2020

Development of advanced methods for large-scale transcriptomic profiling and application to screening of metabolism disrupting compounds

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES
AND
COLLEGE OF ENGINEERING

Dissertation

DEVELOPMENT OF ADVANCED METHODS FOR LARGE-SCALE TRANSCRIPTOMIC PROFILING AND APPLICATION TO SCREENING OF METABOLISM DISRUPTING COMPOUNDS

By

ERIC R. REED

B.S., University of Massachusetts Boston, 2010
M.S., University of Massachusetts Amherst, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader

Stefano Monti, Ph.D.
Associate Professor of Medicine

Second Reader

Paola Sebastiani, Ph.D.
Professor of Biostatistics

DEDICATION

To Rev. Patricia Kathleen O'Keefe Reed and Stephen Douglas Reed for exemplifying unconditional love and support, challenging my worldview, and giving me the space to find a career that suits my talents and values.

To Matthew Ramey Reed and Rev. Melissa O'Keefe Reed for watching out.

To Zachary Joseph Baker.

To Boston Bike Polo for the contusions.

ACKNOWLEDGMENTS

Thank you, Stefano Monti. You've always taken equal interest in my work and well-being. I've felt advocated for throughout this process. I couldn't have asked for a better mentor.

Thank you, Andrea Foulkes. You took a big chance on me as a first semester biostatistics Master's student. You were integral to me building the confidence I needed to get to where I am now.

Thank you to all of the professors I've been fortunate to collaborate with, especially: Jennifer Schlezinger, Camron Bryant, Evan Johnson, David Sherr, Paola Sebastiani, and Joshua Campbell. Being able to work with so many talented people is part of the reason why I chose bioinformatics as a research discipline.

Last and not least, thank you to my former classmate, Stephanie Kim. Our collaboration on the Adipogen Project was among my most rewarding experiences at Boston University.

DEVELOPMENT OF ADVANCED METHODS FOR LARGE-SCALE TRANSCRIPTOMIC PROFILING AND APPLICATION TO SCREENING OF METABOLISM DISRUPTING COMPOUNDS

ERIC R. REED

Boston University Graduate School of Arts and Sciences,

and College of Engineering 2020

Major Professor: Stefano Monti, Associate Professor of Medicine

ABSTRACT

High-throughput transcriptomic profiling has become a ubiquitous tool to assay an organism transcriptome and to characterize gene expression patterns in different cellular states or disease conditions, as well as in response to molecular and pharmacologic perturbations. Refinements to data preparation techniques have enabled integration of transcriptomic profiling into large-scale biomedical studies, generally devised to elucidate phenotypic factors contributing to transcriptional differences across a cohort of interest. Understanding these factors and the mechanisms through which they contribute to disease is a principal objective of numerous projects, such as The Cancer Genome Atlas and the Cancer Cell Line Encyclopedia. Additionally, transcriptomic profiling has been applied in toxicogenomic screening studies, which profile molecular responses of chemical perturbations in order to identify environmental toxicants and characterize their mechanisms-of-action.

Further adoption of high-throughput transcriptomic profiling requires continued effort to improve and lower the costs of implementation. Accordingly, my dissertation

work encompasses both the development and assessment of cost-effective RNA sequencing platforms, and of novel machine learning techniques applicable to the analyses of large-scale transcriptomic data sets. The utility of these techniques is evaluated through their application to a toxicogenomic screen in which our lab profiled exposures of adipocytes to metabolism disrupting chemicals. Such exposures have been implicated in metabolic dyshomeostasis, which is the predominant cause of obesity pathogenesis. Considering that an estimated 10% of the global population is obese, understanding the role these exposures play in disrupting metabolic balance has the potential to help combating this pervasive health threat.

This dissertation consists of three sections. In the first section, I assess data generated by a highly-multiplexed RNA sequencing platform developed by our section, and report on its significantly better quality relative to similar platforms, and on its comparable quality to more expensive platforms. Next, I present the analysis of a toxicogenomic screen of metabolism disrupting compounds. This analysis crucially relied on novel supervised and unsupervised machine learning techniques which I specifically developed to take advantage of the experimental design we adopted for data generation. Lastly, I describe the further development, evaluation, and optimization of one of these methods, *K2Taxonomer*, into a computational tool for unsupervised molecular subgrouping of bulk and single-cell gene expression data, and for the comprehensive in-silico annotation of the discovered subgroups.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

3'DGE.....3' Digital Gene Expression

ANOVA..... Analysis of Variance

AUC.....Area Under the Curve

BRCA ..... Breast Cancer

cDNA.....Clonal Deoxyribonucleic Acid

CCLE.....Cancer Cell Line Encyclopedia

CI..... Confidence Interval

CMap .....Connectivity Map

CNA..... Copy Number Alteration

CRISPR ..... Clustered Regularly Interspaced Short Palindromic Repeats

DMSO..... Dimethyl Sulfoxide

DNA .....Deoxyribonucleic Acid

ES ..... Enrichment Score

FDR .....False Discovery Rate

GEO.....Gene Expression Omnibus

GO ..... Gene Ontology

GSEA.....Gene Set Enrichment Analysis

GSVA .....Gene Set Variation Analysis

L1000.....Luminex-1000

LUAD .....Lung Adenocarcinoma

LUSC.....Lung Squamous Cell Carcinoma

MAD ..... Median Absolute Deviation

METABRIC ..... Molecular Taxonomy of Breast Cancer International Consortium

METSIM ..... Metabolic Syndrome in Men

PC ..... Principal Component

PCR ..... Polymerase Chain Reaction

RNA ..... Ribonucleic Acid

RNAseq ..... RNA Sequencing

rRNA ..... Ribosomal Ribonucleic Acid

scRNAseq ..... single-cell RNA Sequencing

SD ..... Standard Deviation

SE ..... Standard Error

SFL ..... Sparse Full Length RNA Sequencing

TCGA ..... The Cancer Genome Atlas

TIL ..... Tumor Infiltrating Lymphocyte

TMM ..... Trimmed Mean of M Values

UMI ..... Unique Molecular Identifier

Chapter 1: Introduction

1.1 Motivation

High-throughput transcriptomic profiling is employed to assay the collection of RNA transcripts extracted from a tissue sample or cell culture. Due in part to the revolutionizing role of next generation sequencing[1], over the past decade transcriptomic profiling has emerged as a ubiquitous tool for comprehensively examining molecular patterns across cohorts representative of diverse phenotypic factors, such as cell states, disease states, and experimental conditions. However, the relatively high cost of transcriptomic profiling limits the practicality of implementation, particularly for large-scale projects, i.e., projects which would require the generation of hundreds or thousands of profiles. Thanks to various technological innovations, as well ambitious and well-funded projects, transcriptomic profiling has been integrated in several large-scale studies, notably in the fields of population-based cancer genomics and toxicogenomic screens of environmental chemicals. The generation of such data sets presents exciting new opportunities for biological discovery. However, the feasibility of generating these data must be met with suitable computational tools with which they can be appropriately and exhaustively analyzed. Accordingly, in this dissertation I present work covering both the refinement of protocols for high-throughput transcriptomic data generation, as well as the development of analytical tools for extracting relevant information. The utility of these developments is showcased via their application to transcriptomic profiling data generated as part of a toxicogenomic screen of metabolism disrupting chemicals. Collectively, this work seeks to further biological discovery by transcriptomic profiling

through the contribution of novel methodological approaches and illustration of their application.

1.2 Overview of high-throughput transcriptomic profiling

High-throughput transcriptomic profiling culminated from various practical breakthroughs in genomic science, including: the reverse transcription of RNA to cDNA in 1970[2], DNA sequencing in 1975[3], and polymerase chain reaction (PCR) mediated DNA amplification in 1990[4]. Whole transcriptome quantification protocols were first implemented with the development of hybridization-based microarrays in the mid-1990's[5]. Microarrays measure transcript abundance based on fluorescence generated through the annealing of a sample-derived ribonucleotide sequences to a compendia of oligonucleotide probes[5]. The oligonucleotide probes are generally curated to measure the abundance of a large sets of transcripts. Most often these transcripts pertain to transcribed genes, such that these measurements are referred to as gene expression. Ten years after the development of microarrays, the advent of next generation sequencing[1] lead to the development of high-throughput RNA sequencing (RNAseq) with which the sequences derived from RNA fragments, referred to as reads, are recorded. Transcript abundance is then quantified by mapping these sequences to the annotated genome of the organism from which the sample originated. In the case of gene expression, both microarrays and RNAseq quantify individual genes on the order of tens of thousands, e.g. ~20,000 genes from human samples[6].

Whereas microarray profiling is only implemented to measure transcript abundance[7], the utility of RNAseq profiling is more general, such that additional use

cases include but are not limited to: *de novo* transcriptome assembly[8] and characterization of alternative splicing variants[9]. Furthermore, RNAseq protocols are flexible, facilitating the optimization of data resolution relative to sequencing depth[10]. As a result, the popularity of RNAseq profiling eclipsed that of microarray profiling in the mid-2010's; an estimated 3000 and 1500 publications reported the use of RNAseq and microarrays in 2016, respectively[11].

Despite RNAseq becoming the gold-standard method of high-throughput transcriptomics[12], its high cost relative to other methods has deterred its implementation in large-scale projects. In recent studies performed in our lab, the cost of an RNAseq run was $500, while that of microarray was $350, such that a study of 100 samples would result in a $15,000 price difference between the techniques. Accordingly, many large-scale transcriptomic profiling studies have used microarrays, well after the availability of RNAseq. For example, in 2012 *the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)*, used microarrays to profile breast cancer tumors from ~2000 patients[13]. Moreover, as of 2018, *the Connectivity Map (CMap)*, a toxicogenomic/drug repurposing project, has generated ~1.3 million exposure profiles, forgoing the use of high-throughput platforms altogether by developing their own assay, which only profiles 973 individual genes[14].

The need to lower the per-sample cost of RNAseq lead to the development of multiplexing techniques, in which unique barcodes are attached to cDNA fragments of each sample, thereby enabling pooling and simultaneous sequencing[15]. These methods were quickly adopted for generating RNAseq data for such large-scale projects as *The*

Cancer Genome Atlas (TCGA)[16], a consortium which currently contains 10,558 profiles from patients across 33 cancer types (https://portal.gdc.cancer.gov/) and the Cancer Cell Line Encyclopedia (CCLE), a database of 1,072 cell line profiles[17]. Innovations in barcoding techniques eventually lead to the practicality of performing single-cell RNA sequencing (scRNAseq), in which the cells from a tissue sample are isolated and sequenced individually[18]. Recently, the size of scRNAseq data sets has exploded, commonly consisting of thousands of individual cell profiles from a single tissue sample[19].

When performing multiplexed RNAseq, an important consideration is that pooling reduces number of sequenced reads for each sample, thereby reducing the precision of expression measurements, especially for lowly expressed transcripts[10]. Furthermore, preparing multiplexed RNAseq libraries requires more extensive library preparation protocols, including specific techniques to improve efficiency[20]. Therefore, careful assessment must be made to assess the quality of the data generated by novel protocols. In this dissertation, I present an extensive assessment of a highly-multiplexed RNAseq library preparation protocol developed by collaborators within the Boston University Section of Computational Biomedicine. The performance of this platform was examined relative to full-coverage RNAseq and microarray platforms, as well as a similar, well-established technique, called 3' digital gene expression (3'DGE)[21]. 3'DGE has been utilized in our lab for other projects, one of which I introduce next.

1.3 Toxicogenomic screening and metabolism disrupting compounds

Shortly after the availability of high-throughput transcriptomics platforms, the field of toxicogenomics emerged[22]. Toxicogenomic research is concerned with characterization of molecular perturbation of biological systems following exposure to environmental chemicals in order to identify environmental toxicants and elucidate mechanisms-of-action contributing to chemical toxicity[22]. This mission is daunting given that there are more than 84,000 registered consumer chemicals with ~200 new chemicals being registered each year[23]. Furthermore, effects of chemical exposure is both dose and context dependent, making the task of characterizing chemical toxicants even more difficult. To make a dent in this problem, numerous toxicogenomics screening studies have been performed, which are devised to efficiently expose model organisms or cell lines to panels of environmental chemicals. Examples of previous toxicogenomic screening studies include those previously mentioned, the *CMap* project, as well the *Carcinogenome* project, a toxicogenomic screen performed by our lab, profiling chemical carcinogenicity in two separate human cells lines, HEPG2 (liver) and MCF10A (breast), exposed to respective panels of 330 and 345 individual chemicals[24].

Unlike carcinogenic chemicals, the effect of chemical exposures on human metabolism has been relatively unexamined. In a recent statement, the Endocrine Society implicated the disruption of energy homeostasis as the main cause of obesity pathogenesis, attributable to multiple aspects of modern lifestyle (e.g., high calorie intake, sedentary lifestyle, stable home temperatures, disrupted circadian rhythms)." More recently, evidence for the potential role of chemical exposures has grown[25].

Importantly, obesity is associated with increases the risk of metabolic syndrome (i.e., fat around the waist, elevated blood pressure, high blood sugar, high serum triglycerides, low HDL-cholesterol) and metabolic disease (e.g., type 2 diabetes, cardiovascular disease, nonalcoholic fatty liver disease, and stroke)[26]. Furthermore, the prevalence of global obesity has doubled since 1980, such that an estimated 108 million children and 604 million adults were obese in 2015, i.e. ~10% of the global population[27]. Adipose tissue is a crucial regulator of metabolic homeostasis, functioning as repositories of free fatty acids and releasing hormones that can regulate multiple aspects of metabolic homeostasis from food intake to insulin responsiveness[28]. Identifying chemical exposures that disrupt adipocyte activity and the elucidating the specific biological processes that are perturbed by these chemical exposures presents an opportunity to understand risk factors contributing to this major public health threat.

In this dissertation, I present analyses of 3'DGE transcriptomic profiling data, generated as part of a toxicogenomic screen of chemical exposures to an adipocytes cell line. This panel was curated to include chemicals for which there is published evidence as to their effect or lack thereof on adipocyte activity, as well as chemicals for which previous evidence suggests an effect on adipocyte activity, but have not been rigorously examined in this context. The objectives of this project necessitated the development of novel machine learning methods to take better advantage of the experimental design used in this study. After describing each method and demonstrating their validity for modeling the data in this study, I finish by reporting the further development of one of these

methods and demonstrate its general applicability for unsupervised learning of both bulk and single-cell transcriptomic profiling data.

1.4 Dissertation Aims

1.4.1 Aim 1: Assessment of a cost-effective highly multiplexed RNA sequencing platform

The need to reduce per sample cost of RNAseq profiling for scalable data generation has led to the emergence of highly multiplexed RNAseq. In chapter 2, I report the assessment of one such technique, denoted as sparse full length sequencing (SFL), a ribosomal RNA depletion-based RNA sequencing approach that allows for the simultaneous sequencing of 96 samples and higher. The performance of SFL was examined relative to well established single-sample techniques, including: full coverage Poly-A capture RNAseq, microarrays, as well as 3' digital gene expression (3′DGE), another highly multiplexed technique. Our lab generated data for a set of exposure experiments on immortalized human lung epithelial (AALE) cells in a two-by-two study design, in which samples received both genetic and chemical perturbations of known oncogenes/tumor suppressors and lung carcinogens.

The quality of each data set was evaluated analytically in terms of transcriptomic coverage, as well as the extent to which they captured expected transcriptional differences between exposure groups. To this end, I leveraged published transcriptional signatures of lung carcinogen exposure, as well as gene expression signatures associated with mutations of the genes over-expressed in this data from, derived from the *TCGA Lung Squamous Cell Carcinoma (LUSC)* and *Lung Adenocarcinoma (LUAD)* datasets. Additionally, discussion of the performance of each platform weighs their relative cost.

1.4.2 Aim 2: Transcriptomic profiling of adipocyte activity disrupting chemicals

In chapter 3, I report on transcriptomic profiling of chemical exposures which disrupt adipocyte function by modifying activity of the transcription factor, Peroxisome proliferator activated receptor γ (*PPARγ*), the predominant transcriptional regulator of adipocytes[29]. These chemicals, denoted as adipogens, have been previously shown to activate specific subsets of *PPARγ*'s transcriptional programs to induce adipocytes with distinct phenotypes[30–32].

The two main objectives of this work was to 1) predict *PPARγ* modifying chemicals and 2) group *PPARγ* modifying chemicals based on their effects on the adipocyte transcriptome and downstream metabolic functions. Data was generated using 3'DGE transcriptomic profiling of 3T3-L1 mouse adipocytes following differentiation in the presence of 76 chemicals.

The analytical focus of this chapter encompasses the implementation of two novel machine learning methods I developed to improve modelling of the data. For predicting *PPARγ* modifying chemicals, I developed an amended random forest classification procedure, tailored to account for the presence of replicates, i.e. multiple samples in the data that received the same chemical exposure. The classification performance of this procedure relative to traditional random forest modeling was evaluated by cross validation. To identify subgroups of *PPARγ* modifying chemicals, I developed a novel recursive partitioning algorithm, denoted as *K2Taxonomer*, which utilizes repeated perturbations of the data to estimate robust partitions, recursively, thereby estimating a taxonomy-like subgrouping of the chemical exposures. The resulting taxonomy was then

annotated in-silico based on the patterns of gene expression and pathway enrichment specific to each subgroup. Finally, I developed an interactive web-portal of these results, which is publicly available (https://montilab.bu.edu/adipogenome/). Chapter 3 concludes with a series of experiments to validate the findings of these analyses.

1.4.3 Aim 3: Tool development for characterization of molecular subgroups in bulk and single-cell transcriptomic profiling data

Unsupervised learning methods are commonly employed on high-dimensional data sets with the goal of identifying data-driven molecular subtypes, i.e. groups of observations that had not been distinguished prior to the analysis. In addition to the validity of subtype estimation, the utility of these approaches depends on the level to which the resulting models can be appropriately interpreted. This task is non-trivial considering the breadth of information that was utilized in model estimation.

In chapter 4, I present the further development of *K2Taxonomer* as an R package to perform robust recursive partitioning of large-scale transcriptomic data sets and facilitate the comprehensive exploration of the resulting models. *K2Taxonomer's* generalizability to large-scale transcriptomic data sets was examined via application to simulated data, as well as various publicly available gene expression data collected from breast cancer tissue collected from living patients and airway tissue collected from healthy human subjects. These applications included implementation of *K2Taxonomer* on both bulk and single-cell gene expression data. To assess the validity of *K2Taxonomer*, I include performance comparisons to relative agglomerative hierarchical clustering methods. Chapter 4, concludes with a practical example, in which I applied

K2Taxonomer to an scRNAseq data set of breast cancer tumor infiltrating lymphocytes and utilized the package's post-modeling functionality to characterize subgroups of T cell subtypes. Finally, I leveraged bulk gene expression data generated from breast cancer patient tissue to confirm these findings *in silico*.

Chapter 2: Assessment of a cost-effective highly multiplexed RNA sequencing platform

2.1 Background

Since its inception in 2008, RNA sequencing has become the gold-standard for whole-transcriptome high-throughput data generation[33]. In addition to RNA transcript expression quantification, RNAseq allows for more advanced analyses including de novo transcriptome assembly[8] and characterization of alternative splicing variants[9]. Furthermore, RNAseq is species agnostic, such that the same library preparation technique may be utilized for humans, mouse, rat, kidney bean, etc. These represent clear advantages over hybridization-based microarray platforms in which individual microarray platforms are designed to quantify specific transcripts for a specific species[7]. However, one persistent drawback of RNAseq has been its relatively high cost. The use of classic RNAseq techniques for experimental designs that require profiling of many samples – especially when the marginal information value of each sample is relatively low, such as in medium- and high-throughput screening applications – can thus present a disqualifying cost burden.

Large-scale projects based on transcriptional profiling of chemical exposure experiments include the *Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs)*[34], the *DrugMatrix* database[35], and the Connectivity Map (*CMap*)[14], among others. Both the *TG-GATEs* and the *DrugMatrix* projects used microarrays for expression profiling, which was at the time significantly less costly than full coverage RNAseq, yet still requiring multi-million dollar budgets. Alternatively, the

CMap project utilizes the Luminex-1000 (L1000) profiling platform, a bead-based analog expression assay which quantifies 973 human transcripts, which are used to impute the expression of 11,350 additional transcripts[14]. This technique is among the least expensive expression assays available, but it is restricted to human screens and it directly profiles only a limited panel of genes. Given the flexibility of RNAseq platforms, highly multiplexed techniques represent a viable alternative for generating transcriptional data from exposure screens, as well as from other experiments that require a large sample size. Therefore, evaluation of the technical validity of specific techniques serves to inform research strategies for a variety of biological inquiries.

The need to reduce the per sample cost of RNAseq has led to the adoption of barcoding technologies, where cDNA sequences from individual samples are tagged and their libraries are combined and multiplex sequenced in a single lane[36]. More recently, these techniques have been optimized to allow multiplex sequencing of 96 samples per lane or higher[37,38]. Here, we report the results of our effort at optimizing and evaluating one such technique denoted as sparse full length (SFL) sequencing[38], a ribosomal RNA depletion-based RNA sequencing approach. We offer comparisons to well established single-sample techniques, including: full coverage Poly-A capture RNAseq and microarray, as well as another low-cost highly multiplexed technique known as 3' digital gene expression (3'DGE)[39]. Assessments include comparisons of coverage between the three RNAseq techniques, as well as signal-to-noise and biological recapitulation of gene-level differential signals between treatment groups for the same samples profiled across SFL, microarray, and 3'DGE. For this evaluation study, we generated a set of

exposure experiments on immortalized human lung epithelial (AALE) cells[40] in a two-by-two study design, in which samples received both genetic and chemical perturbations of known oncogenes/tumor suppressors and lung carcinogens (Illustration 2.1). The goal of this report is not only to assess the performance of our optimized highly multiplexed technique, but to inform future research in terms of the strengths and pitfalls of available cost-effective high throughput transcriptomic profiling techniques.

2.2 Methods

2.2.1 Samples

Exposure experiments were performed on immortalized human bronchial epithelial cells (AALE). Cells were exposed to both genotypic and chemical perturbations with three replicates per perturbation combination. Illustration 2.1 describes the specific combinations of perturbations profiled across each platform. Cells were thawed from liquid nitrogen and grown up in SAGM small airway epithelial cell growth media (Lonza, Portsmouth NH). Cells were subcultured using Clonetics ReagentPack subculture reagents (Lonza, Portsmouth, NH). In preparation for exposure, cells were plated into 24-well plates and allowed to reach confluency for 24 hours. Genotypic perturbations included CRISPR knockouts of genes coding for lung tumor suppressor proteins: protocadherin FAT1 (*FAT1*)[41] and cyclin-dependent kinase inhibitor 2A (*CDKN2A*)[42], as well as overexpression of oncogenes: nuclear factor erythroid-derived 2-like 2 (*NRF2*)[43], fibroblast growth factor receptor 1 (*FGFR1*)[44], neuregulin 1 (*NRG1*)[45] and phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA*)[46]. Cells transfected with a pSpCas9-EGFP (*GFP*) plasmid (PX458) in the absence of sgRNAs

were used as controls for the CRISPR perturbations while overexpression of an empty vector containing the reporter HcRed served as control for the overexpression experiments. Cell culture media was then replaced with either dimethylsulfoxide (DMSO) vehicle or one of three lung carcinogens: 24 μg/ml cigarette smoke concentrate (CSC)[47], 173μM benzo[a]pyrene (BaP)[48], or 490μM nicotine-derived nitrosamine ketone (NNK)[49]. NNK and BaP compounds were obtained from Sigma-Aldrich (St. Louis MO) and CSC obtained from Murty Pharmaceuticals (Lexington, KY). Aliquots of total RNA from the same samples were profiled across SFL, microarray, and 3'DGE for a subset of combinations of exposures, though all samples were profiled by SFL. In addition, full coverage poly-A RNAseq was performed on a separate set of samples for a subset of genotypic exposures, including CRISPR knockouts of *FAT1*, as well as overexpression of *NRF2*, *NRG1*, and *PIK3CA*. These samples did not receive any chemical exposures (Illustration 2.1). Note that in a few cases there was not enough material to perform 3'DGE, as indicated by the sample numbers of certain perturbation combinations.

2.2.2 Library preparation

Total RNA was isolated using a standard Qiazol and Qiacube protocol from Qiagen (Valencia, CA). RNA purity was assessed using a NanoDrop spectrophotometer and no samples were excluded from downstream analysis. Library preparation for SFL sequencing was carried out based on the published protocol[38]. The dual-barcoded SFL libraries were pooled from 96 individual samples and then sequenced on the Illumina® NextSeq 550 to generate more than 400 million single-end 75-bp reads. Microarray procedures were performed as described in GeneChip™ WT PLUS Reagent Kit manual

and GeneChip™ WT Terminal Labeling and Controls Kit protocol (Thermo Fisher Scientific). The labeled fragmented DNA was generated from 100 ng of total RNA and was hybridized to the GeneChip™ Human Gene 2.0 ST Array. Microarrays were scanned using Affymetrix GeneArray Scanner 3000 7G Plus. 3'DGE library preparation was performed by Broad Institute, Cambridge, MA, USA, similar to [50]. Final libraries were purified using AMPure XP beads (Beckman Coulter) according to the manufacturer's recommended protocol and sequenced on an Illumina NextSeq 500 using paired-end reads of 17bp (read1) + 46bp (read2). Read1 contains the 6-base well barcode along with the 10-base UMI. Poly-A RNA Sequencing libraries were prepared from total RNA samples using Illumina® TruSeq® RNA Sample Preparation Kit v2 and then sequenced on the Illumina® HiSeq 2500 to generate more than 5 million single-end 50-bp reads per sample. Across all platforms, the number of samples that were successfully profiled per perturbation combination is shown in Illustration 2.1.

2.2.3 Data pre-processing

Affymetrix GeneChip Human Gene 2.0 ST Microarray CEL files were annotated to unique Entrez gene IDs, using a custom CDF file from BrainArray (hugene20st_Hs_ENTREZG_21.0.0) and RMA-normalized. For SFL, adapter sequences were trimmed from raw sequence files using *Cutadapt (v1.12)*. Quality assessment of trimmed SFL sequence files as well as raw full coverage RNAseq sequencing files was performed with *FastQC (v0.11.5)*. Both SFL and RNAseq reads were aligned to human genome (UCSC RefSeq hg19) with STAR v2.5.2b with the non-defulat parameter, "--outSAMtype BAM SortedByCoordinate"[51]. Expression quantification in RefSeq genes

was carried out with *featureCounts (v1.5.0)*[52]. For 3'DGE, pre-quantified gene expression count matrices were obtained from the Broad Institute, Cambridge, MA, USA. These reads had been aligned to the transcriptome (UCSC RefSeq hg19), using *BWA aln (v0.7.10)* with the non-default parameter, "-l 24"[53]. Considering that there are 4^{10} ($\sim 1.05E^6$) possible UMIs and the 3'DGE library sizes are on the order of 10^6 reads, it is highly unlikely for the same UMI to be added to multiple cDNA fragments from the same gene. Therefore, using a custom python program [50], reads with the same UMI and sample barcode were only counted once per gene. All further data processing and analysis were carried out in R.

2.2.4 Coverage assessment

Read coverage across the 82 samples, shared between SFL and 3'DGE, as well as all 18 full coverage RNAseq samples was assessed for library size as well as percentage of the library size that was aligned, uniquely aligned (i.e. reads that only align once in the genome), and counted in the 22,233 genes which were annotated across all three platforms, i.e. the intersection of annotated genes. The full set of counted reads is hereafter referred to as the counted library. Unlike SFL and full coverage RNAseq, 3'DGE reads are aligned directly to mRNA sequences, such that the reported numbers of counted reads and uniquely aligned reads are the same. To assess the relative distribution of reads across the total set of shared genes, we plotted the cumulative proportion of the sum of reads aligning to individual genes per samples ranked by relative expression across all three platforms. Saturation analysis of the estimated minimum percentage of the counted library size to maximize the number of genes quantified by each platform

was performed using a loess fit the gene discovery of 20 subsamplings of the per sample counted libraries. All subsampling analysis was performed using *Subseq (v1.8.0).*

Finally, we assessed the relative induction of noise introduced by subsampling progressively larger proportions of the original counted library sizes in each platform, as measured by the principal component error[54]. In order to compare the three platforms assuming equally sized starting library, we repeated the assessment after first subsampling full coverage RNAseq libraries and 3'DGE libraries to sizes matching that of SFL, the smallest library of the three platforms. This analysis was performed on the 18 samples of like genotypic perturbations, with no chemical treatment in the case of full coverage RNAseq samples and vehicle DMSO treatment in SFL and 3'DGE samples. Reported values reflect means across 20 iterations of the subsampling and principal component error calculation procedure.

2.2.5 Signal-to-noise assessment

Signal-to-noise was compared among SFL, 3'DGE and microarrays based on four-group ANOVA analysis and two-group differential analysis. In order to estimate signal-to-noise as a means for assessing expected performance when applying standard statistical methods to the data, rather than differential gene expression analysis packages, classic ANOVA was performed for each gene using normalized data across all three platforms, using the glm function in R. In this analysis, the signal-to-noise was assessed across like samples undergoing exposure to CSC or DMSO vehicle, as well as genotypic perturbations of *NRF2* overexpression or HcRed control. Thus, the analysis included four independent groups of samples, receiving each combination of chemical (CSC or DMSO)

and genotypic (*NRF2* or HcRed) perturbations, with three replicates in each group. Only genes with mean expression ≥ 1 across all 12 samples in both SFL and 3'DGE were included in the analysis (9,813 total genes). Expression levels across SFL and 3'DGE were normalized via trimmed mean of M values (TMM)[55] scaling and $\log_2$ counts-per-million transformation. Additionally, two-group differential gene expression analysis was performed for each stratified chemical and genotypic perturbation, using *limma (v3.30.7)*. That is, differential expression of CSC- vs. DMSO-treated samples, within either HcRed or NRF2 treatment, as well as differential expression of *NRF2*- vs. HcRed-treated samples, within either DMSO or CSC exposure, was performed. The SFL and 3'DGE count data were transformed for linear modeling based on voom[56]. Following modeling, results were restricted to the top 10,000 genes as ranked by median-absolute-deviation (MAD). This heuristic gene filtering procedure was adopted because quantification-based filtering is not applicable to microarray data. This approach follows recommendations detailed in the *limma* manual[56]. All p-values reported from two-group differential analysis are two-sided. In both ANOVA and *limma* analyses, nominal p-values for each gene were corrected for multiple comparisons using the Benjamini-Hochberg procedure [57].

2.2.6 Biological signal recapitulation

Two-group differential analysis signatures were compared by pre-ranked gene set enrichment analysis (GSEA) to gene sets derived from published signatures of smoking exposure in the airway from healthy volunteers[58,59], as well as to gene sets analytically derived from *The Cancer Genome Atlas (TCGA)* for patients with l*ung squamous cell*

carcinoma (LUSC) or *lung adenocarcinoma (LUAD)*. The two smoking gene sets consist of genes reported as either up- or down-regulated in response to smoking in at least one of the two publications, while *TCGA* gene sets were derived by probing differential expression of individual genes between patients with or without point mutations or copy number alterations (CNA) in genes of interest. These include mutations for the same panel of genes profiled for genotypic perturbations. In addition we include *KEAP1* mutations, a repressor of *NRF2*[60]. Specifically, point mutation signatures were derived from *LUSC* and *LUAD*, independently, by performing differential analysis of subjects with and without point mutations in genes of interest, matched for age, sex, and cancer stage. For *NRF2* and *PIK3CA* point mutations were defined at specific mutation hotspots of along the gene body (Figure A.2)[61]. Likewise, CNA gene signatures were assessed for amplification and deletions of genes of interest by differential analysis, using subjects with zero, one, or two additional copies or deletions of a gene of interest, respectively. All models for mutations and CNA were adjusted for tumor purity, as reported[61]. Differential signatures were derived using *limma*. Genes associated with specific mutations or CNA were defined as those with significance and magnitude of the linear model's genetic alteration coefficient at FDR Q-value < 0.05 and |log2 fold-change| > $\log_2(1.5)$, respectively.

Each of our genotypic perturbation signatures was compared by GSEA to the corresponding *TCGA*-derived gene sets. For example, the *PIK3CA* overexpression signatures were compared to the gene sets derived from *PIK3CA* mutation and copy number alterations in the *TCGA* data. To assess the effect of read counts on gene discovery

and biological recapitulation of each platform, we compared the differential analysis and GSEA results to that derived from subsampled libraries across full coverage RNAseq, SFL, and 3'DGE. Similar to coverage assessment, this analysis was performed starting with full libraries across all three platforms, as well as initially subsampling the full coverage RNAseq and 3'DGE libraries to sizes matching that of SFL. Reported values reflect means from 20 iterations of the subsampling followed by differential analysis and GSEA procedures.

2.3 Results

2.3.1 Coverage assessment

Comparison of coverage of the three sequencing platforms, full coverage poly-A RNAseq, SFL, and 3'DGE, is summarized in Table 2.1, Figure 2.1, and Figure A.1. Comparison between SFL and 3'DGE included 82 samples each, while full coverage poly-A RNAseq included all 18 available samples. None of the three platforms demonstrated differences in the library size variability (total number of assigned reads) across samples, although there was a notably high difference between the largest and smallest library size for the SFL samples, with a fold change of 4.3. Fold changes for full coverage RNAseq and 3'DGE were 1.9 and 2.9, respectively (Table 2.1, Figure 2.1A).

Unsurprisingly, full coverage poly-A RNAseq generated the largest library size, while the SFL and 3'DGE libraries were of comparable size (Figure 2.1A). Furthermore, full coverage poly-A RNAseq yielded the highest percentage of reads aligned to the genome, followed by SFL and 3'DGE (Table 2.1, Figure 2.1Ci, Figure A.1A). The lower mapping rate of 3'DGE is most likely due to the lower read quality scores of 3'DGE

compared to full coverage RNAseq and SFL (Figure A.1B). The mean percentage of reads with Phred quality scores greater than 20 (Q20) was only ~88% for 3'DGE, compared to ~100% for both full coverage RNAseq and SFL. The relative 5'-3' transcript coverage for each sample across all three platforms is shown in Figure A.1f. As expected, reads alignments were skewed towards the 3' end of transcripts for 3'DGE, while we did observe relatively uniform coverage along the transcript for full coverage RNAseq and SFL.

For SFL there was a clear drop-off when going from percentage of aligned reads to percentage of uniquely aligned reads due to ribosomal RNA (rRNA) contamination of the SFL samples (Figure 2.1Cii). The majority of reads aligning to ribosomal regions specifically align to RNA28S (Figure A.3). For 3'DGE, unique UMIs are aligned directly to transcript sequences and not to the whole genome, such that the number of uniquely aligned reads and reads counted in transcripts are the same (Figure 2.1Cii-iii)[62]. The percentage of reads that are counted in transcripts is greatest for full coverage poly-A RNAseq (mean percentage of total library size: 65.2%), followed by 3'DGE (33.3%), and SFL (24.5%). However, while the counted read library size is greater for 3'DGE than for SFL, more genes were quantified by SFL than by 3'DGE (Figure 2.1Civ) (counts > 0 across all samples for 22,233 genes shared across all three platforms,). A median of 60.9% and 50.5% genes were quantified by SFL and 3'DGE, respectively. The number of genes quantified was near the saturation point for each platform, such that this discrepancy is not due to read depth of each platform (FigureS1C). The reason for the low gene discovery of 3'DGE is further illustrated in Figure 2.1B, where it is shown that

the reads are more evenly distributed across the 22,233 genes by SFL than by 3'DGE, with the cumulative distribution of reads counted in individual genes nearly identical in SFL and full coverage poly-A RNAseq.

The principal component (PC) error was estimated for each platform for different subsamples of the full counted library size. The first PC is shown in Figure 2.1D, while the second through the fifth PCs are shown in Figure A.1D. We observe that as the counted library size increases, the PC error decreases at the fastest rate for full coverage RNAseq, followed by SFL, then 3'DGE. Though these differences are more prominent when comparing full coverage RNAseq to either SFL or 3'DGE, we do observe that when down-sampling from 10% to 100% of the counted library size, the PC error decreases at a consistently faster rate for SFL than for 3'DGE. Initially subsampling full coverage RNAseq and 3'DGE to match the full SFL counted library size does not change the results. The same trend is also observed in the cumulative variance explained by each successive PC across full coverage RNAseq, SFL, and 3'DGE (Figure A.1E).

In summary, despite lower overall counted library size due to ribosomal RNA contamination, SFL demonstrates greater coverage in low-to-medium expressed genes than 3'DGE, comparable to full coverage poly-A RNAseq. Consequently, the transcriptional signal captured by the SFL libraries are more robust to subsampling of the data compared to 3'DGE as measured by the principal component error.

2.3.2 Signal-to-noise evaluation

Differential expression models comparing experimental groups of matched samples was performed in SFL, microarray, and 3'DGE and the corresponding signal-to-

noise scores were compared pairwise between platforms (Figure 2.2). Samples shared across the three platforms include three replicates for each of four experimental groups, corresponding to NRF2 overexpression or HcRed vehicle, as well as CSC chemical exposure or DMSO vehicle (Illustration 2.1). Signal-to-noise was assessed by a four-group comparison with classic ANOVA (Figure 2.2A-D), as well as by stratified two-group differential analyses using *limma* (Figure 2.2 E-F).

We compared the $\log_{10}$ F-statistics between ANOVA models across all three platforms (Figure 2.2A). Overall, the distribution of F-statistics is most similar between SFL and microarrays, with a Pearson correlation of 0.291. Though statistically significant ($p < 0.01$), the corresponding mean difference between $\log_{10}$ F-statistics is only 0.026. The mean differences of the $\log_{10}$ F-statistics between SFL and 3'DGE, and between 3'DGE and microarray are 0.328 and 0.302, respectively, and the corresponding Pearson correlations are 0.160 and 0.216, respectively. These results are consistent with the discovery rates estimated for different FDR Q-value thresholds (Figure 2.2B). For example, at the FDR Q-value threshold of 0.05, the discovery rates of SFL and microarray are almost identical, 0.214 (2083 genes), 0.209 (2038 genes), respectively, while the discovery rate of 3'DGE is much smaller 0.032 (310 genes).

Loess regression of the $\log_{10}$ F-statistics as a function of mean gene expression shows that the statistical signal increases with mean normalized expression. This trend is consistently positive for both SFL and 3'DGE, while leveling off at the most highly expressed genes in microarrays (Figure 2.2C). Furthermore, SFL signal is greater than 3'DGE signal at all levels of mean expression (Figure 2.2C). In agreement with the

results from coverage comparison, the distribution of mean normalized expressions in 3'DGE is smaller than that of SFL, while SFL is comparable to that of microarray (Figure 2.2D). Adherence to assumption of normality, assessed through a Shapiro-Wilk test, is also associated with higher mean normalized expression (Figure A.4).

The results of the comparisons of the two-group differential analyses across all three platforms were generally congruous with those of the four-group ANOVA analyses (Figure 2.2E-F, Figure A.5, Figure A.6). In all four two-group comparisons, the correlation of test statistics is closest between microarray and SFL results, followed by 3'DGE versus microarray results, and 3'DGE versus SFL. For example, in the DMSO-stratified, NRF2 versus HcRed analysis, estimates of the Pearson correlations of test statistics are 0.66, 0.45, and 0.43, respectively (Figure 2.2E). The discovery rate of 3'DGE is the lowest across all four differential analyses, while the discovery rate of SFL is higher in three out of four of these analyses (Figure 2.2F, Figure A.5, Figure A.6).

In summary SFL demonstrated greater statistical power than 3'DGE to detect differentially expressed genes, and its results more closely matched those in microarrays.

2.3.3 Biological signal recapitulation evaluation

To evaluate the ability of each platform to recapitulate biologically relevant results, we utilized previously published signatures of smoking exposure in lung[58,59], as well as differential signatures derived from the *TCGA LUSC* and *LUAD* datasets associated with mutations of the genes over-expressed in our experiments. From each of these signatures two gene sets were extracted, one of genes positively associated and one

of genes negatively associated to the variable of interest. These gene sets were then tested via pre-ranked gene set enrichment analysis against each of our differential analysis results (CSC vs. DMSO, stratified by *NRF2* or HcRed perturbation; *NRF2* vs. HcRed, stratified by CSC or DMSO perturbation). The enrichment results with respect to both the smoking exposure signatures and the *TCGA* mutations are summarized in Figure 2.3A, and further detailed in Figure A.5, and confirm the highest sensitivity of microarrays, followed by SFL and 3'DGE.

The set of genes up-regulated in "smokers vs. non-smokers" was found to be significantly (FDR Q-value < 0.05) enriched in all "CSC vs. DMSO" signatures, within both genotypic stratifications for all three platforms. Conversely, the set of down-regulated genes in "smokers vs. non-smokers" was only enriched in the microarray signature of "NRF2 over-expressed; CSC vs. DMSO" (Figure A.7).

The enrichment results of *TCGA*-derived gene sets with respect to differential signatures of genotypic perturbations were in agreement with the gene-level results, in that they consistently demonstrated smaller discovery rates by 3'DGE than by SFL or by microarrays (Figure 2.3A). For example, the significantly enriched gene sets in "DMSO-treated; *NRF2* vs. HcRed" differential signatures across all three platforms are highlighted in Figure A.7. The number of gene sets enriched in microarray, SFL, and 3'DGE platforms are five, three, and zero, respectively.

In addition to comparing which gene sets were significantly enriched in individual differential signatures, we compared the relative statistical signal of these enrichments. To this end, we transformed the permutation-based FDR Q-values by taking the negative

Log_{10} and multiplying by the direction of the enrichment score (ES), –Log10(FDR Q-values)*sign(ES). For each two-platform comparison, we fit a regression model through the origin. Since consistent results across platforms would result in a model fit close to the identity line, y=x, we tested whether the slope coefficient equaled 1 (i.e. $B_1 = 1$). Figure 2.3B shows these results for each of the three comparisons of the *NRF2* and *KEAP1* mutation-based gene sets enrichment against the "DMSO-treated; *NRF2* vs. HcRed" signatures. In all three comparisons, microarrays have the highest measured enrichment signal, followed by SFL and 3'DGE, however the difference between microarray and SFL results is not significant, $B_1 = 0.73$; p-value = 0.2. The coefficients for both of the comparisons to 3'DGE, are highly skewed in favor of microarray and SFL, $B_1 = 0.18$ and 0.14, respectively. Both of these comparisons are highly significant with p-values < 0.01. Comparison of the enrichment results for other differential signatures show similar trends (Figure A.8).

Next, we compared enrichment results with respect to all genotypic perturbation signatures between SFL and 3'DGE (Figure 2.4A; Figure A.9A). Each comparison (i.e., each point in the plot) denotes gene set enrichment results with respect to genotypic perturbations within each of the four chemical exposures, DMSO, CSC, BaP, and NNK. Gene sets were tested for enrichment against concordant differential signatures, e.g., the *PIK3CA* mutation-derived gene set was tested against the "*PIK3CA* vs. HcRed" signatures. As in the previous analysis, the permutation-based enrichment FDR Q-values were transformed by –Log10(FDR Q-values)*sign(ES). In the "DMSO-treated; genotypic perturbation vs. control" signatures, we observe that the gene set enrichment is

generally more significant for SFL than for 3'DGE ($B_1 = 0.63$; p-value < 0.01; Figure 2.4A). The results obtained in CSC- and NNK-treated signatures, demonstrate concordance to these results ($B_1 = 0.65$; p-value = 0.03 and $B_1 = 0.60$; p-value = 0.01, respectively). The BaP-treated results are less comparable since only one genotypic perturbation signature, "*FAT1 vs. GFP*", is available for this stratification (Figure A.9A).

Additionally, we compared our differential signatures to available full coverage poly-A RNAseq genotypic perturbations (Figure A.9B), although these results are considered less comparable because of differences in experimental set-up. In particular, in the full coverage poly-A RNAseq experiments the genotypic perturbations were performed on untreated rather than DMSO-treated cell lines (Illustration 2.1).

The effect on discovery rate by subsampling the data across all three platforms is shown in Figure 2.4B. Generally, we did not observe a plateauing of discovery rate, where the number of detected genes plateaus near full counted library size. When comparing the correlation between GSEA results on subsampled data we observe similar trends across full coverage RNAseq, SFL, and 3'DGE (Figure 2.4C). Initial subsampling of full coverage RNAseq and 3'DGE to the SFL counted library size did not change the analysis results.

In summary, differential analysis of molecular and genotypic perturbations with SFL recapitulates biologically meaningful signals of gene sets derived from high coverage in vivo data sets. This performance is comparable to both 3'DGE and microarray.

2.4 Discussion

The goal of this study was to evaluate the performance of SFL sequencing, a low-cost method for performing highly multiplexed RNAseq, and to compare it to other high-throughput gene expression profiling platforms. The development of such methods would be instrumental to the generation of large-scale perturbation screens based on in-vitro models. The reduction of the cost per profile would make it feasible to significantly increase the number of replicates and conditions to be profiled, including multiple time points, concentrations, and biological models, and thus would support a more in-depth investigation of the heterogeneity of the biological response to different exposures. It would also support the development of more accurate predictive models of the adverse or therapeutic outcomes of various exposures. Finally, insights gained from our study will also inform the design of protocols for single-cell RNAseq (scRNAseq)[63], given their reliance on highly-multiplexed libraries.

In addition to SFL, the platforms included in this analysis were 3'DGE, an alternative highly multiplexed sequencing platform, Affymetrix GeneChip Human Gene 2.0 ST Microarray, an analog expression platform, and full coverage poly-A capture RNAseq. The cost per sample for SFL and 3'DGE was ~$50, a 10-fold decrease from that of full coverage RNAseq, $500, and a 7-fold decrease from that of the microarray, $350 USD. Throughout this analysis we demonstrate comparable performances of SFL and 3'DGE to these more expensive platforms. Furthermore, in this analysis we consistently find evidence that SFL outperforms 3'DGE.

Performance was assessed in terms of coverage, signal-to-noise, and recapitulation of expected biological signal derived from independently generated, publicly available data collected from human subjects. Coverage was assessed by comparing the three digital expression platforms, while signal-to-noise and biological recapitulation was assessed by comparing SFL, 3'DGE, and microarrays. Microarray expression quantification has been shown to be highly correlated with qRT-PCR, especially when processed with updated probe set annotations, utilized in this analysis[64]. Chemical and molecular perturbations were carried out in the same samples, and concurrently profiled by SFL, 3'DGE, and microarrays. We also leveraged previously generated full coverage poly-A RNAseq profiles from similar perturbations of AALE cell lines.

For coverage assessment, performance was evaluated in terms of the distribution of total reads, or library size, that were aligned to the human genome, and further quantified in annotated genes. The best performance was expected in full coverage poly-A RNAseq, given that this is the most well-established technique and has by far the highest sequencing depth. This was confirmed, as full coverage poly-A RNAseq was measured to have the highest per sample library size, percentage of aligned reads, percentage of uniquely aligned reads, and percentage of counted reads (Figure 2.1, Figure A.1). The coverage performance of SFL suffered as a result of rRNA contamination, where as many as 53% of the total library size per sample was assigned to ribosomal regions of the genome (Figure A.3).

3’DGE is a poly-A capture technique, therefore ribosomal depletion is not a possible pitfall. 3’DGE generates short nucleotide tags from transposon-based fragmentation, which are enriched for 3’ adjacent sequences of a given transcript [50]. Since many transcripts of the same gene generate identical sequence tags, unique molecular identifiers (UMIs) are used to distinguish between unique reads and duplicate reads generated from PCR amplification. Although mRNA fragment duplication occurs with any RNAseq protocol, the impact of this artifact on downstream analyses is negligible for techniques, such as SFL, which generate more complex sequence libraries[65].

3’DGE sequences were aligned directly to human mRNAs, rather than the whole genome. Therefore, percentages of reads aligned and reads counted (Figure 2.1Ci,iii) reflect the percentages of these non-unique UMIs that align to at least one gene and the number of unique UMIs that align to only one gene, respectively. We observe that the percentage of counted reads is greater for 3’DGE than SFL, which is explained by a loss of reads to rRNA contamination in SFL. However, we observe notably more genes quantified by SFL than by 3’DGE (Figure 2.1B, Figure 2.1Civ), which indicates that more reads are assigned to fewer genes in 3’DGE compared to SFL, as well as to full coverage RNAseq (Figure 2.1C). Although rRNA contamination is a potential drawback of any ribosomal depletion RNA sequencing technique, the extent of ribosomal contamination is variable, and could be potentially improved by further optimization of the library preparation protocol.

The difference in distribution of reads across shared genes between SFL and 3'DGE likely explains the difference in information retained by subsampling as measured by principal component error. We consistently observe that, as the counted library size increases, the rate of principal component error decreases faster for SFL than 3'DGE (Figure 2.1D, Figure A.1D). This is unsurprising considering that not only are considerably fewer genes quantified by SFL compared to 3'DGE, but there is also no discernable difference between the rate of genes counted as a function of counted library size between the two platforms (Figure A.1C). As we subsample the counted libraries, though we may lose the same number of genes between SFL and 3'DGE, the percent of genes lost, and consequently the information lost, will be greater for 3'DGE than SFL. Furthermore, this more even read distribution likely explains the improved performance of SFL over 3'DGE in statistical signal. In particular, our signal-to-noise evaluation shows consistently higher gene-level statistical signal from SFL and microarray experiments than from 3'DGE experiments (Figure 2.2). These differences appear to be driven by the differences in the relative quantification of genes, given that statistical signal is positively associated with mean gene expression for each platform, and 3'DGE experiments showed lower gene-level quantification than SFL and microarrays (Figure 2.2C-D). We observe similar cross-platform relationships in the two-group differential analyses (Figure 2.2E-F).

The gene set-based enrichment results are consistent with those from signal-to-noise analyses. In every comparison of enrichment scores between SFL and 3'DGE, we observe generally higher gene set enrichment with respect to the SFL-derived signatures

(Figure 2.3, Figure 2.4A, Figure A.8, Figure A.9). The gene sets were selected to represent known biological responses to the profiled perturbations, and thus their enrichment with respect to the perturbation signatures are expected to be true positives.

The enrichment results confirm this expectation. For example, in the signatures of *NRF2* overexpression, we consistently observe enrichment of the gene sets derived from *NRF2* amplifications and *KEAP1* deletions, each of which should increase NRF2 activity (Figure A.7)[60]. Similarly, we observe significant concordant enrichment of the gene sets derived from *NRF2* and *KEAP1*-dysregulated lung tumors in the signature of CSC exposure, suggesting that the *NRF2* pathway is activated by CSC exposure in vitro (Figure A.7), which has been previously reported[66]. Interestingly, these results demonstrate that the activation of the *NRF2* pathway in normal airway epithelial cells *in vitro* (by ectopic expression of the gene or by CSC treatment) is concordant with the activation of *NRF2* by somatic genome alterations in lung tumors, a finding that, to the best of our knowledge, has not been previously observed.

Possible sources of technical variability in this study are the different sequencing platforms, service providers, and read lengths. However, when subsampling the 3'DGE and SFL counted libraries, we generally observe higher discovery rates at all percentages of the full counted libraries, and even more so when the 3'DGE counted libraries are initially subsampled to full SFL counted library sizes (Figure 2.4B), demonstrating that SFL shows improvements independent of the mapping rate. This result confirms previous reports showing that increasing read length above 50-bp does not improve read quantification[67]. Furthermore, similar results have been reported even when the same

sequencing platform is used. A recent study reported a greater number of genes detected, as well as higher differential analysis discovery rates, in conventional RNAseq than in 3'DGE at identical counted library sizes, using the Illumina HiSeq 2500 platform to generate both libraries[68].

In summary, in this study we observe higher performance of SFL than 3'DGE, as measured by coverage, signal-to-noise, and biological recapitulation of known signal, with the performance of SFL often matching that of well-established "gold standards" (full coverage RNAseq or microarrays). On the other hand, the fact that 3'DGE is shown to allocate a large number of reads to relatively fewer, highly expressed genes, makes this platform more suitable for problems where high accuracy in the differential quantification of highly expressed genes is needed. Furthermore, the ready availability of 3'DGE as a core-provided option, which allows for the out-sourcing of library preparation, sequence read pre-processing and gene quantification, is an additional value-added of the platform. Ultimately, the best-suited platform for a specific project will depend on the study goals, design, and availability of different resources. We believe our study presents useful results to make a more informed choice.

The utility of highly multiplexed RNAseq crucially depends on the trade-off between cost and data quality, and on the nature of the experiments for which the platform would be ideally suitable. These will in general be experiments where the marginal information content of a single profile is relatively low, and thus justifies trading-off some data quality for reduced cost.

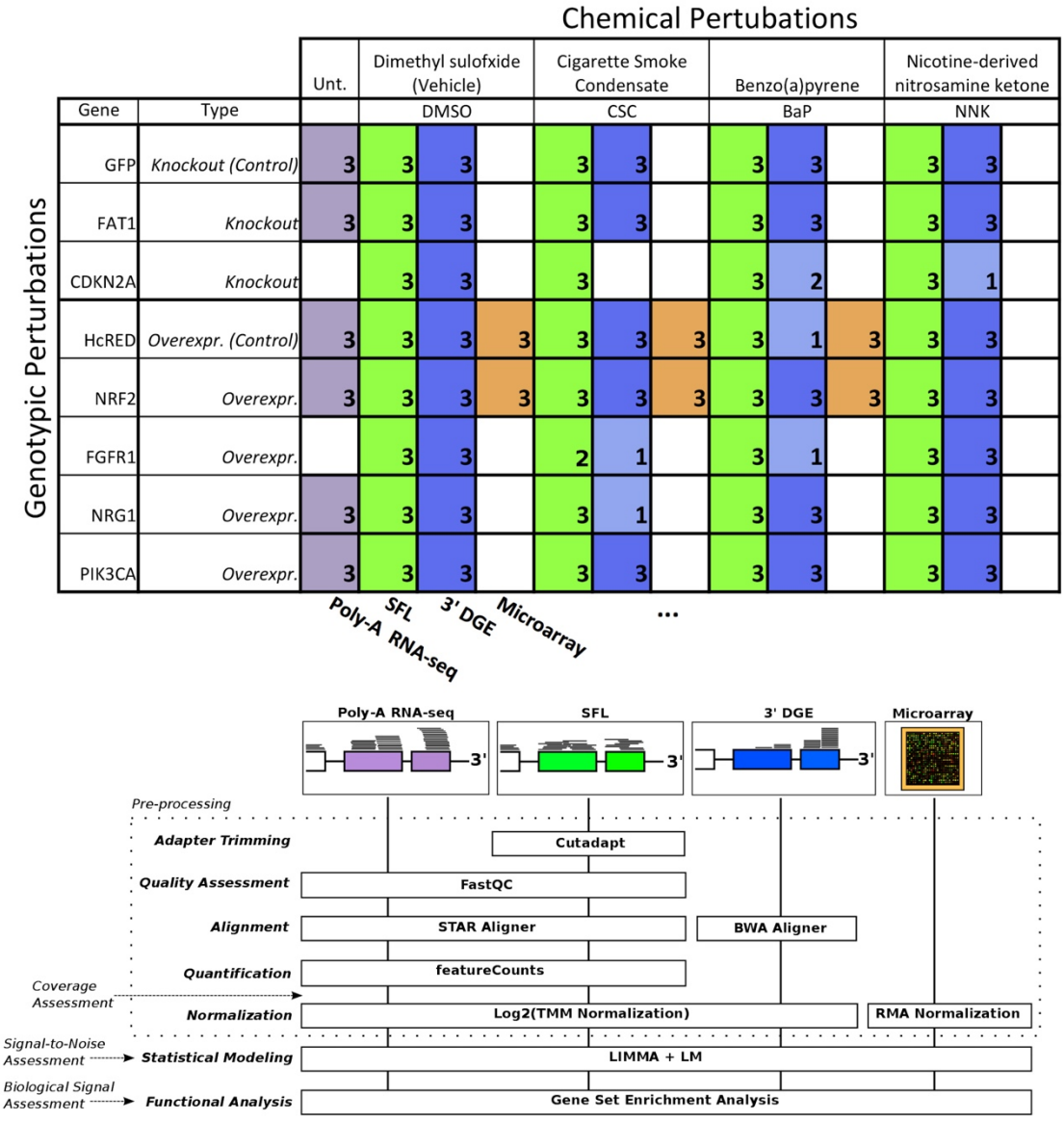


Illustration 2.1: Design of Cross-Platform Experiments and High-throughput Data Processing

Schematic of the number of each pair of genotypic and chemical perturbations, as well as a summary of preprocessing methods used to quantify gene-level expression for each platform. Note that "Unt." is an abbreviation of "untreated", denoting that the RNAseq samples used in this experiment did not receive chemical perturbations. Numbers in each box represent biological replicates of each condition. The color scheme for each platform is consistent throughout this report.

	Poly-A RNA-seq (RNA-seq)							
	Counts (Million)				(Value/Library Size)			
	Mean(SD)	Median	Min	Max	Mean(SD)	Median	Min	Max
Library Size (Total Reads)	13.0(2.3)	12.6	9.3	17.6				
Aligned Reads	12.4(2.2)	12	9	16.9	95.9(1.3)	96	92.4	97.9
Uniquely Aligned Reads	10.8(1.9)	10.3	7.8	14.8	82.9(1.5)	83	79.5	85.3
Counted Reads	8.4(1.5)	8.1	6.4	10.9	65.2(2.7)	64.6	60.5	70.3
	Sparse Full Length Sequencing (SFL)							
	Counts (Million)				Percent (Value/Library Size)			
	Mean(SD)	Median	Min	Max	Mean(SD)	Median	Min	Max
Library Size (Total Reads)	3.8(1.1)	3.5	1.6	6.9				
Aligned Reads	3.3(1.0)	3.1	1.4	5.9	88.5(2.9)	88.8	73	92.5
Uniquely Aligned Reads	1.8(0.6)	1.8	0.7	3.2	48.5(8.0)	46.8	27.6	64.8
Counted Reads	0.9(0.3)	0.9	0.3	1.6	24.5(4.0)	23.8	14.3	31.7
	3' Digital Gene Expression (3'DGE)							
	Counts (Million)				Percent (Value/Library Size)			
	Mean(SD)	Median	Min	Max	Mean(SD)	Median	Min	Max
Library Size (Total Reads)	3.7(0.7)	3.7	1.9	5.6				
Aligned Reads	3.0(0.6)	3	1.5	4.5	80.6(1.6)	81	73.5	82.2
Uniquely Aligned Reads								
Counted Reads	1.2(0.2)	1.2	0.7	1.8	33.3(1.4)	33	30.5	38.6

Table 2.1: Comparison of Read Assignment Between Full Coverage Poly-A RNAseq, SFL, and 3'DGE

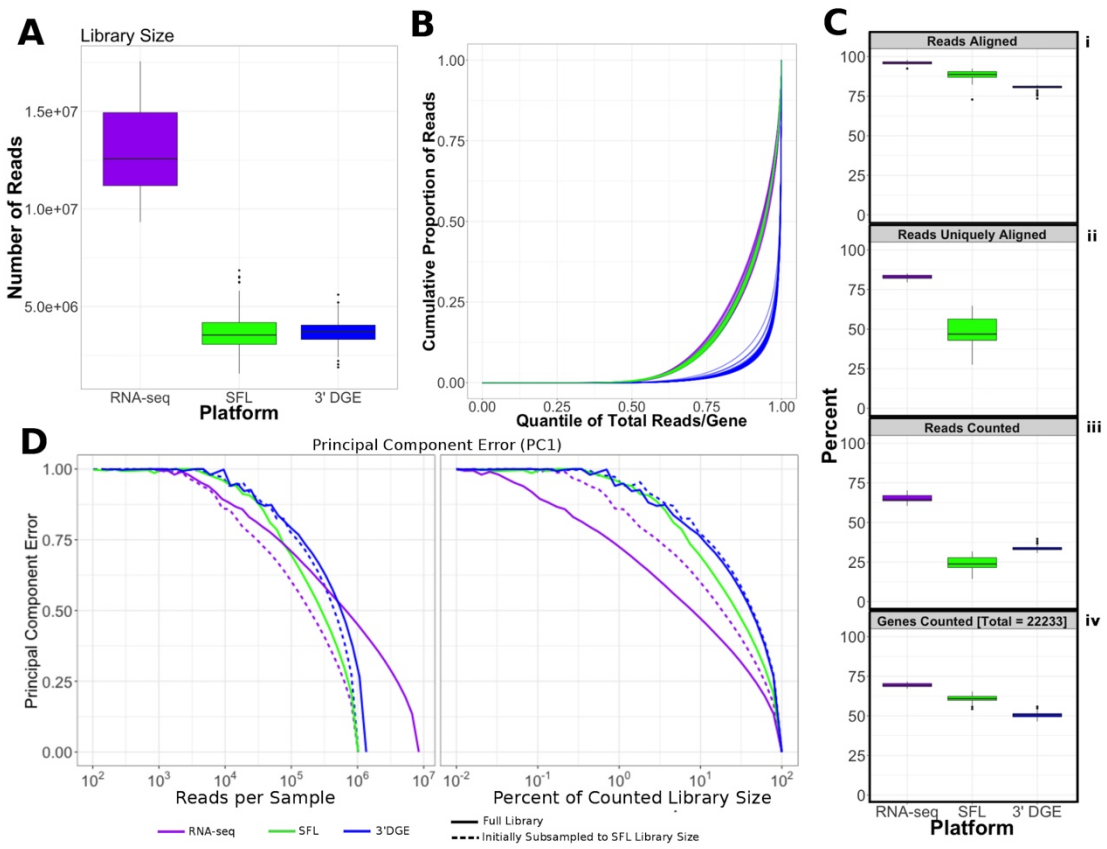


Figure 2.1: Comparison of Coverage Between Poly-A RNAseq, SFL, and 3'DGE

A) Boxplots of distribution of library size for each platform.
B) Cumulative distribution of reads assigned to individual genes per sample. The x-axis indicates the quantile for each gene in terms of ranking by relative expression. The y-axis shows the cumulative proportion of total counted reads assigned to these genes, i.e., the running sum of reads divided by the total number of reads across all genes.
C) The top 3 boxplots show the percentage of reads aligned (i), uniquely aligned (ii), and counted(iii) relative to the total library size for each platform. The bottom boxplot (iv) shows the proportion of genes with counts > 1, for protein-coding genes annotated across all 3 platforms (18,488). For Figure 2.1Cii, "Reads Uniquely Aligned" is not shown for 3'DGE because "Reads Uniquely Aligned" and "Reads Counted" are the same values as a result of the data pre-processing protocol, specific to 3'DGE (see Methods). Counts values for these percentages are given in Figure A.1A.
D) Analysis of the principal component error of subsampled counted library sizes for full coverage poly-A RNAseq, SFL, and 3'DGE for principal component 1. Results for principal component 2-5 is shown in Figure A.1D. Initial subsamples of Poly-A RNAseq and 3'DGE to the SFL library size are also given as dotted lines.

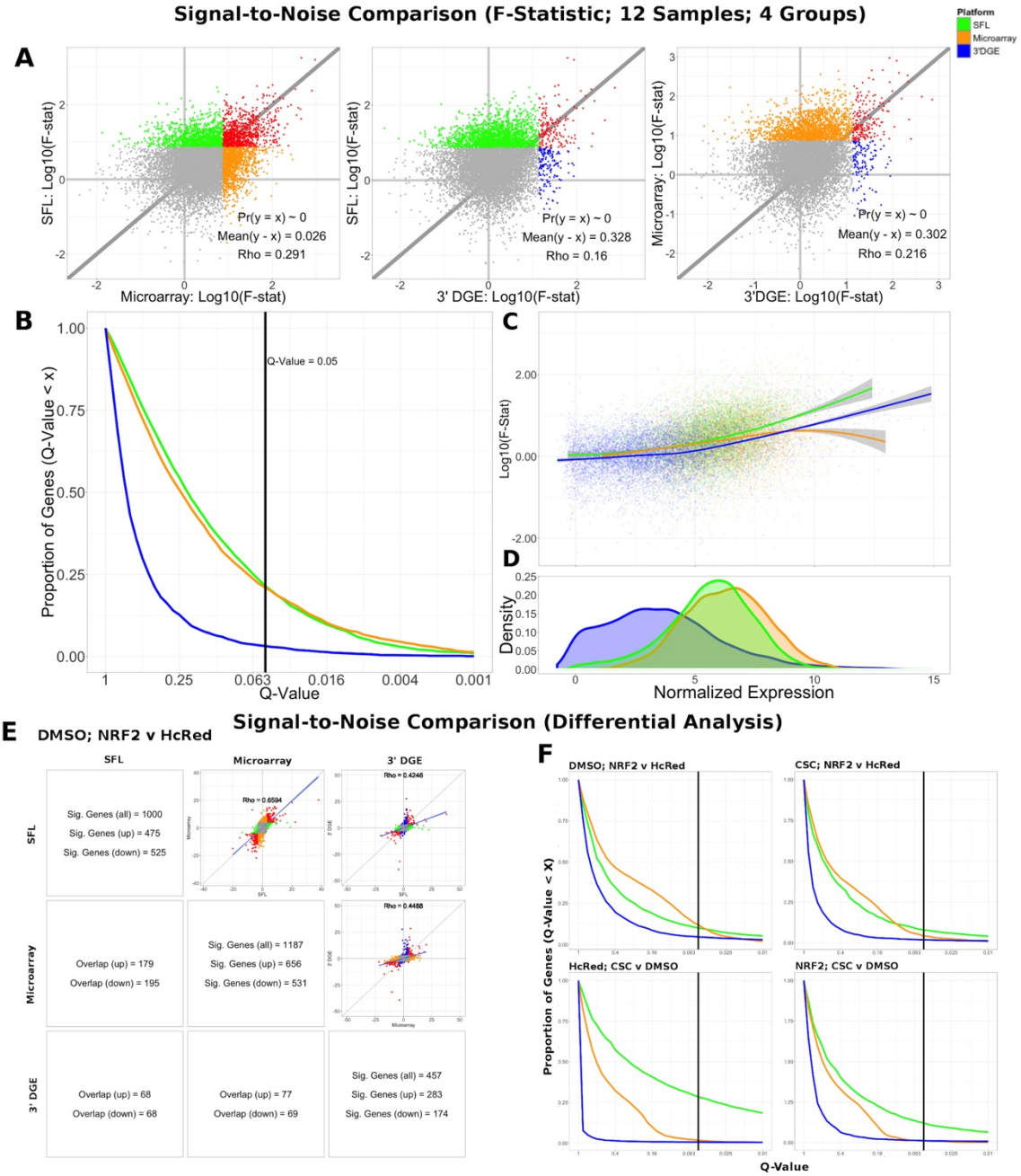


Figure 2.2 Signal-to-Noise Comparison Between SFL, Microarray, and 3'DGE

A) Scatterplots comparing the log10(F-Statistics) from ANOVA models comparing four n=3 groups (HcRed:DMSO, HcRed:CSC, *NRF2*:DMSO, and *NRF2*:CSC). The grey line shows y=x. The platform with the higher mean log10(F-Statistic) is plotted on the y-axis. Also, included are the p-value and difference in mean between each bi-platform comparison from paired t-testing, as well as the squared correlation coefficient. P-values ~ 0 are less than 0.01. Color of indicate genes discovered by

individual platforms (green, orange, or blue), neither platform (grey), and both platforms (red).

B) Plot of the Discovery Rate versus FDR Q-Value from threshold for each platform from four group ANOVA models. The x-axis is plotted on a -log10 scale. The vertical line is indicative of a Q-value threshold of 0.05.

C) Loess fit of the log10(F-Statistic) versus median normalized expression from four group ANOVA models.

D) Distribution of mean normalized expression across all three platforms.

E) Comparison of gene discovery (FDR Q-Value < 0.05) by differential analysis with limma, comparing normalized gene expression between DMSO:*NRF2* and DMSO:HcRed, including the raw discovery rates, discovered gene overlap, and linear fits, comparing test statistics from each platform. Genes that are discovered by more than 1 platform are shown in red in the scatterplots. Additional comparisons are shown in Figure A.5.

F) Plot of the Discovery Rate versus FDR Q-Value from threshold for each platform from two group differential analyses. The x-axis is plotted on a -Log10 scale. The vertical line is indicative of a Q-value threshold of 0.05.

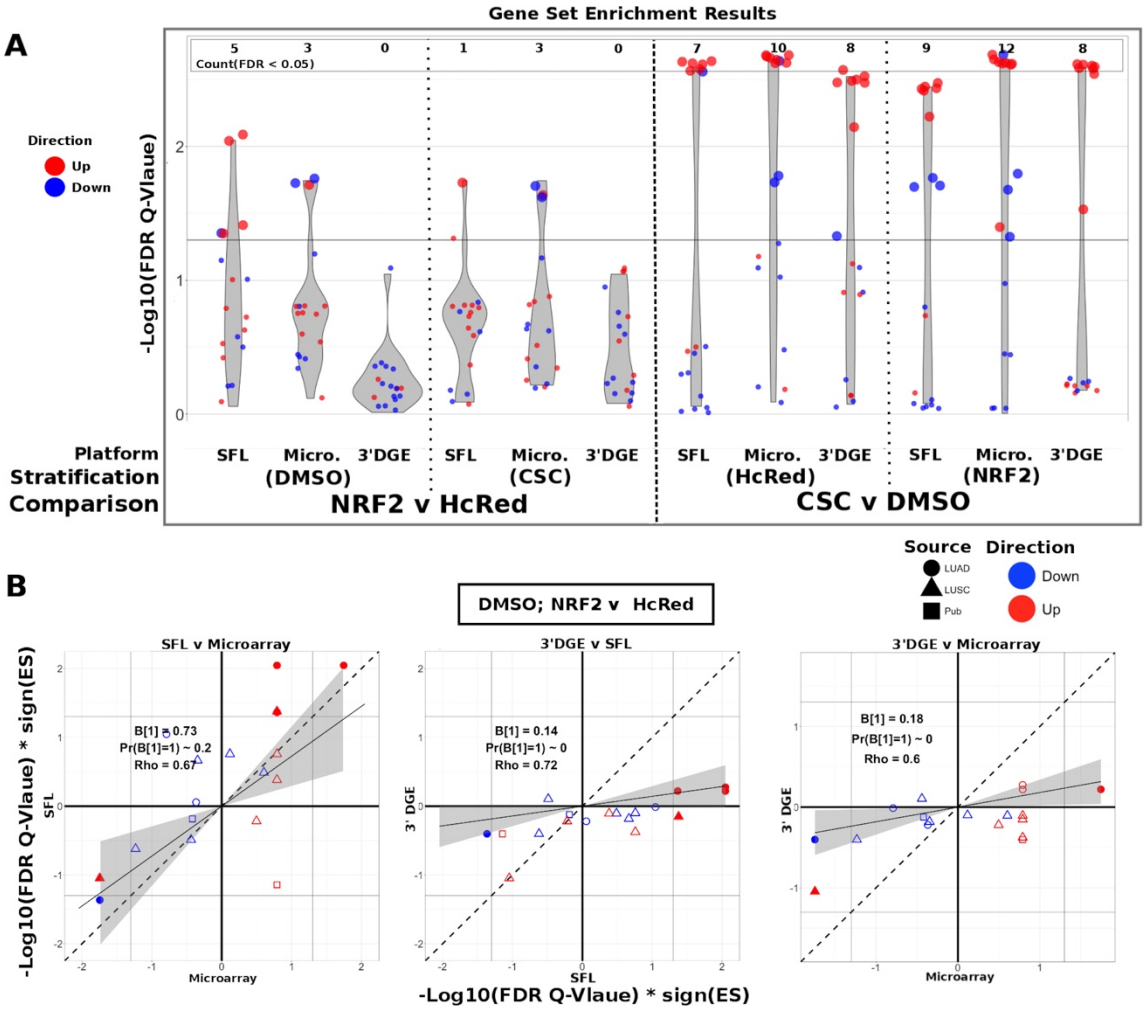


Figure 2.3: Comparison of Gene-set enrichment of Smoking and Gene Mutation Signatures across SFL, 3'DGE and Microarray

A) Violin plots of the –Log10(FDR Q-Value) from gene set enrichment analysis of *TCGA*-derived gene-sets with respect genotypic perturbations (left) and chemical perturbations (right) differential signatures across like samples within SFL, Microarray, and 3'DGE. Each column corresponds to differential signatures comparing genotypic or chemical perturbation groups, stratified by a single chemical or genotypic perturbation group, respectively, e.g. the left-most column shows the enrichment results with respect to the "DMSO-treated; *NRF2 vs.* HcRed" signature within the samples (*stratum*) in SFL data. Specific results for *TCGA*-derived genes sets are shown in Figure A.7.

B) Comparison of the gene set enrichment results between SFL, microarray and 3'DGE with respect to the "DMSO-treated; *NRF2 vs.* HcRed" differential signature. Shown are the transformed FDR Q-values of the *TCGA*-derived gene sets corresponding to mutations of *NRF2* and CNA of *KEAP1*. The |–Log10(FDR Q-Values)|

corresponding to the FDR<0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. Points of gene sets whose enrichment meets this threshold in either of the two platforms are filled in. Colors and shape of points denote direction and source of the gene set, respectively. Additional results for chemical and genotypic perturbation signatures are shown in Figure A.8.

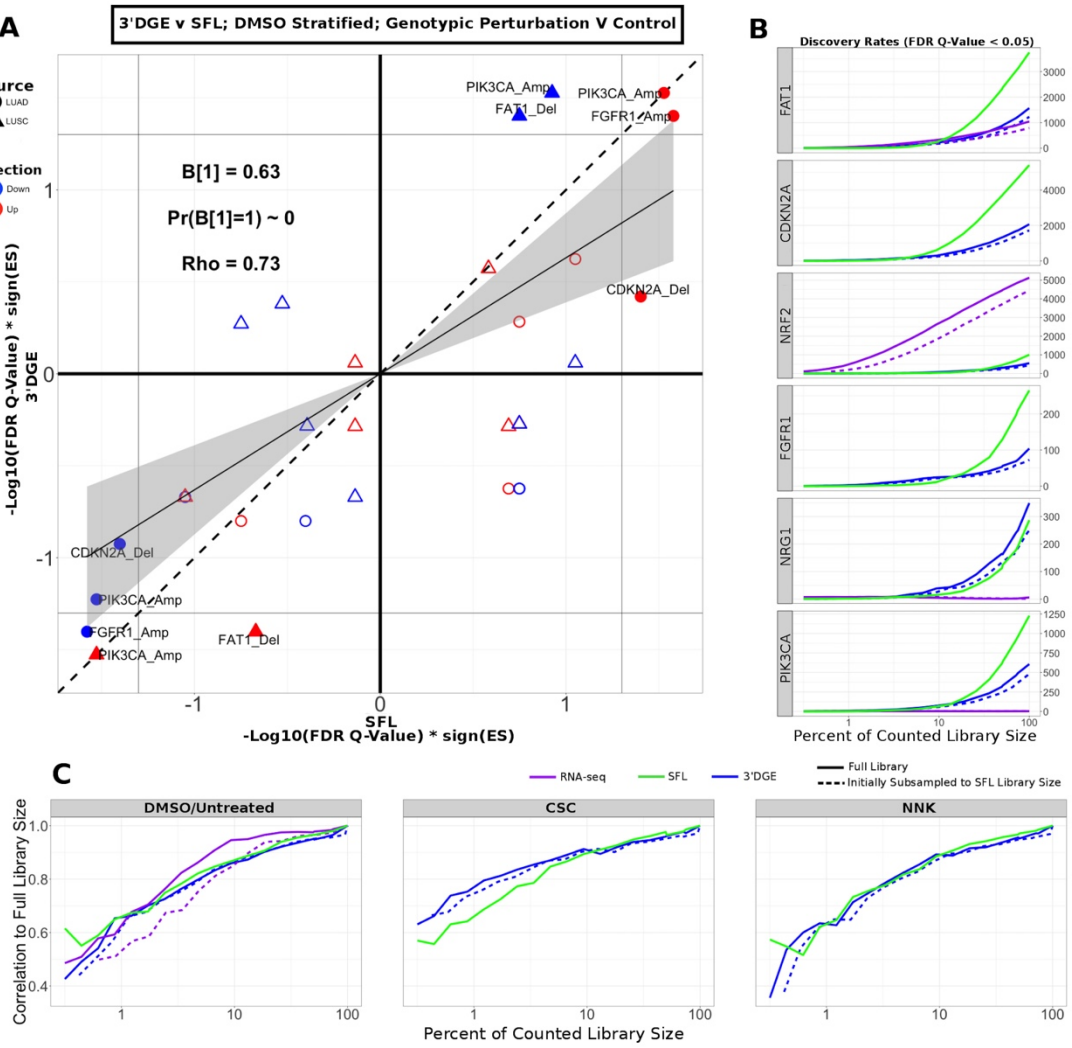


Figure 2.4: Comparison of Gene-set enrichment of Gene Mutation Signatures across SFL and 3'DGE

A) Comparison of the gene set enrichment results between SFL and 3'DGE with respect to the "DMSO-treated; genotypic perturbation *vs.* control" differential signatures. Points indicate gene set enrichment against concordant signatures, e.g., *PIK3CA* mutation and CNA gene sets against the "*PIK3CA vs.* HcRed" differential signatures. Shown are the transformed FDR Q-values from permutation-based testing by pre-ranked GSEA. |–Log10(FDR Q-Values)| corresponding to the FDR=0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. The names of the gene sets whose enrichment meets this threshold in either of the two platforms are shown and their points are filled in. Colors and shape of points denote direction and source of the gene set, respectively. Additional results for CSC, NNK, and BaP stratified genotypic perturbation signatures, as well as

comparisons between full coverage RNAseq and either SFL and 3'DGE are shown in Figure A.9.

B) Discovery rates for genotypic perturbations across full coverage poly-A RNAseq, SFL, and 3'DGE, for chemically untreated (full coverage RNAseq) and DMSO treated (SFL and 3'DGE) samples. Results demonstrate full counted library size, as well as subsampled libraries.

C) Correlation between transformed FDR Q-values from gene set enrichment at different subsamples of each platform and the results from the full counted library size. Shown are the results from genotypic perturbations from untreated (full coverage RNAseq)/DMSO treated (SFL and 3'DGE), CSC, and NNK chemically treated samples.

Chapter 3: Transcriptomic profiling of adipocyte activity disrupting chemicals

3.1 Background

Since 1980, the prevalence of obesity has been increasing globally and has doubled in more than 70 countries. In 2015, it was estimated that a total of 108 million children and 604 million adults were obese worldwide[27]. This poses a major public health threat since overweight and obesity increase the risk of metabolic syndrome, which, in turn, sets the stage for metabolic diseases, such as type 2 diabetes, cardiovascular disease, nonalcoholic fatty liver disease, and stroke[26]. The Endocrine Society's latest scientific statement on the obesity pathogenesis states that obesity is a disorder of the energy homeostasis system, rather than just a passive accumulation of adipose, and that environmental factors, including chemicals, confer obesity risk[25]. The rapid increases in obesity and metabolic diseases correlate with substantial increases in environmental chemical production and exposures over the last few decades, and experimental evidence in animal models demonstrates the ability of a broad spectrum of various environmental metabolism-disrupting chemicals to induce adiposity and metabolism disruption[69].

Adipocytes are crucial for maintaining metabolic homeostasis as they are repositories of free fatty acids and release hormones that can modulate body fat mass[28]. Adipogenesis is a highly regulated process that involves a network of transcription factors acting at different time points during differentiation[70]. Peroxisome proliferator activated receptor γ (*PPARγ*) is a ligand activated, nuclear receptor and essential regulator of adipocyte formation and function[29], as well as metabolic homeostasis, as all

PPARγ haploinsufficient and KO models present with lack of adipocyte formation and metabolism disruption[71–75].

PPARγ activation regulates energy homeostasis by both stimulating storage of excess energy as lipids in white adipocytes and stimulating energy utilization by triggering mitochondrial biogenesis, fatty acid oxidation and thermogenesis in brite and brown adipocytes. The white adipogenic, brite/brown adipogenic and insulin sensitizing activities of *PPARγ* are regulated separately through post-translational modifications[76–78] and differential co-regulator recruitment[79–82]. Importantly, humans with minimal brite adipocyte populations are at higher risk for obesity and type 2 diabetes[83–85].

Growing evidence supports the hypothesis that environmental *PPARγ* ligands induce phenotypically distinct adipocytes. Tributyltin (TBT) induces the formation of an adipocyte with reduced adiponectin expression and altered glucose homeostasis[30]. Furthermore, TBT fails to induce expression of genes associated with browning of adipocytes (e.g. *Ppara*, *Pgc1a*, *Cidea*, *Elovl3*, *Ucp1*) in differentiating 3T3-L1 adipocytes[31,32]. As a result, TBT-induced adipocytes fail to up-regulate mitochondrial biogenesis and have low levels of cellular respiration[31,32]. The structurally similar environmental *PPARγ* ligand, triphenyl phosphate, also fails to induce brite adipogenesis, and this correlates with an inability to prevent *PPARγ* from being phosphorylated at S273[86].

The EPA developed the Toxicity Forecaster (ToxCast™) program to use high-throughput screening assays to prioritize chemicals and inform regulatory decisions regarding thousands of environmental chemicals[87]. Several ToxCast™ assays can

measure the ability of chemicals to bind to or activate *PPARγ*, and these assays have been used to generate a toxicological priority index (ToxPi) that were expected to predict the adipogenic potential of chemicals in cell culture models[88]. Yet, it has been shown that the results of ToxCast™ *PPARγ* assays do not always correlate well with activity measured in a laboratory setting and that the ToxPi designed for adipogenesis was prone to predicting false positives[89]. Furthermore, the ToxCast/ToxPi approach cannot distinguish between white and brite adipogens[90].

In this study, we investigate differences in cellular response between adipogenic and non-adipogenic compounds, as well as the heterogeneity of response across adipogenic compounds. Our ultimate goal is the identification of potential novel adipogenic compounds, and the taxonomic organization of known and predicted adipogenic compounds based on their divergent transcriptional response. To this end, we generated phenotypic and transcriptomic data from adipocytes differentiated in the presence of 76 different chemicals. We combined the cost-effective generation of agonistic transcriptomic data by 3'Digital Gene Expression, a highly multiplexed RNAseq technology, with a new classification method to predict *PPARγ*-activating and modifying chemicals. Further, we investigated metabolism-related outcome pathways as effects of the chemical exposures. We created a data-driven taxonomy to specifically classify chemicals into distinct categories based on their various interactions with and effects on *PPARγ*. Based on the taxonomy-based predictions, we tested the phenotype (white vs. brite adipocyte functions) of environmental adipogens predicted to fail to induce brite adipogenesis in 3T3-L1 cells and primary human adipocytes.

3.2 Methods

3.2.1 Chemicals

DMSO was purchased from American Bioanalytical (Natick, MA). CAS numbers, sources and catalog numbers of experimental chemicals are provided in Table A.1. Human insulin, dexamethasone, 3-isobutyl-1-methylxanthine (IBMX), and all other chemicals were from Sigma-Aldrich (St. Louis, MO) unless noted.

3.2.2 Cell culture

3T3 L1 (RRID:CVCL_0123, Lot # 63343749) cells were originally derived from a Swiss mouse embryonic fibroblast line. Cells were maintained in high-glucose DMEM (Corning, 10-013-CV) with 10% calf serum (Sigma), 100 U/ml penicillin, 100 μg/ml streptomycin, 0.25 μg/ml amphotericin B. All experiments were conducted with cells between passages 3 and 8. Experimental conditions are outlined in Table 3.1 and Figure A.10A. For experiments, cells were plated in Growth Medium and incubated for 4 days, at which time the cultures are confluent for 2 days. “Naïve” pre-adipocytes were cultured in Growth Medium for the duration of an experiment. On day 0, differentiation was induced by replacing the medium with Differentiation Medium (DMEM,10% fetal bovine serum (FBS, Sigma-Aldrich), 100 U/ml penicillin, 100 μg/ml streptomycin, 250 nM dexamethasone, 167 nM human insulin, 0.5 mM IBMX). Also on day 0, single experimental wells were treated with vehicle (DMSO, 0.2% final concentration), rosiglitazone (100 nM) or test chemicals. Rosiglitazone was included in experiments as a positive control because it is a potent *PPARγ* agonist and metabolic therapeutic. On days

3 and 5 of differentiation, medium was replaced with Maintenance Medium (DMEM, 10% FBS, 167 nM human insulin, 100 U/ml penicillin, 100 μg/ml streptomycin), and the cultures were re-dosed. On Day 7 of differentiation, medium was replaced with Adipocyte Medium (DMEM, 10% FBS, 100 U/ml penicillin, 100 μg/ml streptomycin), and the cultures were re-dosed. On day 10, cytotoxicity was assessed by microscopic inspection, with cultures containing more than 10% rounded cells excluded from consideration (See Table A.1 for information on maximum concentrations tested). Wells with healthy cells were harvested for analysis of gene expression, lipid accumulation, fatty acid uptake, mitochondrial biogenesis, mitochondrial membrane potential, and cellular respiration.

OP9 cells (RRID:CVCL_4398, Lot # 63544739) are a bone marrow stromal cell line derived from newborn calvaria of the (C57BL/6xC3H)F2-op/op mouse. Cells were maintained in αMEM (Gibco, 12-561-056) with 20% FBS, 26 mM sodium bicarbonate, 100 U/ml penicillin, 100 μg/ml streptomycin, 0.25 μg/ml amphotericin B. Cells were plated in 24 well plates at 50,000 cells were well in 500 μl medium and incubated for 4 days. Induction and maintenance of adipogenesis and treatment were as described for 3T3 L1 cells, except that the dexamethasone concentration was 125 nM.

Primary human subcutaneous pre-adipocytes were obtained from the Boston Nutrition Obesity Research Center (Boston, MA) and differentiated as previously described[91]. Experimental conditions are outlined in Table 3.1 and Figure A.10B. The pre-adipocytes were maintained in Growth Medium (αMEM with 10% FBS, 100 U/ml penicillin, 100 μg/ml streptomycin, 0.25 μg/ml amphotericin B). For experiments, human

pre-adipocytes were plated in Growth Medium and grown to confluence (3-5) days. "Naïve" pre-adipocytes were cultured in Growth Medium for the duration of an experiment. On day 0, differentiation was induced by replacing the Growth Medium with Differentiation Medium (DMEM/F12, 25 mM $NaHCO_3$, 100 U/ml penicillin, 100 µg/ml streptomycin, 33 µM d-Biotin, 17 µM pantothenate, 100 nM dexamethasone, 100 nM human insulin, 0.5 mM IBMX, 2 nM T_3, 10 µg/ml transferrin). Also on day 0, single experimental wells also were treated with vehicle (DMSO, 0.1% final concentration), rosiglitazone (positive control, 4 µM) or test chemicals. On day 3 of differentiation, medium was replaced with fresh Differentiation Medium, and the cultures were re-dosed. On days 5, 7, 10, and 12 of differentiation, the medium was replaced with Maintenance Medium (DMEM/F12, 25 mM $NaHCO_3$, 100 U/ml penicillin, 100 µg/ml streptomycin, 3% FBS, 33 µM d-Biotin, 17 µM pantothenate, 10 nM dexamethasone, 10 nM insulin), and the cultures were re-dosed. Following 14 days of differentiation and dosing, cells were harvested for analysis of gene expression, lipid accumulation, fatty acid uptake, mitochondrial biogenesis, and cellular respiration.

3.2.3 Lipid accumulation

3T3-L1 cells or human preadipocytes were plated in 24 well plates at 50,000 cells per well in 0.5 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Medium was removed from the differentiated cells, and they were rinsed with PBS. The cells were then incubated with Nile Red (1 µg/ml in PBS) for 15 min in the dark. Fluorescence (λex= 485 nm, λem= 530 nm) was measured using a Synergy2 plate reader (BioTek Inc., Winooski, VT). The fluorescence in all experimental wells was

normalized by subtracting the fluorescence measured in naïve pre-adipocyte cultures reported as "RFU."

3.2.4 Transcriptome analysis

3T3-L1 cells were plated in 24 well plates at 50,000 cells per well in 0.5 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Total RNA was extracted and genomic DNA was removed using the Direct-zol MagBead RNA Kit and following manufacturer's protocol (Zymo Research, Orange, CA). A final concentration of 5 ng RNA/ul was used for each sample. For each chemical, 3-5 replicates were profiled and carefully randomized across six 96-well plates, including 26 DMSO vehicle controls, and 16 naïve pre-adipocyte cultures. Sequencing and gene expression quantification was carried out by the MIT Technology Lab at Broad Institute (Cambridge, MA). RNA libraries were prepared using a highly multiplexed 3' Digital Gene Expression (3' DGE) protocol developed by [92] and sequenced on an Illumina NextSeq 500, generating between 2.13E8 and 3.87E8 reads and a mean of 3.02E8 reads per lane across 96 samples. All reads containing bases with Phred quality scores < Q10 were removed. The remaining reads were aligned to mouse reference genome, GRCm38, and counted in 21,511 possible transcripts annotations. Only instances of uniquely aligned reads were quantified (i.e. reads that aligned to only one transcript). Furthermore, multiple reads with the same unique molecular identifier (UMI), aligning to the same gene were quantified as a single count.

All processing and analyses of gene expression data were carried out in *R (v 3.4.3)*. The number of counted reads per samples varied widely with a range of 7.90E1 to

2.27E6 (Mean = 2.25E5, SD = 2.94E5). To remove technical noise introduced by low overall expression quantification of individual samples, we performed an iterative clustering-based approach to determine sets of samples which segregate as a result of low total read counts. Each iteration included four steps: removal of low count genes, normalization, plate-level batch correction, and hierarchical clustering. Low count genes, defined as having mean counts <1 across all samples, were removed to reduce statistical noise introduced by inaccurate quantification of consistently lowly expressed transcripts. Normalization was performed using Trimmed Mean of M-values (TMM), the default method employed by *limma (v 3.34.9)*[93]. Batch correction was performed by *ComBat (v 3.26.0)*[94]. Hierarchical clustering was performed on the 3000 genes with the largest median absolute deviation (MAD) score, using Euclidean distance and 1-Pearson correlation as the distance metric for samples and genes, respectively, and Ward's agglomerative method[95]. Clusters of samples clearly representative of low expression quantification were removed. This process was repeated until no such low expression outlier sample cluster was present (four iterations). For the remaining samples, low count genes were removed and samples were normalized and batch corrected by the same procedure.

Following sample- and gene-level quality control filtering, the final processed data set included expression levels of 9,616 genes for each of 234 samples. These 234 samples include 2-4 remaining replicates of each compound, 25 DMSO vehicle controls, and 15 naïve pre-adipocyte cultures.

3.2.5 PPARγ ligand/modifier classification

A classification model was inferred from the *training set* consisting of 38 known *PPARγ*-modifying compounds and 22 known non-*PPARγ* modifying compounds, including vehicle, to predict the label of the *test set* of 17 suspected PPARγ-modifying compounds (Table A.1). The model inference was based on an amended random forest procedure developed to better account for the presence of biological replicates in the data. Specifically, for each classification tree, samples and genes were bagged using sampling techniques consistent with [96]. In particular, samples were bootstrapped (i.e., sampled with replacement), and genes were subsampled by the square root of the number of represented genes. To account for chemical-level variability and to prevent replicates of the same chemical exposure from being separated, we implemented an extra step, which we denote as "bag-merging", to summarize values from replicate samples of the same chemical exposure after bootstrap sampling: within each "bag" of samples, replicates of the same chemical exposure were merged to their mean expression. For prediction of test data, replicates of the same chemical exposure were merged to their mean expression and run through the trained random forest, such that each tree generates a vote of either 0 or 1, *PPARγ*-modifying negative or positive, respectively. The mean of these votes across all trees is a value between 0 and 1, which can be interpreted as the pseudo-probability of the chemical exposure being *PPARγ*-modifying.

Prior to generating the final predictive model, the expected performance of this classification approach for predicting *PPARγ*-modification status with this data set was assessed using 10-fold cross validation on the training set of known *PPARγ*-modification

status. For each fold, samples were stratified at the chemical exposure level, such that each fold included 6 distinct compounds and a different number of samples, and all replicates of the same compound were only included in either the training or the test folds. Next, prior to training each random forest model, gene filtering based on within *vs.* between exposure variance using ANOVA was performed. Genes with an F-statistics associated with an FDR corrected p-value > 0.05 were filtered out. Thresholds for determining class membership based on voting was determined by running the training folds through the random forest and selecting the threshold producing the highest F1-score, i.e., the harmonic mean of precision and sensitivity. Performance was assessed in terms of area under the ROC curve (AUC), as well as precision, sensitivity, specificity, F1-score, and balanced accuracy, i.e., the mean of specificity and sensitivity. All random forests were generated using 2000 decision trees.

The performance of this procedure was compared to three alternative random forest strategies. In the first, which we denote as “pre-merging”, the mean gene expression across replicates was computed, and a classic random forest was applied to the classification of each merged chemical profile. In the second, which we denote as “classic”, replicate samples were treated as independent perturbations and classified based on a classic random forest. Finally, for the third, which we denote as “pooled”, the mean of votes across replicates from the previous strategy were used to estimate class membership per compound. To compare the classifiers’ performance, we repeated the 10-fold CV procedure 10-times to generate a distribution of performance statistics for each strategy.

The final classification model used to predict the unlabeled chemicals was built using the full training set of 59 labelled chemicals, including vehicle, and 1,199 genes after performing the same ANOVA gene filtering approach used for cross validation as described above. The relative importance of each gene in each random forest model was measured using the gini importance measure[96].

3.2.6 PPARγ ligand/modifier clustering

Known and suspected *PPARγ* modifiers were clustered based on their test statistics from univariate analysis comparing each chemical exposure or naïve pre-adipocyte culture to vehicle using *limma (v 3.34.9)*[93]. In order to assess taxonomic differences between different exposure outcomes, a recursive unsupervised procedure, which we denote as *K2Taxonomer*, was developed, whereby the set of *PPARγ* modifying compounds underwent recursive partitioning into subgroups. At each iteration of the procedure, the top 10% genes was selected based on estimation of non-random changes in gene expression across compounds via the sum of squared test statistics across the current set or subset of chemicals. Chemicals were then separated into two clusters based on their Euclidean distance, using Ward's agglomerative method [95]and the *cutree* R function. The procedure was then applied to each of the two subsequent subgroups of chemicals and repeated until the two-cluster split would result in a single chemical in the terminal subgroup. To obtain and measure the most stable clusters, each iteration was bootstrapped 200 times by resampling gene-level statistics with replacement. The most common clusters were used, and the proportion of total bootstrapping iterations that included these identical clustering assignments was reported.

In order to derive gene-signatures of each split, differential analysis was performed to compare chemicals from the two clusters at that split. In these models, biological replicate status was accounted for using the duplicate correlation procedure in the *limma (v 3.34.9)* package. From these models, differential signatures were defined whereby genes were assigned to one of four genesets based on two criteria: their differential expression between the two chemical clusters at the split, and within each chemical cluster – the differential expression between chemicals and vehicle. In particular, for a particular gene, the difference between mean expression between the two chemical clusters must have $|\log2(\text{Fold-Change})| > 1$ (i.e., Fold-Change>2) and an FDR Q-value < 0.1. Each gene was then assigned to one-of-four genesets: 1) genes up-regulated in the "left" chemical group vs. the "right" chemical group, and up-regulated in the left chemical group vs. vehicle; 2) genes up-regulated in the left chemical group vs. the right chemical group, and down-regulated in the left chemical group vs. vehicle; 3) genes up-regulated in the right chemical group vs. the left chemical group, and up-regulated in the right chemical group vs. vehicle; 2) genes up-regulated in the right chemical group vs. the left chemical group, and down-regulated in the right chemical group vs. vehicle. Since the results of direct differential analysis between the two chemical clusters are not indicative of overall up- or down-regulation, these designations were determined based on the aggregate of the comparisons of each chemical or naïve pre-adipocyte culture to vehicle. Specifically, a gene was assigned to a cluster based on maximum absolute value of the mean of the chemical or naïve pre-adipocyte culture versus vehicle derived test statistics used for clustering. Direction of regulation was then

determined based on the sign of the mean of these test statistics. Functional enrichment, comparing these gene sets to independently annotated gene sets was carried out via Fisher's exact test. These gene sets include those of the full set of Gene Ontology (GO) Biological Processes gene set compendia downloaded from *MSigDB* (*c5.bp.v6.2.symbols.gmt)*, as well two gene sets derived from publicly available microarray expression data from an experiment using mouse embryonic fibroblasts to compare wild-type samples with mutant samples that do not undergo phosphorylation of PPARγ at Ser273 (*GEO accession number GSE22033)*[77]. These additional gene sets were comprised of genes, measured to be significantly up- or down-regulated (FDR Q-Value < 0.05) in mutant samples, based on differential analysis of RMA normalized expression with *limma (v 3.34.9)*. In these tables, the nomenclature for each intermediate and terminal subgroup reflects those reported in Figure 3.3.

3.2.7 Human transcriptome analysis

To assess the human relevance of the gene signatures derived from the PPARγ ligand/modifier clustering in chemical-exposed 3T3-L1 cells, we analyzed projections of these signatures onto the transcriptional profiles from 770 subjects who were part of the *Metabolic Syndrome in Men (METSIM)* study. Though comprised of only male subjects, the *METSIM* data set was chosen for this analysis because it is the largest publicly available human subject data set (*GEO accession number GSE70353)*[97], which includes gene expression profiles from subcutaneous adipose tissue, as well as a comprehensive set of cardio-metabolic measurements. Affymetrix Human Genome U219 Array Microarray CEL files were annotated to unique Entrez gene IDs, using a custom CDF file

from *BrainArray (HGU219_Hs_ENSG_22.0.0)*. Each of the four possible gene sets derived from PPARγ ligand/modifier clustering were then projected on each of the 770 human transcriptome profiles using gene set variation analysis (GSVA), using the *GSVA(v 1.30.0)* package [35], resulting in an enrichment score for each gene set and sample.

For each projected gene set, we tested for relationships between single-sample enrichment scores and clinical measurements. Of notice, many of the clinical measurements are correlated with each other, such that confounding is likely to generate many spurious results. To overcome this problem, for each gene set projection, we tested the significance of the partial correlation between single-sample enrichment scores and each of the clinical variables while controlling for the remaining ones, including age, using the *ppcor* (v 1.1) R package. Given that the relationships between single-sample enrichment scores and any one clinical variable are not assumed to be linear, these partial correlations were calculated from Spearman correlation estimates. P-values were adjusted across all combinations of gene set and clinical measurement. Measurements with at least one comparison with a gene set projection yielding an FDR Q-value < 0.1 are reported. To reduce expected redundancies in the measurements, of the 23 initial quantitative measurements included in these data, we ran this analysis on a subset of 12 measurements (Table A.2). Of note, fat free mass % was chosen in this subset of 12 measurements over body mass index because it has been shown to be more associated with risk and presence of metabolic syndrome[98].

3.2.8 Reverse transcriptase (RT)-qPCR

3T3-L1 cells or human preadipocytes were plated in 24 well plates at 50,000 cells per well in 0.5 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Total RNA was extracted and genomic DNA was removed using the 96-well Direct-zol MagBead RNA Kit (Zymo Research). cDNA was synthesized from total RNA using the iScript™ Reverse Transcription System (BioRad, Hercules, CA). All qPCR reactions were performed using the PowerUp™ SYBR Green Master Mix (Thermo Fisher Scientific, Waltham, MA). The qPCR reactions were performed using a 7500 Fast Real-Time PCR System (Applied Biosystems, Carlsbad, CA): UDG activation (50°C for 2 min), polymerase activation (95°C for 2 min), 40 cycles of denaturation (95°C for 15 sec) and annealing (various temperatures for 15 sec), extension (72°C for 60 sec). The primer sequences and annealing temperatures are provided in Table A.3. Relative gene expression was determined using the Pfaffl method to account for differential primer efficiencies [99], using the geometric mean of the Cq values for beta-2-microglobulin (B2m) and 18s ribosomal RNA (Rn18s) for mouse gene normalization and of ribosomal protein L27 (*Rpl27*) and *B2M* for human gene normalization. The Cq value from naïve, pre-adipocyte cultures was used as the reference point. Data are reported as "Relative Expression."

3.2.9 Cell death

3T3-L1 cells or human preadipocytes were plated in 96 well, black-sided plates at 10,000 cells per well in 0.2 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Cells death was measured by treating differentiated cells

will MitoOrange Dye according to manufacturer's protocol (Abcam, Cambridge, MA). Measurement of fluorescence intensity (λex= 485 nm, λem= 530 nm) was performed using a Synergy2 plate reader. The fluorescence in experimental wells was normalized by subtracting the fluorescence measured in naïve pre-adipocyte cultures and reported as "RFU."

3.2.10 Fatty acid uptake

3T3-L1 cells or human preadipocytes were plated in 96 well, black-sided plates at 10,000 cells per well in 0.2 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Fatty acid uptake was measured by treating differentiated cells with 100 µL of Fatty Acid Dye Loading Solution (Sigma-Aldrich, MAK156). Following a 1 hr incubation, measurement of fluorescence intensity (λex= 485nm, λem= 530nm) was performed using a Synergy2 plate reader. The fluorescence in experimental wells was normalized by subtracting the fluorescence in differentiated, negative control wells and reported as "RFU."

3.2.11 Mitochondrial biogenesis

3T3-L1 cells or human preadipocytes were plated in 24 well plates at 50,000 cells per well in 0.5 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Mitochondrial biogenesis was measured in differentiated cells using the MitoBiogenesis In-Cell Elisa Colorimetric Kit, following the manufacturer's protocol (Abcam). The expression of two mitochondrial proteins (COX1 and SDH) were measured simultaneously and normalized to the total protein content via JANUS staining.

Absorbance (OD 600nm for COX1, OD 405nm for SDH, and OD 595nm for JANUS) was measured using a BioTek Synergy2 plate reader. The absorbance ratios of COX/SDH in experimental wells were divided by the ratios in naïve pre-adipocyte cultures and reported as "Relative Mitochondrial Protein Expression."

3.2.12 Oxygen consumption

3T3-L1 cells or human preadipocytes were plated in Agilent Seahorse plates at 50,000 cells per well in 0.5 ml maintenance medium at initiation of the experiment. Dosing is outlined in Table 3.1. Prior to all assays, cell media was changed to Seahorse XF Assay Medium without glucose (1mM sodium pyruvate, 1mM GlutaMax, pH 7.4) and incubated at 37°C in a non-CO_2 incubator for 30 min. To measure mitochondrial respiration, the Agilent Seahorse XF96 Cell Mito Stress Test Analyzer (available at BUMC Analytical Instrumentation Core) was used, following the manufacturer's standard protocol. The compounds and their concentrations used to determine oxygen consumption rate (OCR) included 1) 0.5 μM oligomycin, 1.0 μM carbonyl cyanide-p-trifluoromethoxyphenylhydrazone (FCCP) and 2 μM rotenone for 3T3-L1s; and 2) 5 μM oligomycin, 2.5 μM FCCP, and 10 μM rotenone for the primary human adipocytes. Non-mitochondrial respiration was determined from the minimum rate measurement after injection of rotenone. Basal respiration was determined by subtracting non-mitochondrial respiration from the last rate measurement before the injection of ologomycin. Maximum respiration was determined by subtracting non-mitochondrial respiration from the maximum rate measurement after the injection of FCCP. Spare capacity was determined by subtracting the basal respiration from the maximum respiration.

3.2.13 Statistical analyses

All statistical analyses were performed in *R (v 3.4.3)* and *Prism 7 (GraphPad Software, Inc., La Jolla, CA).* Data are presented as means ± standard error (SE). For 3T3-L1 experiments the biological replicates correspond to independently plated experiments. For human primary preadipocyte experiments the biological replicates correspond to distinct individuals' preadipocytes (3 individuals in all). The Nile Red data and the qPCR data were not normally distributed; therefore, the data were log transformed before statistical analyses. One-factor ANOVAs were performed to analyze the qPCR and phenotypic data and determine differences from vehicle-treated cells. Sequencing data from 3'DGE have been deposited into *GEO (Accession number GSE124564)*.

3.3 Results

3.3.1 Classification of novel taxonomic subgroups of PPARγ ligands/modifiers

Potential adipogens (chemicals that change the differentiation and/or function of adipocytes) were identified by review of the literature and based on reports of PPARγ agonism or modulation of adipocyte differentiation (Table A.1). Our classification groups were labeled as "Yes", "No", or "Suspected", based on the chemical's potential ability to act as a ligand of or modify PPARγ (i.e., to alter its post translational modifications) as noted in the "PPARγ Ligand or Modifier" column in Table A.1.

The classic mouse pre-adipocyte model, 3T3-L1 cells, was differentiated and treated with vehicle (DMSO) or with each of the 76 test chemicals (concentrations are reported in Table A.1). In order to maximize the number of chemicals that could be

characterized, each chemical was tested at a single, maximal, non-toxic dose. We also limited the maximum concentrations to 20 μM because concentrations above this would not be reached in humans and because most (although not all) chemicals are not toxic at or below 20 μM. Lipid accumulation was determined after 10 days. Effects on lipid accumulation spanned significant down-regulation to significant up-regulation (Figure 3.1). Lipid accumulation was highly correlated with expression of adipocyte specific genes (e.g. *Cidec*[36], Figure A.11); therefore, we consider it a biomarker of adipocyte differentiation in this system. Of the 27 chemicals that significantly increased lipid accumulation, 18 were known *PPARγ* ligands/modifiers and 9 were suspected *PPARγ* ligands/modifiers. Mono(2-ethylhexyl) phthalate (MEHP), SR1664, and 15-deoxy-$\Delta^{12,14}$-prostaglandin J2 (15dPGJ2) are *PPARγ* agonists that were expected to increase adipocyte differentiation, but did not. LG268 and TBT are *RXR* agonists that were also expected to significantly increase adipocyte differentiation, but did not. The 3 chemicals that significantly down-regulated lipid accumulation are all known to interact with the retinoic acid receptor. T007 is a *PPARγ* antagonist that was expected to decrease adipocyte differentiation, but did not. Some of the suspected *PPARγ* ligands did not significantly enhance lipid accumulation. Somewhat different results were observed in the late pre-adipocyte model OP9 cells; we have found that these cells are more sensitive to differentiation stimulated by *RXR* ligands. 9-cisRA, LG268, LG754, TBT and TPhT all significantly stimulated adipocyte differentiation (Figure A.13).

Following the analysis of lipid accumulation, RNA was isolated from the cells, and the transcriptome was characterized by RNAseq. The transcriptome data then were

used to develop a classification model to identify *PPARγ* ligands/modifiers. More specifically, the training set of 59 chemicals with "Yes"/"No" labels was used to build the classifier, which was then applied to the prediction of the test set of 17 "Suspected" chemicals. When predicting *PPARγ* ligand/modifier status ("Yes" vs. "No"), the mean AUC, precision, sensitivity, specificity, F1-score, and balanced accuracy from repeated 10-fold cross validation (over the training set) of the random forest with bag merging procedure was 0.89, 0.90, 0.80, 0.85, 0.85, and 0.82, respectively (Figure 3.2A). We observed the most drastic improvement of measured balanced accuracy, precision, and specificity by the bag merging procedure compared to other assessed strategies (Figure A.12). The first two metrics in particular (AUC and precision) reflect expectation of relatively few false positive results compared to the other strategies. In the final model, the voting threshold that produced the highest F1-score was 0.53. When we applied the classifier to the test set of the 17 chemicals of unknown interaction with *PPARγ*, 13 had random forest votes greater than this value (Table 3.2). Of these 13 compounds, four had votes > 0.88. These chemicals included quinoxyfen, tonalide, allethrin, and fenthion. These compounds were predicted as *PPARγ* ligands/modifiers with high confidence and were selected for further functional analyses. Of the 1,215 genes that past ANOVA filtering and were included in the random forest models, ribosomal protein L13 (*Rpl13*) and cell death Inducing DFFA Like Effector C (*Cidec*) had the highest measured Gini Importance (Figure 3.2B) with Rpl13 mostly down-regulated and *Cidec* mostly up-regulated by known *PPARγ* ligands/modifiers (Figure A.13).

3.3.2 Adipogen taxonomy discovery

The taxonomy derived by the *K2Taxonomer* procedure recapitulated many known characteristics shared by *PPARγ* ligands/modifiers included in this study (Figure 3.3). For example, three terminal subgroups were labelled in Figure 3.3 based on their shared characteristics. These include: flame retardants (tetrabromobisphenol A (TBBPA) and triphenyl phosphate (TPhP)), phthalates (MBUP, MEHP, MBZP, and BBZP), and RXR agonists (TBT and LG268). Interestingly, we observed two subgroups containing all of the four thiazolidinediones, with rosiglitazone (Rosig) segregating with the non-thiazolidinedione S26948 and pioglitazone, MCC 555, and troglitazone segregating together.

All of these terminal subgroups fell within a larger module containing 26 chemicals, highlighted by expression patterns consistent with increased adipogenic activity including up-regulation of genes significantly enriched in pathways involved in adipogenesis and lipid metabolism[100]. In addition, these chemicals also demonstrated consistent down-regulation of extracellular component genes. This effect was strongest in cells exposed to thiazolidinediones and flame retardants, two classes of chemicals well-described to be strong *PPARγ* agonists[101–103]. The subgroup of thiazolidinediones, including S26948, was characterized by up-regulation of genes involved in beta-oxidation, the process by which fatty acids are metabolized.

The gene expression profiles of the remaining 17 chemicals, including naïve pre-adipocytes, demonstrate markedly less up-regulation of genes regulated by *PPARγ*. Of these 17 perturbations, a subgroup of 8 chemicals (BADGE, PrPar, 15dPGJ2, SR1664,

METBP, DINP, BuPA, and Fenth) includes the reference vehicle signature. Compared to the next closest subgroup, expression profiles of these compounds was characterized by up-regulation of adipogenesis-related pathways indicative of modest *PPARγ* agonism. Additionally, a subgroup comprised of 9CRA, DBT, LG754, and ATRA exposures and naïve pre-adipocyte signatures was characterized by down-regulation of genes involved in adipogenesis and lipid metabolism, indicating repression of PPARγ activity. Interestingly, both protectin D1 (Prote) and resolvin E1 (Resol) clustered closely in a subgroup with the CDK inhibitor, roscovitine (Rosco), which is known to induce insulin sensitivity and brite adipogenesis[104].

In summary, our top-down taxonomy discovery approach elucidated subgroups of *PPARγ* ligands/modifiers, characterized by differential transcriptomic activity at each split. Annotation of these transcriptomic signatures revealed clear differences in the set and magnitude of perturbations to known adipocyte biological processes by subgroups of chemicals. Membership of these subgroups confirmed many expectations, such as subgroups comprised solely of phthalates, thiazolidinediones, or flame retardants

3.3.3 Relationship between taxonomic subgroups and the human adipose transcriptome

Given that our taxonomic clustering was based on adipogen exposures in a mouse model, we sought to establish its relevance to the relationship between human adipose tissue function and markers of cardio-metabolic health. To this end, we projected the gene signatures derived as either up- or down-regulated gene sets in specific taxonomic subgroups onto a publicly available clinical data set of gene expression profiles from subcutaneous adipose tissue of 770 male subjects and assessed the relationship of these

projections to a set of 12 clinical measurements[97]. Figure 3.4A-B shows the partial correlation between plasma adiponectin and projection of gene sets, either up- or down-regulated in specific subgroups. Figure 3.4C shows these relationships for the remaining measurements which demonstrated a statistically significant partial correlation for at least one projection, FDR Q-value < 0.10. For positive associations, i.e. in the same direction, the up-regulated gene sets have positive partial correlation measurements, while down-regulated gene sets have negative partial correlation measurements. The opposite is true for negative associations.

We observed concordant results for putative markers of metabolic health, plasma adiponectin and fat free mass %, which were positively associated with projection of gene signatures from two terminal subgroups which include all TZD and non-TZD type 2 diabetic drugs, Trogl, MCC555, Piogl, Rosig, and S26948, as well as a terminal subgroup which includes Rosco, Resol, and Protec. Similary, we observed concordant results for putative markers of metabolic dysfunction, interleukin-1 receptor antagonist and fasting plasma insulin, which included a primary subgroup comprised of terminal subgroups characterized by weak *PPARγ* agonism, *PPARγ* modification, or PPARγ activity repression.

Taken together, these results confirm the ability of our mouse-based, in vitro-derived signatures of capturing salient functional aspects of healthy and unhealthy metabolic functions in human subjects.

3.3.4 Investigation of the white and brite adipocyte taxonomy

We aimed to better assess how the distinction between gene expression patterns translated into functional differences in the induced adipocytes. Therefore, we selected chemicals from representative groups related of *PPARγ* ligands/modifiers for genotypic and phenotypic characterization. We compared a strong *PPARγ* therapeutic agonist that also modifies PPARγ phosphorylation (Rosig), a chemical that modifies only PPARγ phosphorylation (Rosco), a weak *PPARγ* agonist and endogenous molecule (15dPGJ2) and two known environmental *PPARγ* ligands (TBBPA and TPhP). 3T3-L1 cells were differentiated and treated with vehicle (DMSO), Rosig (positive control, 20 μM), Rosco (2 μM), 15dPGJ2 (1 μM), TBBPA (20 μM) and TPhP (10 μM). Gene expression and phenotype were determined after 10 days. Analysis of mitochondrial membrane potential confirmed that the concentrations used were not toxic (Figure A.14A), while only Rosig significantly increased cell number (Figure A.14B).

First, we determined if changes in gene expression correlated with expression of genes previously shown to be associated with white and brite adipocytes (Qiang et al. 2012). As expected, all of the *PPARγ* agonists (Rosig, 15dPGJ2, TBBPA, TPhP) significantly increased *Pparγ* expression, while Rosco did not (Figure 3.5A). Similarly, the *PPARγ* agonists induced expression of adipocyte genes common to all adipocytes (*Plin*, *Fabp4*, *Cidec*), while roscovitine did not (Figure 3.5B). In contrast, only the chemicals known to prevent phosphorylation of *PPARγ* at Ser273 (i.e., Rosig and Rosco) induced expression of Pgc1a (Figure 3.5A) and induced expression of brite adipocyte genes (Cidea, Elovl3) (Figure 3.5C). Rosig, Rosco, and 15dPGJ2 induced the expression

of Adipoq (Figure 3.5C). In order for brite adipocytes to catabolize fatty acids and expend excess energy, they must up-regulate expression of β-oxidation genes and mitochondrial biogenesis. In line with their browning capacity, Rosig and Rosco up-regulated expression of Ppara and the mitochondrial marker gene Acaa2 (Figure 3.5D). Furthermore, only Rosig and Rosco strongly up-regulated *Ucp1*, the protein product of which dissociates the H^+ gradient the mitochondrial electron transport chain creates from ATP synthesis (Figure 3.5D).

Next, we determined if changes in gene expression correlated with changes in adipocyte function. 3T3-L1 cells were differentiated and treated as described for the mRNA expression analyses. Fatty acid uptake by adipocytes is necessary for lipid droplet formation and for removal of free fatty acids from circulation. Compared to vehicle-treated cells, all of the adipogens significantly induced fatty acid uptake (Figure 3.6). In order to increase the utilization of fatty acids, mitochondrial number and/or function must increase. Only Rosig and Rosco significantly induced mitochondrial biogenesis, while 15dPGJ2 and the environmental *PPARγ* agonists had no effect (Figure 3.7). Rosig modestly and 15dPGJ2 significantly increased cellular respiration (Figure 3.8, Figure A.15). Rosco, TBBPA and TPhP did not increase cellular respiration.

Overall, Rosig and Rosco, therapeutic *PPARγ* ligand and *PPARγ* modifier, respectively, were most efficacious at inducing gene expression and metabolic phenotypes related to up-regulation of mitochondrial processes and energy expenditure. In comparison, environmental *PPARγ* ligands (TBBPA and TPhP) were not able to induce the gene and phenotypic markers of brite adipocytes.

3.3.5 Identification of novel adipogens that favor white adipogenesis

Quinoxyfen (Quino) and tonalide (Tonal) were two of the environmental chemicals that received the highest *PPARγ* ligand/modifier vote and segregated distinctly from the therapeutic ligands (Table 3.2). Thus, we tested the hypothesis that Quino and Tonal are adipogens that do not induce gene expression or metabolic phenotypes indicative of high energy expenditure or brite adipogenesis. 3T3-L1 cells were differentiated and treated with vehicle (DMSO), rosiglitazone (positive control, 20 μM), Quino (10 μM), or Tonal (4 μM). These concentrations were determined to be non-toxic (Figure A.16A). In 3T3-L1 cells, Quino and Tonal significantly induced lipid accumulation (Figure 3.9A), without increasing cell number (Figure A.16B). They significantly increased expression of the white adipocyte marker gene, *Cidec*. However, Quino failed to significantly increase expression of *Cidea*, the brite adipocyte marker gene, while Tonal significantly suppressed expression of *Cidea* (Figure 3.9B). Accordingly, Quino and Tonal increased fatty acid uptake (Figure 3.9C) but not mitochondrial biogenesis (Figure 3.9D). Quino modestly, but not significantly, increased maximal cellular respiration; Tonal had no effect on cellular respiration (Figure 3.8).

Last, we investigated whether results in our mouse model, 3T3-L1 cells, could be recapitulated in a human model. Primary, human subcutaneous preadipocytes were differentiated and treated with vehicle (DMSO), rosiglitazone (positive control, 4 μM), Quino (4 μM), or Tonal (4 μM). Quino and Tonal significantly induced lipid accumulation (Figure 3.10A). Tonal, but not Quino, increased cell number (Figure A.16C). Both Quino and Tonal failed to induce *CIDEA* expression expression (Figure

3.10B). In contrast to 3T3-L1 cells, Quino and Tonal did not increase fatty acid uptake over that induced by the hormonal cocktail in the differentiated primary human adipocytes (Figure 3.10C). Quino and Tonal also reduced mitochondrial biogenesis (Figure 3.10D).

In summary, the combination of random forest classification voting and gene expression clustering identified two environmental contaminants likely to favor the induction of white adipocytes. Hypothesis testing carried out with functional analyses confirmed that Quino and Tonal induce white, but not brite, adipogenesis in both mouse and human preadipocyte models. Importantly, we demonstrate that hypothesis testing can be conducted with readily available cells lines and analytical reagents.

3.4 Discussion

The chemical environment has changed dramatically in the past 40 years, and an epidemic increase in the prevalence of obesity has occurred over the same time period. Yet, it is still unclear how chemical exposures may be contributing to adverse metabolic health effects. New tools are needed not just to identify potential adipogens, but to provide information on the type of adipocyte that is formed. Here, we have both developed a new analytical framework for adipogen identification and characterization and tested its utility in hypothesis generation. We show that adipogens segregate based on distinct patterns of gene expression, which we used to identify two environmental contaminants for hypothesis testing. Our results support the conclusion that quinoxyfen and tonalide have a limited capacity to induce the health-promoting effects of mitochondrial biogenesis and brite adipocyte differentiation.

3.4.1 Adipogen taxonomy identifies environmental chemicals that favor white adipogenesis

Potential adipogens (chemicals that change the differentiation and/or function of adipocytes) were identified by review of the literature and based on reports of *PPARγ* agonism or modulation of adipocyte differentiation, as well as through querying ToxCast data. Not all of the chemicals identified as *PPARγ* ligands/modifiers induced significant lipid accumulation. We hypothesize that this likely resulted from the fact that we did not apply any chemical above 20 μM (with the exception of fenthion). Concentrations in the 100 μM range have been used in previous studies (e.g. DOSS [105]; parabens[106]; phthalates[107]). The *RXR* agonists LG268 and TBT also were expected to significantly increase adipocyte differentiation but did not. We have previously shown that TBT induces adipogenesis with greater efficacy in OP9 cells rather than 3T3-L1 cell[108] and also show here that OP9 cells more efficaciously respond to *RXR* ligands, in general, likely because they are more committed to adipogenesis as late pre-adipocytes than 3T3 L1 cells, which are early pre-adipocytes.

We identified four compounds as high-confidence *PPARγ* ligands/modifiers: quinoxyfen, tonalide, allethrin, and fenthion. In the final model used for these predictions, biologically informative genes emerge as important for predicting *PPARγ* ligand/modification status, specifically the down-regulation of *Rpl13* and the up-regulation of *Cidec*. *Rpl13* is involved in ribosomal machinery is down-regulated during human adipogensis[109]. *Cidec* is a lipid droplet structural gene, the expression of which is positively correlated with adipocyte lipid droplet size, insulin levels, and glycerol

release[110]. Of these four compounds, quinoxyfen and tonalide are of particular public health concern. Quinoxyfen is among a panel of pesticides with different chemical structures and modes of action (i.e., zoxamide, spirodiclofen, fludioxonil, tebupirimfos, forchlorfenuron, flusilazole, acetamaprid, and pymetrozine) that induce adipogenesis and adipogenic gene expression in 3T3-L1 cells[111]. Quinoxyfen is a fungicide widely used to prevent the growth of powdery mildew on grapes[112]. We chose to test tonalide because it was reported to strongly increase adipogenesis in 3T3-L1 cells, although it was concluded that this response was not due to direct *PPARγ* activation[90]. Our results differ in this regard. Tonalide bioaccumulates in adipose tissue of many organisms including humans, and exposure is widespread because of its common use in cosmetics and cleaning agents[113]. Combined, tonalide and galaxolide constitute 95% of the polycyclic musks used in the EU market and 90% of that of the US market[114].

Our results support the conclusion that quinoxyfen and tonalide are adipogenic chemicals, likely to be acting through *PPARγ*. In clustering analysis, quinoxyfen and tonalide were among the largest subgroup of eight potential strong *PPARγ* agonists. Notably, this cluster includes both synthetic/therapeutic (nTZDpa, tesaglitazar, telmisartan) and environmental compounds (tributyl phosphate and triphenyltin) and is characterized by general up-regulation of pathways of adipogenic activity. However, quinoxyfen and tonalide generate adipocytes that are phenotypically distinct from adipocytes induced by therapeutics such as rosiglitazone. That environmental *PPARγ* ligands can induce a distinct adipocyte phenotype has been shown previously for TBT[30–32] and TPhP[86].

We tested the effect of quinoxyfen and tonalide on human subcutaneous preadipocyte differentiation. There is a strong association between abdominal adiposity and metabolic syndrome[115]. Classically, visceral adipose tissue has been thought to be the driver of metabolic dysfunction; however, there is an alternative explanation that visceral adiposity results secondarily from the dysfunction of subcutaneous adipose tissue in the upper body[116,117]. While humans have greater "browning" potential in their visceral adipose tissue than mice[118], subcutaneous adipose represents 85% of all body fat[119] and thus has a large overall capacity for generating brite adipocytes. Additionally, lack of browning capacity of human subcutaneous adipocytes is associated with insulin resistance[120]. As hypothesized based on the taxonomical analysis, quinoxyfen and tonalide induced white adipocyte functions such as increased lipid accumulation, but in contrast to rosiglitazone, did not induce mitochondrial biogenesis, energy expenditure or brite adipocyte gene expression.

We hypothesize that the differences in adipocyte phenotype that are induced by environmental *PPARγ* ligands (e.g. TBBPA, TPhP, quinoxyfen, tonalide) result from the conformation that *PPARγ* assumes when liganded with these chemicals rather than with therapeutic agents. These differences in conformation not only determine the efficacy to which *PPARγ* is activated but also the transcriptional repertoire[121]. Consistent with this hypothesis, we observed multiple terminal subgroups of *PPARγ* ligands/modifiers of shared properties, specifically TZDs and non-TZD type 2 diabetic drugs, phthalates, and flame retardants. Furthermore, subgroups containing therapeutics share gene expression patterns in human adipose tissue in accordance with positive markers of metabolic health,

specifically plasma adiponectin and fat free mass %, suggesting that these gene expression are related to metabolic health in humans.

Access to post-translational modification sites and coregulator binding surfaces depends upon the structure that *PPARγ* assumes. The white adipogenic, brite/brown adipogenic and insulin sensitizing activities of *PPARγ* are regulated separately through differential co-regulator recruitment[82] and post-translational modifications[77,122], with ligands having distinct abilities to activate each of *PPARγ*'s functions. Suites of genes have been shown to be specifically regulated by the acetylation status of *PPARγ* (*SirT1*-mediated)[78], by the phosphorylation status of *PPARγ* (*ERK/MEK/CDK5*-mediated)[77,104] and/or by the recruitment of *Prdm16* to *PPARγ*[123]. Future work will investigate the connections between the phosphorylation status of *PPARγ* liganded with environmental *PPARγ* ligands such as quinoxyfen and tonalide, the recruitment and release of coregulators, and the ability of *PPARγ* to recruit transcriptional machinery to specific DNA-binding sites. It will be important to determine the metabolic effects of chemicals like quinoxyfen and tonalide in vivo.

3.4.2 Analytical approaches for adipogen characterization

In this study, we performed a high-throughput, cost-effective transcriptomic screening to profile adipocytes formed from 3T3-L1 preadipocytes exposed to a panel of compounds of known and unknown adipogenic impact. Common to toxicogenomic projects, this panel-based study design allows for characterization of the extent to which each chemical modifies differentiation (in this case, adipogenesis as related to the change in lipid accumulation). It also supports the exploration of how subsets of chemicals

influence multiple biological processes that determine the functional status of a cell (in this case, processes that determine white vs. brite adipogenesis). Exploration of these biological processes allows for the prediction of the phenotypic impact of previously unclassified compounds, as well as for the characterization of the heterogeneity of the cellular activity of compounds with similar known phenotypic impact. Here we have performed both types of analyses: first through the implementation and application of random forest classification models to identify potential *PPARγ* ligands/modifers, and second via the recursive clustering of the data to identify and characterize taxonomic subgroup of known and predicted *PPARγ* ligands/modifiers.

For both analyses, we introduced amendments to commonly used machine learning procedures, to improve accuracy and resolution of the acquired result. For the classification task, we amended the random forest algorithm to tailor it to study designs typically adopted in toxicogenomic projects (see Methods). With the addition of an extra step to average the expression across replicates of the bootstrapped samples, we observe consistently higher performance across conventional metrics than with the standard algorithm. For the clustering task, we employ a procedure where we recursively divide sets of chemicals into two subgroups and assess the robustness of each division, as well as annotate transcriptional drivers of each division. As a result, we are not limited to interpreting the clustering results as mutually exclusive groups, but rather as a taxonomy of subgroups where sets of compounds share some transcriptional impact and differ in others, as is expected given the dynamic nature of the modifications by which compounds directly and indirectly affect *PPARγ* activity.

Future work will generalize random forest method to incorporate more complex study designs. To this end, the classification approach adopted in this project is being developed as a random forest software tool soon to be made available as an R package, allowing for the interchanging independent functions at different steps of the algorithm. The strength and utility of this approach extends beyond toxicogenomic studies, and can be used in a variety of applications of high-throughput screening, including drug discovery, such as the Connectivity Map (*CMap*)[124] and longitudinal molecular epidemiology studies, such as the Framingham Heart Study[125].

3.4.3 Adipogen portal

Given the breadth of results generated by this analysis, our description here is far from exhaustive. As such, we have created an interactive website (https://montilab.bu.edu/adipogenome/) to support the interactive exploration of these results at both the gene and pathway-level. The portal is built around a point-and-click dendrogram of the clustering results as in Figure 3.3. Selecting a node of this dendrogram will populate the rest of the portal with the chemical lists, differential analysis, and pathway level hyper-enrichment results for each subgroup defined by a split. For instance, selecting node "H" will show the chemicals in each subgroup to the right (Group 1 = Honokiol, T007907; Group 2 = Prote, Resol, and Rosco), as well as the differential gene signature for each group below. Selecting *Cidec*, the top gene in the Group 2 signature, displays hyper-enrichment results for gene sets which include Cidec and have a nominal p-value < 0.50. The hyper-enrichment results for all genes can be found below this table. Finally, selecting a gene set name will display the gene set

members at the bottom frame of the portal, with gene hits in bold. All tables are query-able and downloadable.

3.5 Conclusions

Emerging data implicate contributions of environmental metabolism-disrupting chemicals to perturbations of pathways related to metabolic disease pathogenesis, such as disruptions in insulin signaling and mitochondrial activity. There is still a gap in identifying and examining how environmental chemicals can act as obesity-inducing and metabolism-disrupting chemicals. Our implementation of novel strategies for classification and taxonomy development can help identify environmental chemicals that are acting on *PPARγ*. Further, our approach provides a basis from which to investigate effects of adipogens on not just the generation of adipocytes, but potentially pathological changes in their function. To this end, we have shown how two environmental contaminants, quinoxyfen and tonalide, are inducers of white adipogenesis.

Table 3.1: Summary of experimental conditions

<table>
<tr><th></th><th>3T3-L1</th><th>HUMAN PREADIPOCYTES</th></tr>
<tr><td>Exposure Period (days)</td><td>10</td><td>14</td></tr>
<tr><td>Number of times doseD</td><td>4</td><td>6</td></tr>
<tr><td>Positive Control</td><td>Rosiglitazone</td><td>Rosiglitazone</td></tr>
<tr><td>Negative Control (Vehicle)</td><td>DMSO</td><td>DMSO</td></tr>
<tr><td>Endpoint</td><td colspan="2">Test Chemicals and Concentrations (M)</td></tr>
<tr><td>Transcriptome Analysis</td><td rowspan="2">Rosiglitazone (100 nM)
See Table A.1</td><td>NP</td></tr>
<tr><td>Lipid Accumulation</td><td rowspan="5">Rosiglitazone (4 μM)
Quinoxyfen (4 μM)
Tonalide (4 μM)</td></tr>
<tr><td>RT-qPCR</td><td rowspan="4">Rosiglitazone (1 or 20 μM)
Roscovitine (2 μM)
15dPGJ2 (1 μM)
TBBPA (20 μM)
TPhP (10 μM)
Quinoxyfen (10 μM)
Tonalide (4 μM)</td></tr>
<tr><td>Cell Death</td></tr>
<tr><td>Fatty Acid Uptake</td></tr>
<tr><td>Mitobiogenesis</td></tr>
<tr><td>Cell Number</td><td rowspan="2">NP</td></tr>
<tr><td>Oxygen Consumption</td></tr>
</table>

NP – not performed

Table 3.2: Amended random forest classification results for 17 compounds suspected to be *PPARγ* Ligands/Modifiers.

CHEMICAL NAME	REFEFRENCE	KNOWN SOURCE/USE	PPARγ LIGAND/MODIFIER (VOTE ± 95% CI)
CHEMICALS ABOVE THE HIGHEST F1-SCORE THRESHOLD			
d-cis,trans-Allethrin	88	Insecticide	0.91 ± 0.01
Tonalide	90	Musk (fragrance)	0.90 ± 0.01
Quinoxyfen	88	Fungicide	0.90 ± 0.01
Fenthion	88	Insecticide	0.88 ± 0.01
2,4,6-Tris(tert-butyl)phenol	88	Antioxidant (industrial)	0.80 ± 0.02
Prallethrin	88	Insecticide	0.78 ± 0.02
Tebuconazole	88	Fungicide	0.78 ± 0.02
Fludioxonil	88	Fungicide	0.77 ± 0.02
Tris(1,3-dichloro-2-propyl) phosphate	38	Flame retardant	0.76 ± 0.02
Cyazofamid	88	Pesticide	0.72 ± 0.02
Perfluorooctanoic acid	126	Fluorosurfactant	0.59 ± 0.02
Triphenyl phosphite	102	Pesticide	0.57 ± 0.02
Tris(1-chloro-2-propyl) phosphate	127	Flame retardant	0.54 ± 0.02
CHEMICALS BELOW THE HIGHEST F1-SCORE THRESHOLD			
Triphenylphosphine oxide	128	Crystallizing aid, byproduct	0.49 ± 0.02
Diphenyl phosphate	129	Metabolite of TPhP	0.47 ± 0.02
Dioctyl sulfosuccinate sodium	105	Surfactant	0.41 ± 0.02
Perfluorooctanesulfonic acid	126	Fluorosurfactant	0.40 ± 0.02

[a] We used the ToxPi designed to identify chemicals in the ToxCast dataset that are likely to be *PPARγ* ligands/modifiers.

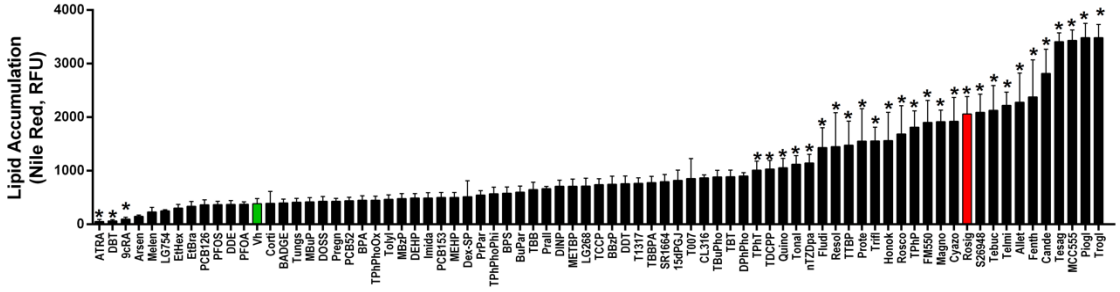


Figure 3.1: Lipid accumulation in differentiated and treated 3T3-L1 pre-adipocytes.

Confluent 3T3 L1 cells were differentiated using a standard hormone cocktail for 10 days. During differentiation, cells were treated with vehicle (Vh, 0.2% DMSO, final concentration), rosiglitazone (positive control, 100 nM) or test chemical (Table A.1). On days 3, 5, and 7 of differentiation, the medium was replaced and the cultures re-dosed. Following 10 days of differentiation and dosing, cells were analyzed for lipid accumulation by Nile Red staining. In Naïve cells (undifferentiated 3T3 -L1 cells) Nile Red staining was 7% of the maximum. In Vh-treated cells (hormone cocktail only) Nile Red staining was 25% of the maximum. Data are presented as mean ± SE (n=4). Statistically different from Vh-treated (highlighted in green) (*$p<0.05$, ANOVA, Dunnett's).

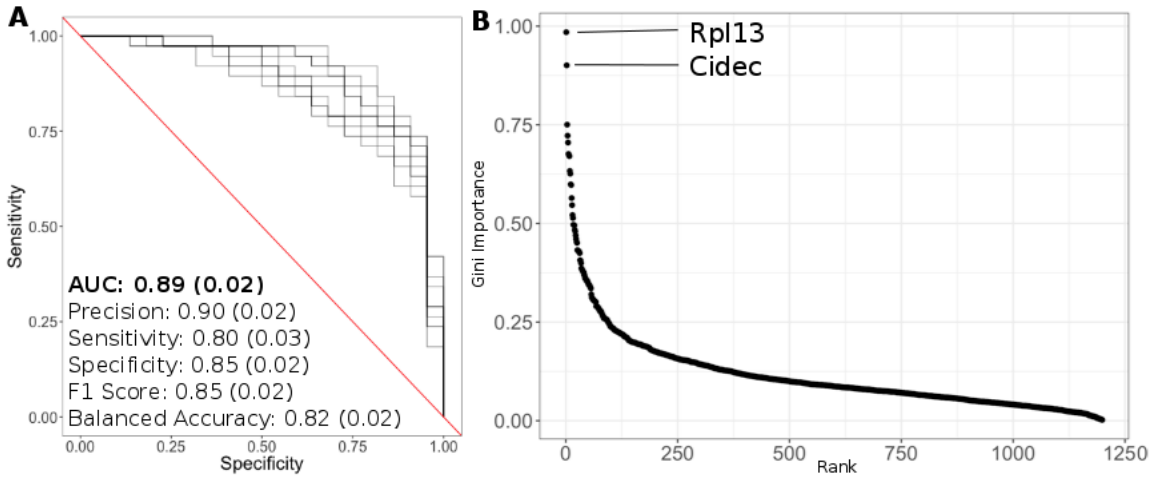


Figure 3.2: Amended random forest classification performance and gene importance of final classification model.

A) Performance of random forest classification procedure based on 10-fold cross validation.
B) Gini Importance versus ranking of genes used in the final random forest model. The names of the top 2 genes are highlighted. Compound-specific gene expression of *Rpl13* and *Cidec* are shown in Figure A.13.

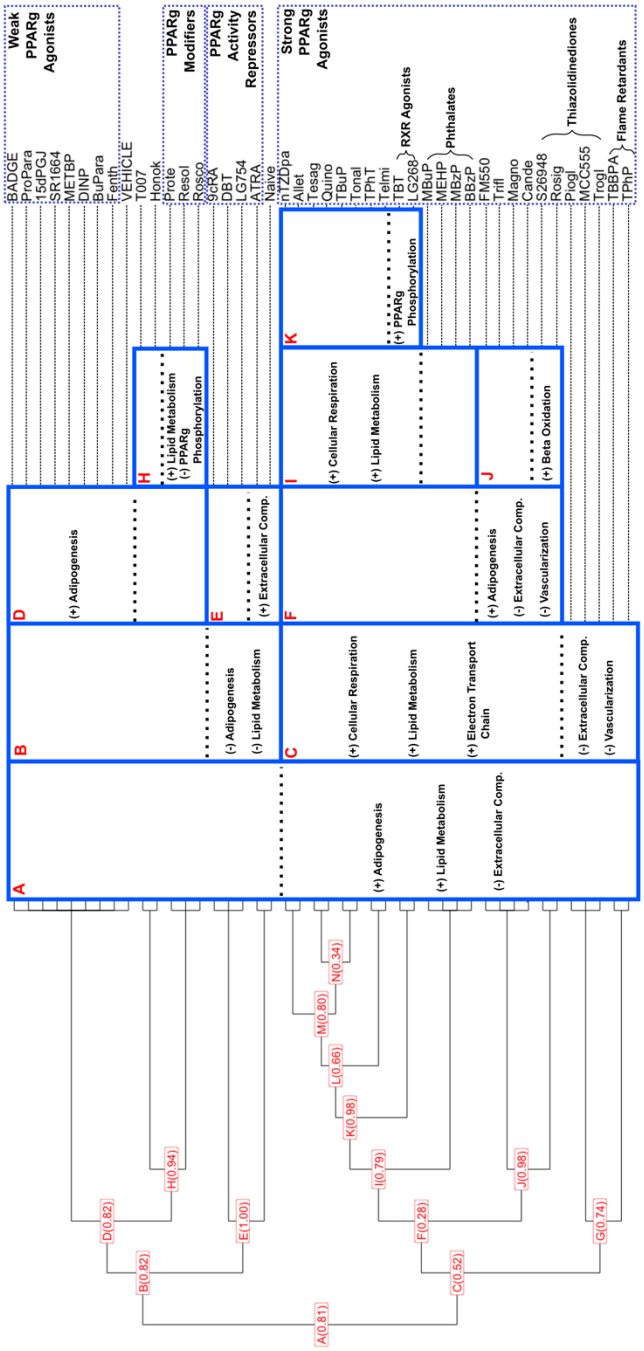


Figure 3.3: Chemical taxonomy of *PPARγ* ligands/modifers based on K2 clustering of the 3'DGE data.

The dendrogram shows the taxonomy-driven hierarchical grouping of test chemical exposures of 3T3-L1 cells or naïve pre-adipocytes. Each split is labeled with a letter, and

the proportion of gene-level bootstraps which produced the resulting split is shown. Highlights of hyper-enrichment of gene ontology (GO) biological processes are shown.

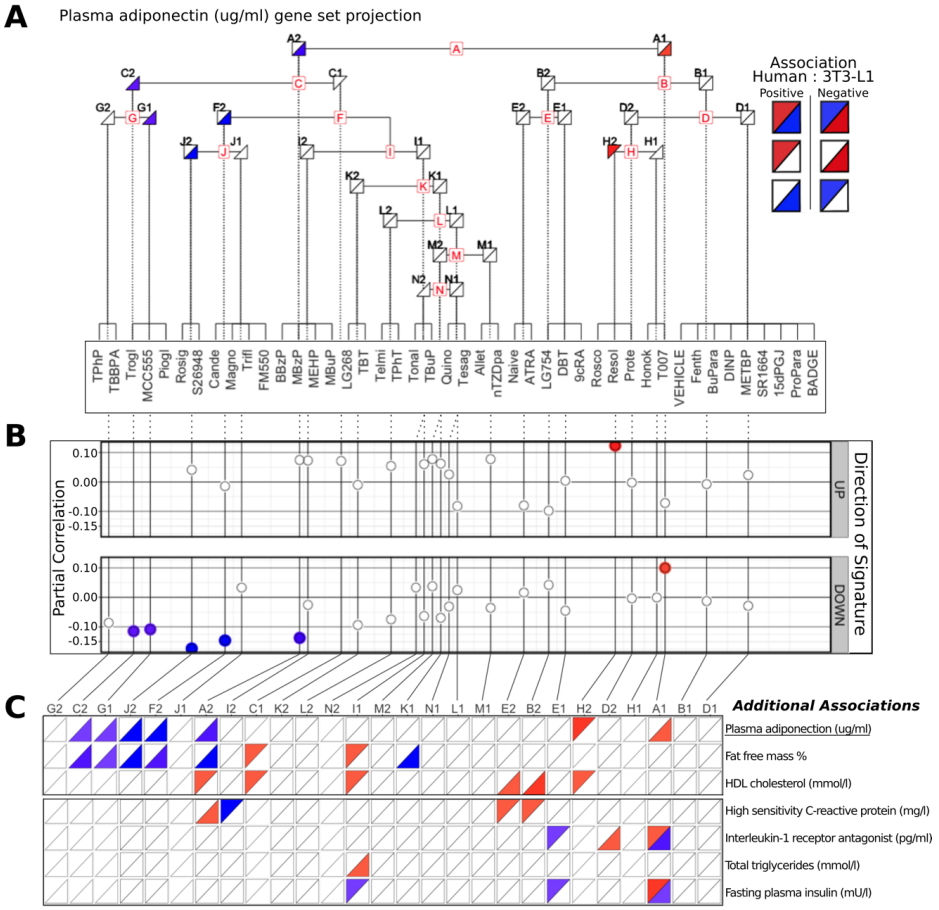


Figure 3.4: Associations between plasma adiponectin levels and projections of the 3T3-L1 derived chemical taxonomy gene signatures onto human adipose tissue gene expression.

A) Each adjoined triangle in the dendrogram denotes a set of genes derived from the 3T3-L1 data, either up-regulated (left) or down-regulated (right) at a given node. Triangles are colored according to the direction of the partial correlation of the projection of these gene sets with plasma adiponectin levels. Interpretation of these associations is in the key on the upper right. All colored triangles reach a significance threshold of FDR Q-value < 0.10.

B) Plots of the partial correlation measurements for gene set projections and plasma adiponectin levels. The top and bottom plots are indicative of gene sets that are either up- or down-regulated in a particular sub-group, respectively. All colored points reach a significance threshold of FDR Q-value < 0.10, as in (A).

C) Results of all partial correlation analyses of projection of taxonomy gene sets with clinical measurements with at least one comparison reaching significance, FDR Q-value < 0.10. The interpretation of these associations is the same as in (A).

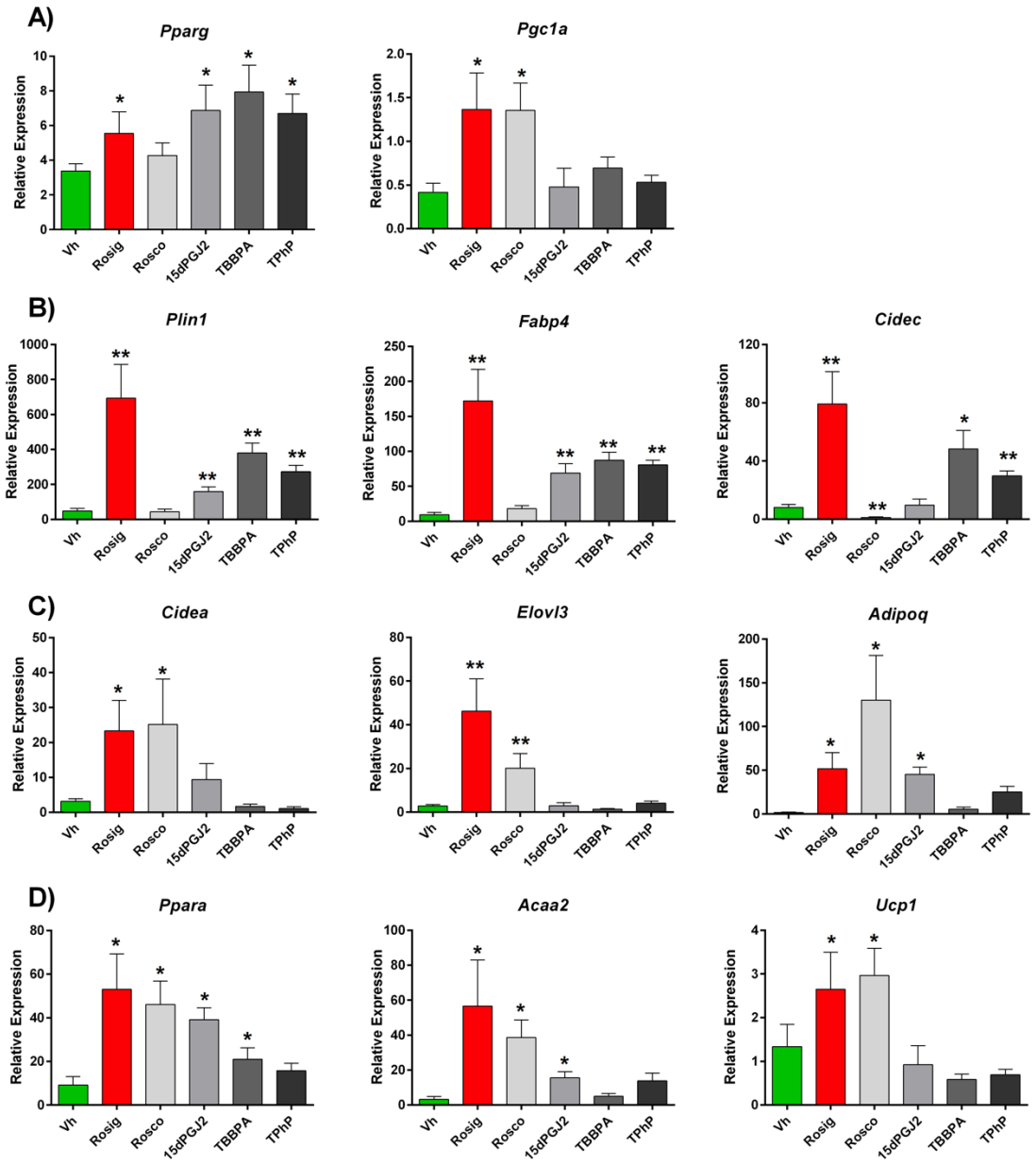


Figure 3.5: White and brite gene expression in differentiated and treated 3T3-L1 adipocytes.

Confluent 3T3 L1 cells were differentiated using a standard hormone cocktail for 10 days. During differentiation, cells were treated with vehicle (Vh, 0.2% DMSO, final concentration), rosiglitazone (Rosig, 1 μM), roscovitine (Rosco, 2 μM), 15dPGJ2 (1 μM), TBBPA (20 μM) and TPhP (10 μM). On days 3, 5, and 7 of differentiation, the adipocyte

maintenance medium was replaced and the cultures re-dosed. Following 10 days of differentiation and dosing, cells were analyzed for gene expression by RT-qPCR.

A) *PPARγ* and coregulator expression.
B) Genes related to white adipogenesis.
C) Genes related to brite adipogenesis.
D) Genes related to mitochondrial biogenesis and energy expenditure.

Data are presented as mean ± SE of n=4 independent experiments. Statistically different from Vh-treated (highlighted in green) (*$p<0.05$, **$p<0.01$, ANOVA, Dunnett's).

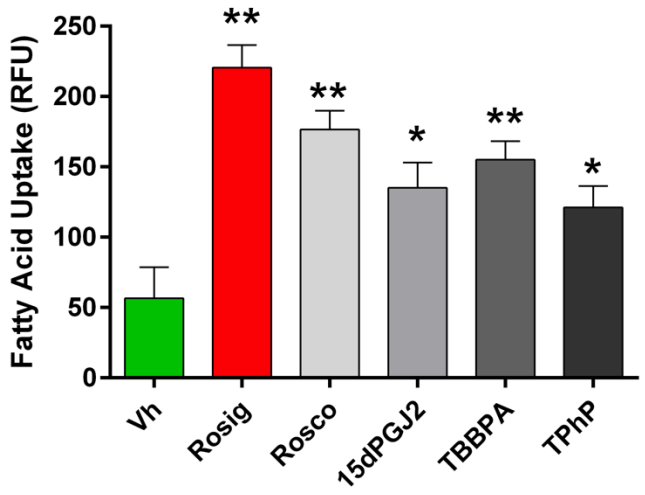


Figure 3.6: Fatty acid uptake in differentiated and treated 3T3-L1 adipocytes.

Differentiation and dosing were carried out as described in Figure 3.5. Following 10 days of differentiation, fatty acid uptake was analyzed using a dodecanoic acid fluorescent fatty acid substrate. Data are presented as means ± SE (n=4). Statistically different from Vh-treated (highlighted in green) (*p<0.05, **p<0.01, ANOVA, Dunnett's).

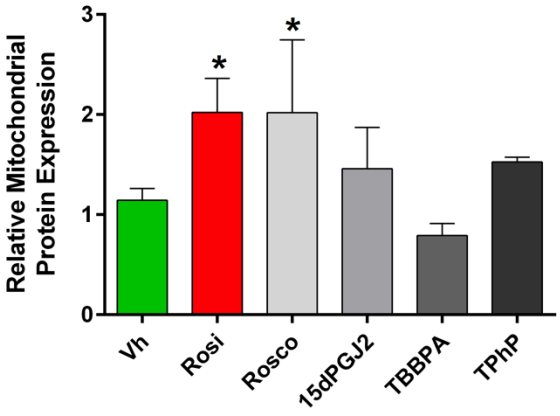


Figure 3.7: Mitochondrial biogenesis in differentiated and treated 3T3-L1 adipocytes.

Differentiation and dosing were carried out as described in Figure 3.5. Following 10 days of differentiation, mitochondrial biogenesis was analyzed by measuring mitochondria-specific proteins. Vehicle, Rosig and TPhP data have been published previously[86]. Data are presented as means ± SE (n=4). Statistically different from Vh-treated (highlighted in green) (*p<0.05, ANOVA, Dunnett's).

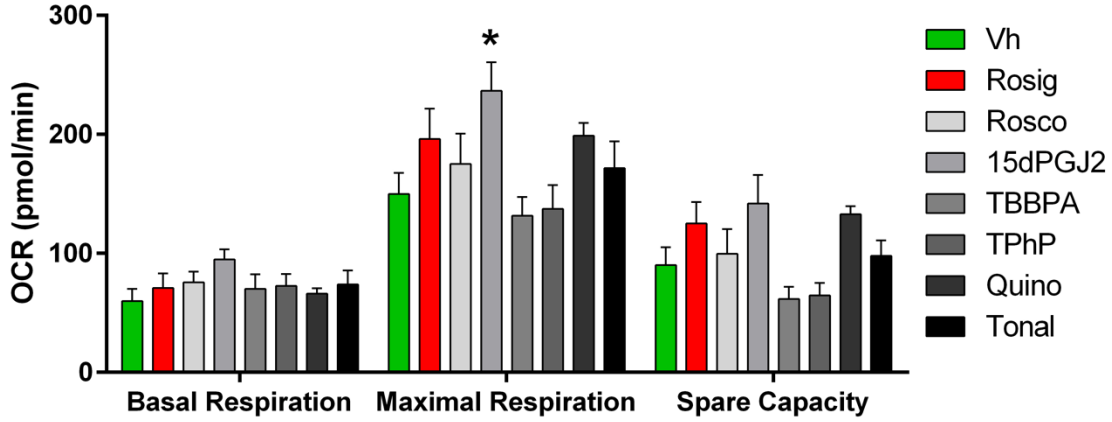


Figure 3.8: Cellular respiration in differentiated and treated 3T3-L1 adipocytes.

Differentiation and dosing were carried out as described in Figure 3.5, with the exception of Rosig (20 µM). Following 10 days of differentiation, mitochondrial respiration was analyzed by Seahorse Assay. Vehicle, Rosig and TPhP data have been published previously[86]. Data are presented as means ± SE (n=4). Statistically different from Vh-treated (highlighted in green) (*$p<0.05$, ANOVA, Dunnett's).

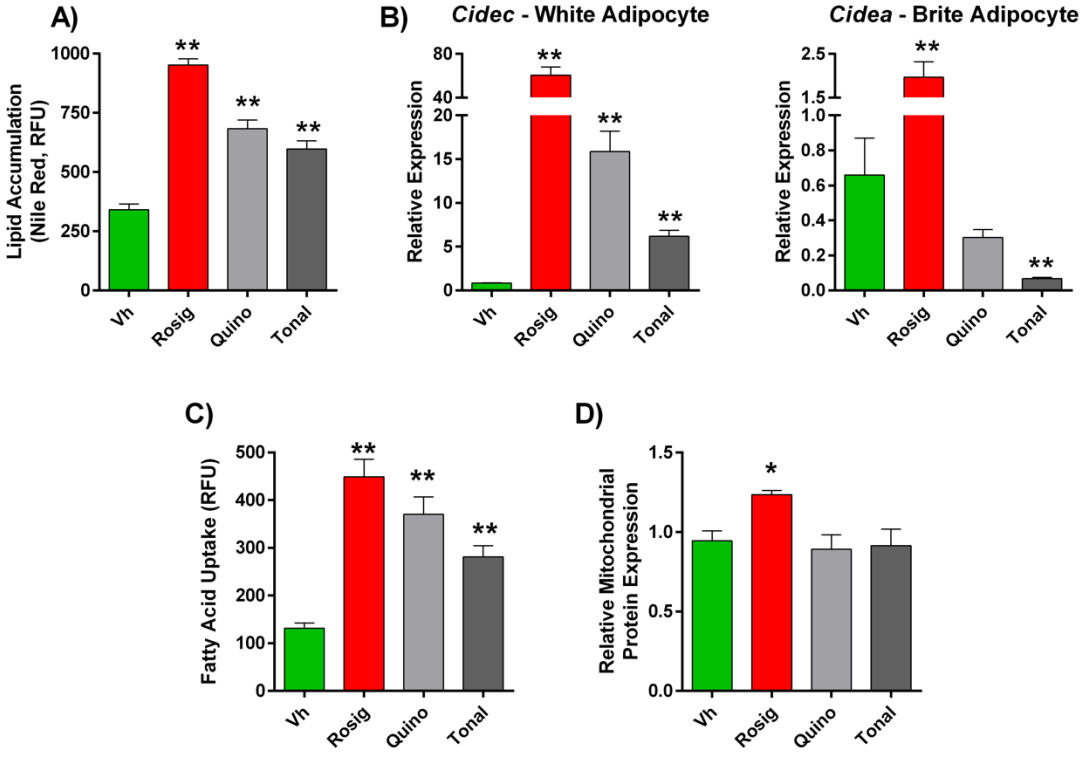


Figure 3.9: Tonalide and quinoxyfen induce white, but not brite, adipogenesis in 3T3-L1 pre-adipocytes.

Confluent 3T3 L1 cells were differentiated using a standard hormone cocktail for 10 days. During differentiation, cells were treated with vehicle (Vh, 0.2% DMSO, final concentration), rosiglitazone (Rosig, 1 µM), quinoxyfen (Quino, 10 µM) or tonalide (Tonal, 4 µM). On days 3, 5, and 7 of differentiation, the adipocyte maintenance medium was replaced and the cultures re-dosed. Following 10 days of differentiation and dosing, cultures were analyzed for

A) adipocyte differentiation
B) white (*Cidec*) and brite (*Cidea*) gene expression
C) fatty acid uptake
D) mitochondrial biogenesis.

Data are presented as means ± SE (n=4). Statistically different from Vh-treated (highlighted in green) (*$p<0.05$, ANOVA, Dunnett's).

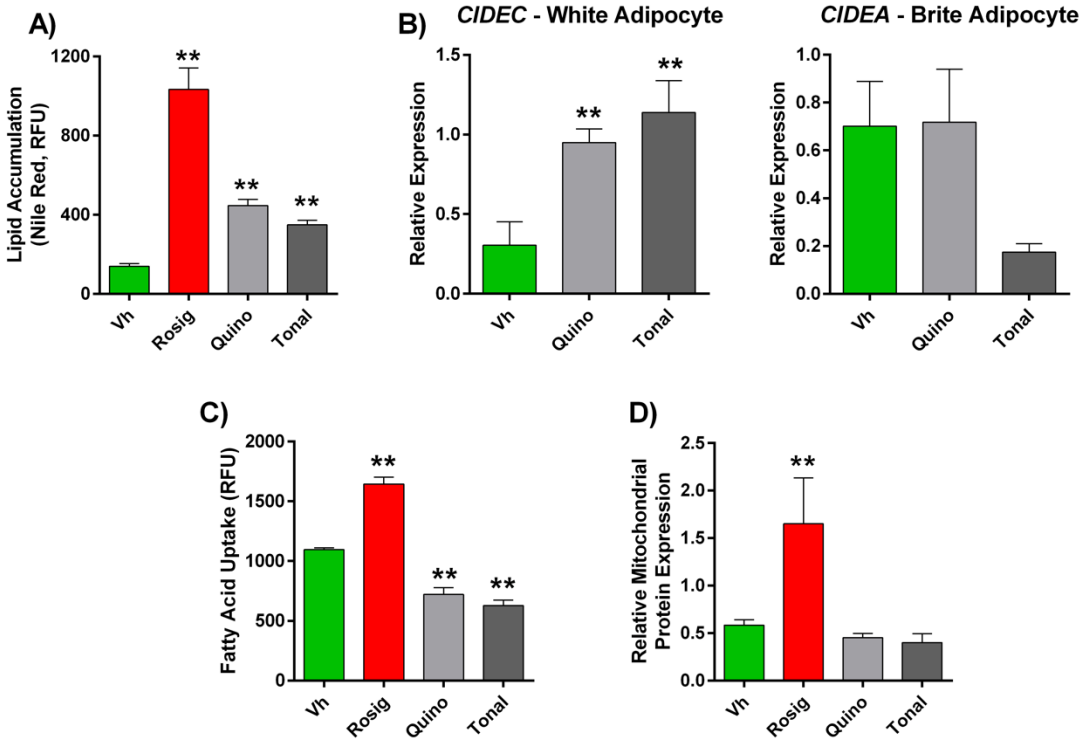


Figure 3.10: Tonalide and quinoxyfen induce white, but not brite, adipogenesis in primary human adipocytes

Confluent primary human preadipocytes were differentiated using a standard hormone cocktail for 14 days. During differentiation, cells were treated with vehicle (Vh, 0.1% DMSO, final concentration), rosiglitazone (Rosig, 4 μM), quinoxyfen (Quino, 4 μM) or tonalide (4 μM). On days 3, 5, 7, 10, and 12 of differentiation, the medium was replace and the cultures re-dosed. Following 14 days of differentiation and dosing, cultures were analyzed for

A) adipocyte differentiation
B) white (*Cidec*) and brite (*Cidea*) gene expression
C) fatty acid uptake
D) mitochondrial biogenesis.

Data are presented as mean ± SE (n=3, each n is from adipocytes from an individual). Statistically different from Vh-treated (highlighted in green) (*$p<0.05$, ANOVA, Dunnett's).

Chapter 4: Tool development for characterization of molecular subgroups in bulk and single-cell transcriptomic profiling data

4.1 Background

As high-throughput transcriptomic assays become more efficient and cost effective, they are being routinely integrated in large-scale biomedical projects[130–133]. Bulk gene expression profiling by RNA sequencing (RNAseq) has been widely adopted in multiple high-throughput genomics studies, the paramount example being *The Cancer Genome Atlas (TCGA)* data commons, which currently include 10,558 bulk RNA sequencing (RNAseq) profiles across 33 cancer types (https://portal.gdc.cancer.gov/). Furthermore, since its first published application in 2009[134], the size of single-cell RNA sequencing (scRNAseq) studies has exploded, such that is now commonplace for studies to generate tens of thousands of profiles[19]. As the scale of these studies and the associated datasets increases, so does their utility as a resource from which biological information can be extracted through the application of machine learning approaches. Common deliverables of these types of analysis include discovery and characterization of molecular subtypes, which is prevalent in both bulk and single-cell gene expression studies. For example, *TCGA* bulk expression data has been utilized to characterize subtypes of numerous cancer, including but not limited to: breast[135], colorectal[136], liver[137], and bladder cancer[138,139]. Similarly, characterization of molecular subtypes is a standard component of the scRNAseq data analysis workflow, insofar as estimation and annotation of subpopulations of cells is one of the primary goals of the assay[140].

The general framework for subtype characterization can be summarized in two steps: 1) estimation of data-driven groups of observations via application of an unsupervised learning procedure, followed by 2) annotation of each group based on the identification of distinct patterns of gene expression relative to other groups. While most approaches focus on discovering a "flat" set of non-overlapping groups or subtypes, in this chapter we present an alternative approach, devised to emphasize "taxonomy-like" hierarchical relationships between observations in order to discover non-mutually exclusive subgroups.

Whereas a wide range of unsupervised learning algorithms is available for the analysis of bulk gene expression data, the considerable sparsity of scRNAseq data has motivated the development of novel methods specifically tailored to the analysis of this type of sparse, high-dimensional data. Popular software packages, such as Seurat[141] and Scran[142], generate "flat" clusters, in which a finite set of mutually exclusive cell types is estimated. In so doing, they fail to capture the "taxonomy-like" hierarchical structure that may exist among subgroups of observations at multiple levels of resolution, driven by transcriptional signatures based on different factors, including but not limited to: shared lineage, cell state, pathway activity, or morphological origin. Complementary methods exist to model such relationships, such as *Neighborhood Joining*[143,144]and more recent single cell *trajectory inference* approaches[145], which estimate "pseudo-temporal" states of individual cells indicative of developmental progression. Given the stringent interpretation of such models, their suitability depends on the assumption that the measured similarity between cell profiles arises solely from temporal progression.

However, relative similarity between cell profiles may be confounded by numerous factors, including: cell cycle progression, spatial patterning, and cell stress[146].

Hierarchical clustering (HC) algorithms at face value address the need for a multi-resolution representation of the relationship among observations, and while originally adopted for the analysis of bulk gene expression data[147], numerous packages have also been developed for scRNAseq analysis, such as *pcaReduce*[148], *ascend*[149], and *BackSPIN*[150]. However, since the number of possible subgroupings increases with the number of observations, robustly identifying such relationships can be challenging. As a result, tree-cutting methods are often applied, ultimately yielding a flat set of non-overlapping clusters. Furthermore, the bottom-up nature of HC's sample aggregation procedure forces the use of the same set of genes/features to drive the agglomeration at all levels of the hierarchy, thus precluding the discovery of nested structures defined by possibly distinct transcriptional programs.

Here we introduce *K2Taxonomer*, a novel taxonomy discovery approach and associated R package for the estimation and in-silico characterization of hierarchical subgroup structures in both bulk and single cell data. An important feature of the approach is that it can analyze both individual samples as well as sample groups such as, but not limited to, those corresponding to scRNAseq cell types. The package employs a recursive partitioning algorithm, which utilizes repeated perturbations of the data at each partition to estimate ensemble-based K=2 subgroups. For scRNAseq analysis, *K2Taxonomer* utilizes the constrained k-means algorithm[151], to estimate partitions of the data at the cell type level, while preserving the influence of each individual cell profile. A

defining feature of the method is that each recursive split of the input data is based on a distinct set of features selected to be most discriminatory within the subset of samples member of the current hierarchy branch. This makes the approach quite distinct from a standard clustering algorithm, and particularly apt to discover nested taxonomies. In addition, the package includes functionalities to comprehensively characterize and statistically test each subgroup based on their estimated stability, gene expression profiles, and a-priori phenotypic annotation of individual profiles. Importantly, all results are aggregated into an automatically generated interactive portal to assist in parsing the results.

In this chapter we assess the performance of *K2Taxonomer* for partitioning both bulk gene expression and scRNAseq data, using both simulated and publicly available data sets, and we compare it to agglomerative clustering procedures. For bulk gene expression data, performance is assessed in terms of unsupervised sorting of breast cancer subtypes and established genotypic markers, using breast cancer patient tumor tissue data from the *Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)*[13] and the *TCGA* compendia. For scRNAseq data, performance is assessed in terms of recapitulation of established relationships between 28 annotated cell types of the airway of healthy subjects[152]. We conclude with a case study where we perform a *K2Taxonomer*-based analysis of breast cancer tumor infiltrating lymphocytes (TILs) profiled by scRNAseq[153]. Our analysis significantly expands upon the previously published results, and identifies a phenotypically diverse subgroup of CD4 and CD8 T cells, characterized by constitutive up-regulation of a subset of translation machinery

genes. We further show that high expression of translational machinery genes in breast cancer tissue bulk expression are associated with better survival, supporting recent findings on the role of the translation machinery assembly in T cell activation[154,155], and demonstrate that this coordinated expression of the translation machinery is pervasive among T cell subpopulations to such a degree high expression of these genes in bulk measurements of tumor tissue is indicative of the presence of immune infiltrating lymphocytes. The complete suite of analysis results is accessible through an automatically generated and publicly accessible portal.

While in this chapter we focus on the analysis of transcriptomics data, we emphasize that our approach is equally applicable to other bulk and single cell 'omics' data, such as those generated by high-throughput proteomics and metabolomics assays.

4.2 Methods

4.2.1 K2Taxonomer algorithm overview

K2Taxonomer implements a recursive partitioning algorithm that takes as input either a set of individual observations or a set of sample groups and returns a top-down hierarchical taxonomy of those samples or groups (Illustration 4.1). To achieve robust model estimation, each partition is defined based on the aggregation of repeated partition estimations from distinct perturbations of the original set. Each of these partition estimates is created in three steps. First, a perturbation-specific data set is generated by bootstrapping features, i.e., sampling features from the original data set with replacement. Next, this perturbation-specific data set undergoes variability-based feature selection filtering. Finally, a K=2 clustering algorithm is run, producing a perturbation-specific

partition estimate. These three steps are repeated, generating a set of perturbation-specific partition estimates, which are aggregated into a cosine similarity matrix.The aggregate partition is then estimated based on hierarchical clustering and a K=2 tree cut. The current implementation of *K2Taxonomer* includes many options for parametrizing steps in this procedure. Additionally, the *K2Taxonomer* package includes functionalities for performing group-level recursive partitions, i.e., partitioning data sets where observations have a priori-assigned group labels, whereby the objective of the *K2Taxonomer* procedure is to identify intermediate relationships between these groups. This functionality was specifically incorporated to enable partitioning and annotations of cell types estimated by scRNA-seq clustering algorithms, but it is applicable to any data set with group-level labels.

4.2.2 K2Taxonomer feature filtering

A distinguishing property of *K2Taxonomer* when compared to other methods, such as traditional agglomerative hierarchical cluster or trajectory inference, is the manner in which feature selection is implemented. Even in large studies of high-throughput data sets, the number of features is typically much larger than the number of observations. This generally requires filtering the data set prior to modelling in order to reduce variance and computational expense of model fit. One way to do this is through feature selection, in which features suspected to contain more information about the relationship between observations are chosen for down-stream analysis. For unsupervised learning, relative information estimation is commonly calculated via variability-based metrics. Assuming the amount of noise is consistent across features, these metrics will

capture the relative magnitude of signal of individual features. Two common choices are standard deviation (SD) and median absolute deviation (MAD), of which the former is more statistically efficient with small sample size and the latter is more robust to outliers[156]. Implementation of these feature selection techniques prior to modelling may be problematic when learning hierarchical models. The magnitude of variability-based metrics is influenced by the frequency of observations for which the signal-to-noise ratio is higher, such that the subset of features is more likely to capture broader relationships between larger sets of observations and less likely to capture relationships between smaller sets of observations. This can obscure important relationships within smaller sets of observations, as when evaluating a sub-group of samples in a hierarchical procedure. In addition, an appropriate choice of the number features to use for modeling is difficult to determine a-priori, and may be obscured by many factors, including: the number of subgroups, number of observations belonging to each subgroup, and the number of features distinguishing individual subgroups.

To overcome these challenges, *K2Taxonomer* produces a model fit for each partition independently, such that feature selection is only performed within the subgroup of observations being evaluated at a given step. In particular, at each recursive step the objective of partition estimation is to split the data based only on the dominant relationship between two subgroups. Since the selected features need only capture one relationship, a much smaller subset of features will be sufficient to discovering this partition. By default, *K2Taxonomer* uses the square root of the total number of features, which is used in a related albeit supervised learning method, random forests[157]. In doing

so, the percentage of filtered features is dependent on the total number of features. For example, if the data set consists of 1,000 or 10,000 features, *K2Taxonomer* will estimate partitions using 3.2% or 1.0% of the total number of features, respectively. The appropriateness of using the square root of the total number of features against fixed percentages is later assessed with simulation-based testing.

The *K2Taxonomer* package includes options to perform both SD and MAD based feature selection. In the case of group-level analysis, *K2Taxonomer* can perform F-statistic based feature selection based on the proportion of between group variability and within group variability implemented by the *limma* R package[93].

4.2.3 K2Taxonomer data partitioning

To estimate each partition, *K2Taxonomer*, performs feature-level bootstrap aggregation, similar to that of consensus clustering[158]. More specifically, each data partition represents the aggregation of a set of partitions estimated from perturbations of the original data set in which features have been sampled with replacement. Feature selection and K=2 clustering is independently performed within each perturbation-specific data set. The final partition estimate is calculated by aggregating the set of perturbation-specific partitions into a *cosine similarity* matrix (defined below), which further undergoes hierarchical clustering, followed by a K=2 tree cut.

K2Taxonomer package implements separate clustering methods tailored to analysis of either observation-level and group-level data input. For observation-level data, the perturbation-specific partitions are estimated via hierarchical clustering of the Euclidean distance matrix, followed by a K=2 tree cut. By default, Ward's agglomerative

method is performed at this step because it has been shown to generally perform well compared to other hierarchical methods[147]. For group-level data, perturbation-specific partitions are estimated via constrained K-means clustering[151]. This algorithm performs semi-supervised clustering, in which group-level information is included as a pairwise "must-link" constraint, preserving relationships between observations from the same group.

To assess the robustness of the partitioning of the aggregated results, hereby referred to as partition stability, as well as to facilitate interpretability, a cosine similarity matrix is computed, with each pairwise cosine similarity measurement functionally equivalent to the Pearson correlation of standardized variables.

Let an "item" denote a single observation or group, depending on whether observation- or group-level analysis is being performed, respectively. The cosine similarity of two items is a measure proportional to the number of times across perturbation iterations that the two items are assigned to the same group in the perturbation-specific dichotomous partitions. It takes its maximum/minimum value when the two items are always/never assigned to the same group.

If we represent with "-1" and "1" the assignments of an item to one or the other group in a dichotomous partition, we can then represent, and compare, the complete set of assignments of any two items across p perturbation-specific partitions as the vectors

$$\begin{aligned} X_i &= (x_{i1},...,x_{ip} \mid x_i \in \{-1,1\}) \\ X_j &= (x_{j1},...,x_{jp} \mid x_j \in \{-1,1\}) \end{aligned},$$

where X_i and X_j represent the i^{th} and j^{th} item, respectively.

We can then define the cosine similarity of X_i and X_j as

$$CS(X_i, X_j) = \frac{X_i \cdot X_j}{\| X_i \| \| X_j \|}.$$

where $X_i \cdot X_j$ represents the dot product, and $\| X_i \| \| X_j \|$ represents the product of the Euclidean norms of X_i and X_j. Next, we prove the equivalence of the cosine similarity and Pearson correlation of two assignment vectors. The cosine similarity can be rewritten as follows:

$$CS(X_i, X_j) = \frac{X_i \cdot X_j}{\| X_i \| \| X_j \|} = \frac{(p - Z) - Z}{p} = 1 - \frac{2Z}{p}.$$

In the above derivation, since X_i and X_j only take values in {-1,1}, their dot product, $X \cdot Y$, is equal to the difference between the number of iterations, *Z*, the two items are assigned to the same group, and the number of iterations, *p*-*Z*, the two items are assigned to different groups. Furthermore, the product of the Euclidean norms of X_i and X_j, $\| X_i \| \| X_j \|$, is equal to p.

Similarly, taking advantage of the relationship between Pearson correlation of standardized variables, $r()$, and Euclidean distance, $d()$, we have

$$r(X_i, X_j) = 1 - \frac{d^2(X_i, X_j)}{2p} = 1 - \frac{(X_i - X_j)^2}{2p} = 1 - \frac{4Z}{2p} = 1 - \frac{2Z}{p},$$

where we used the fact that the squared Euclidean distance of X_i and X_j, $d^2(X_i, X_j)$, is equal to $(X_i - X_j)^2$. Furthermore, for $X_i - X_j$, the difference between mismatched

adjacent elements is 2 and the difference between matched adjacent elements is 0. Therefore, $(X_i - X_j)^2 = 4Z$.

The function, $1 - \frac{2Z}{p}$, is the Hamann similarity index[159]. Consistent with Pearson correlation, the range of possible values for the Hamann similarity is between 1 and -1; these extremes occur if Z is equal to p and 0, respectively, indicating that X_i and X_j are either identical or fully dissimilar. Furthermore, if elements of X_i and X_j share 50% of their matching assignments, then $Z = \frac{p}{2}$ and the Hamann similarity is equal to 0, indicating a lack of a relationship between their perturbation-specific partition estimates. This is not true for the related phi similarity[160], another correlation metric for dichotomous variables, the calculation of which includes adjustment for the marginal distribution of X_i and X_j. In this case the marginal distribution of X_i and X_j is irrelevant because the elements of X_i and X_j are only meaningful in relation to their matching assignments.

4.2.4 K2Taxonomer partition stability

In order to assess the robustness of partition estimates, indicating the consistency of the perturbation-specific partition results, we developed a partition stability metric, which is calculated using the eigen-decomposition of the matrix of pairwise cosine similarities, Q, of dimension, N, the number of items. The eigen-decomposition of Q satisfies

$$QU = diag(\Lambda)U,$$

where U is the matrix of eigenvectors corresponding with Λ, the vector of rank ordered eigenvalues

$$\Lambda = (\lambda_1,...,\lambda_k,...,\lambda_N \mid \lambda_k > \lambda_{k+1}).$$

Each eigenvalue is proportional to the "variance explained" by each eigenvector, such that the cumulative sum of variance explained by the first k eigenvectors, v_k, is given by

$$v_k = \frac{\sum_{l=1}^{k} \lambda_l}{N}.$$

In this context, the variance explained by eigenvectors captures the consistency with which pairs of items received the same or different assignments across perturbation-specific partitions. Therefore, we can summarize this consistency by evaluating the difference between the variance explained by the eigenvectors of the estimated cosine matrix and the variance explained by these eigenvectors *if* there was no consistency across perturbation-specific partition assignments. We denote this deviation as the partition stability, PS, calculated as the maximum difference between v_k and $\frac{k}{N}$, the null value corresponding to all items being linearly independent,

$$PS = \max_k (v_k - \frac{k}{N}).$$

The possible values for the partition stability range between $1 - \frac{1}{N}$ and 0, with the former representing the case in which $\lambda_1 = N$, where every perturbation-specific partition

is identical, i.e., all values of the cosine matrix are either -1 or 1. Conversely, a partition stability of 0 represents the case when the perturbation-specific partition assignments are random, i.e., all values of the cosine matrix are close to 0. The maximum value for a given partition is dependent on the number of items in the partition, approaching 1 when N is large, and equal to 0.5 when $N = 2$. Using the *K2Taxonomer* package, partition stability can be used to set stopping criteria for creating new partitions, thereby serving as a way to control the number of terminal subgroups without prior knowledge.

Finally, partition stability is used as a heuristic for calculating branch heights in dendrogram creation of the *K2Taxonomer* output. For a series of m partitions resulting in a given partition, z_m, the branch height, h_m, is calculated as

$$h_m = log(N_m) + \sum_{l=1}^{m} log(PS_l) + c,$$

where c is a constant added to ensure that the minimum height for a node is equal to 1.

4.2.5 K2Taxonomer R package functionality

In addition to running the recursive partitioning algorithm, the *K2Taxonomer* R package provides functionalities for comprehensive annotation of the estimated subgroups, via subgroup-level statistical analyses, including: differential analysis, gene set enrichment analysis, and phenotypic variable testing. Differential analysis of gene expression is carried out using the *limma* R package, which is well-suited to the analysis of normally distributed data such as microarray gene expression, as well as log transformed and normalized RNAseq data[93]. Gene set enrichment analysis is carried out on a set of user-provided gene sets and implemented in two ways: over representation

analysis based on hypergeometric test, and differential analysis of single sample gene set projections scores based on the *GSVA* R package[161]. Finally, phenotypic variable testing is carried out on user-provided variables labeling individual observations or groups, supporting both continuous and categorical variables. Testing of association between continuous variable and taxonomy subgroups can be performed based on the parametric Student's t-test or the nonparametric Wilcoxon rank-sum test, while categorical testing is carried out using Fisher's exact test. All subgroup-level statistical analyses are corrected for multiple hypothesis testing based on the FDR procedure[162]. The full set of results are compiled into an interactive-web portal for exploration and visualization. Differential analyses comparisons are carried out at the partition-level, i.e., comparing only the two subgroups at a particular node. However, the web portal includes functionality for performing post-hoc differential analysis of any combination of user-selected subgroups. The implementation of *K2Taxonomer* for these analyses was based on *R (v3.6.0)*, *limma (v3.42.2)*, and *GSVA (v1.34.0)*.

4.2.6 Simulated data generation

K2Taxonomer's performance was assessed in terms of its capability to recapitulate induced hierarchical structure in simulated data, and it was further compared to Ward's agglomerative method as a term of reference. Hierarchical structured data was generated by assigning a mutually exclusive set of C labels to N observations. To define hierarchical relationships between labels, the set of C labels was recursively subdivided into intermediate subgroups until the final subgroups contained only a single label. To represent expected hierarchical structures in real data, we allowed for more than two

subgroups to be created at each subdivision. As such, neither *K2Taxonomer* or agglomerative methods are able to recapitulate this structure exactly.

After simulating the hierarchical relationships between labels, data was generated as follows. First, a set of 10,000 features was generated, each from a normal distribution, with mean, 0, and standard deviation, *BN*, denoting the background noise in the data. Second, for each subgroup of labels, we assigned a random set of features for which to add signal. For each feature assigned to a specific subgroup of labels, the value of added signal was sampled from a normal distribution with mean, 0, and standard deviation, 2. To ensure that the value of added signal was the same across all subgroup-specific observations, a feature was only allowed to be assigned to an individual label once. Considering that in real data we expect only a subset of features in a data set contain information of its subgroup structure, prior to assigning random sets of features to each subgroup, the full set of features was subsetted by a given percentage, *DS*, of the total.

4.2.7 Simulated data performance assessment

Following generation of each simulated data set, *K2Taxonomer* and Ward's method were run. Performance of each method was assessed by the relative similarity of the learned structured to the known hierarchy from which the data was generated using Baker's gamma correlation[163]. Baker's gamma correlation is a measure of Spearman correlation between two similarity matrices, where each similarity matrix is calculated from the number of shared partitions between each pair observations in a given hierarchy. Baker's gamma correlation ranges between -1 and 1, with 1 representing the case when two dendrograms induce identical hierarchies. This metric was chosen over *cophenetic*

distance because *cophenetic* pairwise similarities are calculated using branch heights and branch height is not applicable to the known hierarchy.

4.2.8 Observation-level analysis of simulated data

For observation-level analysis, each simulated data set included N=300 observations. We evaluated the following data generation parameters

- Number of labels (C): 5, 10, 20, 30
- Background Noise (BN): 0.5, 1, 2, 3
- Percent features with signal (DS): 2.5, 5, 10, 25, 50.

Especially when the value of DS is small, Ward's method is not expected to perform well when run on the full set of data. Therefore, Ward's method was evaluated on simulated data with the following variability-based prefiltering percentage levels, PF, of the data:

- Pre-filtering percentage level (PF): 5, 10, 25, 5, 1.

On the other hand, *K2Taxonomer* was always run on the square root of the total number of features. It should be emphasized that this design does not allow for a totally unbiased comparison of K2T and Ward, and it favors the latter, since in real settings we would not know the optimal PF to be used.

Standard deviation was used for variability-based feature selection, including partition-specific feature selection by *K2Taxonomer*. For each combination of these parameters, we generated 25 simulated data sets and ran *K2Taxonomer* and Ward's method. For each combination of parameters, the statistical significance of the difference

between the two methods' distributions of Baker's gamma correlation estimates was tested using Wilcox's Rank Sum tests. The resulting p-values were corrected for multiple hypotheses testing using the FDR procedure[162].

As noted, for each of these comparisons, we ran *K2Taxonomer* using the square root of the total number of features as the parameter for partition-specific feature filtering. In order to assess the validity of using this as the default, we performed additional simulations and compared *K2Taxonomer* performance at different partition-specific feature filtering levels, based on the set of percentages of the total number of features: 1%, 2%, 10%, and 20%. These were performed using the same combinations of parameters as the previous analysis and included 25 repetitions.

4.2.9 Group-level analysis of simulated data

For group-level analysis, each simulated data set included N=1000 observations. For data generation, we tested the following sets of data generation parameters

- Number of labels (C): 10, 25, 20, 30
- Background Noise (BN): 0.5, 1, 2, 3
- Percent features with signal (DS): 2.5, 5, 10, 25, 50.

As with observation-level analysis, we performed analyses on simulated data with the following variability-based prefiltering percentage levels, PF, of the data:

- Pre-filtering percentage level (PF): 5, 10, 25, 5, 1.

Linear model-based F-statistics were used for all variability-based feature selection scoring, including partition-specific feature selection by *K2Taxonomer*.

For running each method, the groups were defined by the labels to which each of the observations were assigned. Unlike *K2Taxonomer*, agglomerative methods are not devised to utilize group-level labels for unsupervised learning tasks. Therefore, Ward's method was applied using the Z-score mean value of each group, generated by the same linear model used for feature selection.

4.2.10 METABRIC breast cancer primary tumor bulk gene expression processing

METABRIC breast cancer primary tumor Illumina HT-12 v3 microarray bulk gene expression data was obtained from the *CBioPortal*, https://www.cbioportal.org/datasets[164,165]. The data set includes normalized expression values for 24,360 genes and Pam50 cancer subtype[166] estimations for individual primary breast cancer tumor samples across 1,974 female patients. Additional clinical variables considered for this analysis, included: patient age at diagnosis, survival status, ER-status, PR-status, and HER2-status.

4.2.11 TCGA breast cancer primary tumor bulk gene expression processing

The *cancer genome atlas (TCGA)* breast cancer (BRCA) primary tumor bulk RNAseq data was obtained from *Genomic Data Commons (GDC)*, https://gdc.cancer.gov/access-data/[16]. The data set includes raw gene expression counts for 36,812 genes and Pam50 cancer subtype[166] estimations for individual primary breast

cancer tumor samples across 973 female patients. Additional utilized clinical variables, included: patient age at diagnosis, ER-status, PR-status, and HER2-status.

Raw counts were normalized by the trimmed mean of M-values (TMM) method and log-normalized using *edgeR (v3.28.1)* R package[167] and genes with fewer than 2 reads in more than 90% of samples were removed, resulting in 25,729 genes in the processed data set.

4.2.12 Performance assessment using breast cancer primary tumor bulk gene expression data

K2Taxonomer was evaluated for its ability to recover the Pam50 subtypes[166], as well ER-, PR-, and HER2-status, and the aggregate three-gene genotype of ER-, PR-, and HER2-status, in the *TCGA* and *METABRIC* breast cancer datasets, independently. *K2Taxonomer* was also compared to two agglomerative clustering algorithms, Ward's and average. These specific methods were chosen because they have been previously shown to outperform other common agglomerative methods[147,168]. Given the sensitivity of hierarchical clustering to the level of feature filtering, analyses included individual runs on four filtered data subsets of the total number of features: 100%, 25%, 10%, and 5%, while *K2Taxonomer* was only run on the full set of 100% of the total number of features. This should be kept in mind when comparing performances. Since the best-performing pre-filtering level is not known a-priori, and it is in general dataset dependent. For every pre-filtering level, the median absolute deviation (MAD) score was used for feature selection and Euclidean distance was used to estimate observation-level distance. Performance was assessed as the entropy of each of the phenotypes (e.g., PAM50 labels),

induced by the inferred sample sub-grouping, with lower entropy indicating "purer" subgroups, hence better performance[169]. The different methods were evaluated and compared by the relative decrease in entropy as the number of mutually exclusive clusters, K, increased from 2 to 8 based on tree cuts of the dendrograms produced by each model.

4.2.13 Healthy airway tissue scRNAseq gene expression analysis

Publicly available scRNAseq data of normalized, batch corrected, and log-transformed gene expression estimates from airway tissue of healthy subjects was obtained from the *UCSC Cell Browser* portal, published as a supplement to the original manuscript for which these data were used, https://www.genomique.eu/cellbrowser/HCA/?ds=HCA_airway_epithelium[152]. This data set includes expression estimates for 18,417 genes and 77,969 individual cells from 35 samples across 10 subjects. Multiple samples taken from individual subjects were collected from distinct locations of the human airway including: nasal biopsies, nasal brushings, tracheal biopsies, intermediate bronchial biopsies, and distal brushings. In addition, this data set included cell type estimations for each of the 77,969 cells, and comprised 28 estimated cell types in total. The methods of data processing, as well as distributions of subject-level sample identities and estimated cell types can be found in the original publication[152].

K2Taxonomer partitioning of these 28 estimated cell types was evaluated against the known relationships among the included cell types, and was compared to the partitioning obtained by two agglomerative methods, Ward's and average. Cell type-level

data processing and feature selection were performed consistent with the results of group-level analysis of simulated data.

4.2.14 Breast TIL scRNAseq gene expression analysis

Publicly available scRNAseq gene expression of raw counts from breast cancer TILs of two TNBC patients was obtained from *GEO*, accession number GSE110938[153]. The data was processed in accordance with the original manuscript[153], recapitulating the reported 5,759 individual cells, 4,844 and 915 from either sample, with 15,623 genes passing QC criteria, selection of 1,675 highly variable genes, and 10 latent variables estimated by *ZINB-WaVE* (v1.8.0)[170]. To enable exploration of the data at finer resolution, clustering of the latent variables with Seurat (v1.3.4) was modified by setting the "resolution" argument of "FindClusters()" to 1.1, rather than the default, 0.8[171]. This resulted in 13 estimated cell clusters. Of the 10 cell clusters reported in the original manuscript, two cell clusters, "CD4+ FOXP3+" and "CD4+ IL7R+", were further split into three and two individual clusters, respectively.

K2Taxonomer partitioning of these 13 estimated cell subtypes was performed on the normalized count matrix estimated by *ZINB-WaVE*. According to the developers, *ZINB-WaVE*, normalized count estimates are not recommended for differential analysis[170], hence differential analysis was performed based on drop-out imputed and batch-corrected normalized counts estimated using the *bayNorm (v1.4.14)* R package[172]. Pathway-level analysis was carried out using Reactome gene sets downloaded from *mSigDB (v7.0)*[173]. Signatures of up-regulated genes were derived from each subgroup based on their FDR

corrected p-value (FDR < 1e-10) and minimum subgroup-specific expression, (mean[log2 counts] > 0.5), then restricted to a maximum of 50 genes.

To validate the clinical relevance of signatures of TILs subgroups derived by *K2Taxonomer* we performed survival analysis based on gene signature projection scores, as well as on selected genes in the *METABRIC* breast cancer primary tumor gene expression data set. Gene set projection was carried out using *GSVA*[161]. Multivariate Survival analysis was performed using Cox proportional hazards tests. All models included age and Pam50 subtype as covariates. To account for possible confounding effects of inflammation and proliferation, we generated separate patient-level activity scores for each, using a gene set projections of published signatures of deleterious breast cancer inflammation markers[174] and breast cancer proliferation[175].

4.3 Results

4.3.1 K2Taxonomer discovers hierarchical taxonomies on simulated data

We first evaluated *K2Taxonomer*'s capability to recapitulate hierarchical relationships induced in simulated data, as measured by the Baker's gamma coefficient estimate of similarity between two dendrograms based on the number of partitions separating each pair of items in the data[163]. We evaluated the method's performance both on data where the analysis end-points were single samples (*observation-level*), and groups of samples (*group-level*). The latter correspond to scenarios where the goal is to define a taxonomy over sample groups, such as cell types in single cell experiments, or chemical perturbations profiled in multiple replicates. As a term of reference, we compared *K2Taxonomer*'s performance to Ward's agglomerative method.

For observation-level analysis *K2Taxonomer* significantly outperformed Ward's method in 221 out of the 400 combinations of parameters tested (FDR < 0.05), while Ward's performed better for 17 combinations (Figure 4.1A). For instances in which *K2Taxonomer* method outperformed Ward's method, the differences between Baker's gamma estimates were generally small, such that the median difference in performance was 0.14, compared to 0.04 for instances in which Ward's method performed better. In general, *K2Taxonomer* significantly outperformed Ward's method when the background noise, number of terminal groups, and percent features with signal increased. Remarkably, for group-level analysis *K2Taxonomer* outperformed Ward's method for all 400 combinations of variables tested (Figure 4.1B).

Using the square root of the total number of features as the partition-specific feature filtering parameter for running *K2Taxonomer* demonstrated stable performance. When compared to selecting a fixed percentage of the total number of features (Figure A.17), the square root outperformed larger percentages when the number of features was large, and outperformed smaller percentages when the number of features was small.

4.3.2 K2Taxonomer accurately sorts breast cancer subtypes without pre-filtering of features

We evaluated *K2Taxonomer*'s ability to sort Pam50 subtypes, ER-status, PR-status, and HER2-status from bulk gene expression data from METABRIC and the TCGA BRCA bulk gene expression data, separately. A fourth variable, defined by the Cartesian product of ER-status, PR-status, and HER2-status was also assessed. Performance was assessed in terms of decrease in entropy as the number of cluster

estimates, K, increased from 2 to 8 (Figure 4.2A-B). We also compared *K2T*'s performance to two agglomerative clustering methods, Ward's and average. Since standard hierarchical clustering is sensitive to the level of feature filtering, the comparison was repeated for multiple pre-filtering levels.

In general, *K2Taxonomer* accurately segregated the known sub-types and phenotypes, and performed as well or better than either method (Figure 4.2B). When applied to the *METABRIC* data, *K2Taxonomer* analysis yielded the lowest entropy score compared to all other methods for K=3 and higher with few exceptions. Other methods produced similar entropy measurements at selected higher levels of K. For example, Ward's method resulted in similar entropy scores for Pam50 subtypes and HER2-status at K=4 and K=5, respectively, but for different pre-filtering levels, 5% and 100%, respectively. When applied to the TCGA BRCA data, the difference in performance was less pronounced. *K2Taxonomer* resulted in the lowest entropy score for Pam50 scores, genotype, ER-, and PR-status for K=4. Ward's method at 5% pre-filtering level produced the smallest entropy score for HER2-status for K=4.

4.3.3 K2Taxonomer identifies subgroups of shared progenitors and epithelial cells from healthy airway scRNAseq cell clusters

To assess the capability of *K2Taxonomer* to recapitulate biologically relevant subgroupings of cell types estimated from scRNAseq data, we ran group-level analysis using 29 cell types estimations assigned to 77,969 cells of airway tissue from 10 healthy subjects (Figure 4.3A-B)[24]. *K2Taxonomer* was remarkably accurate in capturing the higher order organization of the 28 cell types. The first partition separated all epithelial

cell subtypes from non-epithelial cell types (Figure 4.3B). Further partitioning of the 17 epithelial cell subtypes yielded two subgroups, characterized by shared morphology, labelled as "Multiciliated" and "Submucosal". The former group was comprised of differentiated multiciliated cells and their precursor, deuterosomal cells[152]. Further partitioning of the 11 non-epithelial cell types yielded three subgroups characterized by shared progenitor cells: myeloid[176], lymphoid[176], and mesenchymal stem cells[177]. In contrast, agglomerative hierarchical clustering of these cell clusters, even if evaluated at multiple F-statistic-based pre-filtering levels, yielded significantly different results poorly reflective of the known taxonomic cell type organization (Figure 4.3C, Figure A.18). While the mesenchymal stem cell subgroup, comprised of fibroblasts, smooth muscle, and pericytes, was identified by Ward's method, and while there were other instances of concordant subgroups, none of these consisted of more than two cell types, and none of the other five subgroups were identified by either methods.

4.3.4 K2Taxonomer identifies subgroups of TILs characterized by differential regulation of TNF signaling, translation, and mitotic activity from BRCA tumor scRNAseq cell clusters

We performed *K2Taxonomer* analysis on scRNAseq data of 13 TIL cell clusters. These 13 cell clusters reflect further subdivisions of the 10 cell types reported in the original publication for these data[153], which was done by reproducing the reported methods[153] with the exception of including a higher resolution parameter when performing clustering with Seuat[141].

The results of *K2Taxonomer* partitioning and annotation of breast cancer cell clusters estimated from scRNAseq data is shown in Figure 4.4A. Biologically informative subgroups, characterized by strongly significant differential expression of gene expression and sample-level pathway enrichment are highlighted and labeled within each boxed sub-dendrogram. Three distinct multi-cell subgroups emerged, labeled as: "Trm All", "CD4+ *CCL5*-", and "Translation+", characterized by consistent up-regulation of PD-1 signaling (Reactome PD-1 signaling, FDR = 1.1e-241), translation (Reactome eukaryotic translation initiation, FDR = 5.6e-137), and TNF signaling (Reactome TNFS bind their physiological receptors, FDR ~ 0.00), respectively (Figure 4.4B). "Trm All" *and* "Treg" subgroups each included a mitotic cell subgroup characterized by high cell cycle activity (Figure 4.4B). Furthermore, the "CD4+ *CCL5*-" subgroup, comprised of the "CD4+ CXCL13+" cell cluster and "Treg" subgroup, is characterized by consistent down-regulation of *CCL5* (FDR ~ 0.00) and up-regulation of *TNFRSF4* (FDR ~ 0.00) (Figure 4.4C-D). Furthermore, additional up-regulation of *TNFRSF4* (FDR = 1.1e-7) and *RGS1* (FDR = 4.3e-55) distinguish non-mitotic "Treg" subgroups (Figure 4.4C). Gene-level markers of the "Translation+" subgroup included numerous ribosomal proteins, epitomized by up-regulation of *RPS27* (FDR = 9.6e-246) (Figure 4.4C-D).

4.3.5 Confounding effects of inflammation and proliferation on association between TIL activity and patient survival

To assess the clinical relevance of *K2Taxonomer* annotation of single-cell immune cell subgroups, we performed survival analysis, via Cox proportional hazards

testing, modeling the relationship between *K2Taxonomer* subgroup gene signature scores and patient survival in the METABRIC breast cancer bulk gene expression data set. For these models, we examined two possible sources of confounding factors. First, inflammation has a well-described paradoxical role in breast cancer progression[178], such that the content of different subpopulations of lymphocytes have been associated with both better and worse prognosis[179]. Given the physiology similarities between different lymphocyte subtypes[180], we hypothesized that expression patterns associated with tumor promoting inflammation could mask those tumor suppressing TILs subsets. Second, we hypothesized that the signatures of the two mitotic T cell subgroups were similar enough to that of tumor-specific proliferation to result in a spurious association between T cell mitosis and worse prognosis. To assess and correct for these confounding effects multivariate survival models were run without and with inclusion of inflammation and proliferation scores as individual covariates. These patient-level scores were estimated by projecting published signatures of inflammation[174] and proliferation[175], each of which associated with poor prognosis in breast cancer.

The results of each of these analyses is summarized in Figure 4.4A. Controlling for inflammation and proliferation scores increased the overall significance of association between subgroup-driven signatures of TILs and improved survival (hazard ratio < 1, FDR < 0.05). Furthermore, signatures of two cell subgroups, “CD8+ mit. Trm” and “Treg mit.”, characterized by increased cell cycle activity (Figure 4.4B), were associated with worse patient survival in models ignoring inflammation and proliferation scores, but were subsequently statistically insignificant in models including these covariates, likely

reflecting the effect of confounding by proliferation activity (Figure 4.4A). This is further illustrated in Figure 4.4E, which shows the 95% confidence intervals of hazard ratios of "marginal" inflammation and proliferation models (left-most), as well as the confidence intervals of hazard ratios of select subgroups of cell subtypes, unadjusted and adjusted for inflammation and proliferation. Controlling for inflammation and proliferation allowed us to disentangle the contribution to survival of different components. For example, in the "CD8+ mit. Trm" subgroup, we observed that the "CD8+ mit. Trm" signature score was highly associated with worse patient survival in the unadjusted model, but the association became insignificant in the full model adjusted for proliferation and inflammation. On the other hand, there were instances where the hazard ratio achieved or improved significance (i.e., patient survival was significantly better) only after controlling for inflammation and proliferation in the full adjusted model, as observed in the "Trm All" subgroup and, to a lesser extent, in the "Translation+" subgroup (Figure 4.4E).

4.3.6 High expression of TNFRSF4, a marker for Treg cell activity is associated with worse survival, when adjusting for CCL5 expression.

TNFRSF4 and *CCL5* were found to be the top two markers constitutively up- and down-regulated, respectively, within Treg subgroups, with *TNFRSF4* the top marker further discriminating between the two non-mitotic Treg subgroups, *Treg TNFRSF4+* and *Treg RGS1+* (Figure 4.4A, C-D). Furthermore, their expression was highly correlated in the METABRIC data set (rho = 0.66, p-value = 3.2e-245) (Figure 4.4F), supporting a pattern of co-expression within TIL microenvironments. To assess whether *TNFRSF4* and *CCL5* expression levels could serve as markers for immunosuppressive

activity of Treg cells, we performed survival analysis of each gene modelled separately and in a combined model (Figure 4.4G). When modelled separately, the expression of *TNFRSF4* is not associated with patient survival (P-Value = 0.32), while *CCL5* is associated with better patient prognosis (P-Value = 1.94E-4). However, in the combined model both genes are associated with patient survival, with *TNFRSF4* associated with worse patient survival (P-Value = 0.015).

4.3.7 Up-regulation of specific translation genes characterizes a subgroup of TILs and is associated with better survival prognosis, independent of inflammation activity

The "Translation+" subgroup was a notable instance for which the subgroup-specific signature projection was associated with better patient survival, regardless of adjustment for inflammation and proliferation score (Figure 4.4A, E). In order to assess the extent to which up-regulation of translation specific genes in this subgroup associated with better patient prognosis, we ran separate survival analysis for each of the 112 genes from the Reactome Eukaryotic Translation Initiation gene set, which were shared between the single-cell BRCA gene set and METABRIC data set. Of the 112 genes, 61 were up-regulated in the "Translation+" subgroup (FDR < 1E-5), including 26 genes within the top 50 marker "Translation+" subgroup signatures (Figure 4.4H-I). The test statistics derived from single gene Cox proportional hazards models were negatively correlated with the corresponding genes' test statistics of their up-regulation in the "Translation+" subgroup (rho = -0.23, p-value = 0.014) (Figure 4.4I). Furthermore, seven of the 112 genes were associated with better patient survival (FDR < 0.1). All of these genes were significantly up-regulated in the "Translation+" subgroup. Of these seven

genes, *RPL36A* had the minimum "Translation+" subgroup associated test statistic (FDR = 1.34e-25) and two (*RPS28* and *RPS27*) were members of the top 50 markers, comprising the "Translation+" subgroup signature. *RPS27* was the top translation gene associated with the "Translation+" subgroup (FDR = 9.6e-246).

4.4 Discussion

In this chapter we presented extensive assessment and practical applications of *K2Taxonomer*, a novel unsupervised recursive partitioning algorithm for taxonomy discovery in both bulk and single-cell high-throughput transcriptomic profiles. An important distinctive feature of the algorithm is that each partition is estimated based on a feature set selected to be most discriminatory within that partition, thus permitting the use of large sets of features to be used as input, without pre-filtering or dimensionality reduction approaches. Additionally, in an effort to minimize generalization error, each partition is based on an ensemble[181] of partition estimates from repeated perturbations of the data. Adoption of an ensemble approach also makes it possible to compute a stability measure for each partition, which can be used to assess the robustness of each partition, as well as a stopping criterion for limiting the number of subgroup estimates.

As we have shown in its multiple applications, *K2Taxonomer* may be applied in a fully unsupervised mode to partition individual level data, or it can take group-level labels as input to estimate inter-group relationships among the known groups. In the latter scenario, partition estimates are based on the constrained K-means algorithm[151], which estimates clusters at the level of known group labels. This approach is perfectly suited to the downstream analysis of scRNAseq data, following the estimation of mutually

exclusive cell types using scRNAseq clustering methods such as Seurat[141] or Scran[142]. To our knowledge, there is no comparable method, in which subgroup relationships are estimated using group-level information. By preserving the single observation information within each group, and by thus being able to tailor the feature set to each of the groups, we expect our approach to outperform methods in which group-level information is summarized into single statistical measures. This conclusion is supported by our simulation analysis, where *K2Taxonomer* was shown to significantly outperform Ward's agglomerative method based on group-level test statistics. Even when adopted for observation-level analysis, where inference was performed on the full set of individual observations, *K2Taxonomer* was still shown to significantly outperform standard agglomerative methods, on both simulated and real data, although not to as large an extent.

In our analysis of healthy airway cell types' annotation[152], we employed *K2Taxonomer* to (re)discover subgroups of cell types characterized by shared lineage. Remarkably, our analysis accurately recapitulated the known taxonomic structure relating the different cell types to an extent not matched by the other methods evaluated. This example illustrates a prototypical use of the tool: in those cases where a data set and its associated cell type estimations are publicly available, *K2Taxonomer* facilitates their immediate repurposing for additional insight and discovery.

It is important to emphasize that in many data sets shared lineage relationships may be non-existent or obscured by phenotype-driven inter-group transcriptional relationships. For example, in primary tumor samples, intratumor heterogeneity includes

phenotypic convergence of ancestrally divergent cell populations. Given that bulk gene expression captures average expression across all cells, identifying dominant transcriptional programs driving phenotypic similarities between subgroups of cell populations offers additional insight to deconvolute the cellular microenvironment of these samples beyond their individual transcriptional signatures.

Phenotypic convergence of cells of disparate lineages is exemplified by subpopulations of CD8+ and CD4+ T cells, each of which exist in various functional states as naïve, effector, and memory subpopulations[182]. Concordant subpopulations of CD8+ and CD4+ T cells share transcriptional signatures that may outweigh those arising from their shared lineage. For example, both CD8+ Trm and CD4+ Trm cells have been reported to express surface molecules, CD69 and CD11a[183]. As well illustrated in our analysis of breast cancer TILs, CD8+ Trm and CD4+ Trm cells segregated into a common subgroup, demonstrating the relative dominance of their shared transcriptional activity. Projection of the expression signature of Trm cell subgroups was associated with better survival in the METABRIC data set. Past studies focusing on CD8+ Trm cell markers have reported similar findings[153,184].

Unlike Trm cells, the presence of immune suppressing Treg cells in the micro-environment has been associated with poor prognosis in breast cancer[185–187]. After identifying *TNFRSF4* as heterogeneously expressed across the Treg cell subgroup, we showed that *TNFRSF4* expression was associated with worse patient survival in the METABRIC data set when adjusted for *CCL5* expression, which was down-regulated among all Treg cells. This supports previous findings that *TNFRSF4*, also known as

OX40, is a marker of high Treg cell immunosuppressive activity[188,189]. The high level of co-expression of *TNFRSF4* and *CCL5* in the METABRIC data set suggests that either gene is associated with immune infiltration in breast cancer tumors. Additionally, this provides a resolution as to why projections of the signature of the Treg cell subgroup was associated with better patient survival, while the signature of Treg cell subset, characterized by high *TNFRSF4* expression was not.

Finally, *K2Taxonomer* identified a diverse subgroup of breast cancer TILs characterized by consistent up-regulation of translational genes. Increased ribosomal biogenesis has been previously implicated in increased tumorigenesis[190–193], but has only recently been described for its roles in T cell activation[154] and expansion[155]. Unlike the majority of other subgroups, the signature of the T cell subgroup overexpressing translational machinery genes was associated with better patient survival in METABRIC patients in cox proportional hazards models regardless of adjustments for inflammation[174] and proliferation[175] signatures. Furthermore, the association of the expression of specific translational genes with better patient survival was significantly correlated to their overexpression in this T cell subgroup. These results suggest that overexpression of these T cell specific translational genes are not masked by tumor specific gene expression and are therefore indicative of CD4+ and CD8+ T cell tumor infiltration.

In summary, *K2Taxonomer* demonstrates a remarkable ability to discover biologically relevant taxonomies when applied to the analysis of both bulk gene expression and scRNAseq data, and to outperform standard agglomerative methods. In multiple practical applications, we showcased the versatility of *K2Taxonomer* to analyze

scRNAseq data toward the characterization of genes and pathways distinguishing specific subgroups, thereby generating hypotheses that were then in-silico validated in independent bulk gene expression data. As noted, while we here focused on the analysis of transcriptomics data, the proposed approach is equally applicable to other bulk and single cell ‘omics’ data, such as those generated by high-throughput proteomics and metabolomics assays.

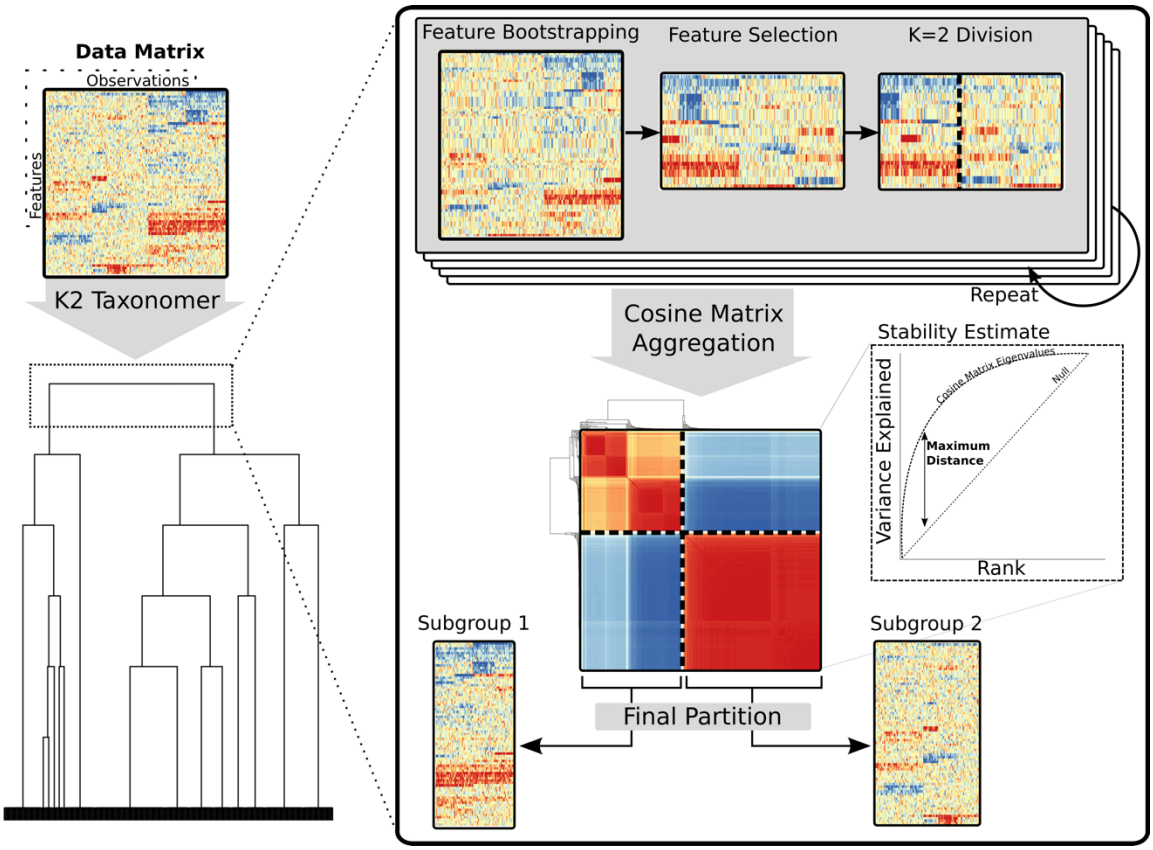


Illustration 4.1: Schematic of the *K2Taxonomer* recursive partitioning algorithm

For each partition, *K2Taxonomer* generates an ensemble of K=2 estimates from the feature bootstrapped data followed by variability-based feature selection. This ensemble is aggregated to a cosine matrix followed by hierarchical clustering and tree cutting. A stability estimate, indicative of the consistency of K=2 estimates, is calculated based on an eigendecomposition of the cosine matrix. *See methods for a more thorough description of the elements of this procedure.*

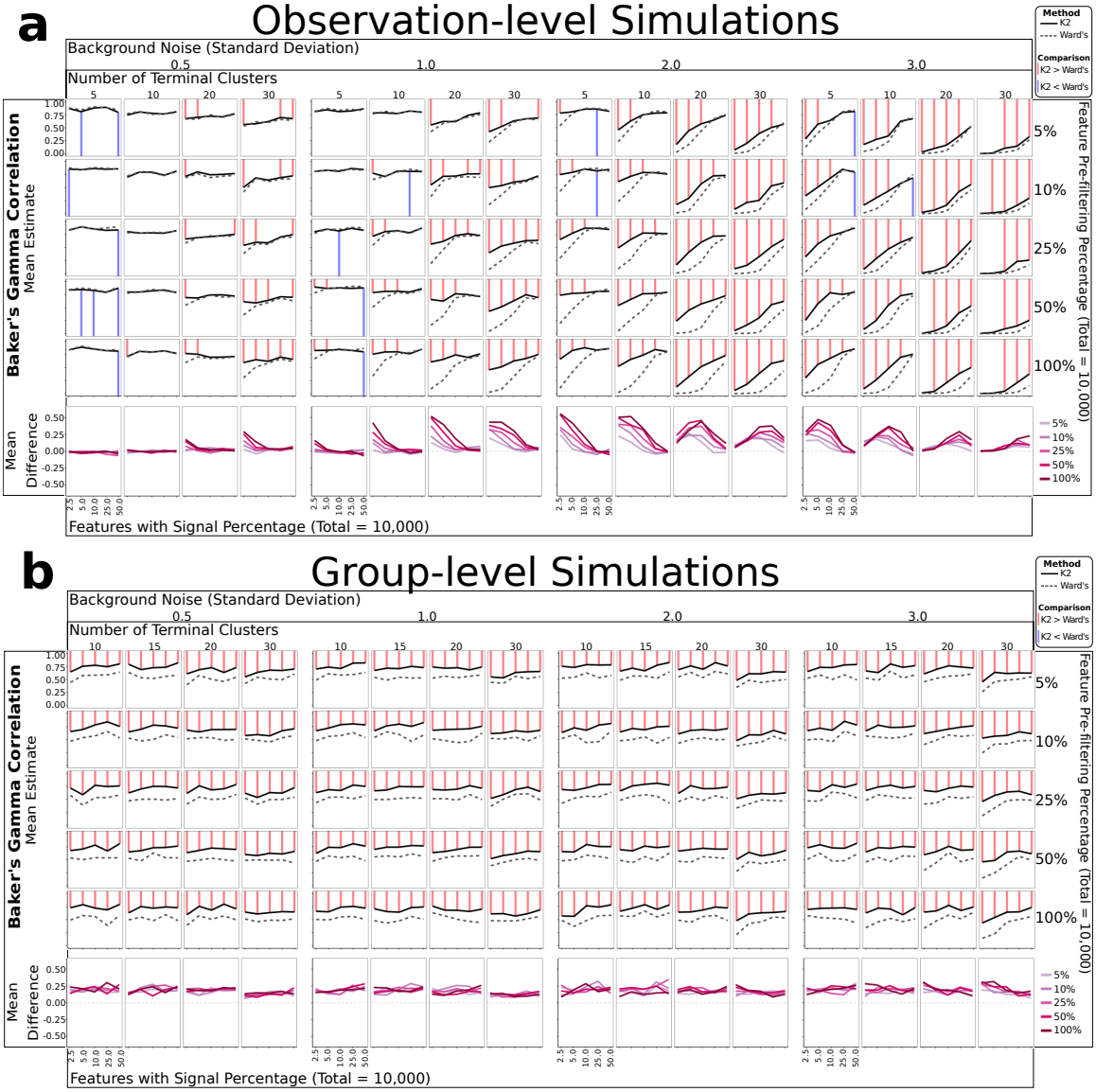


Figure 4.1: Simulation-based performance assessment of *K2Taxonomer* and Ward's agglomerative method

Mean Baker's gamma correlation estimates measuring the similarity of either *K2Taxonomer* and Ward's agglomerative method estimates to the true hierarchy from which the simulated data was generated. Each combination of parameters was simulated 25 times. The red and blue lines are indicative of statistically significant differences between the correlation estimates (FDR < 0.05) based on a Wilcoxon signed-rank test.

A) Observation-level analyses with 300 observations and 10,000 features.
B) Cohort-level analyses with 1,000 observations and 10,000 features.

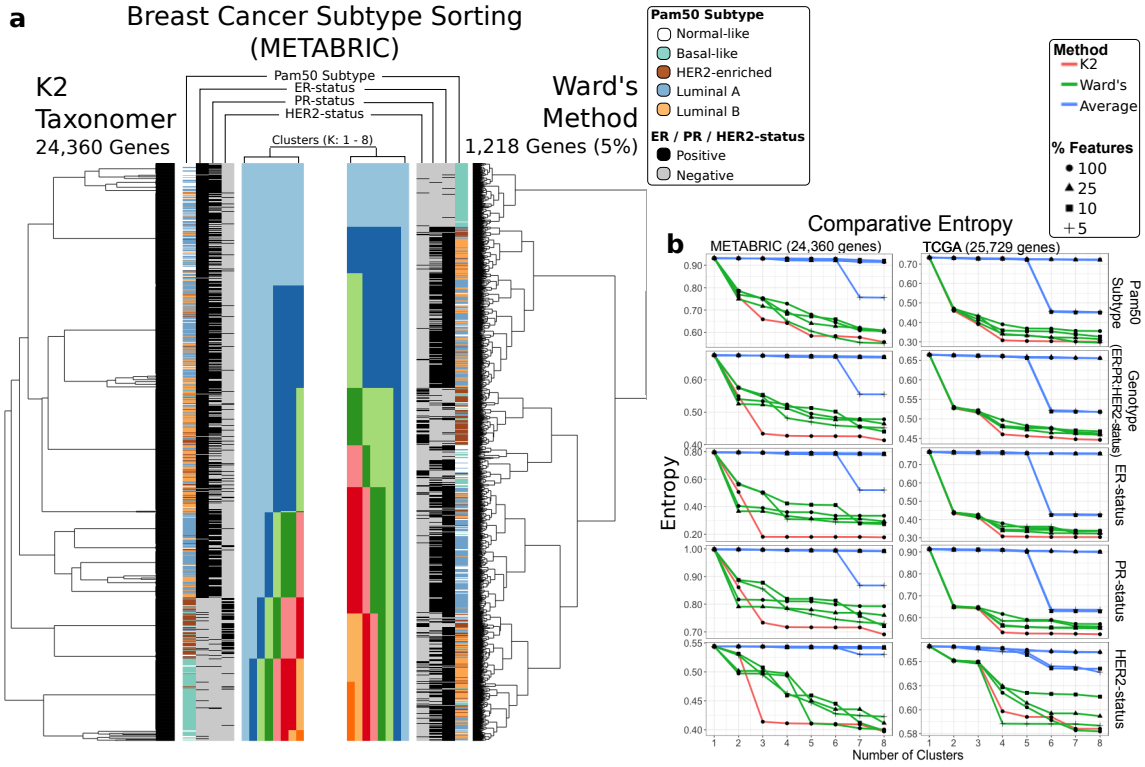


Figure 4.2: Breast cancer subtyping performance assessment of bulk gene expression data

Comparison of sorting of breast cancer Pam50 subtypes and genotypes (ER-, PR-, and ER-status) for two bulk gene expression data sets, *METABRIC* and *TCGA*.An aggregate, three gene genotype status was also included by combining the individual genotypes. Performance was assessed based on reduction of entropy as the number cluster estimate increased based on tree cutting. *K2Taxonomer* was only run on the full set of features, while either agglomerative method, average and Ward's, were run on three additional subset of the data.

A) Illustration of the results generated by *K2Taxonomer* and Ward's method for the *METABRIC* dataset. These results reflect Ward's method run on 5% of the total number of features, which demonstrated the best performance among agglomerative methods.
B) Entropy measurements for each method as K increased across the *METABRIC* and *TCGA* data sets.

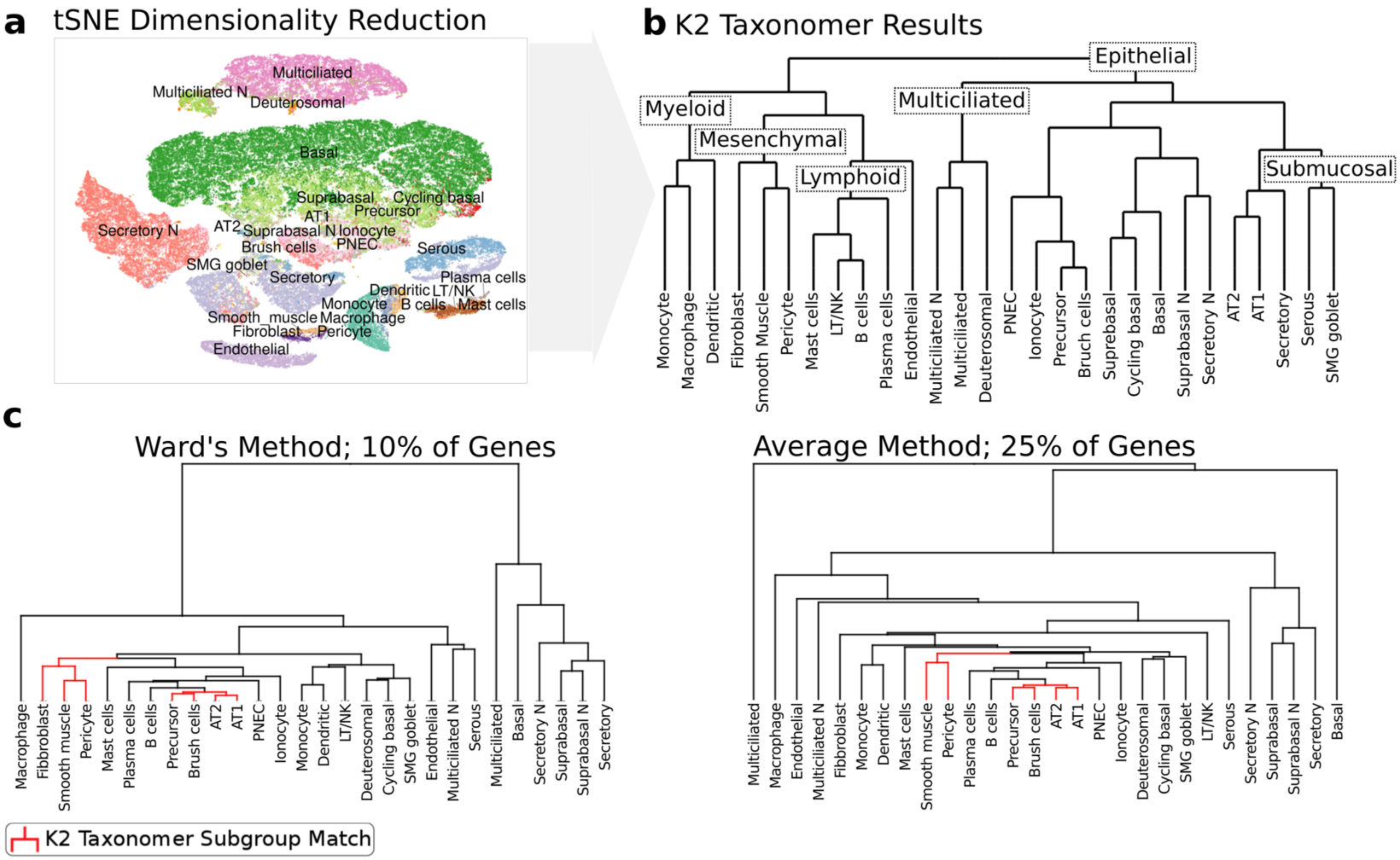

a
tSNE Dimensionality Reduction
Multiciliated
Multiciliated N
Deuterosomal
Basal
Suprabasal
Cycling basal
Precursor
AT1
AT2
Secretory N
Suprabasal N
Ionocyte
Brush cells
PNEC
Serous
SMG goblet
Secretory
Plasma cells
Dendritic
LT/NK
Monocyte
B cells
Mast cells
Smooth_muscle
Macrophage
Fibroblast
Pericyte
Endothelial
b
K2 Taxonomer Results
Epithelial
Myeloid
Multiciliated
Mesenchymal
Lymphoid
Submucosal
Monocyte
Macrophage
Dendritic
Fibroblast
Smooth Muscle
Pericyte
Mast cells
LT/NK
B cells
Plasma cells
Endothelial
Multiciliated N
Multiciliated
Deuterosomal
PNEC
Ionocyte
Precursor
Bruch cells
Suprebasal
Cycling basal
Basal
Suprabasal N
Secretory N
AT2
AT1
Secretory
Serous
SMG goblet
c
Ward's Method; 10% of Genes
Macrophage
Fibroblast
Smooth muscle
Pericyte
Mast cells
Plasma cells
B cells
Precursor
Brush cells
AT2
AT1
PNEC
Ionocyte
Monocyte
Dendritic
LT/NK
Deuterosomal
Cycling basal
SMG goblet
Endothelial
Multiciliated N
Serous
Multiciliated
Basal
Secretory N
Suprabasal
Suprabasal N
Secretory
Average Method; 25% of Genes
Multiciliated
Macrophage
Endothelial
Multiciliated N
Fibroblast
Monocyte
Dendritic
Mast cells
Smooth muscle
Pericyte
Plasma cells
B cells
Precursor
Brush cells
AT2
AT1
PNEC
Ionocyte
Deuterosomal
Cycling basal
SMG goblet
LT/NK
Serous
Secretory N
Suprabasal
Suprabasal N
Secretory
Basal
K2 Taxonomer Subgroup Match

Figure 4.3: Subgrouping of healthy airway cell types from scRNAseq data

A) tSNE dimensionality reduction of healthy airway scRNAseq data with labels for 28 cell types annotated by [152]. Note cell types labelled ending in "N" indicate those which only included cells from nasal samples.
B) *K2Taxonomer* results with six identified lineage subgroups.
C) Ward's (left) and average (right) agglomerative clustering results for selected analyses performed on different subsets of the total number of features. The results for additional feature subsets: 100%, 25%, 10%, and 5%, are shown in Figure A.18. These results were chosen as the best sorting of these data based on lineage for each respective method.

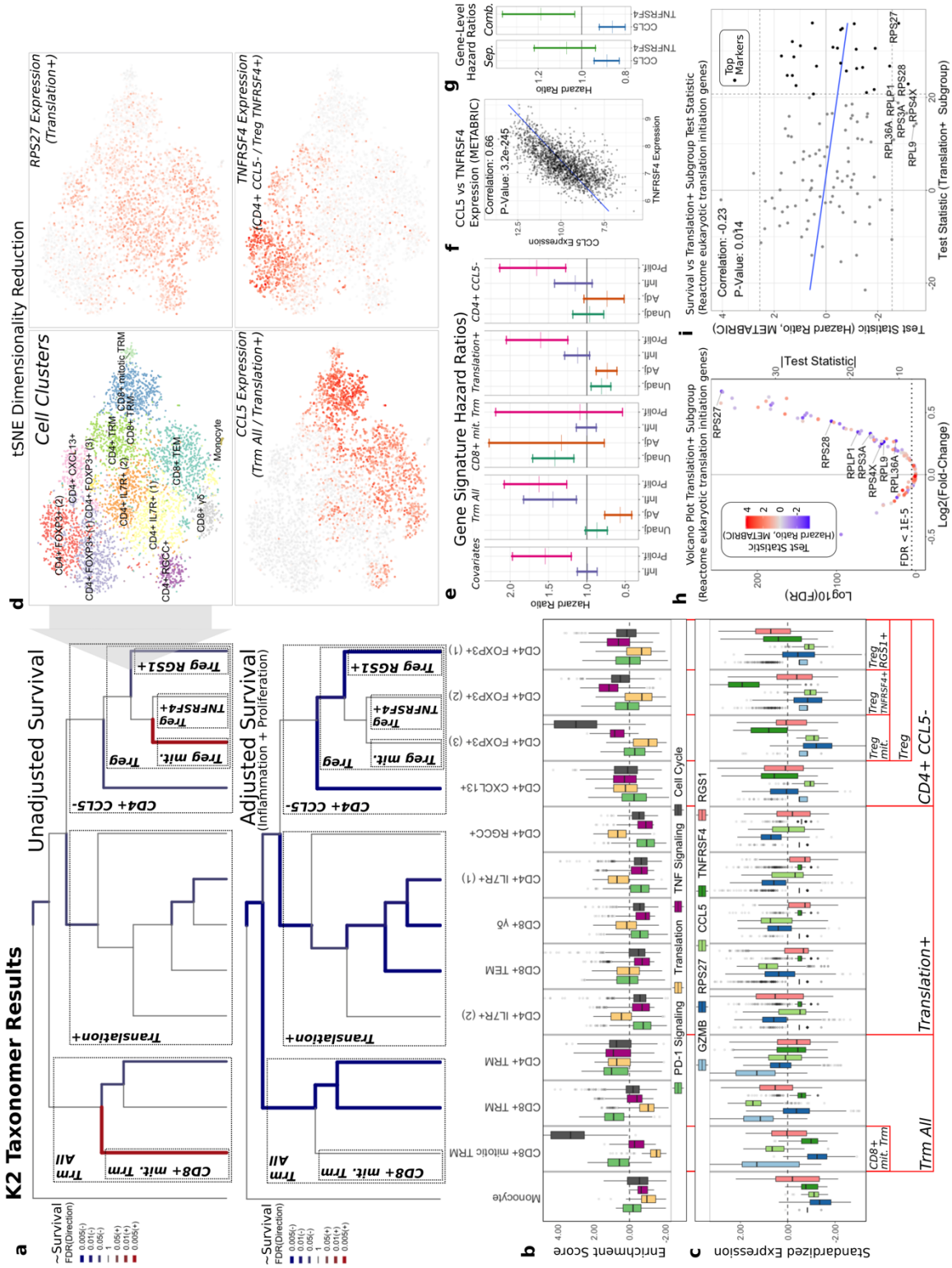

a
K2 Taxonomer Results
Unadjusted Survival
Adjusted Survival
(Inflammation + Proliferation)
~Survival
FDR(Direction)
CD8+ mit. Trm
Trm All
Translation+
CD4+ CCL5-
Treg
Treg mit.
Treg TNFRSF4+
Treg RGS1+
b
Enrichment Score
Monocyte
CD8+ mitotic TRM
CD8+ TRM
CD4+ TRM
CD4+ IL7R+ (2)
CD8+ TEM
CD8+ γδ
CD4+ IL7R+ (1)
CD4+ RGCC+
CD4+ CXCL13+
CD4+ FOXP3+ (3)
CD4+ FOXP3+ (2)
CD4+ FOXP3+ (1)
PD-1 Signaling
Translation
TNF Signaling
Cell Cycle
c
Standardized Expression
GZMB
RPS27
CCL5
TNFRSF4
RGS1
CD8+ mit. Trm
Trm All
Translation+
CD4+ CCL5-
Treg mit.
Treg TNFRSF4+
Treg RGS1+
d
tSNE Dimensionality Reduction
Cell Clusters
RPS27 Expression (Translation+)
TNFRSF4 Expression (CD4+ CCL5- / Treg TNFRSF4+)
CCL5 Expression (Trm All / Translation+)
e
Gene Signature Hazard Ratios)
Covariates
Trm All
CD8+ mit. Trm Translation+
CD4+ CCL5-
Hazard Ratio
Unadj.
Adj.
Inf.
Prolif.
f
CCL5 vs TNFRSF4 Expression (METABRIC)
Correlation: 0.66
P-Value: 3.2e-245
CCL5 Expression
TNFRSF4 Expression
g
Gene-Level Hazard Ratios
Sep.
Comb.
CCL5
TNFRSF4
Hazard Ratio
h
Volcano Plot Translation+ Subgroup
(Reactome eukaryotic translation initiation genes)
|Test Statistic|
Test Statistic (Hazard Ratio, METABRIC)
RPS27
RPS28
RPLP1
RPS5A
RPS4X
RPL9
RPL36A
FDR < 1E-5
-Log10(FDR)
Log2(Fold-Change)
i
Survival vs Translation+ Subgroup Test Statistic
(Reactome eukaryotic translation initiation genes)
Top Markers
Correlation: -0.23
P-Value: 0.014
RPL36A
RPLP1
RPS3A
RPS28
RPS27
RPL9
RPS4X
Test Statistic (Hazard Ratio, METABRIC)
Test Statistic (Translation+ Subgroup)

Figure 4.4: *K2Taxonomer* annotation of scRNA-seq clustering of breast cancer immune cell data and in-silico validation via patient survival on *METABRIC* breast cancer bulk gene expression data set

A) *K2Taxonomer* annotation of 13 cell subtypes of breast cancer immune cell populations. Cell type labels are in accordance to the original publication of these data[153]. Color and thickness of each edge indicates the association between the projected signature of up-regulated genes of each subgroup and patient survival in *METABRIC* breast cancer cohort via Cox proportional hazards testing. The top and bottom dendrograms show the results without and with adjusments of covariates for inflammation and proliferation. Blue and red are indicative of hazard ratio < 1 and hazard ratio > 1, respectively. All models included age and PAM50 subtype as covariates.
B) Boxplots of gene set projection scores of selected REACTOME pathways, enriched in subgroups of immune cells. These pathways include: **PD-1 Signaling**, enriched in the *Trm All* subgroup, **Translation**, enriched in the *Translation+* subgroup, **TNF Signaling**, enriched in the *CD4+ CCL5-* and *Treg TNFRSF4+* subgroups, and **Cell Cycle**, enriched in the *CD8+ mit. Trm and Treg mit. subgroups*,
C) Boxplots of markers constitutively regulated among selected *K2Taxonomer* subgroups. *GZMB* is upregulated in the *Trm All* subgroup. *CCL5* and *TNFRSF4* are up- and down-regulated, respectively, in the *CD4+ CCL5-* subgroup. TNFRSF4 is further up-regulated in the *Treg TNFRSF4+ subgroup,* while *RGS1+* is up-regulated in the *Treg RGS1+* subgroup. Finally, *RPS27* is up-regulated in the *Translation+* subgroup.
D) tSNE dimensionality reduction of the single-cell breast cancer immune cell data, indicating the cell subtype label assignment of every cell, as well as Z-scored expression of selected genes from C.
E) 95% confidence intervals of hazard ratios from Cox proportional hazards testing of gene set projections of cellular subgroups on the *METABRIC* data set. *Covariates* shows the results of the survival model of sample-level inflammation and proliferation scores without a *K2Taxonomer* derived signature. Every other models shows the confidence interval of the subgroup-specific model without and with adjusting for inflammation and proliferation score, as well as the confidence intervals of inflammation and proliferation in the full model. All models included age and Pam50 breast cancer subtype as covariates.
F) Scatterplot of the comparison of the expression of *CCL5* and *TNFRSF4* expression in the *METABRIC* dataset.
G) 95% confidence intervals of hazard ratios from Cox proportional hazards testing of gene-level expression of *CCL5* and *TNFRSF4*, modelled separately, *Sep.,* and combined in a single model, *Comb.*. These models also included age, Pam50 breast cancer subtype, as well as sample-level inflammation and proliferation score as covariates.

H) Volcano plot of differential expression analysis of the *Translation+* subgroup in scRNAseq data of individual genes in the REACTOME eukaryotic translation initiation gene set. An alternative coding of the *y-axis* indicating the absolute value of the test statistic is shown on the right side of the plot. The colors indicates the association of each gene with survival in the *METABRIC* data set. The names of genes, significantly associated with better survival (hazard ratios < 1, FDR < 0.1) are labelled.

I) Comparison of the associations of expression of the REACTOME eukaryotic translation initiation gene set to survival in the *METABRIC* data set and the test statistics indicating up-regulation in the *Translation+* subgroup. Genes that were included as top markers of the *Translation+* subgroup are highlighted. The names of genes, significantly associated with better survival (hazard ratios < 1, FDR < 0.1) are labelled. The blue line indicates the linear fit of these two variables.

Chapter 5: Conclusions and future directions

In this dissertation, I presented the development of methods to advance large-scale transcriptomic profiling, encompassing both data generation and analyses. The strength of these methods was examined through evaluation of their performance relative to other methods, as well as through demonstration of biological insights gained through their application to real studies.

In the first section, I presented my assessment of a novel cost-effective highly multiplexed RNAseq platform, SFL. Throughout unbiased evaluation of the data quality across four different platforms, I showed SFL outperformed the similarly "priced" 3'DGE platform, and had comparable performance to more expensive RNAseq and microarray platforms. Accordingly, our lab is currently working to implement SFL in future projects and the SFL protocol was included in the publication of this work for broader utilization[194]. In addition to the SFL protocol itself, the potential impact of this work extends to the presentation of a rigorous and thorough strategy for assessing data quality of future RNAseq platforms. In short, this strategy includes evaluation of the efficiency of a platform to comprehensively measure the transcriptome relative to library size, the breadth of transcriptional differences that can be statistically identified, and the biological validity of these differences. All of the code I used to implement this strategy was published as a supplement to the manuscript[194].

In the second and third sections, I presented two novel machine learning methods for performing supervised and unsupervised learning on large-scale transcriptomic profiling data. One of the motivating factors for devising these techniques was the lack of

available high-dimensional machine learning tools that appropriately account for known group labels in the data. In the analyses presented in this dissertation, such group labels represented replicates of exposures to metabolism disrupting chemicals (second section) and cell type identifiers of scRNAseq profiles (third section). The knowledge of group labels provides valuable information for generating models that are capable of capturing the within-group and between-group variability. However, including this information complicates the learning procedure, especially for high-dimensional data, and my contributions tackle this challenge.

For classification of *PPARγ* modifying compounds in the toxicogenomic profiles of metabolism disrupting chemicals, I developed an amended random forest procedure that accounts for this group-level information. This procedure outperformed classic random forest modeling procedures run on two different representation of the data: the full set of data ignoring the group-level information, and a processed data set consisting of "meta-profiles" representing the group-level means. The specific addendum to the class random forest procedure was devised to more appropriately accounts for the sources of bias that tree-level bagging is meant to reduce. Bagging prevents overfitting in predictive modeling by equalizing the influence of individual observations, particularly by reducing the over-influence of outliers[195]. In experimental data, the variability associated with individual observations of the same group, e.g., biological replicates, is associated with the error of their group-level mean estimate, while the variability associated with the true group-level means should reflect actual biological differences. Therefore, by taking group-level means *after* performing observation-level bagging, the

amended random forest procedure more appropriately reduces the influence of noise associated with the measurements of each observation. It's important to note the success of this strategy is highly contingent on accurate labeling of groups in the training data. Future work will include further applications of this method to assess the precise conditions in which it may be expected to outperform other methods, as well as additional strategies for modeling other data relationships such as that of dose- and duration-dependent exposures.

The top-down, recursive partitioning approach employed by *K2Taxonomer* utilizes variance reducing properties of ensemble learning, while enabling the flexible modeling of data to include group-level information. In this dissertation, I presented applications of *K2Taxonomer* to three types of data: observational-level bulk gene expression data, toxicogenomic gene expression data with replicates, and single-cell gene expression data. Through numerous performance assessments and comparison to agglomerative clustering techniques, *K2Taxonomer* demonstrated superior performance particularly for modeling single-cell gene expression data. To appropriately model each data set, *K2Taxonomer* was designed so as to be customizable. When fitting data with group-level information, *K2Taxonomer* can be run in two ways. In the case of toxicogenomic screening data, in which the number of replicates is small relative to the number of exposures, *K2Taxonomer* was applied by performing pre-processing of the data set to collapse the replicates of each exposure to a test statistics based on a multivariate linear model. In the case of single-cell gene expression data, in which the number of members of each cell type is large relative to the number of groups,

K2Taxonomer was applied using the constrained K-means semi-supervised algorithm[151] to take advantage of the full set of data. The latter application is especially powerful as few unsupervised methods have been devised to account for group-level information.

In addition to the extensive performance assessment of these methods, this dissertation included several practical applications illustrating their versatility and value in supporting biological discovery. Using the amended random forest technique, I identified four high likelihood *PPARγ* modifying compounds: allethrin, tonalide, quinoxyfen, and fenthion. Identifying tonalide and quinoxyfen as metabolism disruptors has the potential to be particularly impactful as these chemicals remain widely used as artificial scents[196] and fungicides[112], respectively. *K2Taxonomer* subgrouping of PPARγ modifying compounds demonstrated that the transcriptomic profiles of these chemicals' exposures were most similar to those of highly adipogenic environmental compounds. Subsequent experimental validation demonstrated that these chemicals induce white adipogenesis, but not the beneficial brite adipogenesis, induced by the type 2 diabetic drug, rosiglitazone. Furthermore, *K2Taxonomer* characterized profiles of endogenous compounds, protectin D1 and resolvin E1, as being most similar to that of the CDK inhibitor, roscovitine, which has been shown to increase insulin sensitivity and to induce brite adipogenesis[104]. Although the focus of this paper was identifying environmental metabolism disruptors, these findings could inform future studies of the efficacy of protectin D1 and resolvin E1 as metabolic therapeutics. This study represents less than 0.01% of chemicals in use worldwide[23]. The success of these analyses could encourage

more extensive toxicogenomic profiling and consequently inform action to ameliorate exposures contributing to obesity associated pathology.

While there are currently many software packages to cluster RNAseq data, there remains a gap in the availability of tools to comprehensively annotate these clusters and explore relationships between cell types. Through application of *K2Taxonomer* to the analysis of breast cancer immune infiltrating cell types derived from scRNAseq profiles, I presented an elegant way to fill this gap. This analysis enabled the discovery of an immune cell specific translational machinery gene signature of large subgroup of CD4+ and CD8+ T cells associated with improved patient survival in the *METABRIC* breast cancer tissue bulk gene expression data set, suggesting that high expression of this subset of translational genes was distinct from translational gene regulation in tumor cells, thereby indicating disease attenuating immune infiltration. What's most noteworthy about these results is how transcriptional similarities of cell types from disparate lineages can be leveraged to disentangle dominant gene expression patterns of functionally similar subtypes of cells in heterogenous tissue. Whereas other methods have been developed to infer lineage-specific trajectories of single cells[197], our strategy is able to capture instances of phenotypic convergence, which is known to occur in both immune[182] and tumor cells[198]. Accordingly, I am currently working on a similar analysis of head-and-neck cancer tumor scRNAseq data to distinguish coregulatory patterns driving tumor cell states associated with tumor progression, including: epithelial-mesenchymal transition and epithelial differentiation[199].

As exemplified by these analyses, finalization of biological inferences and evaluation of the clinical relevance of findings yielded by machine learning analyses of transcriptomic profiling data typically require independent *in silico* or *in vivo* validation. Furthermore, the wealth of information a *K2Taxonomer*-based analysis yields goes well beyond the hierarchical grouping of input samples, and would be lost if the taxonomic relationship between samples were the only outcome. This is a recurrent challenge faced by machine learning methods for the analysis of high-dimensional data, which all too often represent "black boxes" whose inner working, and the features driving the outcome, remain inaccessible. To overcome this issue, *K2Taxonomer* performs a comprehensive data-driven annotation of each estimated subgroups and automatically generates an interactive web-portal of the full compendia of results . Beside proving pivotal to driving the inferences made throughout the latter sections of this dissertation, these portals are readily publishable so that they can be further explored by other researchers. Therefore, this work highlights opportunities to broaden the impact of one's data and analyses, especially as tools to design these portals, such as *RShiny*, become more popular.

Over the past 25 years[200], such refinements as those presented in this dissertation have facilitated the adoption of high-throughput transcriptomic profiling. The use of these assays for biological discovery have well-known limitations. While most transcripts function as intermediaries between DNA and proteins, transcript abundance and protein abundance are poorly correlated[201]. Additionally, transcriptomic profiling cannot evaluate conformational variants of proteins resulting from post-translational modifications, such as phosphorylation, which are critical to their function[201]. While in many instances high-

throughput proteomics profiling would be better suited for assessing molecular activity, practical challenges limiting its efficiency and precision have persisted and impeded its broader adoption[202]. However, despite the different techniques with which transcriptomes and proteomes are quantified, i.e., sequencing versus mass spectrometry, the processed data generated by each method are comparable[203]. Therefore, if and when proteomic profiling moves to the forefront of high-throughput -omics research, a wealth of analytical resources will be at its disposal, thanks in large part to decades of methodological advances driven by transcriptomic profiling.

Appendix

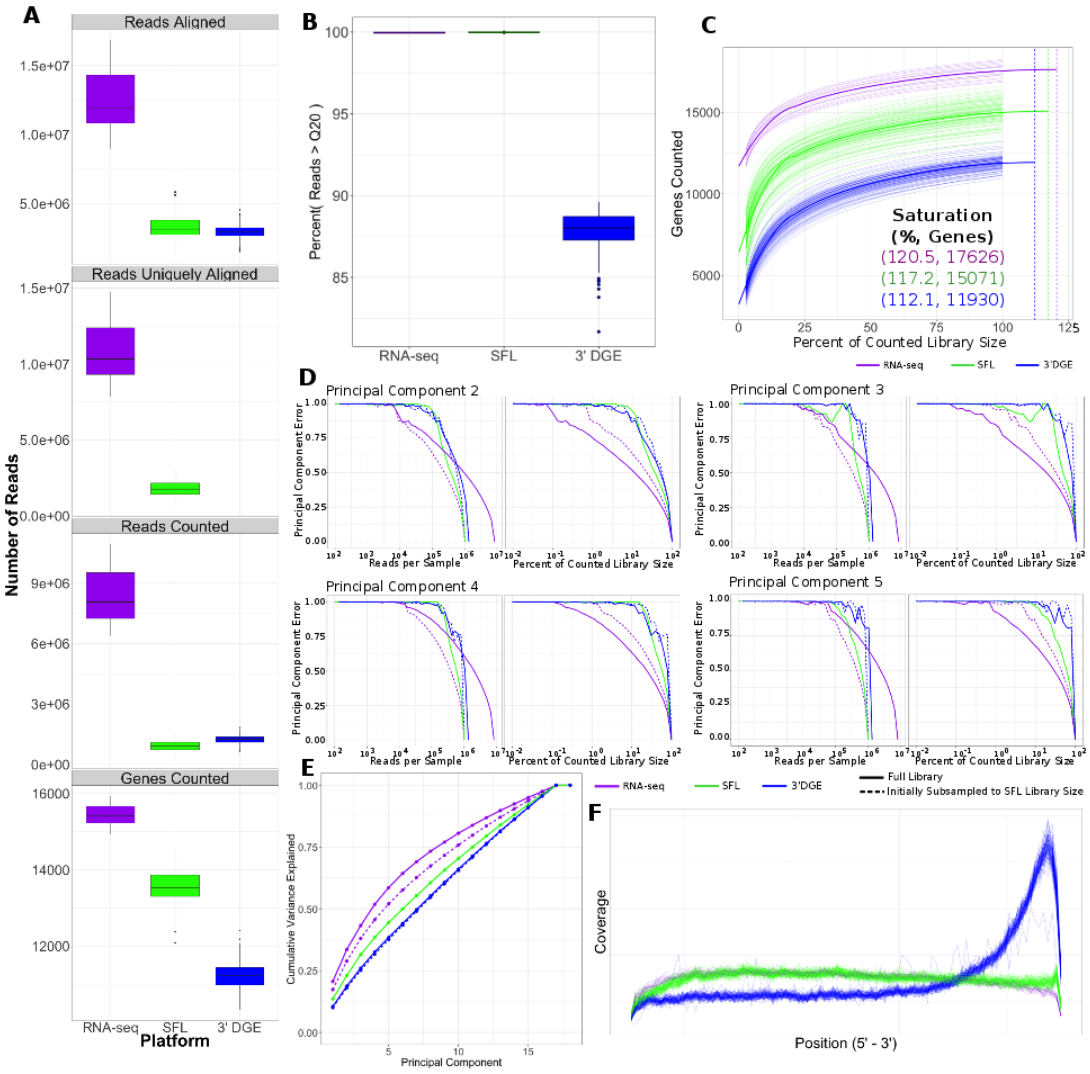


Figure A.1: Additional coverage analyses comparing Full Coverage RNAseq, SFL, and 3'DGE

A) Distribution of read counts at different preprocessing steps, as well as the number of counted genes.

B) Per platform boxplots of the percentage of reads in each sample with Phred Values > 20 (Q20).

C) Saturation analysis showing the number of genes with counts > 0 versus the percent subsampling of the full counted library size. Thick lines show the loess fit of each platform. Vertical lines show the estimated point of saturation, i.e. the

minimum percentage of the full counted library at which the maximum number of genes are counted. This value is also given, as well as the estimated maximum number of counted genes.

D) Principal component error for PC's 2-5 for each platform, including subsampled full coverage RNAseq and 3' DGE to the SFL library size.

E) Cumulative variance explained by each successive principal component.

F) Relative coverage of reads along transcripts from 5' to 3'.

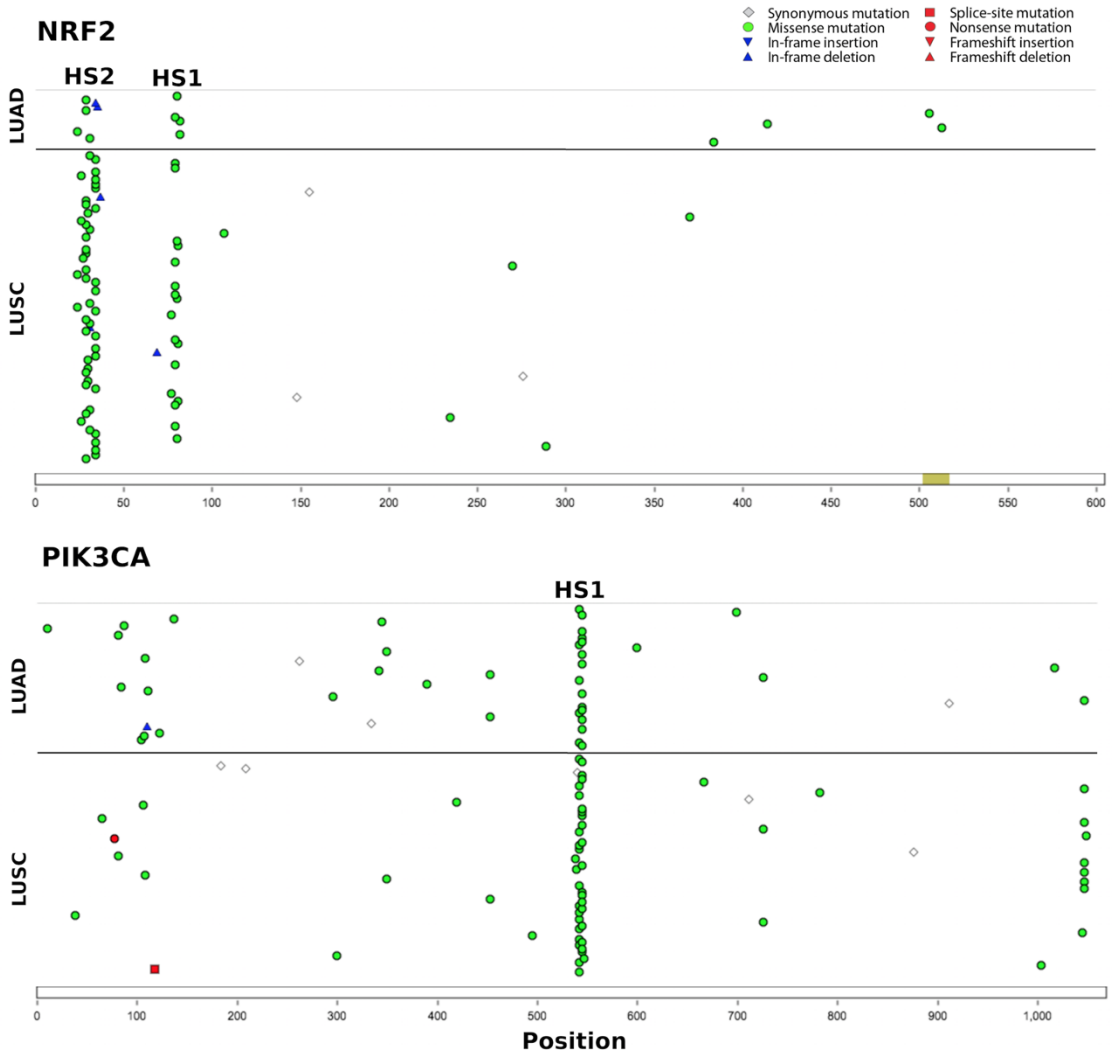


Figure A.2: Locations of mutations hotspots in *NRF2* and *PIK3CA*

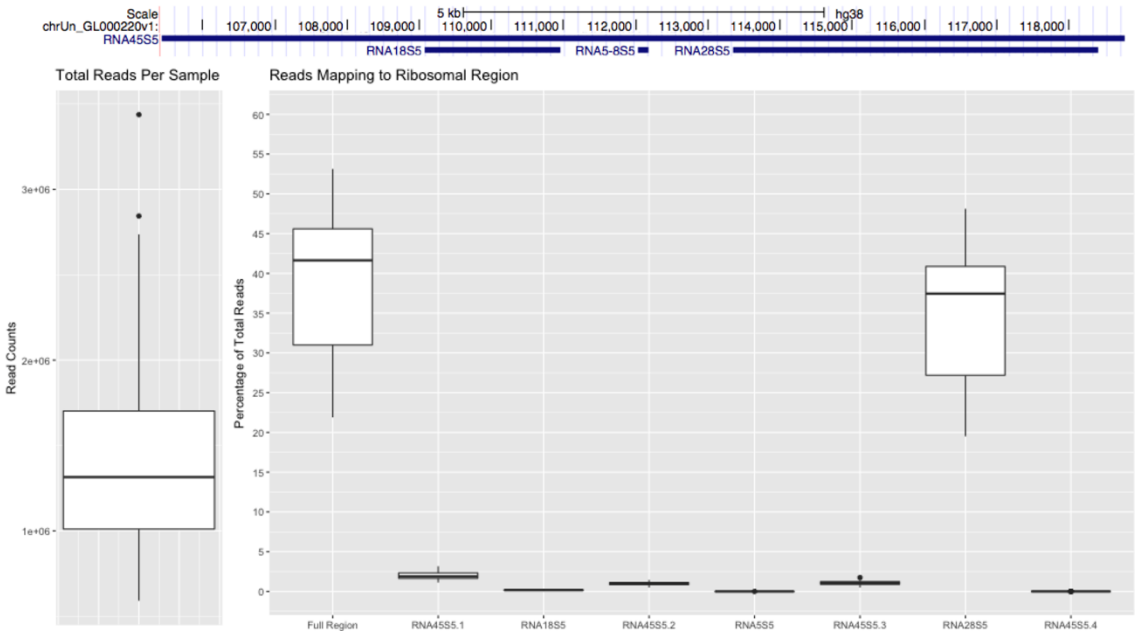


Figure A.3: Summary of rRNA contamination in SFL libraries

The total library sizes for SFL samples (left) and proportion of these reads that align to a ribosomal region (*RNA45S5*) in the genome. The majority of these reads align to *RNA28S5*.

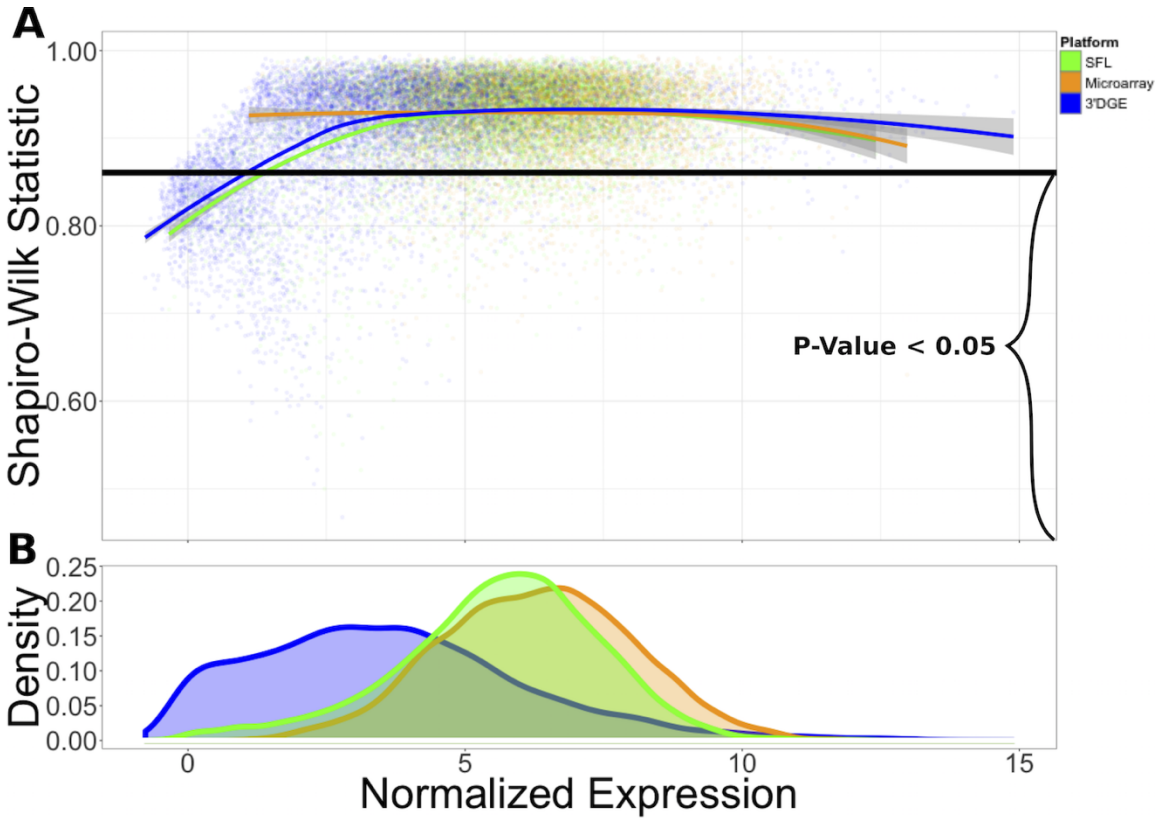


Figure A.4: Shapiro-Wilk Test Statistic VS Normalized Expression for SFL, Microarray, and 3' DGE

A) Loess fit of the Shapiro-Wilks test statistic vs normalized expression across each platform. Values beneath the vertical black line are indicative of test statistics associated with nominal p-values < 0.05.

B) Distribution of mean normalized expression across all three platforms.

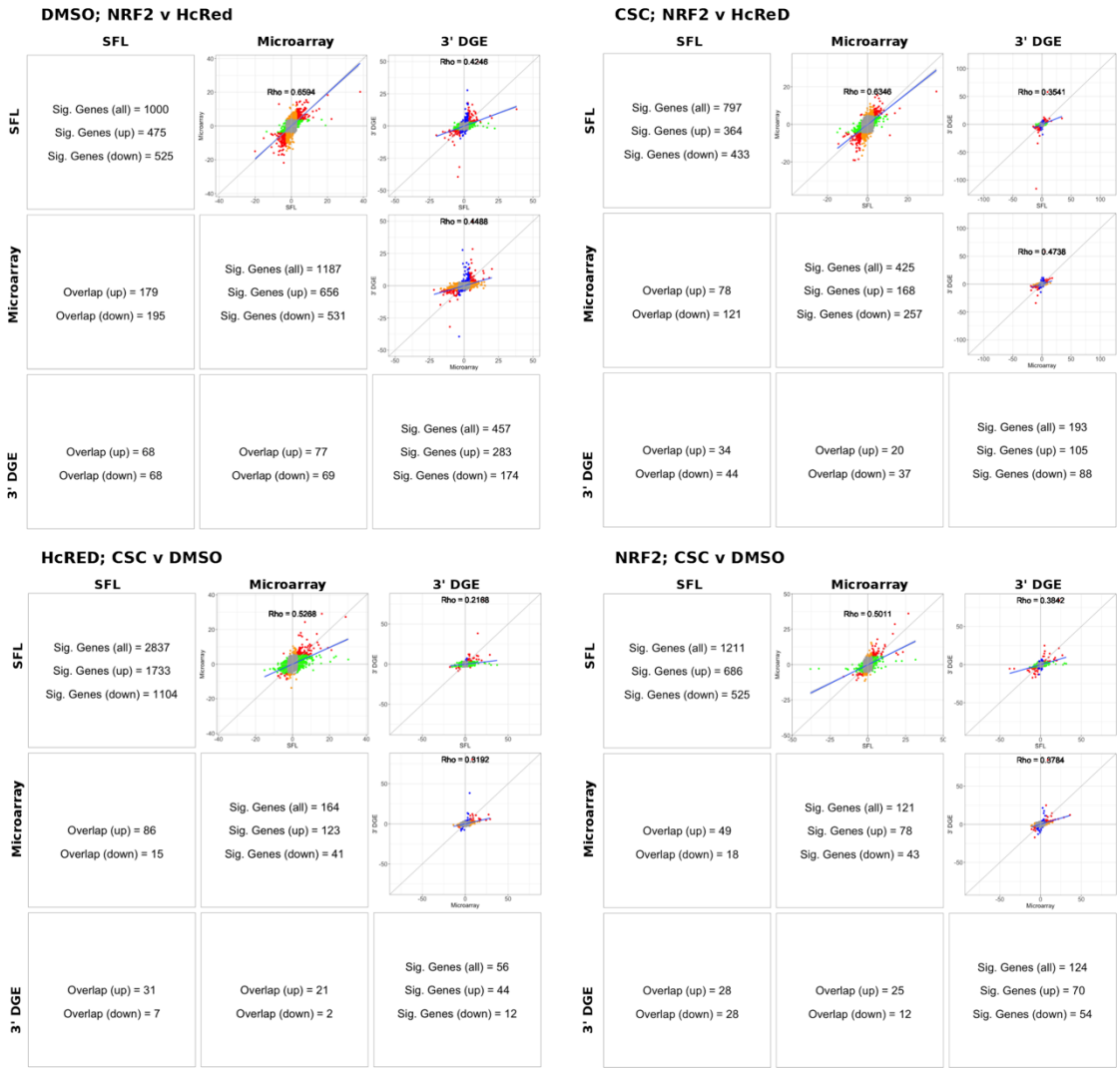


Figure A.5: Additional Differential Analysis Results (SFL; Microarray; 3'DGE)

Comparison of gene discovery (FDR Q-Value < 0.05) by differential analysis with limma, comparing normalized gene expression, including the raw discovery rates, discovered gene overlap, and linear fits, comparing test statistics from each platform. Euler diagrams of these results are shown in Figure A6.

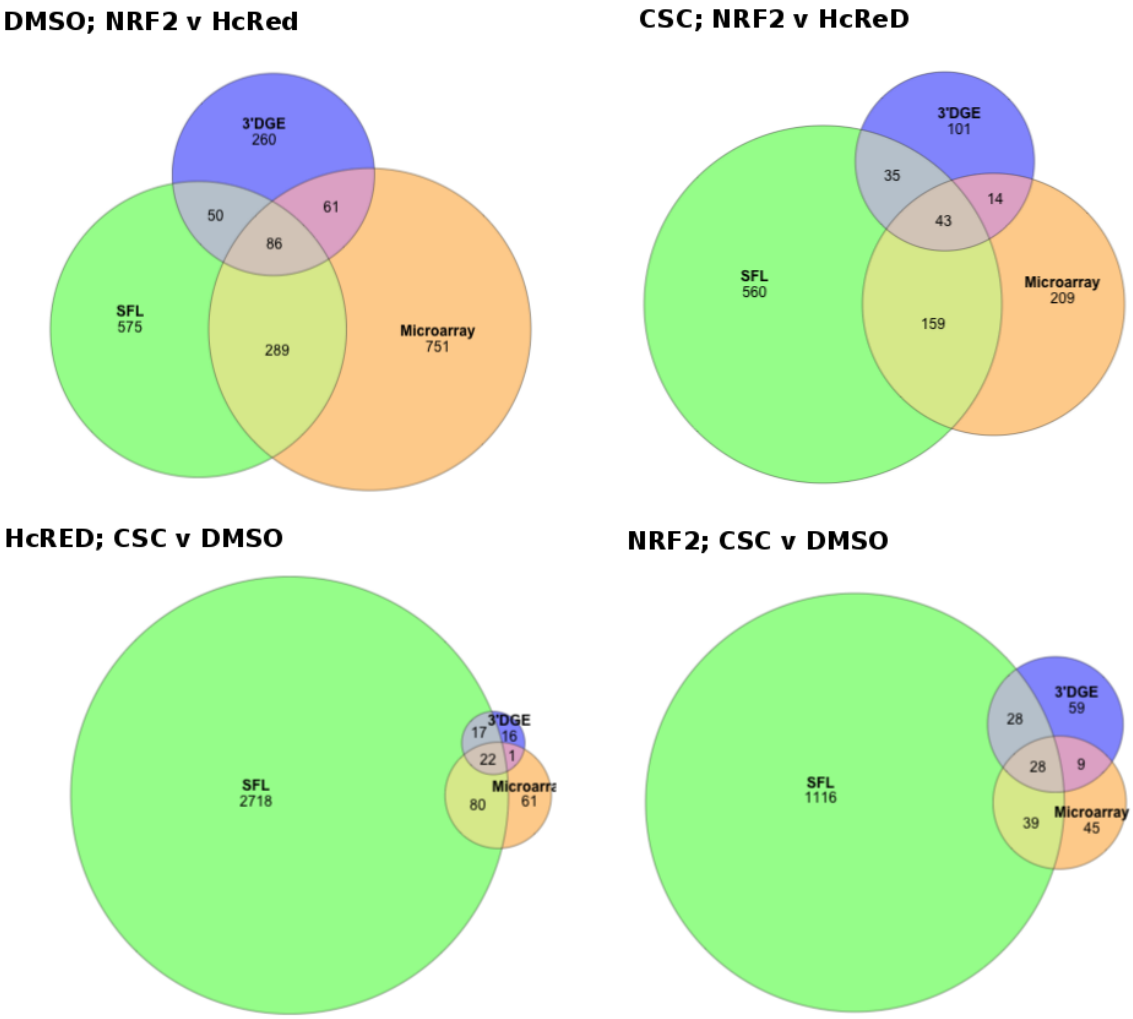


Figure A.6: Venn diagrams of gene discovery from differential analysis (SFL; Microarray; 3'DGE)

Comparison of gene discovery (FDR Q-Value < 0.05) by two-group differential analysis with LIMMA.

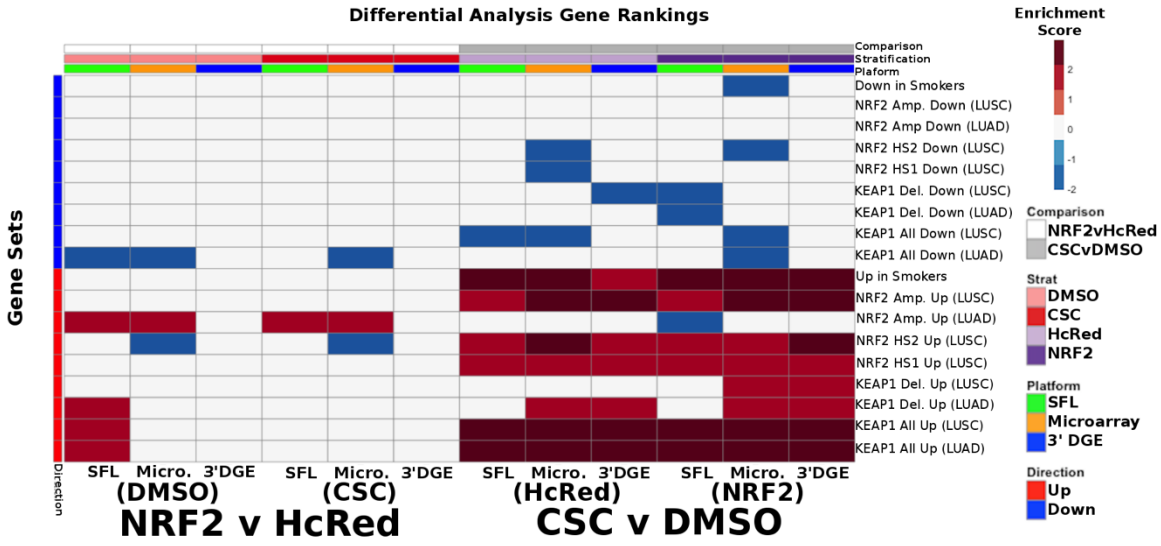


Figure A.7: Gene Set specific results of Smoking and Gene Mutation Signatures across SFL, 3'DGE and Microarray

Gene-set enrichment results with respect to genotypic perturbations (left) and chemical perturbations (right) differential signatures across like samples within SFL, Microarray, and 3' DGE. Columns correspond to differential signatures comparing genotypic or chemical perturbation groups, stratified by a single chemical or genotypic perturbation group, respectively, e.g. the left-most column shows the enrichment results with respect to the "DMSO-treated; NRF2 *vs.* HcRed" signature within the samples (*stratum*) in SFL data. Row labels for the *TCGA*-derived genes sets follow the nomenclature "<GENE> <Mutation Type> <Direction>(<Source>)". Mutation type can be one of: *Amp.* for copy number gain, *Del.* for copy number loss, *HS#* for point mutations in hotspots (See Figure A2), and *All* for collapsed point mutations across the full length of the genes.

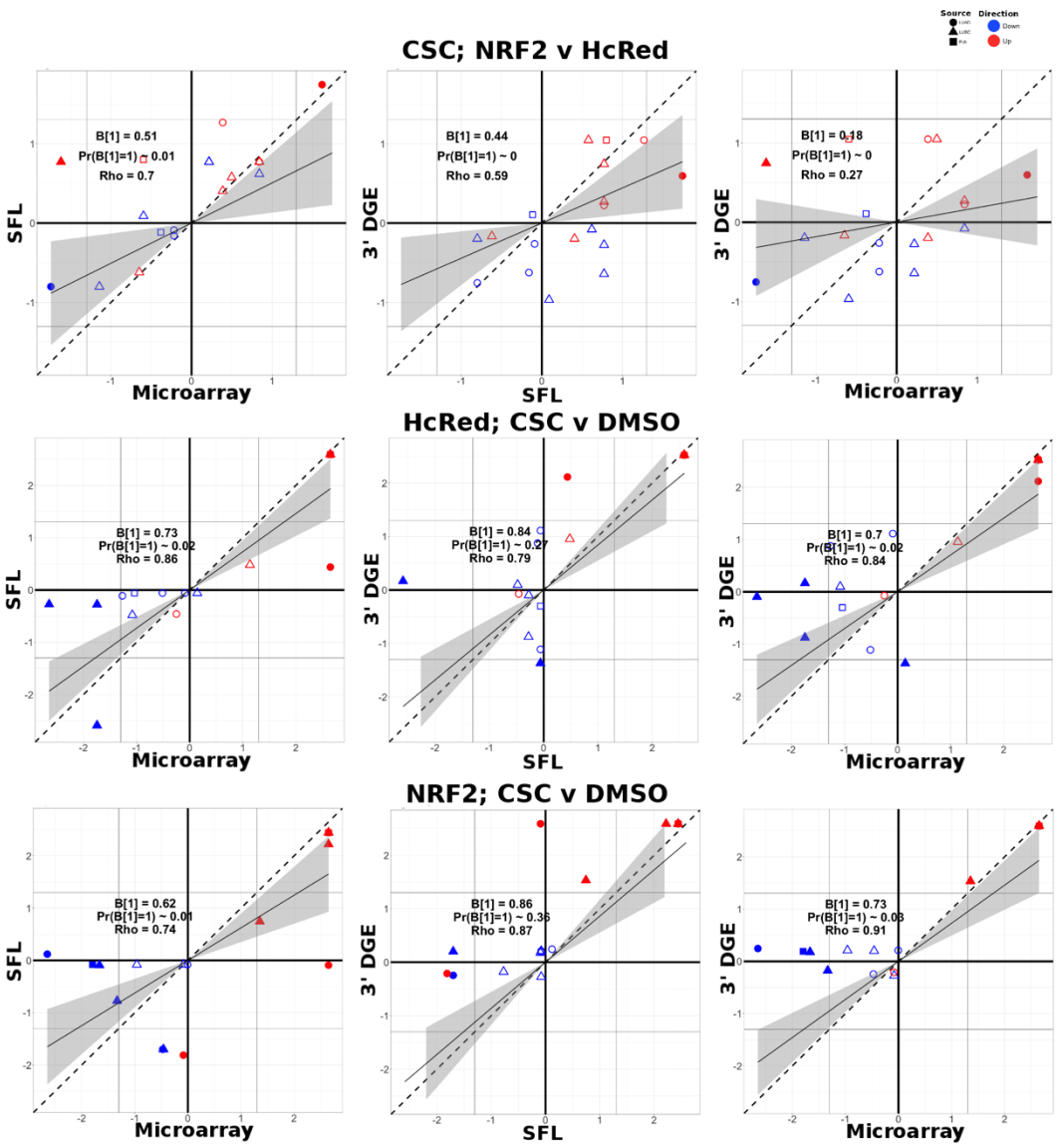


Figure A.8: Additional Biological Recapitulation Comparisons (SFL; Microarray; 3'DGE)

Comparison of the gene set enrichment results between SFL, microarray and 3' DGE with respect to the "CSC-treated; *NRF2 vs.* HcRed", "HcRed-treated; CSC *vs.* DMSO", and "*NRF2*-treated; CSC *vs.* DMSO" differential signature. Shown are the transformed FDR q-values of the *TCGA*-derived gene sets corresponding to mutations of comparable gene sets. The |-Log10(FDR Q-values)| = 0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. Points of gene sets

whose enrichment meets this threshold in either of the two platforms are filled in. Colors and shape of points denote direction and source of the gene set, respectively.

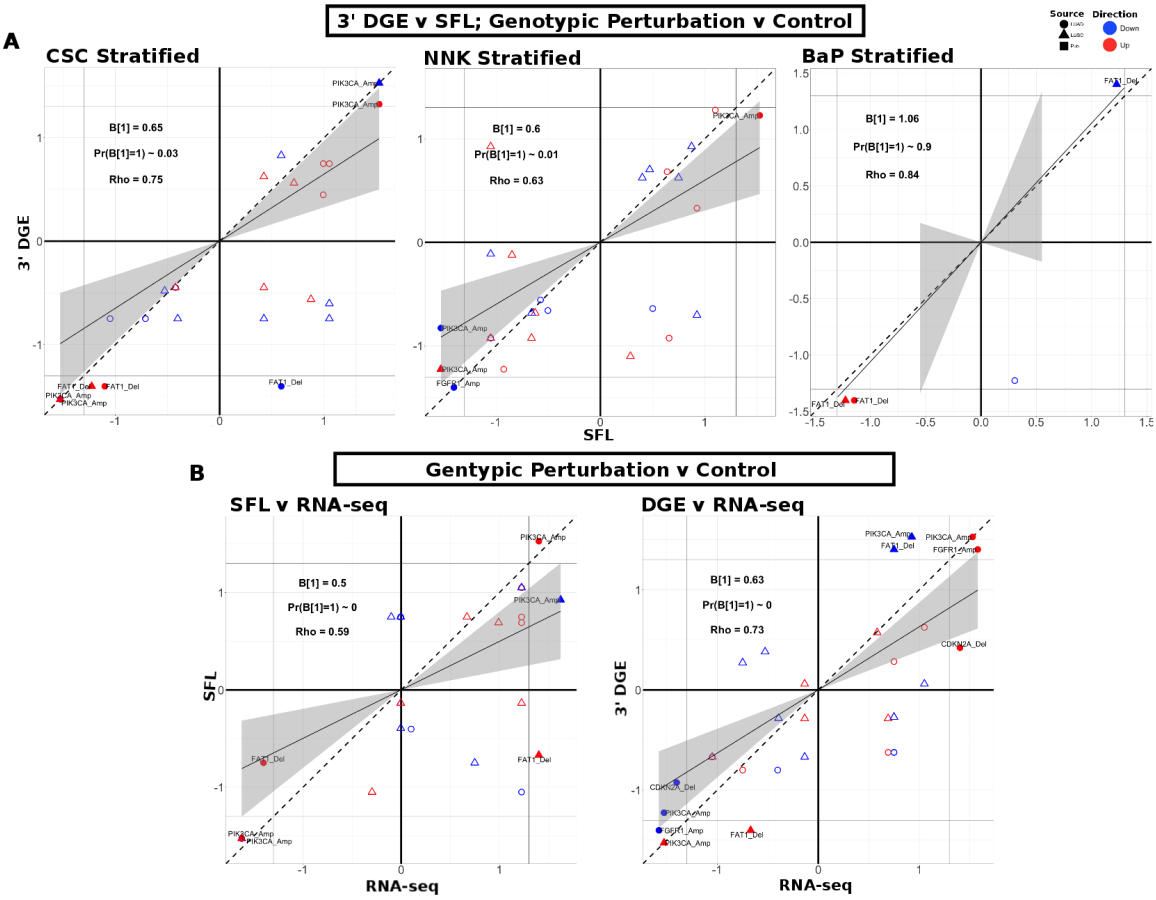


Figure A.9: Additional Biological Recapitulation Comparisons (RNAseq; SFL; 3'DGE)

A) Comparison of the gene set enrichment results between SFL and 3' DGE results for genotypic perturbation v control, stratified for BaP chemical exposure. Points indicate gene set comparisons to concordant signatures, derived from *TCGA*. In this case only FAT1 gene sets are shown because only FAT1 and GFP genotypic perturbations are available for BaP exposed samples. Shown are the transformed FDR Q-values from permutations based testing from preranked GSEA. The |-Log10(FDR Q-values)| = 0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. The names of these gene sets that meet this threshold in either of the two signatures are shown and their points are filled in. Colors and shape of points are indicative of the direction and source of the gene set, respectively.

B) Comparison of the gene set enrichment results between full coverage RNAseq and either SFL or 3' DGE results for genotypic perturbation v control, stratified for DMSO exposures. Points indicate gene set comparisons to concordant signatures, derived from *TCGA*, e.g. *PIK3CA* mutation and CNA gene sets to PIK3CA vs HcRED preranked gene signatures. Shown are the transformed FDR q-values from permutations based testing from preranked GSEA. The |-Log10(FDR Q-values)| = 0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. The names of these gene sets that meet this threshold in either of the two signatures are shown and their points are filled in. Colors and shape of points are indicative of the direction and source of the gene set, respectively.

Table A.1 Chemical information

Chemical Name	Abbr.	CAS #	Supplier	Catalog #	Max. Conc. Tested [M]	Max. Non-Toxic Conc. [M]	PPARγ Ligand or Modifier	Purity	Reference
15-deoxy-Δ12,14-prostaglandin J2	15dPGJ	87893-55-8	Cayman Chemical	18570	1E-6	1E-6	Yes	> 95%	204
2,2',4,4',5,5'-Hexachloro-1,1'-biphenyl	PCB153	35065-27-1	Ultra Scientific	RPC-047	1E-5	1E-5	No	NA*	---
2,2',5,5'-Tetrachloro-1,1'-biphenyl	PCB52	35693-99-3	Sigma Aldrich	35599	1E-5	1E-5	No	> 98%	---
2,4,6-Tris(tert-butyl)phenol	TTBP	732-26-3	Sigma Aldrich	T49409	2E-5	2E-5	Suspected	98%	88
2-ethylhexanol	EtHex	104-76-7	Sigma Aldrich	W315109	1E-5	1E-5	No	> 99%	---
3,3',4,4',5-Pentachloro-1,1'-biphenyl	PCB126	57465-28-8	Ultra Scientific	RPC-102	1E-8	1E-8	No	NA	205
3,3',5,5'-Tetrabromobisphenol A	TBBPA	79-94-7	Sigma Aldrich	330396	2E-5	2E-5	Yes	97%	206
4,4'-Dichloro-diphenyldichloroethylene	DDE	72-55-9	Sigma Aldrich	48679	1E-5	1E-5	No	NA	207
4,4'-dichloro-diphenyltrichloroethane	DDT	50-29-3	Sigma Aldrich	40124	1E-5	1E-5	No	NA	207
4,5,6,7-Tetrabromobenzotriazole	TBB	Synthesized by Asis Chemical			1E-5	1E-5	No	95%	
--9-cis-retinoic acid	9cRA	5300-03-8	Sigma Aldrich	R4643	1E-6	1E-6	Yes	> 98%	208
All-trans retinoic acid	ATRA	302-79-4	Sigma Aldrich	R2625	2E-6	2E-6	Yes	> 98%	209
Benzyl butyl phthalate	BBzP	85-68-7	Sigma Aldrich	36927	1E-5	1E-5	Yes	98%	210
Bisphenol A	BPA	80-05-7	Sigma Aldrich	239658	1E-5	1E-5	No	> 99%	211
Bisphenol A diglycidyl ether	BADGE	1675-54-3	Sigma Aldrich	D3415	1E-5	1E-5	Yes	NA	212
Bisphenol S	BPS	80-09-1	Sigma Aldrich	43034	1E-5	1E-5	No	NA	211
Candesartan	Cande	145040-37-5	Sigma Aldrich	SML0245	2E-5	4E-6	Yes	> 98%	213
CL 316,243	CL316	138908-40-4	Sigma Aldrich	C5976	5E-6	5E-6	No	> 98%	---
Corticosterone	Corti	50-22-6	Sigma Aldrich	27840	2E-6	2E-6	No	> 98%	---
Cyazofamid	Cyazo	120116-88-3	Sigma Aldrich	33874	4E-5	2E-5	Suspected	NA	88
d-cis,trans-Allethrin	Allet	548-79-2	Sigma Aldrich	33396	2E-5	1E-5	Suspected	97%	88
Dexamethasone	Dex-SP	2392-39-4	Sigma Aldrich	D1159	2E-7	2E-7	No	> 98%	---
Di(2-ethylhexyl) phthalate	DEHP	117-81-7	Sigma Aldrich	36735	1E-5	1E-5	No	> 99%	80
Dibutyltin	DBT	683-18-1	Sigma Aldrich	205494	2E-7	2E-7	Yes	96%	214
Diisononyl phthalate	DINP	28553-12-0	Sigma Aldrich	376663	1E-5	1E-5	Yes	> 99%	215
Dioctyl sulfosuccinate sodium	DOSS	577-11-7	Sigma Aldrich	323586	5E-6	5E-6	Suspected	> 97%	105
Diphenyl phosphate	DiPhPho	838-85-7	Sigma Aldrich	850608	1E-5	1E-5	Suspected	99%	129
Ethylene brassylate	EtBra	105-95-3	Sigma Aldrich	W354309	1E-5	1E-5	No	> 95%	---
Fenthion	Fenth	55-38-9	Sigma Aldrich	36552	4E-5	4E-5	Suspected	> 99%	88
Firemaster 550	FM550	Gift from Heather Stapleton, Duke			10 ug/ml	10 ug/ml	Yes	NA	
Fludioxonil	Fludi	131341-86-1	Sigma Aldrich	46102	2E-5	2E-6	Suspected	> 95%	88
Honokiol	Honok	35354-74-6	Sigma Aldrich	H4914	2E-5	4E-6	Yes	> 98%	216
LG100268	LG268	153559-76-3	Sigma Aldrich	SML0279	1E-7	1E-7	Yes	> 98%	217
LG100754	LG754	180713-37-5	Tocris	3831	2E-7	2E-7	Yes	> 99%	217

Magnolol	Magno	528-43-8	Sigma Aldrich	M3445	2E-5	2E-5	Yes	> 95%	[218]
MCC-555	MCC555	161600-01-7	Sigma Aldrich	SML0896	5E-6	5E-6	Yes	98%	[209]
Melengestrol acetate	Melen	2919-66-6	Sigma Aldrich	73248	2E-5	2E-5	No	> 97%	---
Mono-(2-ethyhexyl) tetrabromophthalate	METBP	Synthesized by Asis Chemical			1E-5	1E-5	Yes	95%	
Mono(2-ethylhexyl) phthalate	MEHP	4376-20-9	Sigma Aldrich	CDS010608	1E-5	1E-5	Yes	NA	[80]
Monobenzyl phthalate	MBzP	2528-16-7	Sigma Aldrich	89505	1E-5	1E-5	Yes	95%	[107]
Mono-n-butyl phthalate	MBuP	131-70-4	Sigma Aldrich	30751	2E-5	2E-5	Yes	> 98%	[107]
n-Butylparaben	BuPara	94-26-8	Sigma Aldrich	54680	2E-5	2E-5	Yes	> 99%	[219]
N-nitro-2-imidazolidinimine	Imida	138261-41-3	Sigma Aldrich	37894	1E-5	1E-5	No	> 99%	ToxPi
nTZDpa	nTZDpa	118414-59-8	Sigma Aldrich	SML0616	1E-6	1E-6	Yes	> 98%	[220]
Perfluorooctanesulfonic acid	PFOS	2795-39-3	Sigma Aldrich	77282	4E-5	4E-5	Suspected	NA	[126]
Perfluorooctanoic acid	PFOA	335-67-1	Sigma Aldrich	33824	1E-5	1E-5	Suspected	> 99%	[126]
Pioglitazone hydrochloride	Piogl	112529-15-4	Sigma Aldrich	E6910	1E-5	1E-5	Yes	> 98%	[221]
Prallethrin	Prall	23031-36-9	Sigma Aldrich	32917	1E-5	1E-5	Suspected	> 95%	[88]
Pregnenolone 16α-carbonitrile	Pregn	1434-54-4	Sigma Aldrich	P0543	1E-5	1E-5	No	> 97%	---
Propylparaben	ProPara	94-13-3	Sigma Aldrich	P53357	1E-5	1E-5	Yes	> 99%	[222]
Protectin D1	Prote	871826-47-0	Cayman Chemical	10008128	2E-6	2E-6	Yes	> 98%	[223]
Quinoxyfen	Quino	124495-18-7	Sigma Aldrich	46439	1E-5	1E-5	Suspected	NA	[88]
Resolvin-E1	Resol	552830-51-0	Cayman Chemical	10007848	2E-6	2E-6	Yes	> 95%	[223]
Roscovitine	Rosco	186692-46-6	Sigma Aldrich	R7772	4E-5	4E-6	Yes	> 98%	[104]
Rosiglitazone	Rosig	122320-73-4	Cayman Chemical	71740	1E-6	1E-6	Yes	> 98%	[224]
S26948	S26948	353280-43-0	Sigma Aldrich	SML0510	2E-6	2E-6	Yes	> 98%	[225]
Sodium arsenite	Arsen	7784-46-5	Sigma Aldrich	S7400	4E-7	4E-7	No	> 90%	[226]
Sodium tungstate	Tungs	10213-10-2	Sigma Aldrich	14304	2E-5	2E-5	No	> 99%	[227]
SR1664	SR1664	1338259-05-4	Sigma Aldrich	SML0636	1E-6	1E-6	Yes	98%	[122]
T0901317	T1317	293754-55-9	Sigma Aldrich	T2320	1E-6	1E-6	No	> 98%	---
T0070907	T007	313516-66-4	Sigma Aldrich	T8703	4E-5	8E-6	Yes	> 98%	[228]
Tebuconazole	Tebuc	107534-96-3	Sigma Aldrich	32013	2E-5	2E-5	Suspected	NA	[88]
Telmisartan	Telmi	144701-48-4	Sigma Aldrich	T8949	2E-5	2E-5	Yes	> 98%	[229]
Tesaglitazar	Tesag	251565-85-2	Sigma Aldrich	SML1369	5E-6	5E-6	Yes	> 98%	[230]
Tolylfluanid	Tolyl	731-27-1	Sigma Aldrich	32060	2E-7	2E-7	No	> 99%	[231]
Tonalide	Tonal	21145-77-7	Sigma Aldrich	W526401	4E-6	4E-6	Suspected	> 98%	[222]
Tributyl phosphate	TBuP	126-73-8	Sigma Aldrich	240494	2E-5	2E-5	Yes	> 99%	[102]
Tributyltin	TBT	1461-22-9	Sigma Aldrich	T50202	8E-8	8E-8	Yes	96%	[232]
Triflumizole	Trifl	68694-11-1	Sigma Aldrich	32611	2E-5	2E-5	Yes	NA	[233]
Triphenyl phosphate	TPhP	115-86-6	Sigma Aldrich	241288	2E-5	1E-5	Yes	> 99%	[234]
Triphenyl phosphite	TPhPhi	101-02-0	Sigma Aldrich	T84654	1E-5	1E-5	Suspected	97%	[102]
Triphenylphosphine oxide	TPhPhoOx	791-28-6	Sigma Aldrich	T84603	1E-5	1E-5	Suspected	98%	[128]
Triphenyltin	TPhT	639-58-7	Sigma Aldrich	245712	8E-8	8E-8	Yes	98%	[214]

Trisabbre1,3-dichloro-2-propyl) phosphate	TDCPP	13674-87-8	Sigma Aldrich	32951	2E-5	2E-5	Suspected	NA	[127]
Tris(1-chloro-2-propyl) phosphate	TCCP	13674-87-5	Sigma Aldrich	32952	1E-5	1E-5	Suspected	NA	[127]
Troglitazone	Trogl	97322-87-7	Sigma Aldrich	T2573	5E-6	5E-6	Yes	>98%	[235]

*NA = Not Available

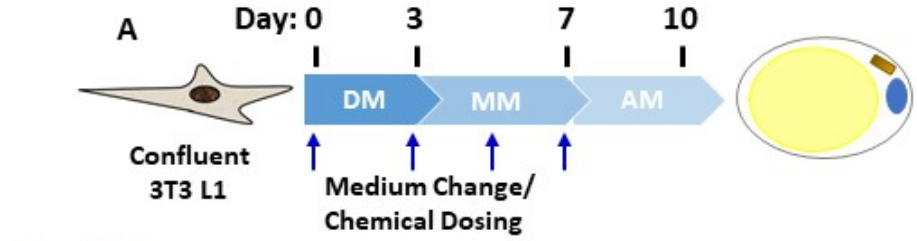
[a] We used the ToxPi designed to identify chemicals in the ToxCast dataset that are likely to be PPARγ ligands/modifiers.

Table A.2: Metabolic parameters included and excluded in human transcriptome analysis.

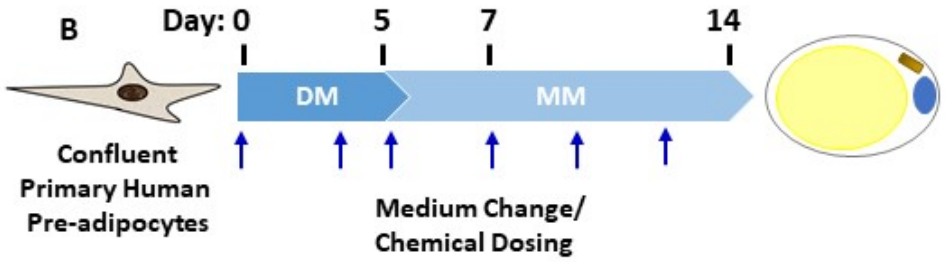
PARAMETERS INCLUDED	PARAMETERS EXCLUDED
Fat free mass %	Body mass index (kg/m2)
Fasting Plasma parameters	Waist-to-hip ratio
Free fatty acid (mmol/l)	Waist circumference (cm)
Total triglycerides (mmol/l)	Hip circumference (cm)
LDL cholesterol (mmol/l)	Plasma total fatty acids (mmol/l)
HDL cholesterol (mmol/l)	Plasma total cholesterol (mmol/l)
Adiponectin (ug/ml)	Matsuda composite insulin sensitivity index
Glucose (mmol/l)	HOMA-IR
Insulin (mU/l)	Systolic blood pressure (mm Hg)
Proinsulin (pmol/l)	Diastolic blood pressure (mm Hg)
Glycated HbA1c (%)	Glomerular filtration rate
High sensitivity C-reactive protein (mg/l)	
Interleukin-1 receptor antagonist (pg/ml)	

Table A.3: Mouse (M) and human (H) primer sequences for reverse transcriptase qPCR.

GENE SYMBOL	FORWARD	REVERSE	ANNEALING TEMP. ° C
M-RN18S	GTAACCCGTTGAACCCCATT	CCATCCAATCGGTAGTAGCG	55
M-B2M	CTGCTACGTAACACAGTTCCACCC	CATGATGCTTGATCACATGTCTCG	55
M-CIDEC	AGGCCCTGTCGTGTTAGCAC	CATGATGCCTTTGCGAACCT	55
M-CIDEA	TGCTCTTCTGTATCGCCCAGT	GCCGTGTTAAGGAATCTGCTG	55
M-ELOVL3	TCCGCGTTCTCATGTAGGTCT	GGACCTGATGCAACCCTATGA	55
M-FABP4	AGCCCAACATGATCATCAGC	TTTCCATCCCACTTCTGCAC	55
M-PLIN1	GGGACCTGTGAGTGCTTCC	GTATTGAAGAGCCGGGATCTTTT	55
M-PGC1A	AACAAGCACTTCGGTCATCCCTG	TTACTGAAGTCGCCATCCCTTAG	55
M-PPARG2	TGGGTGAAACTCTGGGAGATTC	AATTTCTTGTGAAGTGCTCATAGGC	55
M-RIP140	AGAACGCACATCAGGTGGCA	GATGGCCAGACACCCCTTTG	55
M-ADIPOQ	GCACTGGCAAGTTCTACTGCAA	GTAGGTGAAGAGAACGGCCTTGT	55
M-UCP1	ACTGCCACACCTCCAGTCATT	CTTTGCCTCACTCAGGATTGG	55
M-ACAA2	TAACGAGGCTGGCTACTTCAA	AGGGGCGTGAAGTTATGTTTT	55
H-RPL27	GTGAAAGTGTATAACTACAATCACC	TCAAACTTGACCTTGGCCT	58
H-B2M	GCTATCCAGCGTACTCCAAAG	CACACGGCAGGCATACTC	58
H-CIDEC	GGGATACAGTGTTCATGGTCCT	TCAATCTTCTTGGCAGGCTTATG	55
H-CIDEA	GGCAGGTTCACGTGTGGATA	GAAACACAGTGTTTGGCTCAAGA	60



Base Medium: DMEM, 10%FBS, 100 U/ml penicillin, 100 µg/ml streptomycin
Differentiation Medium: Base+250 nM dexamethasone, 167 nM human insulin, 0.5 mM IBMX
Maintenance Medium: Base+167 nM human insulin
Adipocyte Medium: Base+no additions



Base Medium: DMEM/F12, 25 mM $NaHCO_3$, 100 U/ml penicillin, 100 µg/ml streptomycin
Differentiation Medium: Base+33 µM d-Biotin, 17 µM pantothenate, 100 nM dexamethasone, 100 nM human insulin, 0.5 mM IBMX, 2 nM T_3, 10 µg/ml transferrin
Maintenance Medium: Base+3% FBS, 33 µM d-Biotin, 17 µM pantothenate, 10 nM dexamethasone, 10 nM insulin

Figure A.10: Differentiation and dosing protocols for 3T3-L1 cells (A) and primary human preadipocytes (B).

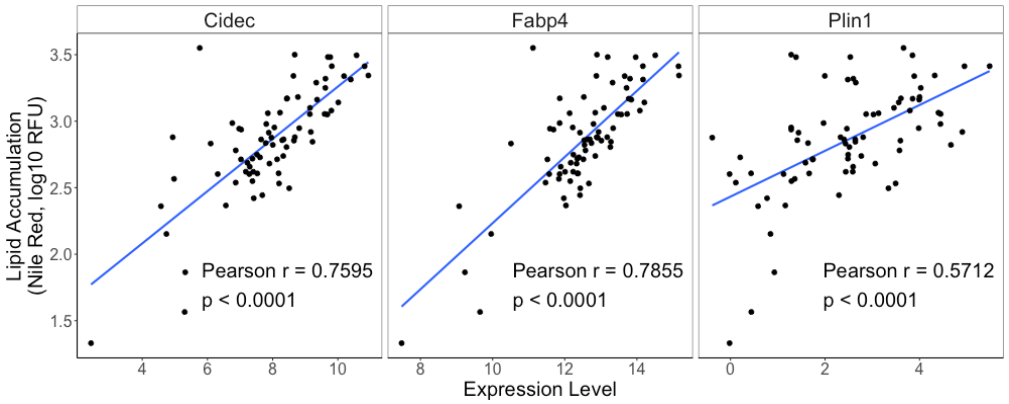


Figure A.11: Correlation of lipid accumulation with *Cidec, Fabp4, and Plin1* expression in differentiated and treated 3T3-L1 pre-adipocytes.

Confluent 3T3 L1 cells were differentiated using a standard hormone cocktail for 10 days. During differentiation, cells were treated with vehicle (Vh, 0.1% DMSO, final concentration) or test chemical (Table A.1). On days 3, 5, and 7 of differentiation, the medium was replaced and the cultures re-dosed. Following 10 days of differentiation and dosing, cells were analyzed for lipid accumulation by Nile Red staining and *gene* expression by 3'DGE. Each point represents the mean data for each chemical, (n=2-4). The least squares linear model estimate is shown in blue.

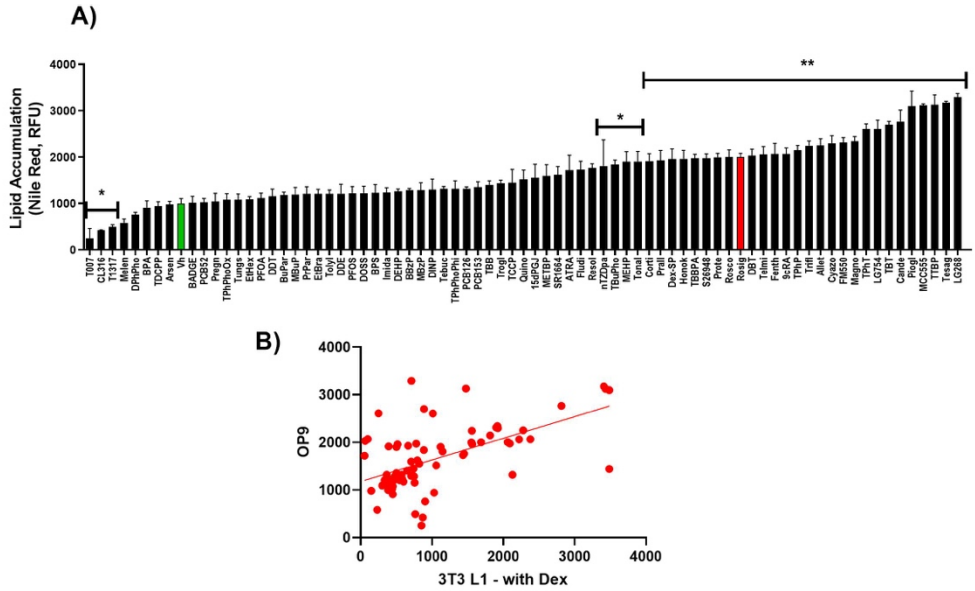


Figure A.12: Lipid accumulation in differentiated and treated OP9 pre-adipocytes.

Confluent OP9 cells were differentiated using a standard hormone cocktail for 10 days, with the exception of using 125 nM dexamethasone. During differentiation, cells were treated with vehicle (Vh, 0.2% DMSO, final concentration), rosiglitazone (positive control, 100 nM) or test chemical (Table A.1). On days 3, 5, and 7 of differentiation, the medium was replaced and the cultures re-dosed. Following 10 days of differentiation and dosing, cells were analyzed for lipid accumulation by Nile Red staining.

A) Nile Red staining induced by individual chemicals. Data are presented as mean ± SE (n=4). Statistically different from Vh-treated (highlighted in green) (*p<0.05, **<p0.01, ANOVA, Dunnett's).
B) Correlation between lipid accumulation induced in 3T3 L1 cells differentiate in the presence of dexamethasone (data are from Figure 3.1) and in OP9 cells. Pearson's r = 0.5768 (p<0.0001).

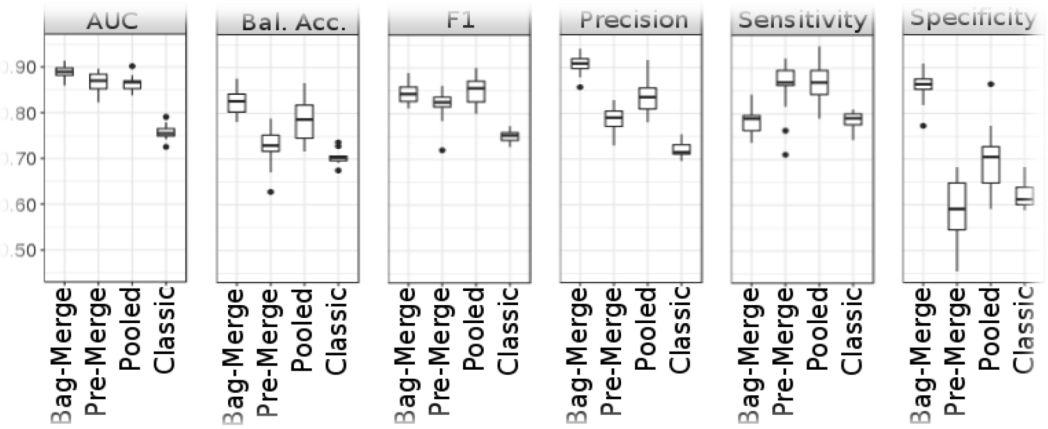


Figure A.13: Performance comparison of random forest methods.

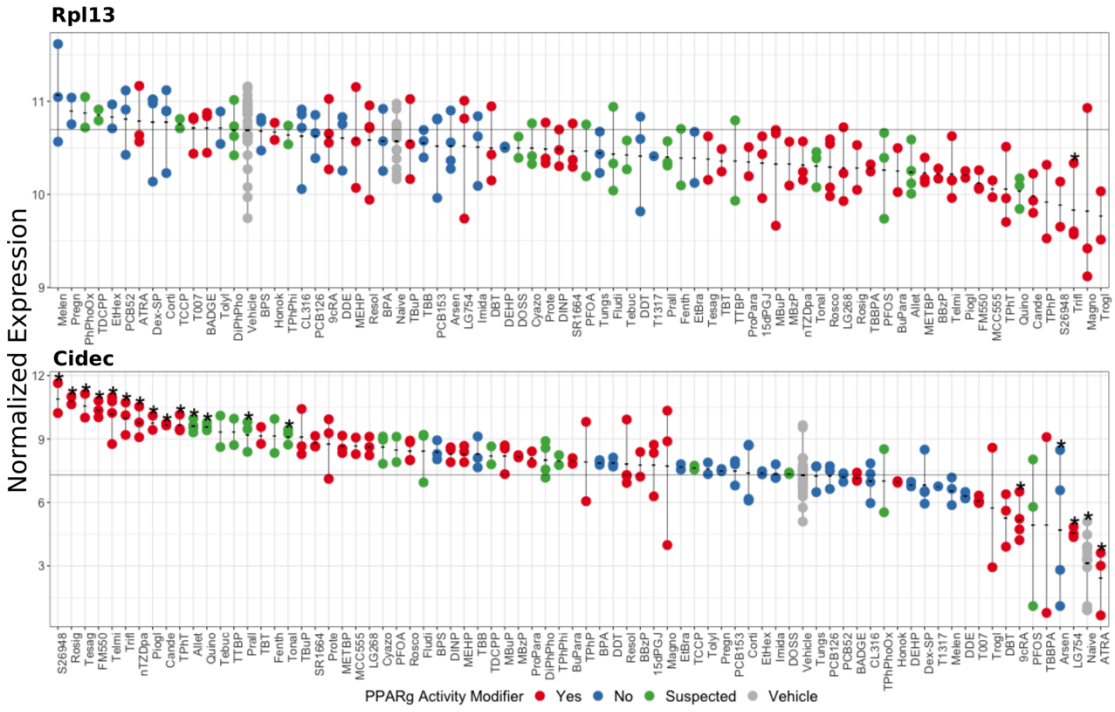


Figure A.14: Classification Results (Distributions of individual genes).

Compound specific normalized gene expression of all compounds for the top two genes (*Rpl13*, *Cidec*). Mean expression across all vehicle samples is shown as a horizontal line spanning the plot. Exposures which have statistically significant different means from vehicle (FDR Q-value < 0.05) are highlighted with an asterisk.

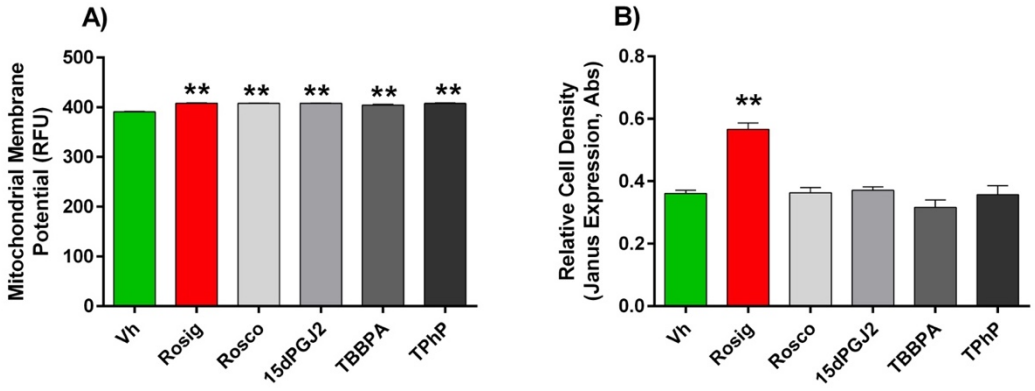


Figure A.15: Mitochondrial membrane potential and cell number analyses in the differentiated and treated 3T3-L1s.

Confluent 3T3 L1 cells were differentiated using a standard hormone cocktail for 10 days. During differentiation, cells were treated with Vh (0.1% DMSO, final concentration), rosiglitazone (Rosig, 1 μM), roscovitine (Rosco, 2 μM), 15dPGJ2 (1 μM), TBBPA (20 μM) and TPhP (10 μM). On days 3, 5, and 7 of differentiation, the adipocyte maintenance medium was replaced and the cultures re-dosed. Cells were incubated for a total of 10 days of differentiation.

A) To assess toxicity, cells were stained with MitoOrange and fluorescence intensity (λex= 485nm/λem= 530nm).
B) To assess cell number, cells were stained with JANUS green stain and absorbance was measured at 595 nM.

Data are presented as means ± SE (n=4). Statistically different from Vh-treated (**$p<0.01$, ANOVA, Dunnett's).

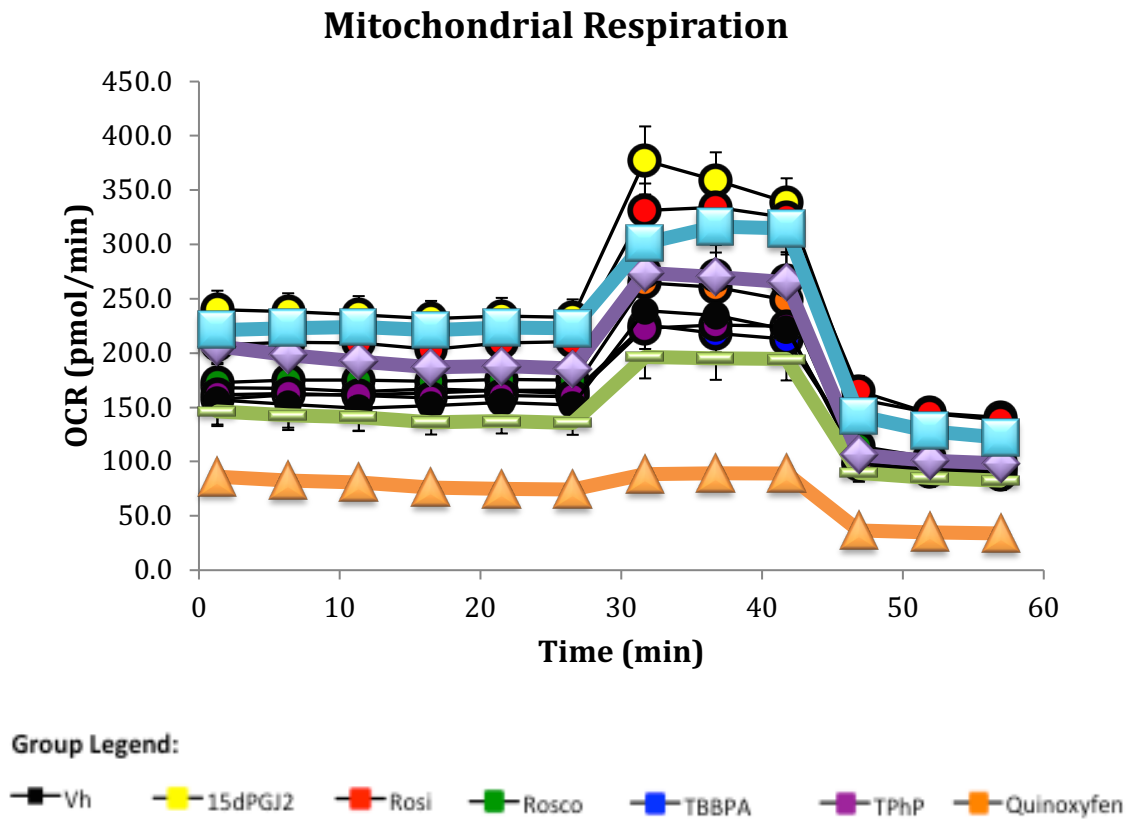


Figure A.16: Seahorse assay for mitochondrial respiration

Confluent 3T3 L1 cells were differentiated using a standard hormone cocktail for 10 days During differentiation, cells were treated with Vh (0.1% DMSO, final concentration), rosiglitazone (Rosig, 20 µM), roscovitine (Rosco, 2 µM), 15dPGJ2 (1 µM), TBBPA (20 µM) and TPhP (10 µM). On days 3, 5, and 7 of differentiation, the adipocyte maintenance medium was replaced and the cultures re-dosed. Following 10 days of differentiation and dosing, cells were analyzed by Seahorse assay for mitochondrial respiration. Representative traces are shown.

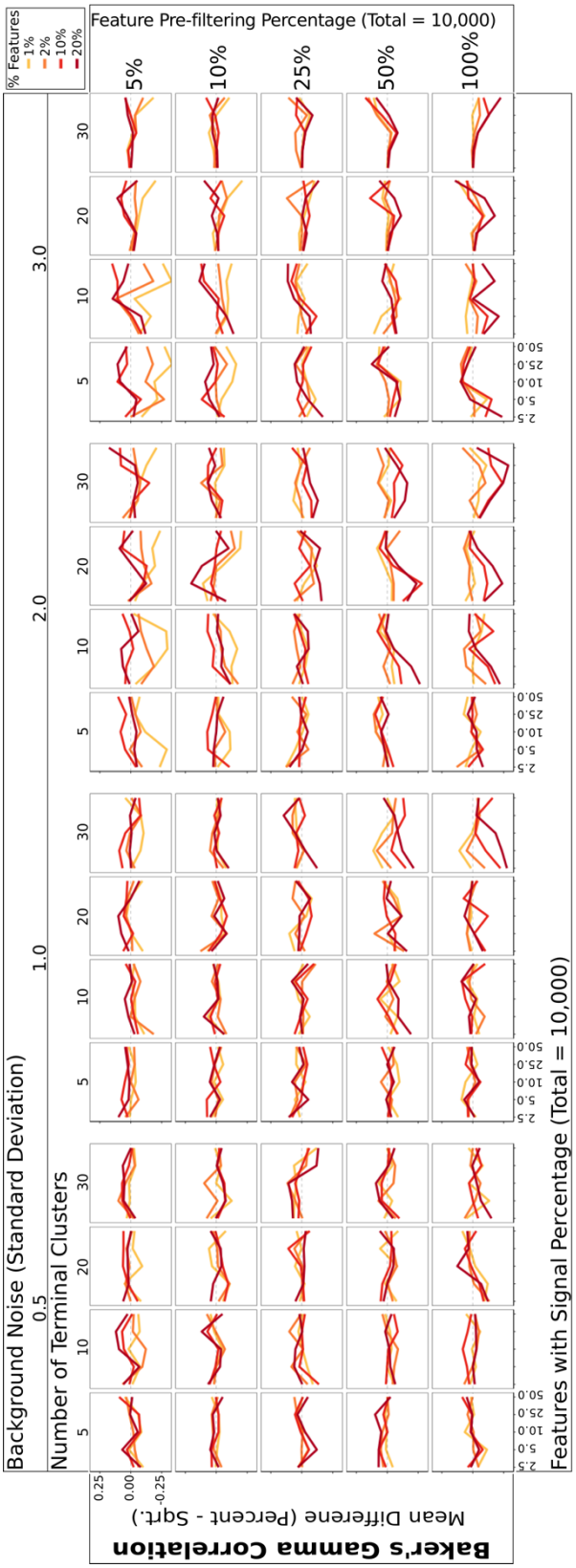

Baker's Gamma Correlation
Mean Differene (Percent - Sqrt.)
Feature Pre-filtering Percentage (Total = 10,000)
% Features
1%
2%
10%
20%
5%
10%
25%
50%
100%
Background Noise (Standard Deviation)
0.5
1.0
2.0
3.0
Number of Terminal Clusters
Features with Signal Percentage (Total = 10,000)

Figure A.17: Simulation-based performance assessment of running *K2Taxonomer* using different partition-specific feature subset sizes of the full data set

Difference of mean Baker's gamma correlation estimates measuring the similarity of *K2Taxonomer* estimates to the true hierarchy from which the simulated data was generated between. Each line shows the difference between a set percentage of the total number of features and the square root of the total number of features. Each combination of parameters was simulated 25 times.

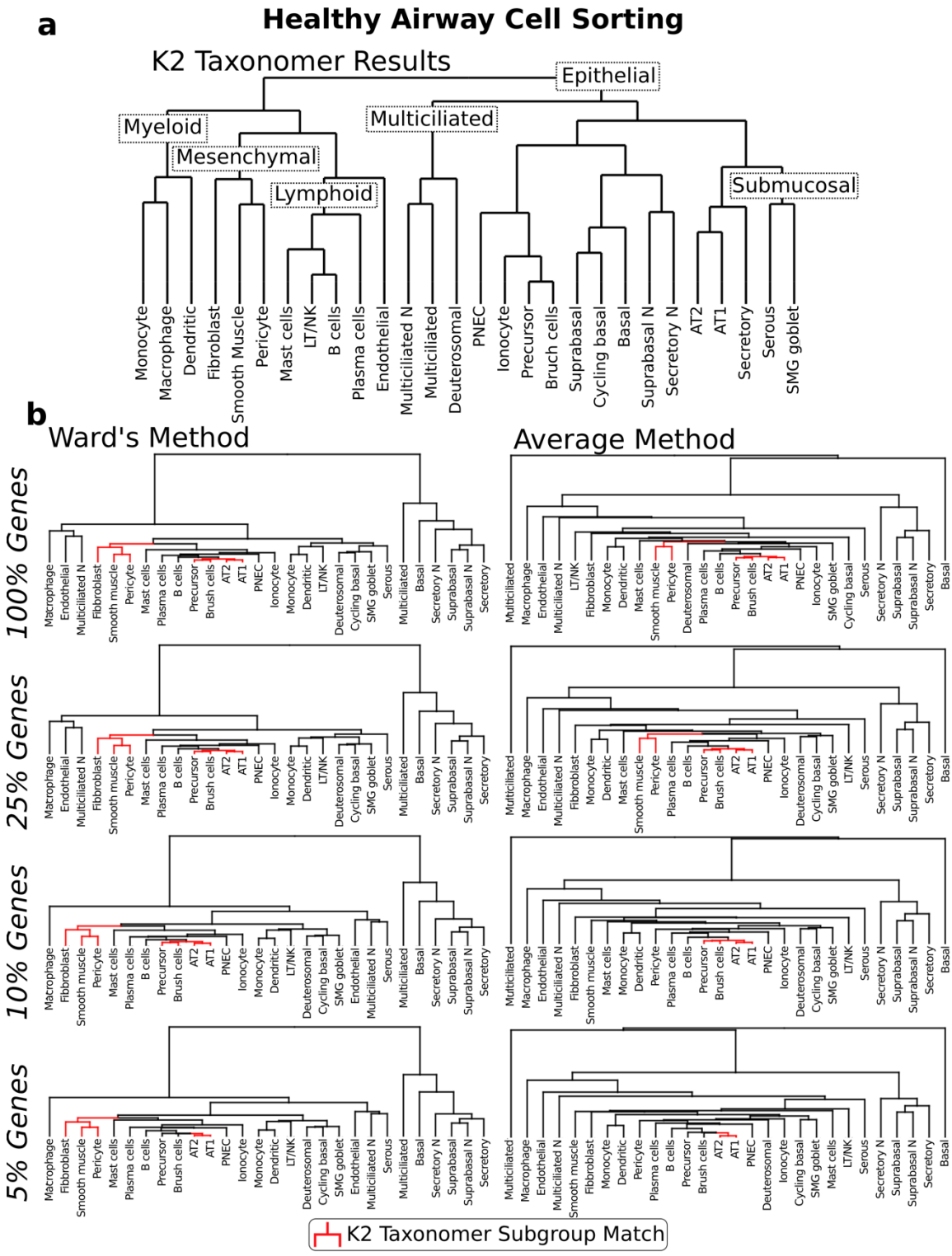

a
Healthy Airway Cell Sorting
K2 Taxonomer Results
Epithelial
Myeloid
Multiciliated
Mesenchymal
Lymphoid
Submucosal
Monocyte
Macrophage
Dendritic
Fibroblast
Smooth Muscle
Pericyte
Mast cells
LT/NK
B cells
Plasma cells
Endothelial
Multiciliated N
Multiciliated
Deuterosomal
PNEC
Ionocyte
Precursor
Bruch cells
Suprabasal
Cycling basal
Basal
Suprabasal N
Secretory N
AT2
AT1
Secretory
Serous
SMG goblet
b
Ward's Method
Average Method
100% Genes
25% Genes
10% Genes
5% Genes
K2 Taxonomer Subgroup Match

Figure A.18: Additional subgrouping results of healthy airway cell types from scRNAseq data

A) *K2Taxonomer* results with six identified lineage subgroups.
B) Ward's (left) and average (right) agglomerative clustering results for selected analyses performed on different subsets of the total number of features.

BIBLIOGRAPHY

1. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

2. Temin, H. M. & Mizutani, S. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature* **226**, 1211–1213 (1970).

3. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441–448 (1975).

4. Brady, G., Barbara, M. & Iscove, N. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods in Molecular and Cellular Biology* 17–75 (1990).

5. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675–1680 (1996).

6. García-Ortega, L. F. & Martínez, O. How Many Genes Are Expressed in a Transcriptome? Estimation and Results for RNA-Seq. *PLoS ONE* **10**, e0130262 (2015).

7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63 (2009).

8. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature Methods* **7**, 909–912 (2010).

9. Bryant, D. W., Priest, H. D. & Mockler, T. C. Detection and Quantification of Alternative Splicing Variants Using RNA-seq. in *RNA Abundance Analysis* (eds. Jin, H. & Gassmann, W.) vol. 883 97–110 (Humana Press, 2012).

10. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems* **2**, 239–250 (2016).

11. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Computational Biology* **13**, e1005457 (2017).

12. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).

13. METABRIC Group *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

14. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

15. Wang, L. *et al.* A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. *PLoS ONE* **6**, e26426 (2011).

16. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

17. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

18. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

19. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**, 599–604 (2018).

20. Shishkin, A. A. *et al.* Simultaneous generation of many RNA-seq libraries in a single reaction. *Nature Methods* **12**, 323–325 (2015).

21. Asmann, Y. W. *et al.* 3’ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics* **10**, 531 (2009).

22. Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C. & Afshari, C. A. Microarrays and toxicology: The advent of toxicogenomics. *Molecular Carcinogenesis* **24**, 153–159 (1999).

23. Scruggs, C. E., Ortolano, L., Schwarzman, M. R. & Wilson, M. P. The role of chemical policy in improving supply chain knowledge and product safety. *Journal of Environmental Studies and Sciences* **4**, 132–141 (2014).

24. Li, A. *et al.* The Carcinogenome Project: In Vitro Gene Expression Profiling of Chemical Perturbations to Predict Long-Term Carcinogenicity. *Environmental Health Perspectives* **127**, 047002 (2019).

25. Schwartz, M. W. *et al.* Obesity Pathogenesis: An Endocrine Society Scientific Statement. *Endocrine Reviews* **38**, 267–296 (2017).

26. Park, Y.-W. *et al.* The Metabolic Syndrome. *Archives of Internal Medicine* **163**, 427 (2003).

27. GBD 2015 Obesity Collaborators *et al.* Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *The New England Journal of Medicine* **377**, 13–27 (2017).

28. Rosen, E. D. & Spiegelman, B. M. Adipocytes as regulators of energy balance and glucose homeostasis. *Nature* **444**, 847–853 (2006).

29. Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPARg2, a lipid-activated transcription factor. *Cell* **79**, 1147–1156 (1994).

30. Regnier, S. M. *et al.* Tributyltin differentially promotes development of a phenotypically distinct adipocyte. *Obesity* **23**, 1864–1871 (2015).

31. Kim, S., Li, A., Monti, S. & Schlezinger, J. J. Tributyltin induces a transcriptional response without a brite adipocyte signature in adipocyte models. *Archives of Toxicology* **92**, 2859–2874 (2018).

32. Shoucri, B. M., Hung, V. T., Chamorro-García, R., Shioda, T. & Blumberg, B. Retinoid X Receptor Activation During Adipogenesis of Female Mesenchymal Stem Cells Programs a Dysfunctional Adipocyte. *Endocrinology* **159**, 2863–2883 (2018).

33. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).

34. Igarashi, Y. *et al.* Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research* **43**, D921–D927 (2015).

35. Ganter, B., Snyder, R. D., Halbert, D. N. & Lee, M. D. Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix ® database. *Pharmacogenomics* **7**, 1025–1044 (2006).

36. Wang, L. *et al.* A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. *PLoS ONE* **6**, e26426 (2011).

37. Hou, Z. *et al.* A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Scientific Reports* **5**, (2015).

38. Shishkin, A. A. *et al.* Simultaneous generation of many RNA-seq libraries in a single reaction. *Nature Methods* **12**, 323–325 (2015).

39. Asmann, Y. W. *et al.* 3’ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics* **10**, 531 (2009).

40. Lundberg, A. S. *et al.* Immortalization and transformation of primary human airway epithelial cells by gene transfer. *Oncogene* **21**, 4577–4586 (2002).

41. Morris, L. G. T. *et al.* Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nature Genetics* **45**, 253–261 (2013).

42. Zhao, R., Choi, B. Y., Lee, M.-H., Bode, A. M. & Dong, Z. Implications of Genetic and Epigenetic Alterations of CDKN2A (p16 INK4a) in Cancer. *EBioMedicine* **8**, 30–39 (2016).

43. Zimta, A.-A. *et al.* The Role of Nrf2 Activity in Cancer Development and Progression. *Cancers* **11**, 1755 (2019).

44. Yang, Y., Lu, T., Li, Z. & Lu, S. FGFR1 regulates proliferation and metastasis by targeting CCND1 in FGFR1 amplified lung cancer. *Cell Adhesion & Migration* **14**, 82–95 (2020).

45. Jones, M. R. *et al.* Successful targeting of the NRG1 pathway indicates novel treatment strategy for metastatic cancer. *Annals of Oncology* **28**, 3092–3097 (2017).

46. Karakas, B., Bachman, K. E. & Park, B. H. Mutation of the PIK3CA oncogene in human cancers. *British Journal of Cancer* **94**, 455–459 (2006).

47. Mur, C. Cigarette smoke concentrate increases 8-epi-PGF2a and TGFb1 secretion in rat mesangial cells. *Life Sciences* **75**, 611–621 (2004).

48. Hardonnière, K. *et al.* The environmental carcinogen benzo[a]pyrene induces a Warburg-like metabolic reprogramming dependent on NHE1 and associated with cell survival. *Scientific Reports* **6**, 30776 (2016).

49. Xue, J., Yang, S. & Seng, S. Mechanisms of Cancer Induction by Tobacco-Specific NNK and NNN. *Cancers* **6**, 1138–1156 (2014).

50. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, (2014) doi:10.1101/003236.

51. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

52. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

54. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems* **2**, 239–250 (2016).

55. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25 (2010).

56. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).

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

58. Beane, J. *et al.* Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biology* **8**, R201 (2007).

59. Spira, A. *et al.* 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**, 10143–10148 (2004).

60. Kansanen, E., Kuosmanen, S. M., Leinonen, H. & Levonen, A.-L. The Keap1-Nrf2 pathway: Mechanisms of activation and dysregulation in cancer. *Redox Biology* **1**, 45–49 (2013).

61. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics* **48**, 607–616 (2016).

62. Morrissy, A. S. *et al.* Next-generation tag sequencing for cancer gene expression profiling. *Genome Research* **19**, 1825–1835 (2009).

63. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nature Methods* **11**, 25–27 (2014).

64. Sandberg, R. & Larsson, O. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics* **8**, 48 (2007).

65. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports* **6**, (2016).

66. Adair-Kirk, T. L. *et al.* Distal Airways in Mice Exposed to Cigarette Smoke: Nrf2-Regulated Genes Are Increased in Clara Cells. *American Journal of Respiratory Cell and Molecular Biology* **39**, 400–411 (2008).

67. Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology* **16**, (2015).

68. Xiong, Y. *et al.* A Comparison of mRNA Sequencing with Random Primed and 3′-Directed Libraries. *Scientific Reports* **7**, (2017).

69. Heindel, J. J. *et al.* Metabolism disrupting chemicals and metabolic disorders. *Reproductive Toxicology* **68**, 3–33 (2017).

70. Farmer, S. R. Transcriptional control of adipocyte formation. *Cell Metabolism* **4**, 263–273 (2006).

71. Gumbilai, V. *et al.* Fat Mass Reduction With Adipocyte Hypertrophy and Insulin Resistance in Heterozygous PPARg Mutant Rats. *Diabetes* **65**, 2954–2965 (2016).

72. He, W. *et al.* Adipose-specific peroxisome proliferator-activated receptor knockout causes insulin resistance in fat and liver but not in muscle. *Proceedings of the National Academy of Sciences* **100**, 15712–15717 (2003).

73. Jiang, Y., Berry, D. C., Tang, W. & Graff, J. M. Independent Stem Cell Lineages Regulate Adipose Organogenesis and Adipose Homeostasis. *Cell Reports* **9**, 1007–1022 (2014).

74. O'Donnell, P. E. *et al.* Lipodystrophy, Diabetes and Normal Serum Insulin in PPARg-Deficient Neonatal Mice. *PLOS ONE* **11**, e0160636 (2016).

75. Zhang, J. *et al.* Selective disruption of PPAR 2 impairs the development of adipose tissue and insulin sensitivity. *Proceedings of the National Academy of Sciences* **101**, 10703–10708 (2004).

76. Banks, A. S. *et al.* An ERK/Cdk5 axis controls the diabetogenic actions of PPARg. *Nature* **517**, 391–395 (2014).

77. Choi, J. H. *et al.* Anti-diabetic drugs inhibit obesity-linked phosphorylation of PPARg by Cdk5. *Nature* **466**, 451–456 (2010).

78. Qiang, L. *et al.* Brown Remodeling of White Adipose Tissue by SirT1-Dependent Deacetylation of Pparg. *Cell* **150**, 620–632 (2012).

79. Burgermeister, E. *et al.* A Novel Partial Agonist of Peroxisome Proliferator-Activated Receptor-Gamma (PPARg) Recruits PPARg-Coactivator-1a, Prevents Triglyceride Accumulation, and Potentiates Insulin Signaling in Vitro. *Molecular Endocrinology* **20**, 809–830 (2006).

80. Feige, J. N. *et al.* The Endocrine Disruptor Monoethyl-hexyl-phthalate Is a Selective Peroxisome Proliferator-activated Receptor Gamma Modulator That Promotes Adipogenesis. *Journal of Biological Chemistry* **282**, 19152–19166 (2007).

81. Ohno, H., Shinoda, K., Spiegelman, B. M. & Kajimura, S. PPARg agonists Induce a White-to-Brown Fat Conversion through Stabilization of PRDM16 Protein. *Cell Metabolism* **15**, 395–404 (2012).

82. Villanueva, C. J. *et al.* Adipose Subtype-Selective Recruitment of TLE3 or Prdm16 by PPARg Specifies Lipid Storage versus Thermogenic Gene Programs. *Cell Metabolism* **17**, 423–435 (2013).

83. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine* **373**, 895–907 (2015).

84. Sidossis, L. & Kajimura, S. Brown and beige fat in humans: thermogenic adipocytes that control energy and glucose homeostasis. *Journal of Clinical Investigation* **125**, 478–486 (2015).

85. Timmons, J. & Pedersen, B. The Importance of Brown Adipose Tissue. *The New England Journal of Medicine* **361**, 415–421 (2009).

86. Kim, S. *et al.* Triphenyl phosphate is a selective PPARg modulator that does not induce brite adipogenesis in vitro and in vivo. *bioRxiv* (2019) doi:10.1101/626390.

87. Kavlock, R. *et al.* Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management. *Chemical Research in Toxicology* **25**, 1287–1302 (2012).

88. Auerbach, S. *et al.* Prioritizing Environmental Chemicals for Obesity and Diabetes Outcomes Research: A Screening Approach Using ToxCast^{TM} High-Throughput Data. *Environmental Health Perspectives* **124**, 1141–1154 (2016).

89. Janesick, A. S. *et al.* On the Utility of ToxCast^{TM} and ToxPi as Methods for Identifying New Obesogens. *Environmental Health Perspectives* **124**, 1214–1226 (2016).

90. Pereira-Fernandes, A. *et al.* Toxicogenomics in the 3T3-L1 Cell Line, a New Approach for Screening of Obesogenic Compounds. *Toxicological Sciences* **140**, 352–363 (2014).

91. Lee, M.-J. & Fried, S. K. Optimal Protocol for the Differentiation and Metabolic Analysis of Human Adipose Stromal Cells. in *Methods in Enzymology* 49–65 (Elsevier, 2014). doi:10.1016/b978-0-12-800280-3.00004-9.

92. Xiong, Y. *et al.* A Comparison of mRNA Sequencing with Random Primed and 3′-Directed Libraries. *Scientific Reports* **7**, 14626 (2017).

93. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47 (2015).

94. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

95. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58**, 236–244 (1963).

96. Breiman, L. Random Forests in Machine Learning 5–32 (Springer, 2001).

97. Civelek, M. *et al.* Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. *The American Journal of Human Genetics* **100**, 428–443 (2017).

98. Liu, P., Ma, F., Lou, H. & Liu, Y. The utility of fat mass index vs. body mass index and percentage of body fat in the screening of metabolic syndrome. *BMC Public Health* **13**, 629 (2013).

99. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research* **29**, 45e–45 (2001).

100. Soukas, A., Socci, N. D., Saatkamp, B. D., Novelli, S. & Friedman, J. M. Distinct Transcriptional Profiles of Adipogenesisin Vivoandin Vitro. *Journal of Biological Chemistry* **276**, 34167–34174 (2001).

101. Berger, J. *et al.* Thiazolidinediones produce a conformational change in peroxisomal proliferator-activated receptor-gamma: binding and activation correlate with antidiabetic actions in db/db mice. *Endocrinology* **137**, 4189–4195 (1996).

102. Fang, M., Webster, T. F., Ferguson, P. L. & Stapleton, H. M. Characterizing the Peroxisome Proliferator-Activated Receptor (PPARg) Ligand Binding Potential of Several Major Flame Retardants, Their Metabolites, and Chemical Mixtures in House Dust. *Environmental Health Perspectives* **123**, 166–172 (2015).

103. Riu, A. *et al.* Peroxisome Proliferator-Activated Receptor γ Is a Target for Halogenated Analogs of Bisphenol A. *Environmental Health Perspectives* **119**, 1227–1232 (2011).

104. Wang, H., Liu, L., Lin, J. Z., Aprahamian, T. R. & Farmer, S. R. Browning of White Adipose Tissue with Roscovitine Induces a Distinct Population of UCP1+ Adipocytes. *Cell Metabolism* **24**, 835–847 (2016).

105. Temkin, A. M. *et al.* Effects of Crude Oil/Dispersant Mixture and Dispersant Components on PPARg Activity in Vitro and in Vivo : Identification of Dioctyl Sodium Sulfosuccinate (DOSS; CAS #577-11-7) as a Probable Obesogen. *Environmental Health Perspectives* **124**, 112–119 (2016).

106. Hu, P. *et al.* Effects of Parabens on Adipocyte Differentiation. *Toxicological Sciences* **131**, 56–70 (2012).

107. Hurst, C. H. & Waxman, D. J. Activation of PPAR and PPAR by Environmental Phthalate Monoesters. *Toxicological Sciences* **74**, 297–308 (2003).

108. Kassotis, C. D. *et al.* Characterization of Adipogenic Chemicals in Three Different Cell Culture Systems: Implications for Reproducibility Based on Cell Source and Handling. *Scientific Reports* **7**, (2017).

109. Marcon, B. H. *et al.* Downregulation of the protein synthesis machinery is a major regulatory event during early adipogenic differentiation of human adipose-derived stromal cells. *Stem Cell Research* **25**, 191–201 (2017).

110. Ito, M., Nagasawa, M., Hara, T., Ide, T. & Murakami, K. Differential roles of CIDEA and CIDEC in insulin-induced anti-apoptosis and lipid droplet formation in human adipocytes. *Journal of Lipid Research* **51**, 1676–1684 (2010).

111. Janesick, A. S. & Blumberg, B. Obesogens: an emerging threat to public health. *American Journal of Obstetrics and Gynecology* **214**, 559–565 (2016).

112. Duncan, H., Abad-Somovilla, A., Abad-Fuentes, A., Agulló, C. & Mercader, J. V. Immunochemical rapid determination of quinoxyfen, a priority hazardous pollutant. *Chemosphere* **211**, 302–307 (2018).

113. Kannan, K. *et al.* Polycyclic musk compounds in higher trophic level aquatic organisms and humans from the United States. *Chemosphere* **61**, 693–700 (2005).

114. HERA. Human & Environmental Risk Assessment on ingredients of Household Cleaning Products Polycyclic musks AHTN (CAS 1506-02-1) and HHCB (CAS 1222-05-05). (2004).

115. Cornier, M.-A. *et al.* The Metabolic Syndrome. *Endocrine Reviews* **29**, 777–822 (2008).

116. Jensen, M. D. Role of Body Fat Distribution and the Metabolic Complications of Obesity. *The Journal of Clinical Endocrinology & Metabolism* **93**, s57–s63 (2008).

117. Lee, J. J. *et al.* Upper Body Subcutaneous Fat Is Associated with Cardiometabolic Risk Factors. *The American Journal of Medicine* **130**, 958–966.e1 (2017).

118. Zuriaga, M. A., Fuster, J. J., Gokce, N. & Walsh, K. Humans and Mice Display Opposing Patterns of "Browning" Gene Expression in Visceral and Subcutaneous White Adipose Tissue Depots. *Frontiers in Cardiovascular Medicine* **4**, (2017).

119. Frayn, K. N. & Karpe, F. Regulation of human subcutaneous adipose tissue blood flow. *International Journal of Obesity* **38**, 1019–1026 (2013).

120. Yang, X., Enerbäck, S. & Smith, U. Reduced Expression of FOXC2 and Brown Adipogenic Genes in Human Subjects with Insulin Resistance. *Obesity Research* **11**, 1182–1191 (2003).

121. Chrisman, I. M. *et al.* Defining a conformational ensemble that directs activation of PPARg. *Nature Communications* **9**, (2018).

122. Choi, J. H. *et al.* Antidiabetic actions of a non-agonist PPARg ligand blocking Cdk5-mediated phosphorylation. *Nature* **477**, 477–481 (2011).

123. Seale, P. *et al.* Transcriptional Control of Brown Fat Determination by PRDM16. *Cell Metabolism* **6**, 38–54 (2007).

124. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

125. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet* **383**, 999–1008 (2014).

126. Takacs, M. L. & Abbott, B. D. Activation of Mouse and Human Peroxisome Proliferator–Activated Receptors (a, b/d, g) by Perfluorooctanoic Acid and Perfluorooctane Sulfonate. *Toxicological Sciences* **95**, 108–117 (2006).

127. Fang, M., Webster, T. F. & Stapleton, H. M. Effect-Directed Analysis of Human Peroxisome Proliferator-Activated Nuclear Receptors (PPARg1) Ligands in Indoor Dust. *Environmental Science & Technology* **49**, 10065–10073 (2015).

128. Hiromori, Y. *et al.* Ligand Activity of Group 15 Compounds Possessing Triphenyl Substituent for the RXR and PPARg Nuclear Receptors. *Biological & Pharmaceutical Bulletin* **39**, 1596–1603 (2016).

129. Cano-Sancho, G., Smith, A. & Merrill, M. A. L. Triphenyl phosphate enhances adipogenic differentiation, glucose uptake and lipolysis via endocrine and noradrenergic mechanisms. *Toxicology in Vitro* **40**, 280–288 (2017).

130. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580–585 (2013).

131. Gao, G. F. *et al.* Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Systems* **9**, 24-34.e10 (2019).

132. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

133. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* (2019) doi:10.1101/563866.

134. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

135. Raj-Kumar, P.-K. *et al.* PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Scientific Reports* **9**, 7956 (2019).

136. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**, 1350–1356 (2015).

137. Ma, X. *et al.* Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection. *Scientific Reports* **9**, 7073 (2019).

138. Wang, W.-H., Xie, T.-Y., Xie, G.-L., Ren, Z.-L. & Li, J.-M. An Integrated Approach for Identifying Molecular Subtypes in Human Colon Cancer Using Gene Expression Data. *Genes* **9**, 397 (2018).

139. Aine, M., Eriksson, P., Liedberg, F., Sjödahl, G. & Höglund, M. Biological determinants of bladder cancer gene expression subtypes. *Scientific Reports* **5**, 10957 (2015).

140. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**, 1141 (2018).

141. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

142. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).

143. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425 (1987).

144. Hughes, A. L. & Friedman, R. A phylogenetic approach to gene expression data: evidence for the evolutionary origin of mammalian leukocyte phenotypes. *Evolution & Development* **11**, 382–390 (2009).

145. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37**, 547–554 (2019).

146. Tritschler, S. *et al.* Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, dev170506 (2019).

147. Hossen, B. Methods for Evaluating Agglomerative Hierarchical Clustering for Gene Expression Data: A Comparative Study. *Computational Biology and Bioinformatics* **3**, 88 (2015).

148. Zurauskienė, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140 (2016).

149. Senabouth, A. *et al.* ascend: R package for analysis of single-cell RNA-seq data. *GigaScience* **8**, giz087 (2019).

150. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).

151. Wagstaff, K., Cardie, C., Rogers, S. & Schrödl, S. Constrained K-means Clustering with Background Knowledge. in *International Conference on Machine Learning* 577–584 (2001).

152. Deprez, M. *et al. A single-cell atlas of the human healthy airways*. *bioRxiv* (2019) doi:10.1101/2019.12.21.884759.

153. Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab) *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nature Medicine* **24**, 986–993 (2018).

154. Ricciardi, S. *et al.* The Translational Machinery of Human CD4+ T Cells Is Poised for Activation and Controls the Switch from Quiescence to Metabolic Remodeling. *Cell Metabolism* **28**, 895-906.e5 (2018).

155. Araki, K. *et al.* Translation is actively regulated during the differentiation of CD8+ effector T cells. *Nature Immunology* **18**, 1046–1057 (2017).

156. Rousseeuw, P. J. & Croux, C. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association* **88**, 1273–1283 (1993).

157. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

158. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91–118 (2003).

159. Hamann, U. Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen. *Willdenowia* 639–768.

160. Yule, G. U. On the methods of measuring the association between two variables. *Journal of the Royal Statistical Society* **75**, 576–642 (1912).

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

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

163. Baker, F. B. Stability of Two Hierarchical Grouping Techniques Case 1: Sensitivity to Data Errors. *Journal of the American Statistical Association* **69**, 440 (1974).

164. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications* **7**, 11479 (2016).

165. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery* **2**, 401–404 (2012).

166. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* **27**, 1160–1167 (2009).

167. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

168. Erman, N., Korosec, A. & Suklan, J. Performance of Selected Agglomerative Hierarchical Clustering Methods. *Innovative Issues and Approaches in Social Sciences* **8**, 180–204 (2015).

169. Kim, H. & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495–1502 (2007).

170. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9**, 284 (2018).

171. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

172. Tang, W. *et al.* bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics* btz726 (2019) doi:10.1093/bioinformatics/btz726.

173. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

174. Winslow, S., Leandersson, K., Edsjö, A. & Larsson, C. Prognostic stromal gene signatures in breast cancer. *Breast Cancer Research* **17**, 23 (2015).

175. Venet, D., Dumont, J. E. & Detours, V. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLOS Computational Biology* **7**, e1002240 (2011).

176. Torang, A., Gupta, P. & Klinke, D. J. An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets. *BMC Bioinformatics* **20**, 433 (2019).

177. Pittenger, M. F. *et al.* Mesenchymal stem cell perspective: cell biology to clinical progress. *Nature Partner Journals Regenerative Medicine* **4**, 22 (2019).

178. DeNardo, D. G. & Coussens, L. M. Inflammation and breast cancer. Balancing immune response: crosstalk between adaptive and innate immune cells during breast cancer progression. *Breast Cancer Research* **9**, 212 (2007).

179. Zhang, S.-C. *et al.* Clinical Implications of Tumor-Infiltrating Immune Cells in Breast Cancer. *Journal of Cancer* **10**, 6175–6184 (2019).

180. Fang, P. *et al.* Immune cell subset differentiation and tissue inflammation. *Journal of Hematology & Oncology* **11**, 97 (2018).

181. Christiansen, B. Ensemble Averaging and the Curse of Dimensionality. *Journal of Climate* **31**, 1587–1596 (2018).

182. Golubovskaya, V. & Wu, L. Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy. *Cancers* **8**, 36 (2016).

183. Teijaro, J. R. *et al.* Cutting Edge: Tissue-Retentive Lung Memory CD4 T Cells Mediate Optimal Protection to Respiratory Virus Infection. *Journal of Immunology* **187**, 5510–5514 (2011).

184. Wang, Z.-Q. *et al.* CD103 and Intratumoral Immune Response in Breast Cancer. *Clinical Cancer Research* **22**, 6290–6297 (2016).

185. Dziobek, K. *et al.* Analysis of Treg cell population in patients with breast cancer with respect to progesterone receptor status. *Contemporary Oncology* **22**, 236–239 (2018).

186. Merlo, A. *et al.* FOXP3 Expression and Overall Survival in Breast Cancer. *Journal of Clinical Oncology* **27**, 1746–1752 (2009).

187. Bates, G. J. *et al.* Quantification of Regulatory T Cells Enables the Identification of High-Risk Breast Cancer Patients and Those at Risk of Late Relapse. *Journal of Clinical Oncology* **24**, 5373–5380 (2006).

188. Piconese, S., Timperi, E. & Barnaba, V. 'Hardcore' OX40 $^{+}$ immunosuppressive regulatory T cells in hepatic cirrhosis and cancer. *OncoImmunology* **3**, e29257 (2014).

189. Aspeslagh, S. *et al.* Rationale for anti-OX40 cancer immunotherapy. *European Journal of Cancer* **52**, 50–66 (2016).

190. Barna, M. *et al.* Suppression of Myc oncogenic activity by ribosomal protein haploinsufficiency. *Nature* **456**, 971–975 (2008).

191. Mendillo, M. L. *et al.* HSF1 Drives a Transcriptional Program Distinct from Heat Shock to Support Highly Malignant Human Cancers. *Cell* **150**, 549–562 (2012).

192. Santagata, S. *et al.* Tight Coordination of Protein Translation and HSF1 Activation Supports the Anabolic Malignant State. *Science* **341**, 1238303–1238303 (2013).

193. Pelletier, J., Thomas, G. & Volarević, S. Ribosome biogenesis in cancer: new players and therapeutic avenues. *Nature Reviews Cancer* **18**, 51–63 (2018).

194. Reed, E. *et al.* Assessment of a Highly Multiplexed RNA Sequencing Platform and Comparison to Existing High-Throughput Gene Expression Profiling Techniques. *Frontiers in Genetics* **10**, 150 (2019).

195. Grandvalet, Y. Bagging Equalizes Influence. *Machine Learning* **55**, 251–270 (2004).

196. Balk, F. & Ford, R. A. Environmental risk assessment for the polycyclic musks, AHTN and HHCB. *Toxicology Letters* **111**, 81–94 (1999).

197. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, 381–386 (2014).

198. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).

199. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

200. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**, 467–470 (1995).

201. Feder, M. E. & Walser, J.-C. The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology* **18**, 901–910 (2005).

202. Li, H. *et al.* Current trends in quantitative proteomics - an update: Current trends in quantitative proteomics. *Journal of Mass Spectrometry* **52**, 319–341 (2017).

203. Lualdi, M. & Fasano, M. Statistical analysis of proteomics data: A review on feature selection. *Journal of Proteomics* **198**, 18–26 (2019).

204. Forman, B. M. *et al.* 15-Deoxy-Delta. *Cell* **83**, 803–812 (1995).

205. Gadupudi, G., Gourronc, F. A., Ludewig, G., Robertson, L. W. & Klingelhutz, A. J. PCB126 inhibits adipogenesis of human preadipocytes. *Toxicology in Vitro* **29**, 132–141 (2015).

206. Riu, A. *et al.* Peroxisome Proliferator-Activated Receptor Gamma Is a Target for Halogenated Analogs of Bisphenol A. *Environmental Health Perspectives* **119**, 1227–1232 (2011).

207. Kim, J. *et al.* 4,4'-Dichlorodiphenyltrichloroethane (DDT) and 4,4'-dichlorodiphenyldichloroethylene (DDE) promote adipogenesis in 3T3-L1 adipocyte cell culture. *Pesticide Biochemistry and Physiology* **131**, 40–45 (2016).

208. Szeles, L. *et al.* Research Resource: Transcriptome Profiling of Genes Regulated by RXR and Its Permissive and Nonpermissive Partners in Differentiating Monocyte-Derived Dendritic Cells. *Molecular Endocrinology* **24**, 2218–2231 (2010).

209. Schwarz, E. J., Reginato, M. J., Shao, D., Krakow, S. L. & Lazar, M. A. Retinoic acid blocks adipogenesis by inhibiting C/EBPbeta-mediated transcription. *Molecular and Cellular Biology* **17**, 1552–1561 (1997).

210. Yin, L., Yu, K. S., Lu, K. & Yu, X. Benzyl butyl phthalate promotes adipogenesis in 3T3-L1 preadipocytes: A High Content Cellomics and metabolomic analysis. *Toxicology in Vitro* **32**, 297–309 (2016).

211. Ahmed, S. & Atlas, E. Bisphenol S- and bisphenol A-induced adipogenesis of murine preadipocytes occurs through direct peroxisome proliferator-activated receptor gamma activation. *International Journal of Obesity* **40**, 1566–1573 (2016).

212. Chamorro-García, R. *et al.* Bisphenol A Diglycidyl Ether Induces Adipogenic Differentiation of Multipotent Stromal Stem Cells through a Peroxisome Proliferator–Activated Receptor Gamma-Independent Mechanism. *Environmental Health Perspectives* **120**, 984–989 (2012).

213. Erbe, D. V. *et al.* Molecular activation of PPARg by angiotensin II type 1-receptor antagonists. *Vascular Pharmacology* **45**, 154–162 (2006).

214. Yanik, S. C., Baker, A. H., Mann, K. K. & Schlezinger, J. J. Organotins Are Potent Activators of PPARg and Adipocyte Differentiation in Bone Marrow Multipotent Mesenchymal Stromal Cells. *Toxicological Sciences* **122**, 476–488 (2011).

215. Pomatto, V. *et al.* Plasticizers used in food-contact materials affect adipogenesis in 3T3-L1 cells. *The Journal of Steroid Biochemistry and Molecular Biology* **178**, 322–332 (2018).

216. Atanasov, A. G. *et al.* Honokiol: A non-adipogenic PPARg agonist from nature. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1830**, 4813–4819 (2013).

217. Cesario, R. M. *et al.* The Rexinoid LG100754 Is a Novel RXR:PPARg Agonist and Decreases Glucose Levelsin Vivo. *Molecular Endocrinology* **15**, 1360–1369 (2001).

218. Wang, L. *et al.* Natural product agonists of peroxisome proliferator-activated receptor gamma (PPARg): a review. *Biochemical Pharmacology* **92**, 73–89 (2014).

219. Hu, P. *et al.* Effects of Parabens on Adipocyte Differentiation. *Toxicological Sciences* **131**, 56–70 (2012).

220. Berger, J. P. *et al.* Distinct Properties and Advantages of a Novel Peroxisome Proliferator-Activated Protein Gamma Selective Modulator. *Molecular Endocrinology* **17**, 662–676 (2003).

221. Gimble, J. M. *et al.* Peroxisome proliferator-activated receptor-gamma activation by thiazolidinediones induces adipogenesis in bone marrow stromal cells. *Molecular Pharmacology* **50**, 1087–1094 (1996).

222. Pereira-Fernandes, A. *et al.* Evaluation of a Screening System for Obesogenic Compounds: Screening of Endocrine Disrupting Compounds and Evaluation of the PPAR Dependency of the Effect. *PLoS ONE* **8**, e77481 (2013).

223. Muralikumar, S., Vetrivel, U., Narayanasamy, A. & Das, U. N. Probing the intermolecular interactions of PPARg-LBD with polyunsaturated fatty acids and their anti-inflammatory metabolites to infer most potential binding moieties. *Lipids in Health and Disease* **16**, (2017).

224. Lehmann, J. M. *et al.* An Antidiabetic Thiazolidinedione Is a High Affinity Ligand for Peroxisome Proliferator-activated Receptor Gamma (PPARg). *Journal of Biological Chemistry* **270**, 12953–12956 (1995).

225. Carmona, M. C. *et al.* S 26948: a New Specific Peroxisome Proliferator Activated Receptor Modulator With Potent Antidiabetes and Antiatherogenic Effects. *Diabetes* **56**, 2797–2808 (2007).

226. Wauson, E. M. Sodium Arsenite Inhibits and Reverses Expression of Adipogenic and Fat Cell-Specific Genes during in Vitro Adipogenesis. *Toxicological Sciences* **65**, 211–219 (2002).

227. Carmona, M. C. *et al.* Dual effects of sodium tungstate on adipocyte biology: inhibition of adipogenesis and stimulation of cellular oxygen consumption. *International Journal of Obesity* **33**, 534–540 (2009).

228. Lee, G. *et al.* T0070907, a Selective Ligand for Peroxisome Proliferator-activated Receptor Gamma, Functions as an Antagonist of Biochemical and Cellular Activities. *Journal of Biological Chemistry* **277**, 19649–19657 (2002).

229. Yamagishi, S. & Takeuchi, M. Telmisartan is a promising cardiometabolic sartan due to its unique PPAR-Gamma-inducing property. *Medical Hypotheses* **64**, 476–478 (2005).

230. Ljung, B. *et al.* AZ 242, a novel PPARa/g agonist with beneficial effects on insulin resistance and carbohydrate and lipid metabolism in ob/ob mice and obese Zucker rats. *Journal of Lipid Research* **43**, 1855–1863 (2002).

231. Regnier, S. M. *et al.* Dietary Exposure to the Endocrine Disruptor Tolylfluanid Promotes Global Metabolic Dysfunction in Male Mice. *Endocrinology* **156**, 896–910 (2015).

232. Grün, F. *et al.* Endocrine-Disrupting Organotin Compounds Are Potent Inducers of Adipogenesis in Vertebrates. *Molecular Endocrinology* **20**, 2141–2155 (2006).

233. Li, X., Pham, H. T., Janesick, A. S. & Blumberg, B. Triflumizole Is an Obesogen in Mice that Acts through Peroxisome Proliferator Activated Receptor Gamma (PPARg). *Environmental Health Perspectives* **120**, 1720–1726 (2012).

234. Pillai, H. K. *et al.* Ligand Binding and Activation of PPARg by Firemaster ® 550: Effects on Adipogenesis and Osteogenesis in Vitro. *Environmental Health Perspectives* **122**, 1225–1232 (2014).

235. Lambe, K. G. & Tugwood, J. D. A Human Peroxisome-Proliferator-Activated Receptor-gamma is Activated by Inducers of Adipogenesis, Including Thiazolidinedione Drugs. *European Journal of Biochemistry* **239**, 1–7 (1996).

CURRICULUM VITAE