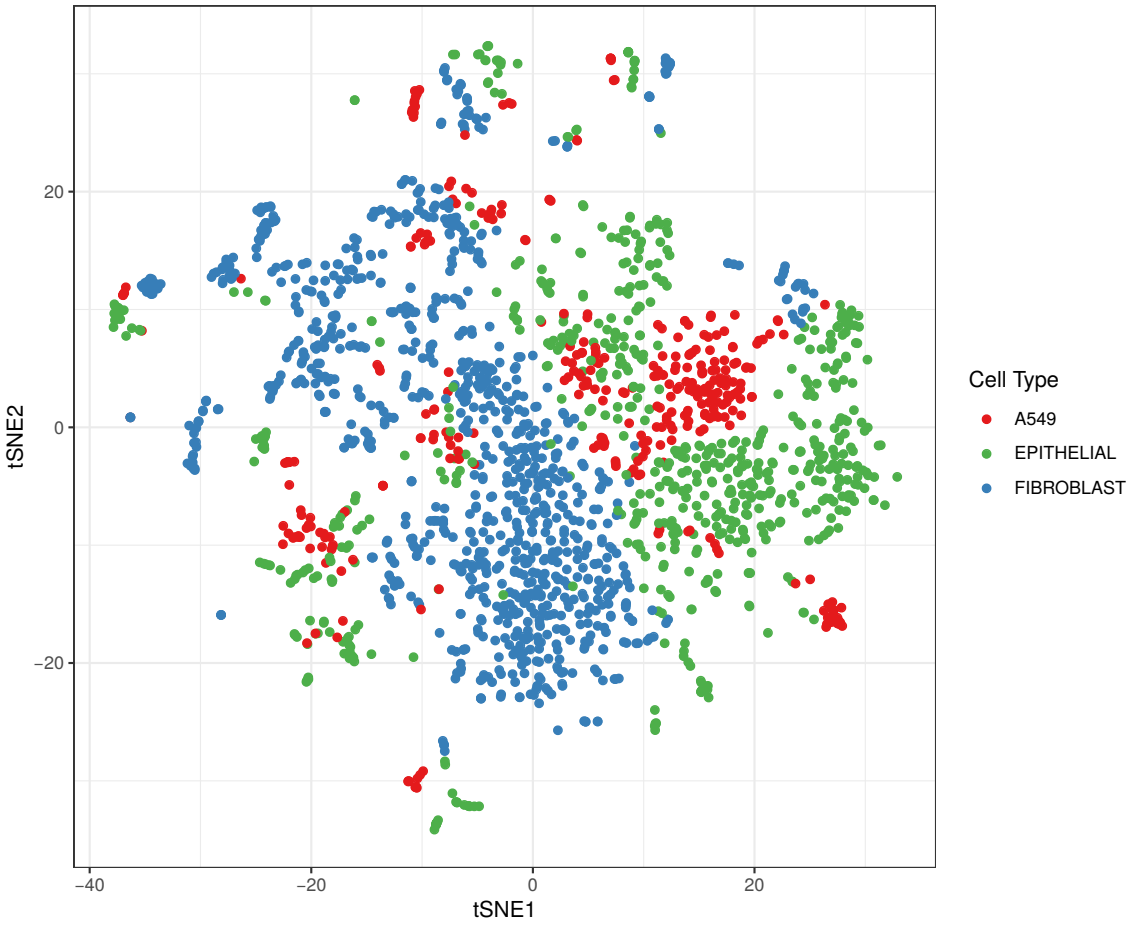
all LCMap perturbagens yielded a tau score >= 90, and 93% of all LCMAP perturbagens yielded a tau score > 0, suggesting that most of the perturbagens selected for LCMap yielded a similar signature between the datasets despite being generated in effectively different batches (Figure 4.2).

We additionally performed STR profiling of the LCMap A549s to ensure genetic similarity with the reference STR profile for this cell line. The STR profile

determined for the LCMap A549s exactly matched the ATCC reference profile for this cell line, suggesting that the LCMap and CMap A549s should be comparable from a genetic standpoint.

4.3.3 Overview of L1000 Gene Expression Signatures in LCMap Perturbations.

Figure 4.3

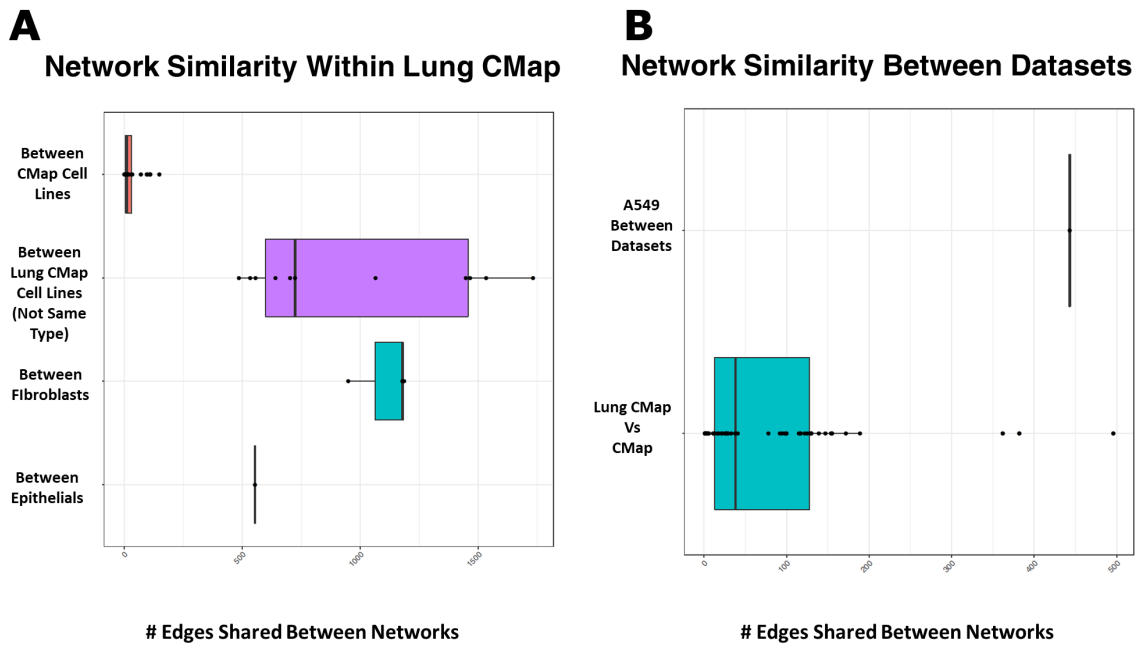


To capture a high-level summary of all of the signatures generated in LCMap and their relative similarity to each other, we visualized all of the L1000 z-score expression data from LCMap using t-SNE (Figure 4.3). Points visualized closely

in t-SNE space represent perturbational signatures which are similar. I observed groups of perturbations (indicated) which induced similar gene expression signatures between all three of the cell types (fibroblast, epithelial and A549) in LCMap. These sets of perturbations tended to be comprised of perturbagens with a single mechanism of action, such as HDAC or HSP inhibition. perturbagens with these mechanisms of action were previously found to reproduce consistent transcriptional responses in the Touchstone CMap cell lines (Subramanian et al. (2017)). However, the majority of perturbations clustered within these cell types.

4.3.4 Comparison of Cell Line Network Structural Similarities

Figure 4.4

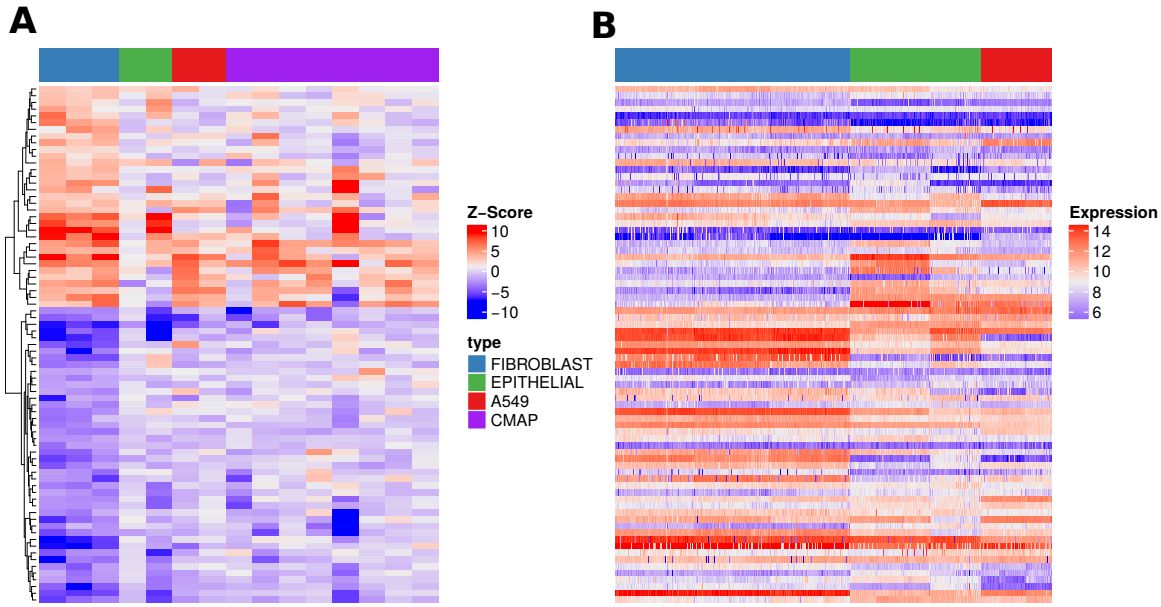


To gauge how similar responses to perturbation was across cell lines, I calculated the number of edges drawn between the same pairs of compounds for each pair of cell line networks (Figure 4.4). I observed that fibroblast cell lines

had more network edges in common than the two epithelial cell lines, but that in some cases cross-cell-type network comparisons yielded more shared network edges than within-cell-type comparisons (LUNGvsLUNG comparisons). Cell line network comparisons amongst the CMap cell lines, or between LCMap and CMap cell lines, tended to have very few shared edges, if any. The networks representing the LCMap and CMap A549 cell lines demonstrated fewer shared edges than within-cell-type comparisons (EPITHELIAL and FIBROBLAST), though these networks had more in common than almost all cross-database or within-CMap network comparisons.

4.3.5 Identification of PHA-665752, a c-MET inhibitor Demonstrating Fibroblast Associated Transcriptional Activity

Figure 4.5



I identified a list of perturbagens which both had a strong transcriptional effect (TAS > 0.2), and a "neighborhood" of connected perturbagens which consis-

tently appeared within lung fibroblasts that did not appear within any subset of a comparable number of the CMap cell lines. PHA-665752 was the perturbagen with the largest consistent neighborhood within fibroblasts which did not appear in other cell lines (33), suggesting it has transcriptional effects specific to that cell type which were not observed in other assayed cell types.

Through the differential expression approach above, I identified 72 42 genes ($q < 0.05$) associated with fibroblasts perturbed by PHA-665752, as well as in unperturbed fibroblasts compared to all oter unperturbed cell lines. Functional enrichment of these genes yielded significant enrichment of genes associated with IMR90 cells, as well as genes associated with fibroblast, myofibroblast, myoblast, and fetal lung cell types. .

4.4 DISCUSSION

The Lung Connectivity Map (LCMap) represents a novel compendium of gene expression signatures of molecular perturbation of primary, non-cancer lung cell lines. It represents the first such dataset generated through the CMap pipeline using primary, non-cancer cell lines, and serves as a relevant proof-of-concept for future studies intending to perform high throughput gene expression measurements on the L1000 platform.

Importantly, the identification of PHA-665752 provides an example of how LCMap contains responses to perturbagen perturbation that are not only unique to a single cell line, but are associated specifically with several cell lines of the same cell type and tissue of origin. This perturbagen has previously been shown to impair wound healing in lung fibroblasts *in vitro*(Ito et al., 2014), and has been implicated in inducing apoptosis in mouse-derived lung adenocarcinoma cell lines (Yang et al., 2008). Because this compound most significantly modulates genes

associated with fibroblasts, fetal lung fibroblasts, and myofibroblasts, it is likely that this perturbagen displays a fibroblast-associated activity not because it has off-target effects, but because it modulates genes associated with the baseline expression of these cell types. This represents a discernible biological response that was not represented in the original CMap dataset, and suggests that perturbations whose effects primarily modulate the baseline program of gene expression in a cell type not found in a given dataset will be missed. This could have important downstream implications towards CMap's ability to identify potential therapeutics with effects specific to these missing cell types, and argue in favor of the inclusion of an increase in the amount of cell types, with each ideally represented by multiple cell lines, in future expansions of CMap.

Several limitations are relevant in interpreting these results. Because CMap and LCMap were generated at different time points, by different teams of experimentalists, and because each of these datasets has some degree of inherent batch effects due to the assay design, it is possible that lung-cell-type associated responses to perturbation I identified can be attributed to a batch effect. However, the similarity of A549 perturbation signatures despite variance introduced by these technical effects (Figure 4.2) mitigate the notion that there is an uninterpretably large batch effect between these two datasets. Another important limitation to note is the selection of compounds used in LCMap, which was performed at random. Selection of compounds with known lung cell type associated activity would have proven useful for generation of testable positive control perturbational signatures. Our random compound selection also did not account for the variation in transcriptional activity of compounds in the CMap library, and thus LCMap's comparatively smaller compound library has a proportionately small amount of transcriptionally active compounds for which we may have been able to discern cell-type-

specific responses. Finally, LCMap only profiled a total of 3 fibroblast cell lines and 2 epithelial cell lines. Inclusion of an increased number of each type would have enabled us to assert with more statistical certainty the size of the effect of the compounds we detected.

In this chapter, I described the Lung Connectivity Map (LCMap), which served as a proof-of-concept experiment demonstrating the feasibility of the CMap L1000 protocol in non-cancer cell lines, as well as the identification of lung cell-type-associated responses to compound perturbation.

CHAPTER 5

General Conclusions

The work represented in this thesis represents the first whole-transcriptome characterization of ECIG exposure in the airway epithelium, as well as novel methods towards capturing cell-type associated response to perturbation. The important implications of these chapters are: - That ECIGs induce a distinct transcriptional response in the airway epithelium relative to cigarettes, and that these products do not demonstrate concordant changes in cigarette-associated gene expression pathways, - That CELDA, a novel Bayesian hierarchical model for identifying cell subpopulations in scRNA-seq data, is able to capture subtle transcriptional differences between cell subpopulations in a hetreogeneous sample and that these differences are biologically meaningful, and - That LCMap, a platform for high-throughput profiling of compound perturbation in primary non-cancer lung cell lines, is able to identify lung cell-type associated responses to perturbation that are not observable in the comparable perturbations of the CMap.

Together, this work demonstrates both approaches for contextualizing novel exposures as well as methods for elucidating cell-level heterogeneity of response to perturbation.

.1 DIFFERENTIAL EXPRESSION OF AFFYMETRIX HUGENE ST 1.0 PROBES BY ECIG/TCIG USE

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
10626_at	TRIM16	2.219424596	0.038351337	0.298211024	1a
11025_at	LILRB3	3.578273779	0.020962622	0.170036215	1a
11314_at	CD300A	3.277573637	0.047565898	0.170036215	1a
1240_at	CMKLR1	3.368680029	0.014669825	0.170036215	1a
133_at	ADM	3.701768324	0.031057938	0.170036215	1a
146862_at	UNC45B	2.903400653	0.016773967	0.188887324	1a
2204_at	FCAR	2.467840841	0.014408564	0.251379895	1a
23584_at	VSIG2	2.816391539	0.027169378	0.201770016	1a
2692_at	GHRHR	3.245444274	0.03097562	0.170036215	1a
2832_at	NPBWR2	2.531253354	0.046359813	0.251379895	1a
283316_at	CD163L1	2.486693448	0.000765082	0.251379895	1a
283487_at	LINC00346	2.44082676	0.000376328	0.253675577	1a
284422_at	C19orf77	2.107186858	0.038139243	0.316983141	1a
3239_at	HOXD13	3.29219042	0.036865416	0.170036215	1a
3448_at	IFNA14	3.035384087	0.024689875	0.185624053	1a
390142_at	OR5D13	2.03123265	0.04499981	0.329413667	1a
401498_at	TMEM215	2.988787639	0.033828091	0.185674591	1a
406925_at	MIR135A1	2.382000105	0.047857608	0.26443596	1a
406967_at	MIR192	2.957188257	0.013732293	0.185674591	1a
407051_at	MIR9-3	2.742462632	9.90E-05	0.221135746	1a
4284_at	MIP	2.366321205	0.022970835	0.26624587	1a
5140_at	PDE3B	3.188394953	0.046017972	0.170036215	1a
6860_at	SYT4	3.158541966	0.038102979	0.170036215	1a
7033_at	TFF3	2.605843249	0.012837301	0.245155909	1a
81030_at	ZBP1	2.874717238	0.044647337	0.191195978	1a
83850_at	ESYT3	2.611480094	0.039609735	0.244808963	1a
9071_at	CLDN10	2.121535666	1.48E-06	0.316834501	1a
10101_at	NUBP2	2.721863041	0.014225616	0.224301998	1b
10105_at	PPIF	2.085592971	0.03246803	0.320068531	1b
10189_at	ALYREF	2.055427335	0.015372282	0.323140358	1b
10212_at	DDX39A	2.076783983	0.00022851	0.320068531	1b
10226_at	PLIN3	2.878498222	0.017366239	0.191195978	1b
1028_at	CDKN1C	2.196341943	0.028978208	0.304290138	1b
1050_at	CEBPA	2.515710159	0.040563797	0.251379895	1b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
1052_at	CEBPD	2.020094957	0.000566915	0.334642082	1b
10567_at	RABAC1	3.024018772	0.020480066	0.185624053	1b
11076_at	TPPP	2.27424894	0.000682279	0.287842533	1b
11267_at	SNF8	2.334999791	0.039164421	0.274186022	1b
113828_at	FAM83F	2.307921188	0.033553838	0.283732234	1b
114599_at	SNORD15B	2.328966678	1.83E-05	0.276152816	1b
1163_at	CKS1B	3.79384594	0.024054178	0.170036215	1b
116832_at	RPL39L	2.445152626	0.016962819	0.253675577	1b
124222_at	PAQR4	2.310992596	0.032926517	0.283686448	1b
126695_at	C1orf172	2.710072212	0.029453508	0.22567235	1b
131177_at	FAM3D	2.098307516	0.007420718	0.317664773	1b
1318_at	SLC31A2	3.131236438	0.016877204	0.170036215	1b
137797_at	LYPD2	3.546718595	0.002569086	0.170036215	1b
14_at	AAMP	2.362164314	0.003637023	0.26624587	1b
148170_at	CDC42EP5	3.478542711	0.006789838	0.170036215	1b
150368_at	FAM109B	2.733595002	0.035612432	0.222282838	1b
1535_at	CYBA	2.737758938	0.001146815	0.222078246	1b
153768_at	PRELID2	2.38360267	0.028083494	0.26443596	1b
1572_at	CYP2F1	2.099794342	0.004069373	0.317664773	1b
1613_at	DAPK3	2.434813508	0.004530209	0.253675577	1b
161424_at	NOP9	2.36827557	6.63E-05	0.26624587	1b
162515_at	SLC16A11	2.100276922	0.025529036	0.317664773	1b
170463_at	SSBP4	2.129529921	0.027696798	0.316834501	1b
1849_at	DUSP7	3.141323027	0.000651953	0.170036215	1b
192111_at	PGAM5	3.404151352	0.001818552	0.170036215	1b
200916_at	RPL22L1	3.954679445	0.010363901	0.170036215	1b
2193_at	FARSA	2.23417058	0.005036449	0.296179358	1b
2194_at	FASN	2.274869161	0.000950938	0.287842533	1b
222_at	ALDH3B2	2.455346733	0.035845223	0.253675577	1b
2517_at	FUCA1	2.19231834	0.027067168	0.304290138	1b
25851_at	TECPR1	3.103451426	0.040018236	0.170036215	1b
26090_at	ABHD12	2.172998025	0.033548272	0.309207233	1b
26155_at	NOC2L	2.09717682	0.000211972	0.317664773	1b
26769_at	SNORD81	2.084826413	0.005859334	0.320068531	1b
26780_at	SNORA68	2.26299305	0.001008125	0.291569404	1b
26788_at	SNORD60	2.752262503	0.002262557	0.219616464	1b
26799_at	SNORD50A	2.132274152	0.007297209	0.316834501	1b
26813_at	SNORD36C	2.731336184	0.002330297	0.222282838	1b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
26817_at	SNORD34	2.843750475	0.00746714	0.197182753	1b
27183_at	VPS4A	2.080631881	0.002855167	0.320068531	1b
283871_at	PGP	3.149225457	0.000211972	0.170036215	1b
284361_at	EMC10	2.433742619	0.002037044	0.253675577	1b
28989_at	NTMT1	2.766348308	0.048612199	0.219616464	1b
29890_at	RBM15B	2.770792135	0.001313423	0.219616464	1b
29988_at	SLC2A8	2.968421027	0.028397743	0.185674591	1b
29997_at	GLTSCR2	2.535882966	0.001962602	0.251379895	1b
3397_at	ID1	2.0505585	0.006038245	0.325571539	1b
3430_at	IFI35	2.374575843	0.000377703	0.26624587	1b
3505_at	IGHGP	2.593754058	0.047089906	0.245155909	1b
362_at	AQP5	2.270696891	0.000922025	0.288894618	1b
374659_at	HDDC3	2.071939789	0.008098326	0.321630104	1b
4046_at	LSP1	2.927116588	0.040133366	0.188887324	1b
4141_at	MARS	2.506114108	0.001381876	0.251379895	1b
440093_at	H3F3C	3.459693496	0.000362719	0.170036215	1b
4543_at	MTNR1A	2.193683156	0.018926581	0.304290138	1b
4595_at	MUTYH	2.403666575	0.040955881	0.262870805	1b
4641_at	MYO1C	2.527672244	0.014511024	0.251379895	1b
4669_at	NAGLU	2.360280217	0.01249207	0.26624587	1b
4708_at	NDUFB2	3.155595242	0.000376328	0.170036215	1b
489_at	ATP2A3	2.54318362	0.003903415	0.251379895	1b
51073_at	MRPL4	2.448916543	0.013623778	0.253675577	1b
51147_at	ING4	2.24572767	0.032103139	0.296179358	1b
51421_at	AMOTL2	2.56828591	0.011009323	0.251379895	1b
51588_at	PIAS4	2.320093492	0.000552167	0.280425141	1b
517_at	ATP5G2	2.480033059	0.026364751	0.251379895	1b
5184_at	PEPD	2.172388884	0.000264818	0.309207233	1b
53904_at	MYO3A	3.106126832	0.04698365	0.170036215	1b
54442_at	KCTD5	2.277062166	0.032039724	0.287842533	1b
54461_at	FBXW5	2.087621667	0.013336998	0.320068531	1b
54531_at	MIER2	2.200492263	0.002298365	0.304290138	1b
54555_at	DDX49	2.864052858	0.035477411	0.191195978	1b
54854_at	FAM83E	2.495750133	0.002183432	0.251379895	1b
54929_at	TMEM161A	2.168571145	0.006202367	0.309207233	1b
54958_at	TMEM160	2.205028119	0.048176165	0.303564458	1b
55111_at	PLEKHJ1	2.395559007	0.000902159	0.263474197	1b
55357_at	TBC1D2	2.748901204	0.019651606	0.219616464	1b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
55684_at	RABL6	2.891227158	0.000298285	0.190087964	1b
56147_at	PCDHA1	2.078000606	0.026030575	0.320068531	1b
56662_at	VTRNA1-3	2.838353828	0.001942753	0.197182753	1b
56901_at	NDUFA4L2	2.632374885	0.010447149	0.240360882	1b
57109_at	REXO4	3.264074818	0.001146823	0.170036215	1b
57140_at	RNPEPL1	2.506043732	0.000498668	0.251379895	1b
57165_at	GJC2	2.520815592	0.022766659	0.251379895	1b
57787_at	MARK4	3.117841539	0.00293243	0.170036215	1b
6079_at	SNORD15A	2.137665933	0.000306823	0.316834501	1b
6090_at	RNY5	2.465067203	0.000454725	0.251379895	1b
6134_at	RPL10	2.110947163	0.040787395	0.316834501	1b
6175_at	RPLP0	2.418946189	0.013042854	0.259203624	1b
619505_at	SNORA21	3.132241037	0.014196426	0.170036215	1b
619570_at	SNORD95	2.314633083	0.000362719	0.282654107	1b
6203_at	RPS9	2.694909928	9.52E-05	0.22567235	1b
6208_at	RPS14	2.163028444	0.010253328	0.31056148	1b
6275_at	S100A4	2.074469753	0.037851992	0.320970662	1b
6277_at	S100A6	2.38250723	0.024566749	0.26443596	1b
63875_at	MRPL17	2.594504862	0.040106568	0.245155909	1b
64101_at	LRRC4	2.098095726	0.001495884	0.317664773	1b
64115_at	C10orf54	3.002147234	0.025604294	0.185674591	1b
64131_at	XYLT1	2.407084234	0.026282935	0.261773777	1b
64207_at	IRF2BPL	2.470108763	0.011630567	0.251379895	1b
64782_at	AEN	2.125435521	0.040545876	0.316834501	1b
64928_at	MRPL14	3.407073293	0.000892674	0.170036215	1b
64979_at	MRPL36	2.1747446	0.037538792	0.309207233	1b
65094_at	JMJD4	2.839609588	0.009834576	0.197182753	1b
65249_at	ZSWIM4	3.778892014	0.005414934	0.170036215	1b
65263_at	PYCRL	2.35959631	0.048338804	0.26624587	1b
653784_at	MZT2A	2.443051207	0.04698365	0.253675577	1b
65996_at	MGC2752	2.627114939	0.001778805	0.240360882	1b
677773_at	SCARNA23	2.625709611	0.01660373	0.240360882	1b
677777_at	SCARNA12	2.237887879	0.008157399	0.296179358	1b
677798_at	SNORA9	2.655310102	0.000482158	0.235385224	1b
677809_at	SNORA24	2.12578282	0.034104765	0.316834501	1b
677819_at	SNORA37	2.105201845	0.004166548	0.317374895	1b
677850_at	SNORD1C	2.084031894	0.003933011	0.320068531	1b
6794_at	STK11	2.380305491	0.000235544	0.26443596	1b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
6810_at	STX4	2.555231976	0.006109774	0.251379895	1b
692072_at	SNORD5	2.307074376	0.031586711	0.283732234	1b
692073_at	SNORA16A	3.019993893	0.002539261	0.185624053	1b
692158_at	SNORA57	2.110804199	0.005474256	0.316834501	1b
692196_at	SNORD76	2.484103748	0.010302503	0.251379895	1b
692227_at	SNORD104	2.038854608	0.000589431	0.329350388	1b
718_at	C3	2.057177708	0.002262557	0.323140358	1b
7264_at	TSTA3	2.383658521	0.0119419	0.26443596	1b
7791_at	ZYX	2.87054891	0.000498668	0.191195978	1b
79050_at	NOC4L	3.614469342	0.022959015	0.170036215	1b
79058_at	ASPSCR1	2.237019055	0.014144554	0.296179358	1b
79180_at	EFHD2	2.426308775	0.043292724	0.256137389	1b
79623_at	GALNT14	2.366717119	0.026654818	0.26624587	1b
79697_at	C14orf169	2.056873167	0.01006997	0.323140358	1b
79713_at	IGFLR1	2.616284324	0.04546082	0.243390026	1b
79751_at	SLC25A22	2.693678838	0.013500763	0.22567235	1b
79803_at	HPS6	2.527931422	0.012446717	0.251379895	1b
80725_at	SRCIN1	3.16026783	0.024476569	0.170036215	1b
80851_at	SH3BP5L	2.10228355	0.029449039	0.317664773	1b
81570_at	CLPB	2.670789051	0.006567661	0.234038296	1b
81628_at	TSC22D4	2.549128785	0.020282261	0.251379895	1b
81844_at	TRIM56	2.077585063	0.003255749	0.320068531	1b
8192_at	CLPP	2.023561758	0.023277955	0.33354661	1b
81926_at	FAM108A1	3.237476933	0.032356821	0.170036215	1b
8294_at	HIST1H4I	2.094254147	0.031794798	0.317874618	1b
83481_at	EPPK1	2.275862132	0.020178951	0.287842533	1b
84262_at	PSMG3	2.093650619	0.000201901	0.317874618	1b
84895_at	FAM73B	2.700474809	0.001416811	0.22567235	1b
84954_at	MPND	2.668342561	0.032926517	0.234038296	1b
8568_at	RRP1	2.120532488	0.001877242	0.316834501	1b
8677_at	STX10	2.599565902	0.034104765	0.245155909	1b
8815_at	BANF1	3.049012597	7.23E-05	0.184409427	1b
89790_at	SIGLEC10	3.338542723	0.03890325	0.170036215	1b
90850_at	ZNF598	2.286640379	0.023415052	0.287842533	1b
9123_at	SLC16A3	2.895706164	0.001384229	0.190087964	1b
9144_at	SYNGR2	2.273900016	0.021505933	0.287842533	1b
9150_at	CTDP1	2.220383498	0.035224838	0.298211024	1b
91582_at	RPS19BP1	2.333436219	0.01055089	0.274219368	1b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
92002_at	FAM58A	2.085846097	0.009104471	0.320068531	1b
92305_at	TMEM129	2.359632772	0.048309109	0.26624587	1b
92609_at	TIMM50	2.291374466	0.002031069	0.287842533	1b
9277_at	WDR46	2.146598153	0.025973697	0.316296017	1b
92822_at	ZNF276	2.21390778	0.004387997	0.299992096	1b
9299_at	SNORD30	2.351350561	0.000779907	0.269540789	1b
9301_at	SNORD27	2.061736251	0.031794798	0.323140358	1b
9304_at	SNORD22	2.344882058	1.05E-05	0.269759597	1b
93100_at	NAPRT1	3.276167237	0.028312916	0.170036215	1b
93210_at	PGAP3	3.199924467	0.033038059	0.170036215	1b
9711_at	KIAA0226	2.124263388	0.042087404	0.316834501	1b
9894_at	TELO2	2.08208617	0.02359369	0.320068531	1b
100286979_at	ANAPC1P1	-3.432469849	0.0041276	0.170036215	2a
10253_at	SPRY2	-2.066384946	0.024463709	0.323029894	2a
11138_at	TBC1D8	-2.262582814	0.009573588	0.291569404	2a
1185_at	CLCN6	-2.842854898	0.000992326	0.197182753	2a
1287_at	COL4A5	-2.115233431	0.004387997	0.316834501	2a
130733_at	TMEM178A	-2.867065305	0.000173514	0.191195978	2a
1352_at	COX10	-2.278340539	0.013677169	0.287842533	2a
152195_at	NUDT16P1	-2.180999447	0.007125346	0.30734603	2a
164395_at	TTLL9	-2.064155592	0.04391438	0.323140358	2a
164832_at	LONRF2	-3.463772379	0.040563797	0.170036215	2a
219_at	ALDH1B1	-3.11851669	0.036757607	0.170036215	2a
2200_at	FBN1	-2.130856934	0.046260973	0.316834501	2a
221806_at	VWDE	-3.239261083	0.023559067	0.170036215	2a
23331_at	TTC28	-2.357476954	2.38E-05	0.266609559	2a
23462_at	HEY1	-2.415859886	0.034750964	0.259203624	2a
26284_at	ERAL1	-2.337321507	0.035683118	0.273662465	2a
26580_at	BSCL2	-2.517508191	0.02876799	0.251379895	2a
2737_at	GLI3	-2.152110037	0.002701645	0.314592809	2a
284402_at	SCGB2B2	-3.230457697	0.048426576	0.170036215	2a
284459_at	HKR1	-2.754564098	0.044167316	0.219616464	2a
3131_at	HLF	-2.989268023	0.027265091	0.185674591	2a
4134_at	MAP4	-2.555573713	0.044302729	0.251379895	2a
440104_at	TMEM198B	-2.319479179	0.000376328	0.280425141	2a
5087_at	PBX1	-2.957890659	0.019224202	0.185674591	2a
5218_at	CDK14	-2.133353063	0.010211341	0.316834501	2a
54820_at	NDE1	-2.298495369	0.00011834	0.286529697	2a

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
55084_at	SOBP	-2.696122914	0.0488244	0.22567235	2a
552_at	AVPR1A	-3.518344743	0.032174751	0.170036215	2a
55742_at	PARVA	-2.305142595	0.045864004	0.284037781	2a
55841_at	WWC3	-2.763308313	0.001495884	0.219616464	2a
57099_at	AVEN	-2.484759634	0.019740444	0.251379895	2a
57118_at	CAMK1D	-2.045165989	0.005174803	0.326733736	2a
57478_at	USP31	-2.110563161	0.027285536	0.316834501	2a
57496_at	MKL2	-3.22450952	0.044901435	0.170036215	2a
57507_at	ZNF608	-2.902160704	0.03057192	0.188887324	2a
57835_at	SLC4A5	-2.879581935	0.029013305	0.191195978	2a
5793_at	PTPRG	-2.035148239	0.036339979	0.329350388	2a
58492_at	ZNF77	-2.131191084	0.000672034	0.316834501	2a
64427_at	TTC31	-2.626568611	0.002284973	0.240360882	2a
647135_at	SRGAP2B	-2.943237601	0.030686256	0.186626283	2a
6830_at	SUPT6H	-2.222915722	0.022480866	0.297595538	2a
7436_at	VLDLR	-2.408306402	0.006691129	0.261773777	2a
80206_at	FHOD3	-2.109583808	0.030516463	0.316834501	2a
83450_at	LRRC48	-2.301933527	0.040536801	0.285203289	2a
8440_at	NCK2	-3.123064026	0.036032892	0.170036215	2a
84436_at	ZNF528	-2.232771464	0.04421411	0.296231304	2a
9037_at	SEMA5A	-2.077296454	0.001775861	0.320068531	2a
93233_at	CCDC114	-2.044907664	0.002740324	0.326733736	2a
9353_at	SLIT2	-2.6948711	0.024160058	0.22567235	2a
9656_at	MDC1	-2.924733708	0.029071142	0.188887324	2a
9863_at	MAGI2	-2.234236514	0.043478653	0.296179358	2a
9898_at	UBAP2L	-2.478881043	0.045265494	0.251379895	2a
100506564_at	THEGL	-2.748201482	0.011553247	0.219616464	2b
100529855_at	ZNF625-ZNF20	-2.032900204	0.002225906	0.329350388	2b
100652824_at	LOC100652824	-2.700241532	0.047555742	0.22567235	2b
10283_at	CWC27	-2.057409966	0.006265269	0.323140358	2b
10350_at	ABCA9	-2.227722835	0.006644644	0.29715543	2b
10427_at	SEC24B	-2.226848464	0.030425255	0.29715543	2b
10428_at	CFDP1	-2.212970604	0.002295337	0.299992096	2b
10600_at	USP16	-2.121513437	0.003288224	0.316834501	2b
10826_at	C5orf4	-3.690437737	0.002624937	0.170036215	2b
10927_at	SPIN1	-2.279850189	0.021025888	0.287842533	2b
10943_at	MSL3	-3.184952109	0.000235544	0.170036215	2b
11147_at	HHLA3	-3.290019183	0.000174064	0.170036215	2b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
11280_at	SCN11A	-4.594369629	0.000922025	0.055826959	2b
113263_at	GLCCI1	-2.19699154	0.000438803	0.304290138	2b
114883_at	OSBPL9	-2.962219413	0.000580534	0.185674591	2b
114932_at	MRFAP1L1	-2.250449683	0.00746714	0.295162685	2b
119710_at	C11orf74	-2.535959428	0.01025149	0.251379895	2b
123016_at	TTC8	-2.159656522	0.000684994	0.312095691	2b
123624_at	AGBL1	-2.151113672	0.027663249	0.314592809	2b
126859_at	AXDND1	-2.281065074	0.021572541	0.287842533	2b
127003_at	C1orf194	-2.241117366	0.040162523	0.296179358	2b
128710_at	C20orf94	-2.346431428	0.004608558	0.269743641	2b
129831_at	RBM45	-2.168874404	0.024378556	0.309207233	2b
130940_at	CCDC148	-2.438659151	0.001361143	0.253675577	2b
132851_at	SPATA4	-2.254504836	0.03155723	0.293323097	2b
134728_at	IRAK1BP1	-2.242333131	0.037485591	0.296179358	2b
136332_at	LRGUK	-2.434835882	0.029453508	0.253675577	2b
139212_at	PIH1D3	-2.069643466	0.035101382	0.322183541	2b
140733_at	MACROD2	-2.53314007	0.005036449	0.251379895	2b
145447_at	ABHD12B	-2.20397918	0.038302494	0.303564458	2b
145482_at	PTGR2	-2.167233104	0.009644289	0.30929771	2b
148268_at	ZNF570	-2.035893393	0.009285223	0.329350388	2b
150864_at	FAM117B	-2.584637172	0.000211811	0.249338199	2b
150967_at	PKI55	-3.157495648	0.04359861	0.170036215	2b
151827_at	LRRC34	-2.22877499	0.033737262	0.29715543	2b
152110_at	NEK10	-2.361754225	0.034750964	0.26624587	2b
153643_at	FAM81B	-2.236652102	0.00035033	0.296179358	2b
154091_at	SLC2A12	-2.483369353	0.000444838	0.251379895	2b
1558_at	CYP2C8	-2.914165575	0.003692569	0.188887324	2b
160140_at	C11orf65	-2.130750108	0.001204622	0.316834501	2b
160857_at	CCDC122	-2.054972385	0.019456215	0.323140358	2b
161394_at	SAMD15	-2.640682211	0.000829754	0.240360882	2b
161835_at	FSIP1	-2.623745691	0.003559464	0.240360882	2b
163081_at	ZNF567	-2.225527875	0.005232433	0.29715543	2b
164684_at	WBP2NL	-3.576972637	0.022703758	0.170036215	2b
166824_at	RASSF6	-2.063462594	0.036750054	0.323140358	2b
1740_at	DLG2	-2.661334541	0.004385992	0.23498643	2b
196527_at	ANO6	-3.065905516	0.000241631	0.181529502	2b
1982_at	EIF4G2	-2.238278267	0.000959834	0.296179358	2b
200373_at	PCDP1	-2.640221536	0.005934007	0.240360882	2b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
2005_at	ELK4	-2.888295238	0.010095237	0.190087964	2b
203523_at	ZNF449	-2.113051888	0.018112483	0.316834501	2b
205717_at	KIAA2018	-2.755861407	0.015160983	0.219616464	2b
2066_at	ERBB4	-2.523712899	0.024612831	0.251379895	2b
2070_at	EYA4	-3.00726748	0.025737637	0.185674591	2b
221458_at	KIF6	-2.134034501	0.038301609	0.316834501	2b
222255_at	ATXN7L1	-2.435941401	0.000416807	0.253675577	2b
222611_at	GPR111	-3.361404704	0.011540751	0.170036215	2b
2257_at	FGF12	-4.840850632	0.00041628	0.049772345	2b
22887_at	FOXJ3	-2.687330722	9.52E-05	0.227686466	2b
22920_at	KIFAP3	-2.114105597	0.035190646	0.316834501	2b
23168_at	RTF1	-2.279700625	0.036210658	0.287842533	2b
23248_at	RPRD2	-2.5186269	0.017930674	0.251379895	2b
23318_at	ZCCHC11	-2.041759865	0.028779803	0.328276871	2b
23360_at	FNBP4	-2.631651185	0.022199861	0.240360882	2b
23460_at	ABCA6	-2.471879568	0.014282727	0.251379895	2b
23505_at	TMEM131	-2.712558286	0.023924016	0.22567235	2b
23710_at	GABARAPL1	-2.119254372	0.001863592	0.316834501	2b
253724_at	GNN	-2.464259968	0.002707933	0.251379895	2b
255082_at	CASC2	-2.349597581	0.018609779	0.269677992	2b
258010_at	SVIP	-2.401526189	0.004623632	0.263162717	2b
25827_at	FBXL2	-2.109495072	0.006497388	0.316834501	2b
2620_at	GAS2	-2.538300734	0.026654818	0.251379895	2b
26343_at	OR5E1P	-3.199598547	0.018384267	0.170036215	2b
26355_at	FAM162A	-2.978833662	0.0488244	0.185674591	2b
26471_at	NUPR1	-2.10926221	0.002112646	0.316834501	2b
26515_at	FXC1	-2.607011958	0.020475897	0.245155909	2b
26716_at	OR2H1	-2.055363625	0.049868167	0.323140358	2b
27130_at	INVS	-3.127792003	0.004352767	0.170036215	2b
27332_at	ZNF638	-2.049449859	0.035477411	0.325641885	2b
284697_at	BTBD8	-2.693571719	0.026563117	0.22567235	2b
285195_at	SLC9A9	-2.484787381	0.023993003	0.251379895	2b
285962_at	FLJ40852	-2.080453623	9.77E-05	0.320068531	2b
286187_at	PPP1R42	-2.683996116	0.017581795	0.228017264	2b
288_at	ANK3	-2.128790527	0.039796902	0.316834501	2b
29081_at	METTL5	-2.928806312	0.004118798	0.188887324	2b
29843_at	SENP1	-2.017318984	0.000528139	0.335764372	2b
29958_at	DMGDH	-3.265534479	0.000983149	0.170036215	2b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
2998_at	GYS2	-2.479569864	0.024671632	0.251379895	2b
29980_at	DONSON	-2.661675863	0.002961197	0.23498643	2b
317671_at	RFESD	-2.594988659	0.003594693	0.245155909	2b
3223_at	HOXC6	-2.34715369	0.019960311	0.269743641	2b
338323_at	NLRP14	-2.473412129	0.00746714	0.251379895	2b
339883_at	C3orf35	-2.047414613	0.018112483	0.326378326	2b
3423_at	IDS	-2.028883555	1.05E-05	0.330398324	2b
3467_at	IFNW1	-2.439173515	0.00196917	0.253675577	2b
3488_at	IGFBP5	-3.159729005	0.013064313	0.170036215	2b
3708_at	ITPR1	-2.032307624	0.000173514	0.329350388	2b
375189_at	PFN4	-3.792463788	0.016276739	0.170036215	2b
401027_at	C2orf66	-3.03821236	0.001316844	0.185624053	2b
4131_at	MAP1B	-2.547576481	0.000498668	0.251379895	2b
414328_at	IDNK	-2.469105829	0.002938574	0.251379895	2b
4205_at	MEF2A	-2.465353262	0.00591468	0.251379895	2b
4212_at	MEIS2	-2.240789554	0.027458852	0.296179358	2b
4281_at	MID1	-2.530922652	0.026108924	0.251379895	2b
4345_at	CD200	-2.657495541	0.030777191	0.235385224	2b
442903_at	MIR331	-2.198675153	0.00969633	0.304290138	2b
4983_at	OPHN1	-2.511719577	0.024540604	0.251379895	2b
50484_at	RRM2B	-2.178329386	0.0046693	0.308355698	2b
51028_at	VPS36	-2.060753994	0.030828696	0.323140358	2b
51082_at	POLR1D	-3.209590221	0.002024504	0.170036215	2b
51095_at	TRNT1	-2.021101529	0.036823323	0.334624098	2b
51315_at	KRCC1	-2.125011438	0.018384267	0.316834501	2b
51317_at	PHF21A	-2.983494492	0.02953761	0.185674591	2b
51375_at	SNX7	-2.310106975	0.049882665	0.283686448	2b
51454_at	GULP1	-2.098590331	0.001121967	0.317664773	2b
51473_at	DCDC2	-2.14182993	0.042509388	0.316834501	2b
51538_at	ZCCHC17	-2.186573046	0.006487958	0.304314898	2b
51562_at	MBIP	-2.106851306	0.000821338	0.316983141	2b
51601_at	LIPT1	-2.084389632	0.012837301	0.320068531	2b
5165_at	PDK3	-2.194691644	0.049772181	0.304290138	2b
5332_at	PLCB4	-2.189280901	0.012200347	0.304314898	2b
5411_at	PNN	-2.960492557	0.003713195	0.185674591	2b
54462_at	FAM190B	-2.132732744	0.002527762	0.316834501	2b
5451_at	POU2F1	-2.95286609	0.0248102	0.185674591	2b
54762_at	GRAMD1C	-2.067002461	0.011765192	0.323029894	2b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
54764_at	ZRANB1	-3.774424439	0.005515451	0.170036215	2b
54845_at	ESRP1	-2.150170852	0.030088298	0.314592809	2b
548645_at	DNAJC25	-2.037260443	0.029478926	0.329350388	2b
54873_at	PALMD	-2.215948222	0.001938769	0.299715443	2b
54940_at	OCIAD1	-2.033063071	0.004365146	0.329350388	2b
54969_at	C4orf27	-2.602639562	0.000990389	0.245155909	2b
55081_at	IFT57	-2.373495123	0.003077639	0.26624587	2b
55216_at	C11orf57	-2.187341666	0.006567661	0.304314898	2b
5523_at	PPP2R3A	-2.059210396	0.035224838	0.323140358	2b
55252_at	ASXL2	-3.02060857	0.00173658	0.185624053	2b
55553_at	SOX6	-3.763513031	0.011344005	0.170036215	2b
55740_at	ENAH	-2.256749798	0.000305897	0.293323097	2b
55769_at	ZNF83	-2.256418572	0.016490837	0.293323097	2b
55777_at	MBD5	-2.827510392	0.010689344	0.201076251	2b
55779_at	WDR52	-3.051027634	0.018457014	0.184409427	2b
5587_at	PRKD1	-3.549257184	0.028222191	0.170036215	2b
5602_at	MAPK10	-2.499105672	0.000454725	0.251379895	2b
56776_at	FMN2	-2.816012815	0.002744253	0.201770016	2b
56906_at	THAP10	-2.171001942	0.006628085	0.309207233	2b
56913_at	C1GALT1	-2.095170235	0.029453508	0.317874618	2b
56987_at	BBX	-2.032657139	0.036743682	0.329350388	2b
57038_at	RARS2	-2.155935679	0.001866944	0.313885859	2b
57393_at	TMEM27	-4.070405968	0.002887668	0.170036215	2b
5747_at	PTK2	-2.542251152	0.028397743	0.251379895	2b
57501_at	KIAA1257	-2.273653376	0.004352767	0.287842533	2b
57509_at	MTUS1	-2.292942089	0.026086553	0.287842533	2b
57562_at	KIAA1377	-2.071459151	0.031794798	0.321630104	2b
57683_at	ZDBF2	-2.77139806	0.002298365	0.219616464	2b
57698_at	KIAA1598	-2.537085762	0.004792365	0.251379895	2b
58486_at	ZBED5	-2.517289159	0.040487309	0.251379895	2b
5865_at	RAB3B	-2.921961331	0.005494999	0.188887324	2b
59084_at	ENPP5	-2.19950253	0.019530252	0.304290138	2b
5915_at	RARB	-2.380787996	0.012515352	0.26443596	2b
5955_at	RCN2	-2.241858804	0.006497388	0.296179358	2b
6101_at	RP1	-2.44666495	0.002684573	0.253675577	2b
6311_at	ATXN2	-2.392722423	0.013364557	0.263474197	2b
6430_at	SRSF5	-2.787608787	0.027032515	0.215511447	2b
64518_at	TEKT3	-2.911567737	0.035618413	0.188887324	2b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
646600_at	C3orf65	-2.465519049	0.027799328	0.251379895	2b
646813_at	LOC646813	-2.224859488	0.025647967	0.29715543	2b
64864_at	RFX7	-2.450063744	0.040018236	0.253675577	2b
6655_at	SOS2	-2.823375348	0.002210836	0.201453355	2b
6760_at	SS18	-2.276519933	0.006862242	0.287842533	2b
6767_at	ST13	-2.29035328	0.015166835	0.287842533	2b
692087_at	SNORD49B	-3.099421055	0.028095711	0.170036215	2b
7029_at	TFDP2	-3.164048745	0.029695208	0.170036215	2b
7569_at	ZNF182	-2.266906691	0.042574648	0.290517051	2b
7750_at	ZMYM2	-2.394485016	0.01788293	0.263474197	2b
7879_at	RAB7A	-2.48422188	0.000534095	0.251379895	2b
79048_at	SECISBP2	-2.25555294	0.007075435	0.293323097	2b
7913_at	DEK	-2.530581662	0.023135751	0.251379895	2b
79618_at	HMBOX1	-2.280659673	0.015129364	0.287842533	2b
79663_at	HSPBAP1	-2.165919393	0.020282261	0.309373096	2b
79895_at	ATP8B4	-2.442719847	0.044155259	0.253675577	2b
79925_at	SPEF2	-2.485811577	0.008157399	0.251379895	2b
8028_at	MLLT10	-2.115668348	0.049222823	0.316834501	2b
80298_at	MTERFD3	-2.723059798	0.013653483	0.224301998	2b
81602_at	CDADC1	-2.122066219	0.036854559	0.316834501	2b
81617_at	CAB39L	-2.414552778	0.033536427	0.259203624	2b
83468_at	GLT8D2	-2.635796208	0.016770638	0.240360882	2b
83657_at	DYNLRB2	-2.966622292	0.000580534	0.185674591	2b
83939_at	EIF2A	-2.532091134	0.004436618	0.251379895	2b
8404_at	SPARCL1	-3.095267705	9.52E-05	0.170036215	2b
84144_at	SYDE2	-2.361778509	0.046319328	0.26624587	2b
84223_at	IQCG	-2.531261839	0.002740324	0.251379895	2b
84343_at	HPS3	-2.48114736	0.005591977	0.251379895	2b
84460_at	ZMAT1	-2.470604952	0.041337846	0.251379895	2b
84529_at	C15orf41	-2.431314064	0.022959015	0.254105496	2b
84911_at	ZNF382	-2.115892438	0.044269855	0.316834501	2b
84946_at	LTV1	-2.191576206	0.046359434	0.304290138	2b
85313_at	PPIL4	-2.168902792	0.021187928	0.309207233	2b
8550_at	MAPKAPK5	-2.754019818	0.020282261	0.219616464	2b
8621_at	CDK13	-2.560794137	0.021047075	0.251379895	2b
8814_at	CDKL1	-2.1538659	0.013739582	0.314507301	2b
89765_at	RSPH1	-3.599913468	0.00791382	0.170036215	2b
89782_at	LMLN	-2.393035412	0.019276188	0.263474197	2b

Table 1 – continued from previous page

Probe ID	Gene Symbol	T-Statistic	ANCOVA Q-Value	LIMMA Q-Value	Cluster
8999_at	CDKL2	-2.470496809	0.029020002	0.251379895	2b
9113_at	LATS1	-2.436668881	0.027185871	0.253675577	2b
9166_at	EBAG9	-2.393906964	0.00192803	0.263474197	2b
9208_at	LRRFIP1	-3.011172925	0.045265494	0.185674591	2b
92211_at	CDHR1	-3.263620673	0.04391438	0.170036215	2b
92737_at	DNER	-2.379812139	0.042105671	0.26443596	2b
9342_at	SNAP29	-2.186683234	0.010895867	0.304314898	2b
93587_at	TRMT10A	-2.414883361	0.024689875	0.259203624	2b
93986_at	FOXP2	-2.903989893	0.002255842	0.188887324	2b
9419_at	CRIPT	-2.492286837	0.014289023	0.251379895	2b
9481_at	SLC25A27	-2.576298175	0.041281382	0.251379895	2b
9657_at	IQCB1	-2.567958478	0.001638099	0.251379895	2b
9693_at	RAPGEF2	-2.138890668	0.04664695	0.316834501	2b
9813_at	KIAA0494	-2.208171115	0.024678074	0.302428144	2b
9851_at	KIAA0753	-2.949410472	0.041392395	0.185674591	2b
9913_at	SUPT7L	-3.300390215	0.005928194	0.170036215	2b

LIST OF JOURNAL ABBREVIATIONS

Am J Physiol ……..	American Journal of Physiology
Environ Health Perspect ……..	Environmental Health Perspectives
Genome Biol ……..	Genome Biology
IEEE Trans Pattern Anal Mach Intell..	IEEE Transactions on Pattern Analysis and Machine Intelligence
J Machine Learn Res……..	Journal of Machine Learning Research
J Mol Med……	Journal of Molecular Medicine
Mol Med……..	Molecular Medicine
N Engl J Med……..	New England Journal of Medicine
Nat Commun …….	Nature Communications
Nat Methods……...	Nature Methods
Nicotine Tob Res……	Nicotine & Tobacco Research
Physiol Genomics……..	Physiological Genomics
PLoS Comp Biol…….	PLoS Computational Genomics
Proc Natl Acad Sci USA……	Proceedings of the National Academy of Sciences of the United States of America
Respir Res……...	Respiratory Research
Toxicol Sci……...	Toxicological Sciences

WIRES Data Mining Knowl Discov…	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery

BIBLIOGRAPHY

(????). DNA sequencing costs: Data. https://www.genome.gov/27541954/dna-sequencing-costs-data/. Accessed: 2019-3-18.

(2016). Products - Data Briefs - Number 217 - October 2015. http://www.cdc.gov/nchs/data/databriefs/db217.htm (Accessed:2016-06-23).

American Thoracic Society (ATS) (2015). Electronic cigarette flavorings alter lung function at the cellular level. *Science Daily*.

Beane, J., Sebastiani, P., Liu, G., Brody, J. S., Lenburg, M. E., & Spira, A. (2007a). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.*, *8*(9), R201.

Beane, J., Sebastiani, P., Liu, G., Brody, J. S., Lenburg, M. E., & Spira, A. (2007b). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biology*, *8*(9), R201.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*(Jan), 993–1022.

CDCMMWR (2017). QuickStats: Percentage of Adults Who Ever Used an E-cigarette and Percentage Who Currently Use E-cigarettes, by Age Group National Health Interview Survey, United States, 2016. *MMWR. Morbidity and Mortality Weekly Report*, *66*.

Chari, R., Lonergan, K. M., Ng, R. T., MacAulay, C., Lam, W. L., & Lam, S. (2007). Effect of active smoking on the human bronchial epithelium transcriptome. *BMC Genomics*, *8*(1), 297.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013a). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, 128.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Maayan, A. (2013b). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, 128.

Chishimba, L., Thickett, D. R., Stockley, R. A., & Wood, A. M. (2010). The vitamin D axis in the lung: a key role for vitamin D-binding protein. *Thorax*, *65*(5), 456–462.

Coleman, B. N., Rostron, B., Johnson, S. E., Ambrose, B. K., Pearson, J., Stanton, C. A., Wang, B., Delnevo, C., Bansal-Travers, M., Kimmel, H. L., Goniewicz, M. L., Niaura, R., Abrams, D., Conway, K. P., Borek, N., Compton, W. M., & Hyland, A. (2017). Electronic cigarette use among US adults in the Population Assessment of Tobacco and Health (PATH) Study, 20132014. *Tobacco Control*, *26*(e2), e117–e126.

Dicpinigaitis, P. V., Lee Chang, A., Dicpinigaitis, A. J., & Negassa, A. (2016). Effect of e-Cigarette Use on Cough Reflex Sensitivity. *Chest*, *149*(1), 161–165.

Didon, L., Zwick, R. K., Chao, I. W., Walters, M. S., Wang, R., Hackett, N. R., & Crystal, R. G. (2013). RFX3 Modulation of FOXJ1 regulation of cilia genes in the human airway epithelium. *Respiratory Research*, *14*(1), 70.

Dinakar, C., & O'Connor, G. T. (2016). The Health Effects of Electronic Cigarettes. *N. Engl. J. Med.*, *375*(14), 1372–1381.

Enache, O. M., Lahr, D. L., Natoli, T. E., Litichevskiy, L., Wadden, D., Flynn, C., Gould, J., Asiedu, J. K., Narayan, R., & Subramanian, A. (2017). The GCTx format and cmap{Py, R, M} packages: resources for the optimized storage and integrated traversal of dense matrices of data and annotations.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, *6*(6), 721–741.

Gerashchenko, M. V., Lobanov, A. V., & Gladyshev, V. N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.*, *109*(43), 17394–17399.

Gohlmann, H., & Talloen, W. (2009). *Gene expression studies using Affymetrix microarrays*. Chapman and Hall/CRC.

Gould, N. S., Min, E., Gauthier, S., Martin, R. J., & Day, B. J. (2011). Lung glutathione adaptive responses to cigarette smoke exposure. *Respir. Res.*, *12*, 133.

Graham, H., Chandler, D. J., & Dunbar, S. A. (2019). The genesis and evolution of bead-based multiplexing. *Methods*.

Hackett, N. R., Butler, M. W., Shaykhiev, R., Salit, J., Omberg, L., Rodriguez-Flores, J. L., Mezey, J. G., Strulovici-Barel, Y., Wang, G., Didon, L., & Crystal, R. G. (2012). RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics*, *13*, 82.

Harvey, B.-G., Heguy, A., Leopold, P. L., Carolan, B. J., Ferris, B., & Crystal, R. G. (2007). Modification of gene expression of the small airway epithelium in response to cigarette smoking. *J. Mol. Med.*, *85*(1), 39–53.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, *14*, 7.

Hübner, R.-H., Schwartz, J. D., De Bishnu, P., Ferris, B., Omberg, L., Mezey, J. G., Hackett, N. R., & Crystal, R. G. (2009). Coordinate control of expression of Nrf2-modulated genes in the human small airway epithelium is highly responsive to cigarette smoking. *Mol. Med.*, *15*(7-8), 203–219.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249–264.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127.

Kalisky, T., & Quake, S. R. (2011). Single-cell genomics. *Nat. Methods*, *8*(4), 311–314.

King, B. A., Alam, S., Promoff, G., Arrazola, R., & Dube, S. R. (2013). Awareness and ever-use of electronic cigarettes among U.S. adults, 2010-2011. *Nicotine Tob. Res.*, *15*(9), 1623–1627.

King, B. A., Patel, R., Nguyen, K. H., & Dube, S. R. (2015). Trends in Awareness and Use of Electronic Cigarettes Among US Adults, 20102013. *Nicotine & Tobacco Research*, *17*(2), 219–227.

Lee, H.-W., Park, S.-H., Weng, M.-W., Wang, H.-T., Huang, W. C., Lepor, H., Wu, X.-R., Chen, L.-C., & Tang, M.-S. (2018). E-cigarette smoke damages DNA and reduces repair activity in mouse lung, heart, and bladder as well as in human lung and bladder cells. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(7), E1560–E1569.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Comput. Biol.*, *13*(5), e1005457.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, *9*(Nov), 2579–2605.

Martin, E. M., Clapp, P. W., Rebuli, M. E., Pawlak, E. A., Glista-Baker, E., Benowitz, N. L., Fry, R. C., & Jaspers, I. (2016). E-cigarette use results in suppression of immune and inflammatory-response genes in nasal epithelial cells similar to cigarette smoke. *American Journal of Physiology. Lung Cellular and Molecular Physiology*, *311*(1), L135–144.

Martínez-Sánchez, J. M., Fu, M., Martín-Sánchez, J. C., Ballbè, M., Saltó, E., & Fernández, E. (2015). Perception of electronic cigarettes in the general population: does their usefulness outweigh their risks? *BMJ Open*, *5*(11).

Mirkin, B. (2011). Choosing the number of clusters. *WIREs Data Mining Knowl Discov*, *1*(3), 252–260.

Moses, E. (2017). Characterizing the airway epithelium following chemical exposure: molecular alterations and their potential utility in the treatment of lung disease. https://open.bu.edu/handle/2144/23410.

Moses, E., Wang, T., Corbett, S., Jackson, G. R., Drizik, E., Perdomo, C., Perdomo, C., Kleerup, E., Brooks, D., O'Connor, G., Dubinett, S., Hayden, P., Lenburg, M. E., & Spira, A. (2017). Molecular Impact of Electronic Cigarette Aerosol Exposure in Human Bronchial Epithelium. *Toxicol. Sci.*, *155*(1), 248–257.

National Academies of Sciences, Engineering, and Medicine (2018). *Public Health Consequences of E-Cigarettes*. Washington, D.C.: National Academies Press.

Onoufriadis, A., Shoemark, A., Schmidts, M., Patel, M., Jimenez, G., Liu, H., Thomas, B., Dixon, M., Hirst, R. A., Rutman, A., Burgoyne, T., Williams, C., Scully, J., Bolard, F., Lafitte, J.-J., Beales, P. L., Hogg, C., Yang, P., Chung, E. M. K., Emes, R. D., O'Callaghan, C., Bouvagnet, P., & Mitchison, H. M. (2014). Targeted NGS gene panel identifies mutations in RSPH1 causing primary ciliary dyskinesia and a common mechanism for ciliary central pair agenesis due to radial spoke defects. *Human Molecular Genetics*, *23*(13), 3362–3374.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., & Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, *344*(6190), 1396–1401.

Pierrou, S., Broberg, P., O'Donnell, R. A., Pawowski, K., Virtala, R., Lindqvist, E., Richter, A., Wilson, S. J., Angco, G., Möller, S., Bergstrand, H., Koopmann, W., Wieslander, E., Strömstedt, P.-E., Holgate, S. T., Davies, D. E., Lund, J., &

Djukanovic, R. (2007). Expression of Genes Involved in Oxidative Stress Responses in Airway Epithelial Cells of Smokers with Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, *175*(6), 577–586.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47.

Rivera, G. M., Antoku, S., Gelkop, S., Shin, N. Y., Hanks, S. K., Pawson, T., & Mayer, B. J. (2006). Requirement of Nck adaptors for actin dynamics and cell migration stimulated by platelet-derived growth factor B. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(25), 9536–9541.

Robbins, R. A., Nelson, K. J., Gossman, G. L., Koyama, S., & Rennard, S. I. (1991). Complement activation by cigarette smoke. *Am. J. Physiol.*, *260*(4 Pt 1), L254–9.

Rostom, R., Svensson, V., Teichmann, S. A., & Kar, G. (2017). Computational approaches for interpreting scrna-seq data. *FEBS letters*, *591*(15), 2213–2225.

Schmittgen, T. D., & Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C_t method. *Nature Protocols*, *3*(6), 1101–1108.

Schweitzer, K. S., Chen, S. X., Law, S., Demark, M. J. V., Poirier, C., Justice, M. J., Hubbard, W. C., Kim, E. S., Lai, X., Wang, M., Kranz, W. D., Carroll, C. J., Ray, B. D., Bittman, R., Goodpaster, J., & Petrache, I. (2015). Endothelial disruptive pro-inflammatory effects of nicotine and e-cigarette vapor exposures. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, (p. ajplung.00411.2015).

Silvestri, G. A., Vachani, A., Whitney, D., Elashoff, M., Porta Smith, K., Ferguson, J. S., Parsons, E., Mitra, N., Brody, J., Lenburg, M. E., & Spira, A. (2015). A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *N. Engl. J. Med.*, *373*(3), 243–251.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*, Article3.

Smyth, G. K. (2005a). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.) *Bioinformatics*

and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health, (pp. 397–420). New York, NY: Springer New York. https://doi.org/10.1007/0-387-29362-0_23.

Smyth, G. K. (2005b). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, (pp. 397–420). Springer New York. http://link.springer.com/chapter/10.1007/0-387-29362-0_23 (Accessed:2016-05-26).

Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., & Brody, J. S. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(27), 10143–10148.

Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E., & Brody, J. S. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, *13*(3), nm1556.

Sridhar, S., Schembri, F., Zeskind, J., Shah, V., Gustafson, A. M., Steiling, K., Liu, G., Dumas, Y.-M., Zhang, X., Brody, J. S., Lenburg, M. E., & Spira, A. (2008). Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics*, *9*, 259.

Steiling, K., Ryan, J., Brody, J. S., & Spira, A. (2008). The Field of Tissue Injury in the Lung and Airway. *Cancer Prevention Research*, *1*(6), 396–403.

Steiling, K., van den Berge, M., Hijazi, K., Florido, R., Campbell, J., Liu, G., Xiao, J., Zhang, X., Duclos, G., Drizik, E., Si, H., Perdomo, C., Dumont, C., Coxson, H. O., Alekseyev, Y. O., Sin, D., Pare, P., Hogg, J. C., McWilliams, A., Hiemstra, P. S., Sterk, P. J., Timens, W., Chang, J. T., Sebastiani, P., O'Connor, G. T., Bild, A. H., Postma, D. S., Lam, S., Spira, A., & Lenburg, M. E. (2013). A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *American Journal of Respiratory and Critical Care Medicine*, *187*(9), 933–942.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O. M., Piccioni, F., Berger, A. H., Shamji, A., Brooks, A. N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Gray,

N. S., Clemons, P. A., Silver, S., Wu, X., Zhao, W.-N., Read-Button, W., Wu, X., Haggarty, S. J., Ronco, L. V., Boehm, J. S., Schreiber, S. L., Doench, J. G., Bittker, J. A., Root, D. E., Wong, B., & Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. http://biorxiv.org/content/early/2017/05/10/136168 (Accessed:2017-05-11).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Thum, T., Erpenbeck, V. J., Moeller, J., Hohlfeld, J. M., Krug, N., & Borlak, J. (2006). Expression of xenobiotic metabolizing enzymes in different lung compartments of smokers and nonsmokers. *Environ. Health Perspect.*, *114*(11), 1655–1661.

Vrijheid, M. (2014). The exposome: a new paradigm to study the impact of environment on health. *Thorax*, *69*(9), 876–878.

Wang, C., & Blei, D. M. (2009). Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.) *Advances in Neural Information Processing Systems 22*, (pp. 1982–1989). Curran Associates, Inc.

Wang, T. W., Vermeulen, R. C. H., Hu, W., Liu, G., Xiao, X., Alekseyev, Y., Xu, J., Reiss, B., Steiling, K., Downward, G. S., Silverman, D. T., Wei, F., Wu, G., Li, J., Lenburg, M. E., Rothman, N., Spira, A., & Lan, Q. (2015). Gene-expression profiling of buccal epithelium among non-smoking women exposed to household air pollution from smoky coal. *Carcinogenesis*, *36*(12), 1494–1501.

Yu, T., Li, Y. J., Bian, A. H., Zuo, H. B., Zhu, T. W., Ji, S. X., Kong, F., Yin, D. Q., Wang, C. B., Wang, Z. F., Wang, H. Q., Yang, Y., Yoo, B. C., & Cho, J. Y. (2014). The Regulatory Role of Activating Transcription Factor 2 in Inflammation. *Mediators of Inflammation*.

Zhang, X., Sebastiani, P., Liu, G., Schembri, F., Zhang, X., Dumas, Y. M., Langer, E. M., Alekseyev, Y., O'Connor, G. T., Brooks, D. R., Lenburg, M. E., & Spira, A. (2010). Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. *Physiol. Genomics*, *41*(1), 1–8.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D.,

Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., & Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, *8*, 14049.

CURRICULUM VITAE

CURRICULUM VITAE