Integrating transcriptomics with genomics for improved discovery of biological pathways of disease

by

H. A. Townsend

B.S., Miami University, 2022

M.S., University of Colorado, 2025

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Molecular, Cellular, & Developmental Biology

2026

Committee Members:

Ed Chuong, Chair

Robin Dowell

Townsend, H. A. (Ph.D., Molecular, Cellular, & Developmental Biology)

Integrating transcriptomics with genomics for improved discovery of biological pathways of disease

Thesis directed by Prof. Robin Dowell

Translation of genetic risk of disease to treatment development benefits from also considering functional data like that capturing gene activity across diverse cell types and conditions, such as transcriptomics. However, this integration is complicated by technical differences between assays, uncertainty in which cell types drive disease pathology, and our limited ability to interpret the regulatory impact of noncoding variants. This thesis develops computational frameworks that address these challenges, enabling stronger mechanistic interpretation of genetic risk. First, I establish methods to integrate heterogeneous transcriptomic datasets, allowing construction of larger, more informative disease cohorts for genomic-transcriptomic integration. This enables more robust predictive modeling and biomarker discovery in ovarian cancer, providing a reliable foundation for downstream genetic analyses. I then benchmark approaches for linking genetic variants to disease-relevant cell types and genes using single-cell RNA sequencing. I show that while disease-relevant cell types can be robustly identified, inferred gene mechanisms from noncoding variants vary widely, highlighting the need for improved regulatory interpretation. Finally, I develop computational methods to link regulatory genetic variants to transcriptional networks in disease-relevant conditions, allowing noncoding variants to be linked to potential drug targets. Applied to lung diseases, these approaches uncover shared and disease-specific regulatory programs, prioritize functional variants, and identify candidate regulatory drivers, several of which were experimentally validated. Together, this work provides an integrated framework for connecting genetic variation to regulatory mechanisms across cell types, advancing our understanding of complex disease and supporting translational discovery.

Dedication

To my family who have supported me despite not understanding what the heck I'm doing.

To my parents-in-law who have adopted me as their own.

To my mom who is one of the strongest women I know.

To my dad who has shown me what it means to trust and rejoice in God, always.

To my little sibling who has given my life meaning beyond science.

To my twin who has stepped through each day with me and who I am so proud of.

To my husband who helps ground and protect me even when I think everything is falling apart.

Above all, to God who enabled me to do this work, created a world for me to explore, and ironically changed a lot of my plans so that I even ended up coming to this University in the first place. Thank you that all of my work and sacrifice pales to the one that you gave for mankind.

Acknowledgements

Thank you to the people who advised me in undergrad and made today possible: Dr. Luis Actis, Dr. Mariah Squire, Dr. Xin Wang, Drs Rebecca and Mitchell Balish and Dr. Abshire. Thank you to the people working behind the scenes at the University who made my graduation possible: Biofrontiers IT, Stephanie Rauscher, Kristin Powell, Katie Larson.

Thank you to the graduate mentors: Dr. Fan Zhang, Dr. Ryan Layer, my committee (Dr. Marrisa Ehringer, Dr. Ed Chuoung, Dr. Ken Krauter, Dr. Aaron Clauset).

Thank you to my collaborators who made computational analyses actually meaningful to the biological and medical community: Dr. Sarah Sasse, Dr. Arnav Gupta, Dr. Tony Gerber, Dr. Ben Bitler.

Thank you to the members of the DnA Lab who truly made my work fun, even when the science wasn't cooperating. Thank you for serving both as mentors and friends. Thank you to Dr. Mary Allen who served as a co-mentor and always made science exciting.

Finally, thank you to Dr. Robin Dowell who treated me as a person she truly cared about. Thank you for helping me explore all the things I wanted, even when it was harder for you. Thank you for guiding me when my brain went 1000mph faster than my mouth or writing. And thank you for showing what it means to consider people first, science second, and politics last (and only because we have to).

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

The advent of high-throughput biological assays combined with improved drug access has brought us to a new age of biomedical research: precision medicine. This type of medicine acknowledges that patients often show significantly different responses to a treatment regimen, even when having the same diagnosed disease and symptoms[168]. This variability can be largely explained by differences in the patient's biology as measurable from high-throughput assays [262, 282, 155, 124]. For example, a single disease may have multiple completely different etiologies, or a patient can have distinct biological baselines leading to adverse or inappropriate drug reactions[168]. In its ideal state, precision medicine uses a patient's biology, measured by high-throughput assays, combined with statistical analyses or machine learning to identify an optimal treatment or alternative drug targets.

Next-generation sequencing has become particularly essential as high-throughput measurements of biology for precision medicine. The full spectrum of DNA and RNA inside a person can now be both identified and quantified for precise, downstream interpretation[167]. Genetics combined with environmental signals, whether at patient- or cellular-scale, determine which genetic sequences are transcribed and thereby activated. After transcription, the RNA itself, or the protein built from the RNA, enables cellular activity and signaling that culminate towards a cellular and eventually organismal phenotype. Therefore, understanding both the genetics and transcription patterns associated with different diseases and the success of treatment has become a key research focus for precision medicine.

1.1 Transcription and Regulatory Networks

Transcription is one of the first regulatory steps by which cells respond to their environment and maintain physiological states (e.g. cell types). Therefore, understanding the underlying transcriptional regulation conferring a response or phenotype can pinpoint the biological mechanisms/pathways of disease, enabling subsequent efforts to disrupt or fix the process [76].

1.1.1 Transcriptional Regulatory Networks

Such transcriptional regulation is often summarized as transcriptional regulatory networks, where a gene is linked to the biological features that regulate it's transcription and possibly the other genes with which it is co-regulated. A gene's transcription levels are established from mainly two biological features: transcription factors (TFs) and enhancers. A TF is a protein that either helps recruit (activator) or prevent (repressor) transcriptional machinery from accessing and subsequently transcribing a gene [200, 108, 114]. A gene's promoter is the genetic sequence where some TFs and transcriptional machinery binds, thereby allowing baseline transcription of a gene[114] (Figure 1.1A). Other genetic sequences, enhancers, largely mitigate this transcription level based on microenvironment or perturbations to confer responses like differentiation and stress-response[284, 215]. The exact mechanism of enhancer-mediated transcription varies, but the general consensus is that an enhancer is a non-promoter genetic sequence that helps recruit or stabilize transcription factors and transcriptional machinery to a gene's promoter for its transcription [146, 207, 186, 302] (Figure 1.1B). One of the more recent (2024) reviews of enhancer mechanisms can be found at [273]. Unlike the promoter sequence which is consistently at the start of a gene, enhancers range broadly in location: within a gene itself or hundreds of kilobases away from a gene [284]. Additionally, multiple enhancers can work cooperatively or antagonistically to regulate transcription of the same or different genes [61, 55]. A classic but simple transcriptional network will therefore, consider the transcription factors activating the enhancer and/or gene, and the enhancers activating the gene [223]. Therefore, a transcriptional regulatory network can be described

as successful linkage of appropriate transcription factor(s) and enhancer(s) to a given gene.

Multiple factors complicate the study of transcriptional regulatory networks. First, a gene's relevant enhancer(s) and transcription factor(s) can significantly change depending on the microenvironment and condition. For example, activation of the same transcription factor in different cell types can lead to the same genes being activated but through completely different enhancers [108, 5, 7] (Figure 1.1B). Therefore, one must clearly define the experimental space (cell type, condition) when identifying these networks. Some population-level studies have been used to identify genetic loci associated with certain gene's expressions (expression quantitative trait loci), but have been largely biased towards promoter rather than enhancer sequences, and are rarely done in disease-relevant cell types and conditions [170]. Therefore, experiments performed in the specific cell type and condition of interest more specifically clarify the relevant transcriptional networks. In this process used to predict networks, different assays measuring RNA levels (summarized as transcriptomics assays) have clear weaknesses and strengths, as described in the next section.

1.1.2 Methods for measuring transcription

Since the advent of next-generation sequencing, there have been several transcriptomics assays developed and this work will focus on some of the most popular contributors to this space. Some of the most commonly used assays include RNA-seq (and its single-cell counterpart scRNA-seq), NanoString, and Microarray. Briefly, NanoString and Microarray use probes targeted towards specific sequences of RNA to allow highly-specific and accurate measurements [70, 40]. Unlike all other transcriptomics approaches described in this work, NanoString does not perform cDNA synthesis from RNA and similar post RNA-extraction steps (e.g. amplification) known to introduce technical biases[262, 70, 40, 172, 155, 44]. This limits such biases, but also means that NanoString can only consider up to 1000 probes in one panel, in contrast to Microarray [293, 70, 220]. Instead of relying on probes, RNA-seq sequences the full set of available RNA, before mapping the sequences back to a genome for quantification[262]. This approach allows more flexible full-genome analysis at the possible cost of lower precision from mapping biases and lower sensitivity from

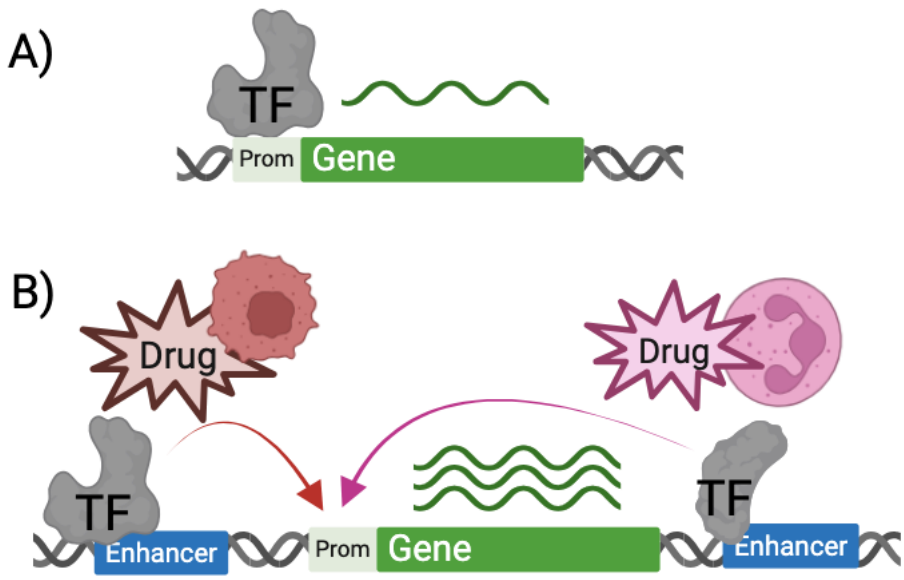

Figure 1.1: **Transcriptional regulatory networks are based on the transcription factors and enhancers that regulate a gene's activity across conditions/cell types**. **A.** A gene has baseline transcription from a transcription factor (TF) binding to its promoter region (Prom). **B.** A gene has increased expression via either the enhancer and transcription factor (TF) on the left with Drug and Celltype A (red), or with the enhancer and TF on the right with Drug and Celltype B (pink). This figure was made in Biorender.

sampling/RNA proportion biases[220]. Finally, scRNA-seq provides sequencing at the single-cell level for downstream clustering of cells with similar transcription profiles, but can only accurately consider the most highly expressed genes [172].

All of these assays are steady-state; measurements represent a balance of both increased RNA from active transcription and decreased RNA from degradation. Mapping out transcriptional-regulatory networks is best performed when we can distinguish between these two processes[223]. For the sake of simplicity, from this point on "expression" will refer to the measurement of steady-state RNA (balance between RNA transcription and degradation) and "transcription" to the active transcription of a gene. Nascent run-on sequencing approaches, like precision run-on (PRO-seq) or global run-on (GRO-seq) sequencing, capture RNA actively being transcribed by RNA polymerase II and have been considered ideal for measuring active transcription levels [155, 44, 278].

Both of these approaches use labeled nucleotides to mark the RNA that is actively transcribed by RNA polymerase at a given timepoint by incorporating biotin-NTPs (PRO-seq)[155] or Br-UTPs (GRO-seq)[44]. Metabolic labeling approaches have similarly attempted to capture nascent transcription, like transient transcriptome sequencing (TT-seq) using 4-thiouridine[216]. However, incubation time with the marked nucleotide is often longer than that of run-on approaches (usually 5-20min for TT-seq compared to 5-min for GRO-seq, so that RNA labeled is less nascent and more a combination of nascent and processed RNA[216, 249]. Therefore, nascent run-on RNA sequencing approaches tend to more consistently capture the accurate dynamics at which regions and nucleotides are actively transcribed in response to perturbations, even between minute-level time frames[60].

Nascent run-on assays like PRO-seq have one additional advantage in predicting transcriptional regulatory networks. Most RNA in a cell is "steady-state" RNA, having undergone RNA-processing (post-transcription) and ready for potential degradation or usage, rather than RNA being actively transcribed. By instead enriching for this nascent RNA, very lowly transcribed and quickly degraded RNAs previously overwhelmed by more stable RNA, can be better considered. Enhancers emit these less stable RNAs and are therefore very poorly (or not at all) captured in steady-state assays, yet well captured in nascent run-on ones [278, 5]. Because of the technical challenge of nascent run-on assays, non-transcriptomics assays have often been used as proxies of enhancer activity like epigentic modifications (H3K27ac ChIP-seq) or open (and therefore likely transcribed) chromatin (ATAC-seq) [145, 164]. Transcription, however, has been shown to be the measurement most strongly associated with an enhancer's activity, with nascent run-on RNA sequencing being particularly adept at pinpointing enhancer transcription [278, 292, 304]. A summary of all of these assays and their pros/cons is listed in Table 1.1. The accurate capture of enhancer functionality means that nascent run-on sequencing cannot only capture enhancer activity but also their target gene's transcription within highly dynamic scenarios. Therefore, recent work has shown great promise in using the correlation of enhancers and genes across perturbations and datasets as a highly informative proxy for enhancer-target gene linkages [223, 5, 138, 132]. In fact, I contributed

to an effort to build a large scale database of nascent run-on transcription which could be used to link enhancers to genes[223].

Table 1.1: Summary of different transcriptomics assays

Element – Impact	NanoString	Microarray	RNA-seq	scRNA-seq	PRO/GRO-seq
Probes – Specific, but pre-defined regions	X	X	-	-	-
Post-processing of RNA – Introduces additional biases (e.g. amplification, cDNA)	-	X	X	X	X
Steady-state (Expression) – RNA levels in the cell (considering both transcription and degradation)	X	X	X	X	-
Transcription – Allows improved capture of enhancer RNAs and measurements for regulatory networks	-	-	-	-	X
Single-cell – Higher granularity but can only confidently consider highly expressed genes	-	-	-	X	-

1.2 Genetic Variation and Linkage to Disease

1.2.1 Measuring genetic variance

Despite transcription varying significantly across tissues, and even spatially adjacent cells, the genomes of cells are almost universally identical within an individual [282]. Instead, genetic variance is found primarily at the level of patient to patient comparison. Genetic variants can include deletions, insertions, complete rearrangements of DNA (e.g. inversion), and finally the primary variant focus for this thesis: single nucleotide changes (single nucleotide polymorphism (SNP)) [282]. This last variant is one of the most common, although this commonality may also be due to it generally being the easiest to identify [47].

Millions of SNPs can be discovered and genotyped in an individual using a wide range of

approaches. SNP arrays probe at a pre-defined set of common SNPs, and then use genetic context to predict the changes of nearby variants[251, 26]. SNPs are rarely inherited independent of one another; instead, alleles at nearby loci are commonly inherited together with recombination patterns and population history creating blocks of linkage disequilibrium (LD)[284, 39]. These LD patterns allow a subset of variants to act as tag SNPs that capture information about other linked variants, reducing the genotyping burden while preserving genomic resolution[251, 26, 213]. Hence, SNP arrays limit costs by only considering common SNPs in a population that are decently representative of the surrounding variation based on LD. In contrast, whole genome sequencing considers the full genome and can include both rare and common SNPs, but at a far greater resource cost [251]. After identifying genetic variants, these SNPs can be tested for association with a given phenotype.

1.2.2 Linking SNPs to disease

One of the most common approaches for identifying variants associated with a phenotype is genome-wide association studies. These studies commonly include hundreds to thousands of individuals and genotype hundreds of thousands to millions of SNPs[251, 26]. For a detailed overview on genome-wide association studies, I suggest [26]. Briefly, each cohort contains a number of cases (with the phenotype) and controls (without the phenotype) or ranges of quantitative trait samples. Then, a variant's allele frequency or genotypic distribution is modeled relative to the phenotype of interest, with regression-based tests typically evaluating association[213, 204, 26]. Due to the large numbers of SNPs/individuals tested and high correlation across tests, extremely stringent significance cutoffs are used (e.g. adjusted p-values $< 5 \times 10^{-8}$). After these tests, to simplify downstream analysis and comparison across datasets, the variant with the strongest statistical association signal within an LD block, referred to as the lead SNP, is focused on [213]. Importantly, the lead SNP is not necessarily the causal variant itself, as is exemplified in Figure 1.2. Additionally, the linkage disequilibrium across SNPs is largely dependent on ancestry (i.e. genetic similarity), so that differing lead and causal SNPs across ancestries may mask a biological feature's relevance[251, 26]. Therefore, functional prediction and mapping of all SNPs to biological sequence

units (e.g. gene) are routinely used to prioritize functional candidates within marked LD-blocks [284, 213]. A review on statistical approaches at identifying causal SNPs in LD-blocks can be found at [213]. SNPs that change amino acid sequences are highly prioritized as this change will often lead to a structural/functional difference to a protein and therefore phenotype(Figure 1.2A). Noncoding SNPs might instead fall within a regulatory element (i.e. enhancer) that, for example, when mutated no longer effectively induces a gene's transcription in response to a drug(Figure 1.2B). Overall, genome-wide studies have proven to be essential for identifying possible biological linkages to disease, but are burdened by weak physiological interpretability, statistical noise, and difficult compatibility across people with largely different genetics (e.g. different ancestries).

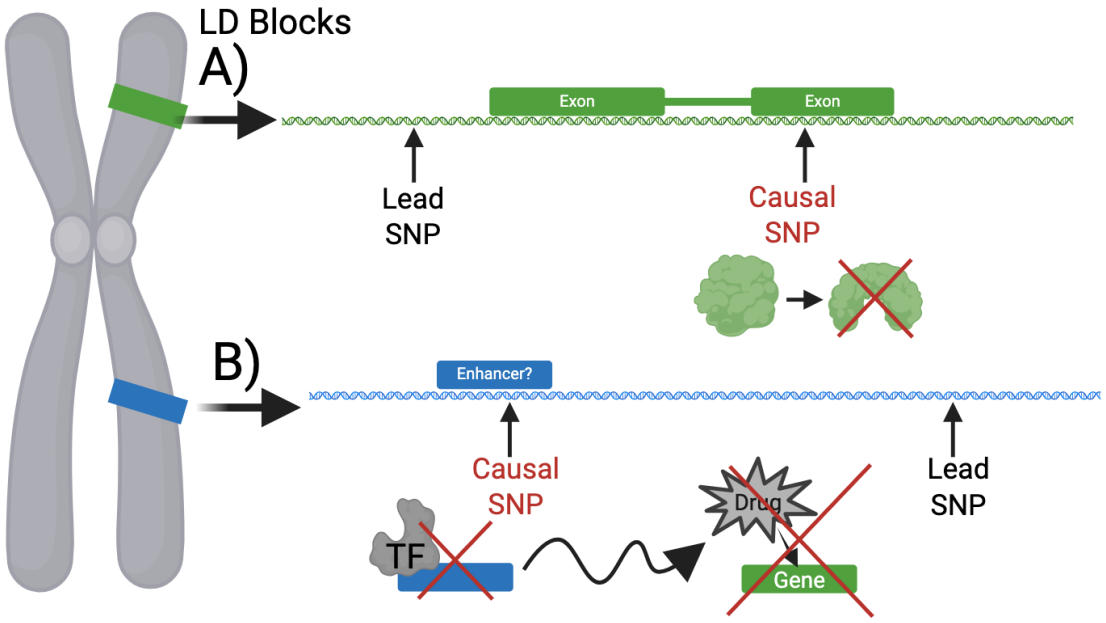

Figure 1.2: **Linkage disequilibrium requires that SNPs are fine-mapped to identify the most probable causal SNPs**. **A.** The causal SNP in a linkage disequilibrium (LD) block is far from the lead SNP and encodes a change in amino acid sequence impacting protein function. **B.** The causal SNP in a second LD block is far from the lead SNP and disrupts the transcription factor binding to an enhancer. This enhancer can therefore no longer effectively induce gene transcription with a drug response. This figure was made in Biorender.

1.3 Motivation for integrating genomics with transcriptomics

Transcriptomics and genomics data provide largely complementary information for understanding disease pathways relevant to developing novel treatments or assigning optimal treatment regimens to a patient.

1.3.1 Transcription clarifies physiological context and relevance

Transcriptomics data helps determine the physiological role and context of genes in a high-throughput approach for downstream mechanistic understanding of disease. Genetics and environment operate together to produce a phenotype, and transcriptomics is often the biological readout of this cooperation. For example, organoids, mice, or cells can be perturbed with different drugs to assess the genes and/or regulatory networks relevant to treatment response (Figure 1.3 Transcriptomics)[210, 5, 99, 106, 76]. Similarly, scRNA-seq can clarify if subgroups of cells exhibit changed transcriptional signal or if the proportions of certain cell types change with disease [286].

Despite providing physiological context, translating transcriptomics to clinical application has many challenges. Transcriptomics can vary largely across environmental factors, time, and in different tissues; this diversity means considering the relevant experimental conditions is essential for direct translation to disease [223, 164, 145]. The high heterogeneity based on tissue and condition also exacerbates the challenges of small cohorts and poor consistency across study findings [224]. Finally, transcriptional patterns associated with a given disease are not necessarily causal of the disease; for example, it is highly difficult to assess whether high levels of inflammatory genes cause autoimmune diseases or are simply consequences [187]. Overall, transcriptomics data can clarify physiology, but only after determining the relevant biological system and condition to consider, and with limited clarity on causality.

1.3.2 Genetics more simply addresses assumptions for causal inference

Genomics data addresses many of the limitations for disease pathway discovery found in transcriptomics, while facing its own challenges. Unlike transcription, the same genome is generally shared across all cells in an individual and unlikely to change as a consequence of disease[282]. Therefore, it is simpler both to capture genetic information (without concern for correct tissue consideration), and argue that the association of a genetic variant with disease is more likely causal rather than simply consequential[204]. Indeed, most clinically successful drugs target genes or proteins whose genetic sequences have been associated with the disease [175, 201].

The massive number of genetic variants and LD, however, has limited researchers to focus on only SNPs achieving extremely strict statistical cutoffs and with obvious functional impacts[26, 213]. This improves chances of a causal SNP being identified and the exact mechanism that can be addressed in drug development (Figure 1.3 Genomics). Such annotation requirements have previously limited focus to variants found in protein-coding regions where the variant's impact on amino acid sequence can be effectively predicted[213]. As necessary as this limitation is for feasible interpretation and target development, it removes a great number of variants that are functional but not easily interpretable, or that are disease-relevant but either rare or have lower effect sizes[234]. Indeed, an estimated 70-90% of disease-associated variants are not considered because they fall in the regulatory elements (e.g. enhancers) discussed previously [159, 284, 39]. Consequently, many diseases still have limited treatment options despite multiple genome-wide association studies [284]. For example, only considering variants passing strict statistical significance cutoffs leaves 95% of the heritability of certain lung-diseases unexplained[41, 298]. Therefore, being able to predict the functionality and biological mechanism of disease-associated variants, whether in coding or non-coding regions, is key to both further clarify which variants are most likely clinically relevant, and allow downstream treatment development.

While genomics can clarify the most relevant variants, it does not provide functional annotations, specifically of noncoding regions, that transcriptomics can. Enhancers are not annotated

across the genome and the relevance of a sequence change within an enhancer is rarely known. Even if the variant changes a nucleotide found within a predicted TF-binding motif, enhancers are enriched in tens and sometimes hundreds of motif instances, and only a small number of these instances might be relevant to a given disease (i.e. condition and cell type) [200, 108]. Since only a small number of these motif instances are used in a cell type and condition context, identifying the functionally active TFs can clarify the function of a disease-associated variant. Similarly, proteins have more historical guidance as drug targets, so linking an enhancer (and therefore SNP) to a gene not only clarifies the SNP's impact on physiology but might support an easier treatment target. As noted previously, nascent run-on RNA sequencing allows transcriptional regulatory networks to be identified and within the actual context of the disease system. Therefore, using nascent run-on RNA sequencing to study the activity of these enhancers (as well as helping to predict the TFs activating them) could be key to filtering and understanding the relevance of these genetic variants. Overall, transcriptomics can clarify the physiological context of SNPs associated with disease while genomics-based studies more succinctly consider causality (Figure 1.3 Combined).

1.4 Addressing key challenges for integrating genetics and transcriptomics

Despite the clear motivation for integrating genomics and transcriptomics to identify biomarkers/pathways of disease, there are several challenges. I address three main challenges for this process in my thesis.

1.4.1 Challenge I: Assay heterogeneity

First, using high-throughput data (e.g. genomics and transcriptomics) to identify pathways or genes associated with disease requires large cohorts of patients. A large cohort helps ensure that biologically relevant subtypes are observed at a high enough number that we can fully characterize the subtypes rather than designate them as noise. These larger cohorts are especially important in transcriptomics data where there is already a high level of heterogeneity simply from cell type proportion (in bulk assays), tissue microenvironment, and environmental conditions. Even when

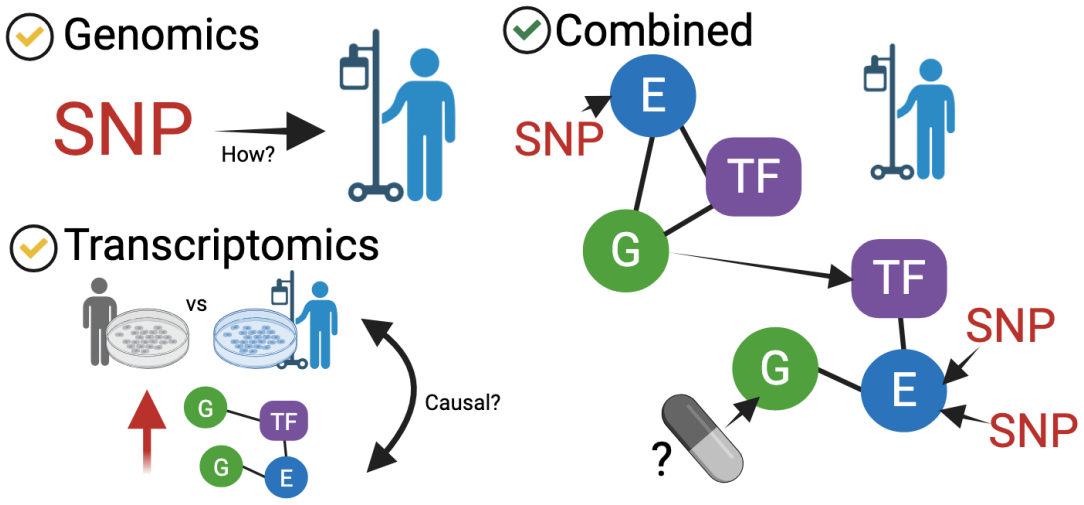

Figure 1.3: **Combining genomics and transcriptomics hypothetically improves discovery of disease pathways**. **Genomics:** SNPs are linked to disease based on genome-wide association studies, but these studies do not tell us what the SNP does. **Transcriptomics:** Physiological differences between disease and non-diseased patients can indicate changes in transcriptional regulatory networks, but this does not indicate causality. **Combined:** Disease-associated SNPs can be linked to enhancers or genes in the transcriptional networks determined in a disease-relevant condition to predict alternative drug targets or mechanisms of disease. This figure was made in Biorender.

considering bulk RNA-seq from the same general tissue and for decently homogeneous diseases, an estimated minimum of 200 samples are needed for machine learning models to predict disease status[224]. Sources of biological heterogeneity are key to consider for proper prediction of biomarkers or using gene expression in a model to predict disease. Transcriptomics data, however, also suffers from large sources of technical heterogeneity that are solely the consequence of different assays being used rather than biological signal [293, 197, 196, 36]. Ideally, smaller cohorts from separate studies can be combined for a larger-scale, and more powerful analysis. However, systematic technical biases can readily overwhelm signal, which has limited meta-analysis that use different assays. Therefore, there is great need to specify the strengths, weaknesses, and systematic technical biases of assays, and perhaps propose methodologies for directly integrating datasets across assays to build larger cohorts. I address these exact needs for transcriptomics data in Chapter 2, demonstrating the feasibility with predictive modeling in ovarian cancer.

1.4.2 Challenge II: Identifying relevant cell types and microenvironment

As hinted at previously, identifying the cell type (or tissue) and condition/microenvironment most relevant to a disease is key to pinpointing the appropriate regulatory networks. Transcriptomics data provides key functional annotation and understanding of a gene's relevance but only if the gene's network under study is reflective of the disease (i.e. proper tissue and condition). Use of patient samples can address disease-relevance, but these samples still contain a mixture of cell-types, all of which might have different regulatory networks for a given gene [222, 164, 145]. Additionally, it is highly difficult to assess causality from transcriptomics data alone. Instead, genome wide association studies identify SNPs with strongest association with a disease and that are more likely to have a causal connection to the disease. Therefore, significant work has gone into integrating disease-associated genetic variants with scRNA-seq to clarify which cell types display the most relevant functionality for possibly causal mechanisms of a given disease [102, 291, 259, 199]. In Chapter 3, I benchmark these methods and identify key weaknesses and growth areas in the field. During this effort, I show both the high variability across approaches linking SNPs to genes, and the impacts of this step on predicting disease mechanisms. Therefore, my last two chapters address the limitations of linking SNPs to molecular mechanisms.

1.4.3 Challenge III: Predicting underlying molecular mechanisms of SNPs

Predicting the molecular mechanisms of genetic variants is key to pinpointing functional variants within an LD block and to link the variant to a druggable target [201, 59]. As hinted at previously, integrating transcriptional regulatory network prediction (e.g. TF-enhancer-gene) from nascent RNA-sequencing could serve as a promising way of doing this, not only for coding but also noncoding variants. As proof of concept, previous work with the Dowell Lab used PRO-seq to identify enhancers responding to clinically relevant perturbations (i.e. drugs) in disease-relevant cell types – thus allowing effective disease-associated variant filtering and annotation for downstream mechanism prediction [68, 209]. However, this process, and the general field, still face

key limitations.

First, the process is heavily reliant on a currently weak ability to annotate and characterize enhancers to link SNPs to changing enhancers in a disease-relevant perturbation. Enhancers are unannotated and commonly overlap one another, so that their annotation has relied on either ChIP peaks (which are historically noisy and large) or more precisely, bidirectional transcription from PRO/GRO-seq[223, 164, 145]. These enhancer midpoints are highly enriched in transcription factor binding sites, and are predicted to be the primary region impacting transcription factor activity and initiation of enhancer activity [223, 145, 164, 120]. However, a focus on midpoint regions means enhancer regions have been misrepresented, minimizing the region over which we count for quantifying expression or to which we map SNPs. The fact that the RNAs of enhancers themselves are not well captured by midpoints was initially deemed less relevant because these transcripts were believed to be nonfunctional. We now know that some full transcripts are key to gene-specific regulation and their SNPs can cause disease phenotypes[171, 161, 233, 89, 135, 214, 177, 275, 18, 190]. Therefore, accurately linking SNPs to enhancers and accurately quantifying which enhancers are expressed under conditions requires that this coordinate space is refined while still addressing overlapping transcription of enhancers. My work with Dr. Jacob Stanley to finalize a Bayesian model predicting gene coordinate space from nascent RNA run-on sequencing provided a framework that might address this complication[232].

Beyond enhancer quantification and mapping, assigning enhancers and genes to their active upstream regulators (e.g. TFs) also faces unique challenges. For example, multiple methods have been developed to predict transcription factors responding to a given perturbation from nascent RNA run-on sequencing or ATAC-seq, but the GC-rich sequence bias of enhancers and many transcription factors consistently leads to false-positive predictions[200, 108]. Introducing additional metrics independent of GC-content bias may improve such analyses to clarify which transcription factor binding motifs in which a SNP is found are most relevant to a given cell type and disease.

To address these complications, in chapter 4, I introduce several new algorithms that optimize how we can link variants to enhancers and identify transcriptonal regulatory networks, partic-

ularly responding enhancers and their TFs, from enhancer-focused assays (e.g. PRO-seq, GRO-seq, ATAC-seq). Then, in chapter 5, I apply these algorithms to identify functionally relevant Asthma- and COPD-associated genetic variants and possible disease mechanisms. My collaborators Drs. Arnav Gupta (National Jewish Health) and Sarah Sasse (University of Kentucky) were able to verify several of these hypotheses through experimental validation.

Chapter 2

Addressing transcriptomic assay heterogeneity for predictive modeling in cancer

2.1 Note to Readers:

I plan to have this work on BiorXiv by April 1, 2026 which will also have all the supplemental tables included. I can send them to you if requested, but in the final thesis I will refer readers to the BiorXiv link as done in Chapter 4.

The "cite lucy paper" refers to a paper in prep that well be uploaded to BioRxiv **before** final thesis submission.

2.2 Contribution Statement

For this work, I was advised by Drs. Robin Dowell, Benjamin Bitler, and Aaron Clauset. Dr. Bitler collected the samples, and processed the samples for sequencing which was then done by external/commercial labs noted below. I performed all downstream computational analyses and found the literature support. I wrote the original draft that was then edited by Drs. Robin Dowell, Benjamin Bitler, and Aaron Clauset.

2.3 Abstract

The clinical heterogeneity of cancer poses a major challenge for precision medicine, particularly when limited cohort sizes and evolving assay platforms impede robust biomarker discovery. Here, we systematically evaluate how to integrate bulk RNA sequencing (RNA-seq), NanoString,

microarray, and single-cell RNA-seq (scRNA-seq) for predictive modeling in cancer. We focus on high-grade serous ovarian carcinoma (HGSOC), as a highly heterogeneous cancer both in terms of biology and assay data.

We first show that using log2 fold-change in patients with matched pre- and post-neoadjuvant chemotherapy samples reduces inter-patient and -assay variability, but is insufficient to overcome platform-specific biases. Microarray and scRNA-seq exhibit systematic biases, while RNA-seq and NanoString show the most promise for direct integration into a single training cohort. To overcome inter-assay limitations, we generate a new HGSOC dataset profiled by both bulk RNA-seq and NanoString, and identify limits of detection and optimal harmonization strategies. Our approaches enable robust integration of independent cohorts for separate and combined RNA-seq- and NanoString-based predictive modeling (test-set AUROCs of 0.81 and 0.86), validated with an external microarray cohort.

We further leverage single-cell data and RNA-seq network-based analyses to provide mechanistic context of predictive model genes. Notably, our models indicate *GBP4* expression is a key predictor of treatment response and biomarker of immune remodeling towards cytotoxicity. Finally, we provide an interactive web portal to facilitate exploration of the datasets and results. Together, this work guides cross-assay harmonization of transcriptomic data and enables improved predictive modeling in heterogeneous cancers.

Statement of Significance: We present a framework for integrating RNA-seq, NanoString, microarray, and single-cell transcriptomic data for predictive modeling, enabling robust biomarker discovery in heterogeneous cancers and identifying *GBP4* as a marker of immune remodeling.

2.4 Introduction

The heterogeneity of diseases like cancer means that a single treatment regimen is often not broadly successful[98]. Precision medicine aims to assign the most promising treatment regimen based on a patient's specific pathology (e.g. tumor microenvironment) as measured in biological data[109]. Machine learning models can then be trained on these data to predict clinical endpoints,

including disease recurrence, therapeutic response, and survival. These models serve a dual purpose: in addition to predicting treatment success, the genes used in the model highlight potential biological mechanisms underlying disease progression and therapy resistance.

These predictive models assume that the data on which they train contain sufficient samples to represent the full heterogeneity of a given disease. However, this is often not the case, particularly for rare diseases and in small datasets[224]. Therefore, analyses that combine smaller cohorts into a single study have often been favored to identify biomarkers less confounded by spurious correlations or batch effects that bias single-cohort studies[98, 148, 202]. While a large, integrated cohort (a mix of smaller cohorts) is highly effective, it requires that all cohorts use comparable measurements (e.g. use the same high-throughput sequencing assays).

Instead, continual technological development has led to data being dispersed across several different assays that measure RNA expression levels of genes in methodologically distinct ways: RNA-sequencing (single-cell (sc) or bulk), microarray, and NanoString. For instance, microarray was one of the first and most successful high-throughput approaches, enabling highly specific quantification even with degraded RNA by using molecular probes built to hybridize to a predetermined subset of genes [40]. Conversely, RNA-seq has become much more popular in recent years, where a sample of RNA is sequenced and then mapped back to a genome for quantification, allowing consideration of any genes and isoforms at the potential cost of specificity and sampling bias[262]. As a cheaper intermediate often favored by clinics, NanoString is a probe-based assay avoids technical biases from post-RNA extraction processing steps at the cost of only considering up to 1,000 genes at a time [293, 70]. While microarray often has larger data suites compared to RNA-seq and NanoString, it has also been shown to have limited dynamic range due to its reliance on fluorescence intensity[297]. Finally, scRNA-seq is the most recent of the four assays, providing insight into cell-state granularity while being limited to confident analysis for only the most highly expressed genes[172]. The technical diversity of these assays has prevented datasets of different assays from being combined for more powerful predictive modeling. Previous work has evaluated the large-scale consistency of these assays in differential expression trends and found reasonable

overlap, but similar studies have not been done for direct cohort combination or predictive modeling [293, 197, 196, 36]. Overall, the question of how to retain as much data as possible for predictive modeling with the rapid advancement of technologies remains unanswered.

Here, we address this question by directly interrogating assay-based biases in predictive modeling for a disease that presents several challenges inherent to effectively answering this question. High-grade serous ovarian carcinoma (HGSOC) is the most common and aggressive subtype of epithelial ovarian carcinoma, and is the fifth leading cause of cancer-death in women [306]. Unfortunately, it is typically diagnosed at advanced stages (III or IV), at which point the 5-year survival rate ranges from 20% to 40%[306]. Despite these sobering statistics, HGSOC has largely standard treatment, including the use of neoadjuvant chemotherapy (NACT), where patients are treated with chemotherapeutics (e.g., carboplatin/cisplatin and paclitaxel) with the goal of reducing the tumor for an increased likelihood of optimal surgical debulking. Given the poor survival rate, several studies have attempted to identify biomarkers of prognosis with NACT treatment, but with limited success and overlap of results[109, 1, 290]. This difference in results can be largely attributed to the known heterogeneity of the disease, further augmented by cohorts being split across multiple transcriptomics assays: NanoString, RNA-seq (sc and bulk), and microarrays. [109, 8, 147, 290, 126, 99, 100, 1]. In addition to this heterogeneity, HGSOC has several datasets of patients with matched measurements pre and post treatment. Such data is valuable as using expression changes after treatment compared to before might alleviate simple scale differences from both subject- and assay-specific biases. Previous studies have shown that even when using such pre-post data, HGSOC patient heterogeneity is still high enough to require increased sample size by combining datasets across assays [cite lucy paper]. Therefore, using pre/post data in HGSOC not only allows clearer evaluation of systematic assay-specific biases but might also provide a solution to combine data across assays for HGSOC.

In this work, we evaluate how to use pre/post data across different transcriptomics assays for predictive modeling in cancer, focusing on their limitations and strengths both individually and combined. First, we assess the limitations and comparability of HGSOC results of these assays

independently, deciphering if pre-post data can alleviate biases. Although cleaner than expression data alone, we show that pre-post data is still unable to fully resolve assay-specific biases. Unlike microarray and scRNA-seq, NanoString and RNA-seq do not show systemic technical biases (e.g., low dynamic range or sensitivity). Therefore, second, we predict limits of detection and introduce preprocessing steps to optimally combine NanoString and RNA-seq. Third, we successfully combine NanoString and RNA-seq for predictive modeling of HGSOC survival, indicating consistent biomarkers also verifiable in both scRNA-seq and microarray data. Finally, we highlight how to use newer, higher-throughput assays RNA-seq and scRNA-seq to predict the mechanistic context of biomarkers with network analyses. Overall, this work provides critical insights and data on harmonizing data between transcriptomics assays for predictive modeling, with a focus on HGSOC.

2.5 Materials and Methods

For the sake of brevity, additional methods and more detailed information can be found in Supplemental Methods.

2.5.1 Cohorts

2.5.1.1 Longitudinal, single-assay combined cohort

Given the ability of longitudinal (e.g. pre/post) data to reduce the impact of patient-specific biases relative to using single data points, we hypothesized pre- and post-NACT data could similarly address assay-specific biases. We predicted that although exact expression levels differ between assays, the change in expression – measured by log2 fold-change (log2FC) – between two samples using the same assay would be consistent, since assay-specific biases would be accounted for by the ratios. Therefore, we collected transcriptomics data for paired pre- and post-NACT samples (i.e., public or in-house) sequenced with microarray, NanoString, or RNA-seq (both bulk and single-cell) for which clinical data were available (Figure 2.1A). Summary statistics for the combined and individual mini-cohorts can be found in Table 2.1. We found one pre/post cohort for microarray

(N=28)[106], and three each for RNA-seq and NanoString (Ns ranging from 15 to 35) [1, 126, 290, 100, 99, 147, 109, 183, 181][cite lucy paper]. Therefore, for downstream analyses, these studies were combined into one mini-cohort labeled EGA. Zhang et al[290] also published paired pre- and post-NACT scRNA-seq samples for 11 patients, 5 of whom matched with bulk RNA-seq. We ended with a combined transcriptomics cohort of 174 patients with platinum-free survival (PFS) or platinum-free interval (PFI), with PFS defined as the time interval between the completion of upfront chemotherapy and disease recurrence via CA125 or imaging (Table 2.1). We then sought the most appropriate approaches to integrate data from the different assays to enable a single, combined cohort for predictive analysis of PFS (Figure 2.1A). This single, combined cohort will from now on be referred to as the "Single-assay cohort." Similarly, a mini-cohort refers to the data coming from one study/assay (i.e. Bitler-NanoString, Manso-NanoString, James-NanoString, Adzib-RNAseq, Jav-RNAseq, EGA-RNAseq, Jim-Sanchez-Microarray, Zhang-scRNAseq).

2.5.1.2 New double-assay post-NACT cohort

To directly interrogate assay biases of RNA-seq and NanoString, we collected a new dataset of 24 FFPE HGSOC post-NACT tumor samples, in which sequential FFPE tumor sections were used for either RNA-seq or NanoString. From now on, samples from this dataset will be referred to as the "Double-assay cohort."

2.5.1.3 Non-longitudinal microarray cohorts

Most HGSOC data and public resources are available as microarray data at a single time point from each patient. This includes two publicly available portals for Ovarian (CSIOVDB) or Serous Ovarian (KMPlot) cancer[240, 80]. These datasets, however, are not specific to HGSOC and did not easily allow us to compare cohorts. Therefore, we also re-analyzed four microarray HGSOC cohorts with overall survival information: Bonome (N=185)[17], Crijns (N=415)[46], Mok (N=53)[162], Yoshishara (N=260)[281].

Table 2.1: **Summary of Key Clinical Data of Longitudinal Cohorts.** Med=Median. Full=Full dataset, N+R=NanoString and RNA-seq dataset combined. Some variables are not available for all patients, so the number of patients with the data and their summary statistics are included and highlighted in grey. Micro=Microarray. PFS = Platinum Free Survival. PFI = Platinum Free Interval. OS = Overall Survival. CRS=Chemo Resistance Score. Articles for the different experiments are found in Methods.

	Full*	**N+R**	**Nanostring**			**RNA-seq**			**Micro**	**scRNA-seq**
			Bitler	Manso	James	Adzib	Jav	EGA	Jim-Sanchez	Zhang
PFS	**168**	**140**	35	17	31	15	20	22	28	8
Med (mths)	12	11	9	18	10	11	14.3	12.3	14	9.8
PFI	**29**	0	0	0	0	0	0	23	0	11
Med (mths)	6.8							6.8		5.9
OS	**109**	**81**	35	0	0	6	20	21	28	8
Med	35	33	34			36	22.9	30	40.5	30
Age	**136**	**108**	35	0	31	4	20	19	28	11
Med	63	63.5	60		63	62.5	62	68	60	67
CRS	**51**	**51**	0	17	0	0	19	16	0	11
Med	2	2		2			2	2		2
Stage	**101**	**73**	0	0	31	0	20	23	28	11
IIIA	1	1			1		0	0	0	0
IIIB	3	3			3		0	0	0	0
IIIC	68	52			23		15	15	16	3
IV	16	4			4		0	0	12	0
IVA	6	6			0		1	5	0	7
IVB	7	7			0		4	3	0	1

*scRNA-seq (Zhang) includes some patients also found in Piet-Zhang so the Full Dataset numbers account for only considering each patient once.

2.5.2 Modeling

2.5.2.1 Predicted Outcome

We used modeling to predict two outcomes. First, a model predicted platinum-free survival (PFS) categories, split between the median of the cohort: PFS $> 12mths$ or $\leq 12mths$. Second, to see if log2FC data could clarify assay, a model predicted whether a sample was sequenced with NanoString or RNA-seq. In all cases, we used the Single-assay cohort for these predictions.

2.5.2.2 Training, Validation, and Hold-out data

Before modeling, a "hold-out" dataset was removed for final testing, including a minimum of 10 patients or about 10% of the total cohort studied and equally representative of each classification and mini-cohort. To ensure performance on the hold-out set was not biased from the patients held having either extreme PFS (therefore likely easier to separate) or PFS closer to 12 months (therefore likely harder to separate), we selected patients based on the following. For RNA-seq vs NanoString prediction, the patients with the third lowest, third highest, and median PFS were used from each mini-cohort (N=18). For PFS prediction, we selected the patients with the third lowest and third highest PFS values within each mini-cohort (N=12). This left 128 patients left for training/validation. When only considering one assay type (e.g. RNA-seq), to maintain a hold out N of 12 rather than 6, we added patients with the median PFS of each category (PFS $> 12mths$ and $\leq 12mths$) (except for the Manso mini-cohort which only had 4 patients with PFS $> 12mths$).

Due to the small sample sizes, we used leave-one-out cross validation (LOOCV) to determine the final number of genes to include in the model (k). The final model parameters were then fit on the full training/validation dataset.

2.5.2.3 Assessing Model Performance

Area under the Receiver Operating Characteristic Curve (AUROC) was primarily used to compare model performance when using different gene numbers in cross-validation. The AUROCs

for each mini-cohort or cohort combinations were also calculated to ensure that the model wasn't clearly biased towards one mini-cohort. Other assessments of model performance are detailed in Supplementary Methods.

2.5.2.4 Feature Selection

Due to small sample sizes, including the full feature list of a shared 770 or all expressed 18,000+ genes would likely cause overfitting. Additionally, assays have different numbers of genes, which could lead to biases in model performance simply due to different dimensionalities across assays.

Therefore, we performed an initial feature selection to confine the feature space considered by a model to a maximum of 25 genes, as is often done in high-dimensional datasets [189, 276]. When predicting PFS, genes were first filtered based on the predicted limits of detection in expression and log2FC, as detailed in Supplementary Methods.

Because the integrated dataset comprised multiple mini-cohorts with differing sample sizes and technical contexts, we implemented two complementary feature selection strategies designed to mitigate cohort-specific biases. All strategies were implemented using training data only:

1) Bootstrapping: Bootstrapping resamples from the cohort to estimate the underlying value in a population despite low sample size. We bootstrapped samples from the training/validation set (with equal number of patients within the two categories: either NanoString and RNA-seq or PFS$>$ $12mths$ and PFS$\leq$ $12mths$) using weights inversely proportional to the sample size of each of the mini-cohorts. Therefore, assays or mini-cohorts with lower sample numbers were about equally represented compared to mini-cohorts with higher sample numbers. With each bootstrapped sample, the top 50 or 100 genes with the highest F-scores (from ANOVA F-tests using python's function f_classif) were recorded. At each bootstrap, for each gene, we updated the cumulative proportion of bootstrapped samples that included it in this top ranking list. This value is hereby referred to as predictive frequency. The genes kept for model training/validation were identified based on those with the top predictive frequency values (details in Supplemental

Methods).

2) Consensus: We also saw if explicitly considering genes found within the top k predictive genes across each mini-cohort would work. For example, there were 21 genes ranked in the top 50 predictive genes of NanoString vs RNAseq, shared across four or more datasets. The exact numbers and genes used for all comparisons for training/testing can be found at the jupyter notebooks RNA-PrePost/Modeling/01_NanoRNAseq_Cons.ipynb and RNA-PrePost/Modeling/03_RNAseqFull_BtspCons.ipynb

2.5.2.5 Models Considered

To assess how robust the predictive power of a feature combination was to model parameters/structure, we considered three different model architectures: logistic LASSO regression, logistic Ridge regression, and support vector machine. In all cases, logistic regression performed better or just as well as the support vector machine approach. Since the regression approaches provided more directly interpretable coefficients, we report these models for the final interpretation of results.

2.6 Results

2.6.1 NanoString and RNA-seq show the greatest potential for direct combination of longitudinal data

2.6.1.1 Pre/post data does not address microarray and scRNA-seq biases

Technical biases of assays may be addressable through preprocessing or instead reflect loss of data that makes them no longer directly comparable to other assays. For example, within-assay scaling is often used to align expression data across assays to comparable scales, but assumes that all assays measure similar dynamic ranges (i.e. the lowest and highest values in each assay should be biologically comparable). Longitudinal-based calculations (like Post vs. Pre NACT log2 fold change (log2FC)) would address scale differences but not data-loss between assays.

We therefore asked what the strengths and weaknesses were of each assay, and whether there

was data loss making the assays no longer directly able to be combined. To do so, we considered how data from different assays aligned when summarized in low-dimensional space from expression or log2FC data, using Principal Component Analysis (PCA) (details in Methods). When considering all assays, we use the shared set of 731 genes, i.e., genes in both NanoString and microarray probe sets and in the most recent RefSeq annotation (GRCh38.p14). Limitations of enforcing a shared feature space are noted in the Supplemental Results. In the case of scRNA-seq and bulk RNA-seq, some patients were sequenced with both assays. Based on previous work, we expected weak correlation between expression levels of bulk RNA-seq and scRNA-seq pseudobulk[231], but asked if log2FC might address such poor correlation.

Using log2FC yielded more consistent measurements across the four assays than with expression levels, yet could still not address some underlying data differences. Expression levels led to patients being clustered based on assay, while use of log2FC between the post and pre samples allowed the assays and cohorts to overlap in low-dimensional space(Figure 2.1B). Applying within-assay scaling improved overlap between assays in low-dimensional space (Appendix Figures A.1A and B). However, as found in previous work[297], the microarray data showed evidence of lower dynamic range: lower log2FCs and interquartile ranges of expression compared to all other assays and cohorts (all adjusted p-values < 0.01 and effect sizes (Cliff's deltas) > 0.33)(Figure 2.1C, Appendix Figure A.1A). Thus, it is unsurprising that, even when using log2FC, microarray data contributed little variance across the combined cohort unless using within-assay scaling (Appendix Figure A.1C). We concluded that within-assay scaling can seem to harmonize data but assumes a comparable dynamic range, which is likely not accurate for microarray data. Therefore, in subsequent analyses, we use unscaled log2FC and consider microarray directionality to validate our predictive model after combining other datasets.

When comparing pseudobulked scRNA-seq and bulk RNA-seq, we observed the expected poor correlation using expression levels (N=13), yet saw similar trends in pre/post log2FCs (N=5) (median spearman coefficients (ρ) of 0.3 and 0.2, respectively) (Figure 2.1D). The resolution of scRNA-seq means that only the most highly expressed genes have sufficient read coverage for

confident downstream analyses [90]. Therefore, we hypothesized this poor correlation was from lowly expressed genes, but the trends were common across all levels of gene expression (Appendix Figure A.2). Ultimately, we observed pseudobulked scRNA-seq biases that limited its potential direct combination with bulk datasets.

2.6.1.2 NanoString and RNA-seq have limited biases influencing clinical predictive modeling

NanoString and RNA-seq showed similar dynamic ranges and patterns in low-dimensional space. To test whether these two assays could be effectively combined using pre-post data, we hypothesized that a predictive model would not be able to distinguish between samples sequenced using RNA-seq and NanoString based on Post/Pre-NACT log2FC alone. We also assessed if assay-biases might impact clinical modeling by checking if genes predictive of assay differences were also predictive of PFS. Specifically, we compared the predictive frequency of genes when predicting NanoString vs RNA-seq or PFS category (proportion of bootstrapped samples where each gene was one of the top 100 ranked predictive features).

First, we successfully trained a model that consistently predicted whether a patient was assayed by NanoString or RNA-seq, regardless of cohorts considered (hold-out AUROC was 0.93; Figure 2.1E) (details in Methods). The model found consistent assay-specific biases, as nine of the twelve genes in the model (including the eight with the highest coefficients) showed the trends between assays maintained across all six mini-cohorts (Appendix Figure A.3). We also found evidence that RNA-seq and NanoString had poor concordance in predictive features of PFS. Predictive frequencies when predicting PFS with NanoString or RNA-seq were more comparable to predictive frequencies when distinguishing NanoString from RNA-seq ($\rho = 0.071, \rho = 0.077$) than to each other ($\rho = 0.028$) (Appendix Figure A.4). Several genes predictive of PFS (predictive frequencies above 0.3 in RNA-seq or NanoString cohorts), were also predictive in distinguishing NanoString and RNA-seq samples (predictive frequencies above 0.3) (Appendix Figure A.4 coloration of dots). These findings suggest that calculating log2FC in longitudinal data is likely insufficient to remove

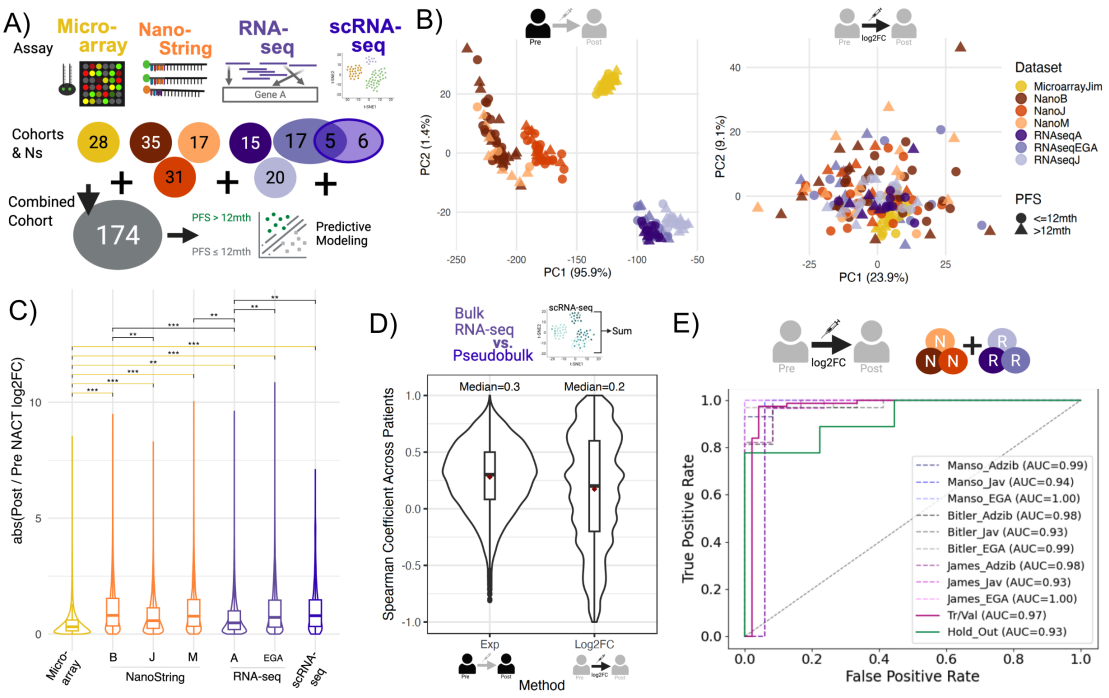

Figure 2.1: **NanoString and RNA-seq show the lowest biases for direction combination of longitudinal cohorts but are still distinguishable**. **A.** Outline of study where we assessed the most appropriate approaches to combine longitudinal data across 4 transcriptomics assays of varying cohort sizes for predictive modeling of PFS. **B.** PCA across all patients/cohorts for the bulk assays using either normalized expression pre-NACT (left) or log2FC between normalized expression post and pre-NACT (right). **C.** Violin plot of absolute values of log2FC between normalized expression post and pre-NACT across the different assays/cohorts. Pairwise comparisons using Wilcoxon-Mann-Whitney tests and minimum effect sizes (measured with Cliff's delta) are shown with those between Microarray highlighted in yellow. Stars indicate comparisons with adjusted p-values below 0.01 (**) and 0.05 (*) with minimum Cliff's deltas of 0.33, or adjusted p-values below 0.001 (***) and a minimum Cliff's delta of 0.474. **D.** Spearman correlation coefficient of gene expression across patients between normalized (CPM) expression levels pre/post NACT or post/pre NACT log2FC for bulk RNA-seq and pseudobulked scRNA-seq sequenced on the same FFPE samples. **E.** Receiver Operator Characteristic Curves of a model trained to predict NanoString samples using log2 Post/Pre NACT fold changes (AUC corresponds to AUROC) from the unmatched dataset. All includes all training/validation samples (N=128) excluding the 24 kept for the Hold out set and remaining cohorts refer to the one of the three NanoString cohorts (i.e. Manso, Bitler, James) combined with one of the three RNA-seq cohorts (i.e. Adzib, Jav, EGA), again excluding the 24 validation samples.

assay-based biases in data.

2.6.2 Differences between NanoString and RNA-seq measurements are largely addressable with data processing

Since NanoString and RNA-seq showed comparable dynamic ranges, we hypothesized that NanoString and RNA-seq could still be effectively combined for predictive modeling by optimizing preprocessing of their data. To test this hypothesis, we collected a new dataset of 24 FFPE HGSOC tumor samples, in which sequential FFPE tumor sections were used for either RNA-seq or NanoString (double-assay cohort) (Figure 2.2A). We evaluated the impact of two preprocessing components: 1) removing lowly expressed genes based on limits of detection, and 2) counting regions in RNA-seq. Normalization approaches had little to no impact and are therefore discussed in Supplemental Results.

In the case of detection limits, a gene might not be expressed sufficiently to be accurately detected, or have such a narrow expression range that there is insufficient variance to accurately estimate covariance (and therefore correlation). Therefore, we gauged the limits of detection of both expression and range for RNA-seq and NanoString harmonization. We estimated cutoffs of RNA-seq and NanoString expression (or ranges) at which we observed a clear (strict) or starting (loose) decline in correlation between RNA-seq and NanoString expression/range (details in Methods; final cutoffs in Supplemental Table 2; Figure 2.2B). For the counting component, an obvious difference between the assays is that while NanoString commonly only considers < 200bp of a gene (via a probe), RNA-seq can consider any gene sequence(Figure 2.1A, Appendix Figure A.5A). Therefore, we evaluated if counting across isoforms (isoform-based counts) or exons (exon-based counts) most closely corresponding to the NanoString probe (the latter expected to resemble probe regions) impacted harmonization (Appendix Figure A.5).

2.6.2.1 Most consistent differences are due to limits of detection

Projection of the double-assay cohort into a low-dimensional space suggested that although absolute values differed between RNA-seq and NanoString, relative gene expression across patients

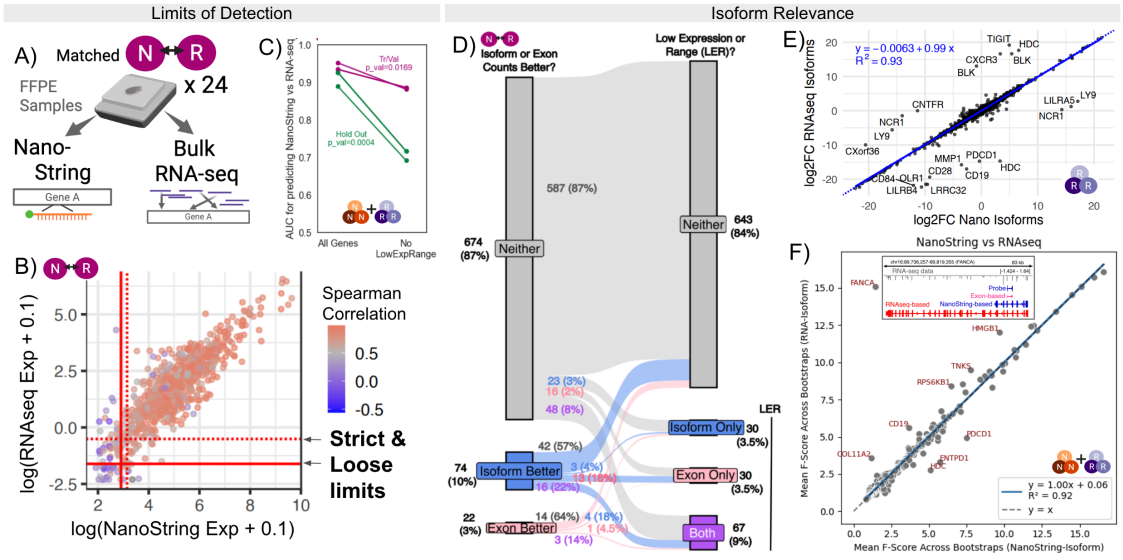

Figure 2.2: **NanoString and RNA-seq assay-specific biases are largely due to limits of detection**. **A.** Sequential samples from 24 FFPE isolations were sequenced with either NanoString and RNA-seq. **B.** Scatter plot of logged Expression from RNA-seq (Isoform-RPKM, y-axis) and NanoString (x-axis) for each gene where the color corresponds to the correlation between the matched datasets. Strict and loose limits of detection are shown as lines. **C.** AUROC of models predicting NanoString vs RNA-seq when allowing consideration of all NanoString genes vs those above the limits of detection for range and expression (No LowExpRange). **D.** Flow graph of the number of genes that first have isoform- or exon-based counting allowing improved harmonization, and second whether the isoform or exon-based counts are below the limits of detection. **E.** Scatterplot of log2FC values of genes using RNA-seq based isoforms (y-axis) vs NanoString based isoforms (x-axis) where genes with differences greater than 5 are noted. **F.** Scatter plot of mean F-score across bootstrapped samples of genes when using RNA-seq based isoforms (y-axis) vs NanoString based isoforms (x-axis). The inset shows the probe-, exon-, NanoString isoform-, and RNaseq isoform- based coordinates for FANCA.

and the dynamic range were maintained (Appendix Figure A.6). Similarly, most genes showed Spearman correlations that were largely unchanged by the using different normalizations or isoform- or exon-based counting (51% $>$ 0.85, 12% $<$ 0.65, 5% 0.65 $<$ x $<$ 0.85), with only 30% of genes showing variability across approaches (details in Methods).

Given counting differences had limited impact on most genes' harmonizations, it is likely that most poor correlations were from technical limitations of detection. Indeed, Spearman correlation coefficients were significantly lower in genes with the lowest expression and/or range quantile (adjusted p-values $<$ 0.001, Appendix Figure A.7). When using cutoffs based on our strict pre-

dicted limits of detection, only 31 genes still had consistent correlation coefficients below 0.7, with *TMEM140*, *RAD50*, *PVRIG*, *HDAC3*, *BAD*, and *ELOB* all with coefficients far below 0.5, possibly partly explained by low expression (< 5 TPM). Removing genes below the limits of detection also restricted the ability to distinguish between NanoString and RNA-seq samples using log2FC (hold-out AUROC from 0.93 to 0.65) (Figure 2.2C). This restricted model seemed to rely primarily on memorizing noise across cohorts rather than learning assay-specific differences, with only two of the twelve model genes exhibiting assay-specific trends (Appendix Figure A.8). Indeed, 10 of the 12 genes included in the original model predicting NanoString vs RNA-seq fell below these limits of detection, whether in the form of expression or range (log2FC below 0.05) (Appendix Figure A.3).

2.6.2.2 Probe-focused counting can harm rather than improve harmonization

There still remained findings indicating that counting approach may impact harmonization. A total of 96 genes (13%) had improved Spearman correlation coefficients between NanoString and RNA-seq when using either exon- or isoform-based counts over the other (increase of > 0.05 in at least two of three normalization approaches) (Figure 2.2D left). Exon-based counts may improve region matching between RNA-seq and NanoString, but they also decrease counts by considering a smaller genomic segment. Indeed, we found that isoform-based counts commonly performed better due to simply increasing data above the limits of detection: 40% of genes for which isoform-based counts are better have expression levels below the limits of detection (Figure 2.2D Isoform-Better to Exon Only and Both). Similarly, genes with better harmonization from exon-based counts have higher exon-based but not isoform-based counts than genes with better isoform-based correlation (adjusted p-value $< 1x10^{-26}$; Appendix Figure A.9). Still, the majority of cases where exon- or isoform-based counting is preferential (64% and 57%, respectively) are not explained by detection limits (Figure 2.2D). However, further evaluation confirmed that probe-focused counting has limited if any improvement in harmonization (Supplemental Results; Appendix Figures A.10, A.11, A.12).

Clinically relevant insights could be lost by restricting RNA-seq analysis to isoforms best matching NanoString probes rather than the most highly expressed ones. However, there was still

high correlation of log2FC in genes when the highest-expressed or NanoString isoform (R^2 of 0.93), even though most genes (456-457) had a different isoform (Figure 2.2E). All outliers (> 5 change in log2FC; labeled in Figure 2.2E) were explained by expression levels changing between values already below our limits of detection and zero (details in Supplemental Results). Similarly, isoform usage did not change NanoString and RNA-seq harmonization; in our double-assay dataset, despite 462/770 genes (60%) having changed isoforms, we observed little to no change in Spearman correlation between NanoString and RNA-seq expression levels (Appendix Figure A.13A). Finally, isoform-based changes in log2FC lead to minimal differences in genes' predictive power of NanoString vs RNA-seq samples or PFS (as measured by mean F-score across bootstraps) (Figure 2.2F and Appendix Figure A.13B, R^2=0.92 and R^2=0.99). The only obvious outlier was the *FANCA* gene whose isoform dramatically changed from the shortest to the longest isoform (increased length by more than 50kb) (subset of Figure 2.2F). Therefore, isoforms used made minor impact on predictive modeling except with extreme changes (e.g. more than 50kb).

2.6.3 Integrating NanoString and RNA-seq pre-post longitudinal data allows generalizable survival prediction

NanoString and RNA-seq showed enough congruence after preprocessing to hypothetically be combined into a single cohort to predict PFS (Figure 2.3A). The extent to which this combination can improve predictive modeling and clarify biomarkers in cancer has not been previously determined. We performed the PFS predictive modeling using a combined cohort of the single-assay NanoString and RNA-seq cohorts, hereby referred to as the "combined cohort." scRNA-seq and microarray data were used as an external verification of modeling results. Because microarray data shows a lower dynamic range, we used longitudinal data to see if the direction of change supported by our model is found in the microarray data. Similarly, scRNA-seq could clarify which cell-states best represent a given gene's bulk signal. Finally, the vast majority of HGSOC data is from non-longitudinal microarray; therefore, we evaluated how well our log2FC-based biomarkers showed association with survival metrics based on expression levels alone, focusing on microarray data.

For the sake of brevity, results showing how our harmonization insights improve upon previous single-assay predictive modeling of PFS can be found in Supplemental Results.

A model that used log2FC of genes to predict patients with PFS $< 12mths$ achieved decently high AUROCs (0.81 for hold-out) across all mini-cohorts and assays (Figure 2.3B). Further results confirming the model learned true biological patterns and feature relevance can be found in Supplemental Results (Appendix Figures A.15, A.16A, A.17). All of the eleven combined model genes except *TMUB2* had explicit literature support of their predictive power for HGSOC survival or differential expression in HGSOC cells with and without chemotherapeutic resistance (Supplemental Table 4). Seven of these genes had more than one source supporting their linkage: *IDO1*, *CD40*, *JAK2*, *ID4*, *ENO1*, *EGR1*, and *RRM2* [87, 289, 174, 133, 62, 169, 66, 239, 118, 206, 258, 178, 20, 131, 156, 205, 136, 294, 274, 3]. Despite the external validation of most model genes in HGSOC specifically, only *JAK2* and *RRM2* showed consistent trends in PFS$> 12mths$ and PFS$\leq 12mths$ across all mini-cohorts: stronger median upregulation and downregulation with $> 12mths$ patients of *JAK2* and *RRM2*, respectively (Figure 2.3C). Based on the DepMap database[94] of CRISPR knockouts and RNA knockdowns, most cancer cell lines show a high dependence on *RRM2*. OVCAR8, an HGSOC cell line characterized by high cisplatin resistance, is classified as *RRM2* essential by DepMap cutoffs (Appendix Figure A.18)[160, 50].

2.6.3.1 Verifying results in external scRNA-seq and microarray

For external verification in scRNA-seq and microarray data, we considered 12 genes: the 11 genes included in the combined model, and *CD274* (protein PD-L1). The latter was included because it was one of the highest and consistently top-ranked genes in our feature selection for both RNA-seq and NanoString, and was also included in a previous PFS model for HGSOC cite lucy (details in Supplemental Results; Appendix Figure A.19). We first assessed if the directionality of pre/post microarray data was consistent with that suggested by our model. Of the 12 genes, 5 (42%) showed the same directionality across all three assays, with a maximum of one mini-cohort within NanoString or RNA-seq disagreeing (Figure 2.3D). In some cases, the microarray data's

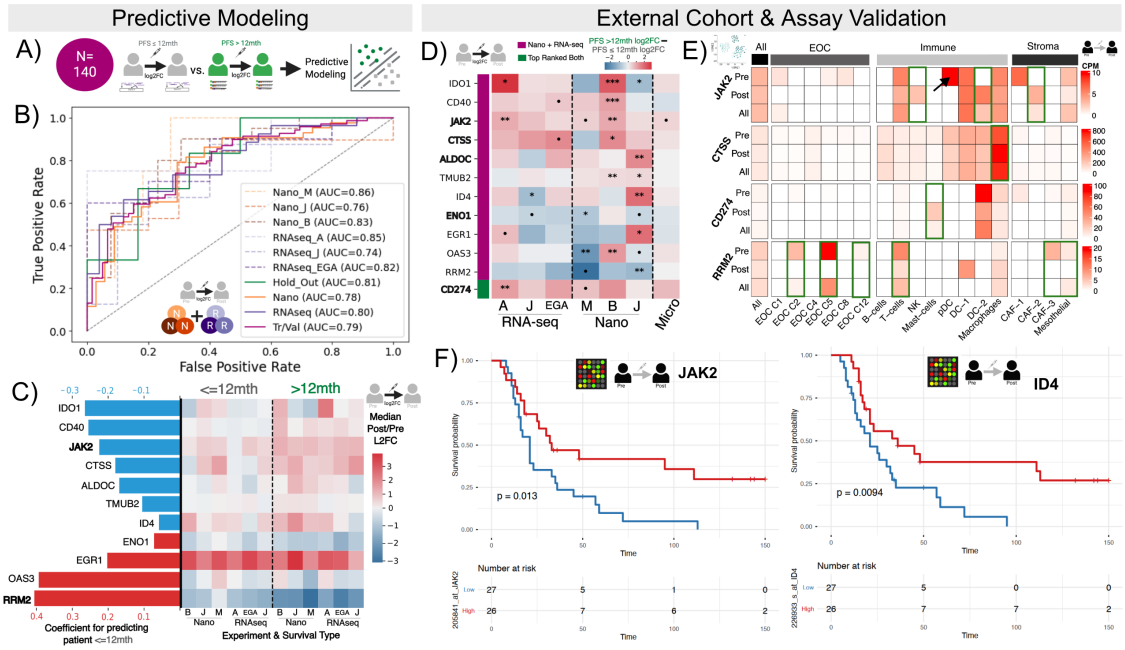

Figure 2.3: **Predictive modeling in combined cohort of NanoString and RNA-seq reveals consistent biomarkers**. **A.** Receiver Operator Characteristic Curves of a model trained to predict if a patient has a PFS$\leq 12 months$ using log2 Post/Pre NACT fold changes (AUC corresponds to AUROC). Tr/Val refers to training/validation samples (N=128) excluding the 12 kept for the Hold out set. **B.** Coefficient values for each of the genes considered in the model where blue indicates a negative and red positive coefficient for the log2FC. Bolded genes have consistent trends across all 6 experiments. Heatmap shows median log2FC in each experiment of patients with PFS $\leq 12 mths$ vs those $> 12 mths$. **C.** Difference between median Log2FC of patients with PFS $> 12 mths$ and PFS $\leq 12 mths$ of model genes found in the NanoString panel across the 7 available experiments with the p-values of t-tests of the log2FC dictated by $\bullet < 0.1, * < 0.05, ** < 0.01, *** < 0.001$. **D.** Heatmap of counts per million (CPM) per pseudobulked cell-state from single-cell data pre-NACT, post-NACT, and combined (All). **E.** Kaplan-Meier curves based on median expression split of *JAK2* and *ID4* in an external Microarray cohort.

directionality might have clarified some initial disagreements between cohorts, supporting most of the other cohorts that higher upregulation of *IDO1* generally corresponds to PFS $> 12 mths$.

scRNA-seq indicated which cell types in which the four genes with the most consistent signals were likely conferring the impact on PFS (results for all genes/cell types are in our Github repository in Methods). While *JAK2* showed the strongest initial signal in pDC cells, it showed increased expression post-NACT in NK and dendritic cells (DC-2) (Figure 2.3E). *CTSS* and *CD274* showed an increase post-NACT in macrophages and mast cells, respectively. Interestingly, the scRNA-seq

data showed the opposite direction for PFI correlation with *CD274* to bulk, instead indicating that decreased post-NACT expression correlated with higher PFI (Appendix Figure A.20A). Finally, *RRM2* showed the clearest decrease in the cancer cell cluster Epithelial Ovarian Cancer (EOC) Cluster 5, which had previously been annotated as proliferative cancer cells (Figure 2.3E)[290]. This finding corresponds with a stronger decline of *RRM2* expression reflecting lower proliferation of cancer cells and therefore serving as a biomarker of treatment success as suggested by our model. *EGR1* and *IDO1* had log2FC significantly correlated (Spearman correlation p-value below 0.05) with the platinum-free interval (PFI) in cancer-associated fibroblasts, but the correlational signal was not recapitulated in patients not previously considered in model training (Appendix Figure A.20A and B).

Finally, all 12 genes showed significant correlation with survival from microarray-based expression alone in much larger cohorts than those used in this study. First, all 12 genes were significantly associated with survival and/or PFS in two publicly available portals for Ovarian (CSIOVDB) or Serous Ovarian (KMPlot) cancer[240, 80] (Supplemental Table 4). When considering HGSOC-specific microarray data, genes *JAK2* and *ID4* showed significant association with survival across multiple cohorts in the same direction, including the cohort with the smallest N (Mok, N=53; Hazards Ratio 95% confidence intervals of [0.23-0.85] and [0.22-0.83]; p-values of 0.01) (Figure 2.3F). Seven of the 12 genes (58%) suggested by our combined cohort analysis showed significant association with survival (padj < 0.05; univariate Cox regression) in at least one of the four non-longitudinal microarray cohorts (Supplemental Table 5): *IDO1*, *JAK2*, *CTSS*, *ID4*, *EGR1*, *OAS3*, *CD274*.

Overall, these findings indicate that data from NanoString and RNA-seq can be effectively combined to identify biomarkers, after optimizing preprocessing steps.

2.6.4 Full genome consideration reveals key genes missing and provides additional mechanistic context

2.6.4.1 Non-NanoString panel genes mark treatment success

Given the successful usage of the combined dataset, we next evaluated the unique insights that RNA-seq could provide for discovery by considering all genes, with NanoString and similar probe-reliant assays used for validation. To this end, we predicted patients with PFS $\leq 12mths$ using the 57 patients with paired RNA-seq and all genes transcribed. Importantly, the feature selection approach we used ensured there was no obvious burden from including more genes (details in Methods).

We achieved higher AUROCs from a model trained on just the RNA-seq data than observed with the combined (NanoString+RNA-seq) model (0.86 vs original 0.81 in hold-out set) (Figure 2.4A). Further results validating the RNA-seq model and feature importance can be found in Supplemental Results (Appendix Figures A.14, A.16B, A.22). Six of the nine genes used in the final RNA-seq model showed consistent median log2FC differences between PFS classes across mini-cohorts (highlighted in 2.4B). This is a higher number than that observed with the combined model, but it is important to consider that we also required fewer mini-cohorts to demonstrate similarity (3 instead of 6). Seven of the nine RNA-seq model genes also had literature support linking them specifically to ovarian cancer or chemoresistance: *GBP4*, *AMOTL2*, *ESTY3*, *HCN3*, *DDX11*, *CLK2*, and *CDKL2*[269, 300, 93, 124, 88, 178, 295, 280, 81, 105] (Supplemental Table 4). Although *GBP4* was the only model gene in the NanoString panel, NanoString panel genes generally showed higher predictive power (as measured by Mean F-score across bootstrapped samples) than other genes (p-value < 0.001, Appendix Figure A.21). Specifically, *CD274* and *GBP4* were consistently within the top 25 genes with highest predictive frequency when only considering NanoString panel genes across the combined cohort.

Like with the combined model, microarray and scRNA-seq data supported our model findings. Six of the nine genes (67%) showed the same log2FC directionality across all available assays, with

a maximum of one NanoString or RNA-seq mini-cohort showing a different directionality (Figure 2.4C). Similarly, all genes showed expression significantly associated with overall survival and/or PFS in at least Ovarian (CSIOVDB) or Serous Ovarian (KMPlot) cancer [240, 80] (Supplemental Table 4). Expression of only three of the model genes (33%, *GBP4*, *DDX11*, and *CLK2*) was significantly associated with survival (padj < 0.05 univariate Cox regression) in a HGSOC-specific cohort (Supplemental Table 5). *GBP4* was significant across multiple cohorts in the same direction, with a Kaplan-Meier curve of the Yoshihara cohort[281] shown in Figure 2.4D (Hazards Ratio 95% confidence interval [0.41-0.85], pvalue=0.004). *GBP4* showed the most consistent, although not by much, results of all guanylate-binding genes in our and other datasets (Supplemental Results; Appendix Figure A.23). Both *GBP4* and *CLK2* also had log2FC correlated with platinum free interval (PFI) in cacner cells according scRNA-seq, in the same directions suggested by the bulk RNA-seq model. *GBP4* showed upregulation in patients with longer PFI and down regulation in those with shorter PFIs in the proliferative cancer cells (EOC_C5); stronger downregulation of *CLK2* was associated with higher PFI in dendritic cells and EOC_C4 (annotated as differentiating cancer cells) (Figure 2.4E, Appendix Figure A.20)[290]. Log2FCs of four additional model genes (total 67%) also correlated with PFI: *HCN3*, *AMOTL2*, *ESYT3*, *GPR173* (Appendix Figure A.20). Overall, being reduced to a gene panel list, even one specific to cancer, can hinder identification of relevant biomarkers.

2.6.4.2 Gene networks clarify mechanistic relevance of *GBP4*

Considering a full gene space also allows us to perform robust network analyses. Therefore, we performed network-based analysis of genes to hypothesize 1) which genes/pathways a gene correlates with and therefore might be represented by that gene in the model, and 2) the contextual role of a gene itself. We focus on *GBP4*, due to its differential signal being highly consistent. Briefly, we reconstituted pre- or post-NACT networks using the topological overlap matrix (TOM) similarity scores (measurement of gene interconnectedness) between genes from WGCNA (RNA-seq) and hdWGCNA (scRNA-seq) (details in Methods, Supplemental Table 6)[128, 166].

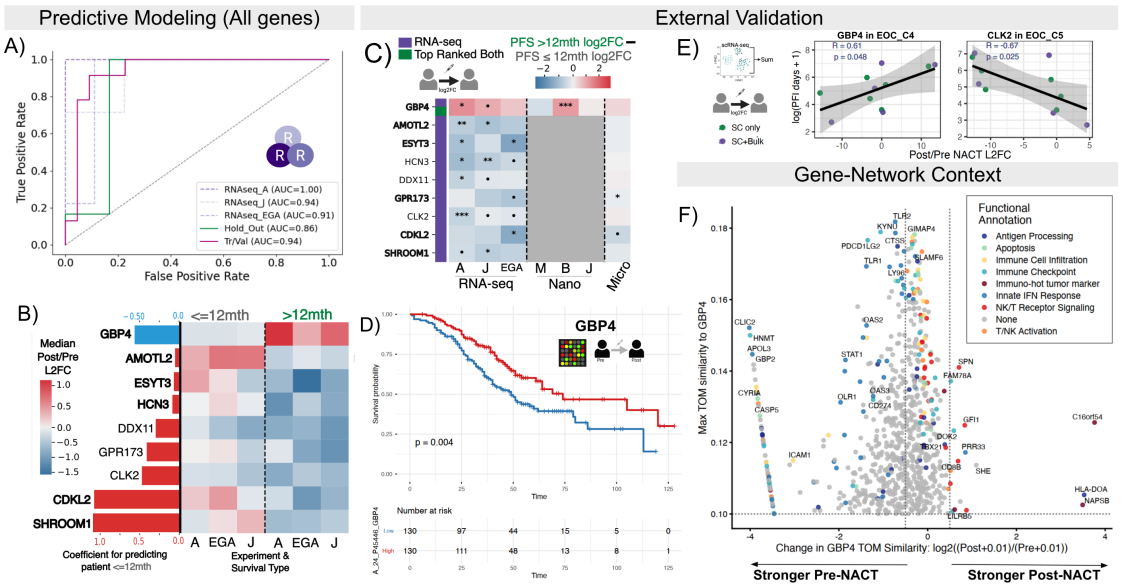

Figure 2.4: **Full genome analysis improves mechanistic understanding of biomarkers**. **A.** Receiver Operator Characteristic Curves of a model trained to predict if a patient has a PFS$\leq$ $12 months$ using log2 Post/Pre NACT fold changes (AUC corresponds to AUROC) using all RNA-seq genes and data. Tr/Val includes all training/validation samples (N=45) excluding the 12 kept for the Hold out set. **B.** Coefficient values for each of the genes considered in the model where blue indicates a negative and red positive coefficient for the log2FC. Bolded genes have consistent trends across all 3 experiments. Heatmap shows median log2FC in each experiment of patients with PFS $\leq$ $12 mths$ vs those $>$ $12 mths$. **C.** Difference between median Log2FC of patients with PFS $>$ $12 mths$ and PFS $\leq$ $12 mths$ of model genes across the experiements considering the gene, with the p-values of t-tests of the log2FC dictated by $\bullet < 0.1, * < 0.05, ** < 0.01, *** < 0.001$. **D.** Kaplan-Meier curve based on median expression split of *GBP4* in an external Microarray cohort. **E.** Scatter plots of Post/Pre NACT Log2FC from pseudobulked scRNA-seq data (N=11, within cell-type) and the log(PFI + 1). Fitted line with R and p-values from a linear regression is shown with standard error highlighted. **F.** Volcano plot of gene's changes in TOM similarity with GBP4 in Pre and Post NACT WGCNA-based networks (x-axis) and the maximum similarity (y-axis). Genes determined as shared or post/pre only are colored based on general immune-based functions, with genes of highly mixed or lesser known functions labeled as None.

We first evaluated how the bulk-based gene-networks of *GBP4* changed based on treatment status (Pre- vs Post-NACT) (details in Methods). Regardless of timing, *GBP4* connected to a core set of immune-relevant genes including immune-checkpoints, lymphocyte regulation, and chemokine signaling, and immune-cell infiltration (Figure 2.4F, Supplemental Table 7). This finding supports that *GBP4* is indeed corresponding to an immune role in the HGSOC context. In Pre-NACT networks, *GBP4* was connected to genes primarily involved in innate immunity: innate

immune sensing, apoptosis, antigen processing, interferon-stimulated genes, metabolic stress, and immune inhibitory checkpoints (including *CD274* [PD-L1], *PDCD1LG2* [PD-L2])) (Figure 2.4F, Supplemental Table 7). Enriched pathways (q value < 0.05) unique to the pre-NACT gene group included viral response, interferon-mediated signaling pathways, and positive regulation of Tumor Necrosis Factor. Conversely, post-NACT *GBP4* connectivity shifted closer to pathways of cytotoxic lymphocytes, with enriched pathways (q value < 0.05) unique to the post-NACT group including thymic T cell selection, T helper (CD4+) cell differentiation and immune response, and regulation of alpha-beta T-cell differentiation. Key post-specific genes direct towards T helper cell fates and expansion, mark cytotoxic T-cell proliferation, increase CD4+ and CD8+ T cell trafficking, and activate T/NK cells (Figure 2.4F, Supplemental Table 7). *HLA-DOA* and *CLEC10A* both mediate antigen loading in macrophages and dendritic cells, thereby influencing T helper responses. Finally, genes *PTPN7*, *NAPSB*, *C16orf54*, and *TMC8* have been independently associated with immune-hot tumor micro environments and serve as predictive markers of cancer prognosis[255, 173, 53, 139] (Figure 2.4F). Overall, these findings indicate that *GBP4* networking may focus on marking a general innate response to one of cytotoxicity.

Single-cell expression and networks support this interpretation. *GBP4*-associated edge weights from bulk RNA-seq most closely matched the scRNA-seq networks of the dendritic cells in which *GBP4* was also most strongly expressed prior to treatment (Appendix Figure A.24A,B). Following NACT, *GBP4* expression then decreased markedly in macrophages, NK cells, B cells, and dendritic cells, while remaining stable in T cells and increasing in innate lymphoid cells and mast cells (Appendix Figure A.24B). *GBP4* network correspondence also shifted from primarily dendritic cells to include the NK cell group (NK+ILC, with 86% NK cells), with the caveat that NK cell numbers above 50 were only found in two patients pre-NACT. Together, these results indicate that *GBP4* expression and network context are both reprogrammed by chemotherapy, so that it's expression patterns may reflect a shift from a myeloid-dominated, immunosuppressive environment toward lymphoid-associated cytotoxic immune programs in our model.

2.7 Discussion

2.7.1 Foundational Guidance on data integration across assays for predictive modeling

In this work, we evaluated how to incorporate the growing technologies in transcriptomics with one another, providing guidelines on how to effectively combine assay data both directly in modeling or for verification of results. We first identified the strengths and weaknesses of each assay, before clarifying necessary steps for harmonizing NanoString and RNA-seq. We ended by performing predictive modeling and network analysis to show how integrating probe-based assays with assays that consider full gene space and single-cell dynamics can powerfully identify biomarkers and their mechanistic insights.

Our work first clarifies that microarray and scRNA-seq contain systematic biases from other transcriptomics assays that should be considered before combining them with other data. Percentile-scaling is commonly used to directly consider microarray with other datasets; however, our and previous work supports that this approach relies on a likely poor assumption that the biological meaning of extremes is shared[297]. Our data indicates that a better solution might come with including a binary variable of whether the assay is microarray in a model trained across all three assays: microarray, NanoString, RNA-seq. Longitudinal data also does not address the known poor correlation between pseudobulked scRNA-seq and bulk RNA-seq. Importantly, we had low sample numbers to effectively address the full extent and reasons for this. The difference might be explained by cell types being preferentially captured for scRNA-seq, thereby changing from the cell type proportion in bulk. Previous studies have shown that smaller, robust cells like immune cells are often more easily captured for scRNA-seq than larger, fragile cells[113].

Our framework and web portal can be used by others to clarify whether modeling/biomarker differences across cohorts are more likely due to assay- or biological-heterogeneity. The greatest limitation to integration across assays was the poor characterization of lowly expressed genes. Therefore, determining the limits of detection for assays, as provided in this work, is essential

not only for accurate integration across but also within one assay. Similarly, our work suggests that RNA-seq data is not commonly sequenced to a high enough depth where probe-focused (e.g. exon-counting) or isoform-specific counting makes a general difference in harmonization between NanoString and RNA-seq. This, however, depends on the isoform difference; for example, a single 100bp probe was not well representative for a gene expressing its largest 100kb+ isoform. Our results can enable researchers to pinpoint why a different cohort/study may have gotten opposing findings by considering exon-counting approaches or differential isoform usage. As a quick-check, researchers can use our portal (in Data Availability) to explore how certain genes harmonize between NanoString and RNA-seq using both isoform-based and exon-based counts.

Finally, we provide guidelines that address how to effectively integrate data as technological innovations continue in single-cell and bulk transcriptomics. RNA-seq and NanoString were consistent enough to allow direct comparison of results, even when considering notoriously variable sample types (e.g. FFPE). While NanoString panels are gene-limited, they provide a simpler, efficient, and often cheaper alternative to RNA-seq for expression profiling. Therefore, we suggest that RNA-seq be used to clarify the most relevant sub-1000 list of genes for a NanoString panel that is then more easily adopted by clinics. NanoString and RNA-seq data can then be effectively combined for more refined predictive modeling, expanding data generation without hindering biomarker selection. Given the current scarcity of scRNA-seq, both RNA-seq and scRNA-seq can be used to further characterize the possible mechanisms underlying biomarkers. For example, we clarified that *GBP4* is not necessarily an effector itself, but a marker of the tumor microenvironment's immune state shifting from immune suppression and generalized inflammatory conditions towards specific cytotoxicity and immune-hot tumor signatures.

2.7.2 Potential biomarkers and models for downstream experimentation in HGSOC

The effectiveness of the above guidelines can be observed by our work presenting several models and biomarkers of treatment response in HGSOC. These results can be used for both

hypothesis generation and focused experimental efforts.

For example, our findings suggested that strong expression of *RRM2* reflects proliferation of cancerous cells and therefore chemoresistance. Indeed, inhibition of ribonucleotide reductase (RNR), whose activity RRM2 directly regulates, has been shown to increase the sensitivity of chemoresistant cancers across multiple clinical trials, including one focused on platinum-resistant ovarian cancer[285, 122]. Since direct RNR inhibition can lead to severe side effects, growing research is focusing on developing expression-level inhibition of RRM2 as a possibly safer alternative[285]. Interestingly, our data suggested that increased expression of *CD274* (PD-L1), and similarly, low initial expression and high post-NACT expression, correlate with increased PFS. Interaction of PD-L1 with its receptor inhibits T-cell activation and cytokine production, so limiting this inhibition has been a key therapeutic target[150]. We provide two key explanations of our finding. First, increased PD-L1 expression is a biomarker of both potential downstream immunosuppression and of a responding immune system, with NF-kB inducing *PD-L1* expression[150]. Indeed, we showed that the immune rather than cancer cells showed expression post-NACT, while decreased expression of *CD274* within immune cell types correlated with higher PFS. Secondly, the bulk RNA-seq expression levels might represent an increased portion of immune cells post-NACT.

Finally, we consistently found that increased *JAK2* and *GBP4* expression post-NACT and generally higher expression correlated with higher PFS. Multiple functional assays in cancer cell lines have indicated decreased expression of *JAK2* instead limits proliferation[66, 239]. Importantly, however, our findings reveal that *JAK2* expression from FFPE cancer samples is primarily in immune cells rather than epithelial cells, indicating the importance of considering a gene within the context of the full tumor microenvironment. Indeed, JAK inhibitors on mice only slightly improved chemosensitvity when using tumor cells already showing higher expression of JAK and STAT genes compared to chemo-sensitive tumors[239]. While a helpful marker, JAK2 is a multifaceted protein across cells and therefore might not be an ideal treatment target. Instead, *GBP4* shows dispersed expression across all cell types, although primarily in immune cells, and has been previously described as a marker of immuno-hot tumors[300]. Indeed, experimental manipulations of *GBP5*

and *GBP1* expression in ovarian cancer cell lines have indicated that these genes are critical to preventing cancerous proliferation[305, 269]. Although GBP4-focused functional assays have not, to our knowledge, been conducted, our finding that *GBP4* more consistently predicts higher PFS than other GBP genes indicates that *GBP4* might be a promising gene for future experimentation.

2.7.3 Limitations and Future work

Finally, the limitations of this study help pinpoint important opportunities for future work in integrating data across assays for predictive modeling. Despite combining across cohorts, the sample size of 140 was still below the 200 ideal for predictive modeling in even simple diseases[224]. Similarly, feature filtering is a necessary step when dealing with a large number of features compared to samples, but can also remove clinically relevant genes. Thus, the ability to integrate cohort data for larger sample sizes, including those across assays, is clear. Finally, while longitudinal data is ideal for addressing patient- and assay-specific biases, it is harder to obtain and perhaps not as clinically applicable as un-matched expression data. The ideal use case for biomarkers and models of treatment success is that they can predict success before a patient receives treatment. Since NACT is a highly standardized treatment for ovarian cancer, our models on log2FC are relevant for enabling immediate action post-treatment, but are still limited. Our findings from our double-assay dataset indicate that NanoString and RNA-seq have comparable dynamic ranges and that, once addressing limits of detection, assay-scaled expression values are directly comparable between samples. Therefore, building a model from the full set of pre-treatment RNA-seq and NanoString data using our suggested preprocessing steps is not only viable but could provide more immediate insight into HGSOC pathology. In all cases, our harmonization insights, models, and the multiple data resources/analyses available on our portal (see Data Availability) will be powerful tools for downstream dataset incorporation and predictive modeling of ovarian cancer and other cancers.

2.8 Data Availability

For NanoString, the Bitler dataset consisted of raw files from GEO GSE319500[109]cite lucy, the Manso dataset from GEO GSE181597[147], and the James dataset from GEO GSE201600[99]. For RNA-seq, the Adzib dataset used raw files from GEO GSE227100[1]. Survival information was provided for a subset of patients by the original authors. The EGA (Piet-Zhang and Piet) dataset was so named since raw files could only be collected from the European Genome Archive (EGAD00001006456). This dataset included consortia of HGSOC studies MUPETFaasi2, HERCULES, and CHEMORESPONSE. The metadata and previous analyses of the data were collected from [183, 294, 181, 129]. The Jav dataset only provided normalized counts (RPKM) as the Supplementary Table 5 ([100]). For Microarray (Jim-Sanchez) data, normalized data were found at GEO=GSE146963[106]. Quality Control counts and metadata for scRNA-seq (Zhang) data from [290] were downloaded from GEO using accession number GSE165897 (GSE165897_cellInfo_HGSOC.tsv and GSE165897_UMIcounts_HGSOC.tsv.gz). Data from the double-platform work can be found at GSE323347 (RNA-seq) and GSE322784 (NanoString).

Users can also explore the data on https://hopetownsend.shinyapps.io/hgsoc-gene-explorer/. This app runs through Shiny apps free portal and therefore users are also encouraged to follow instructions on the github repository (https://github.com/Hope2925/NanoString-RNAseq-HGSOC) to run it locally to avoid server limitations. The same github repository contains all code and data links for this work.

Code for all analyses can be found at https://github.com/Hope2925/NanoString-RNAseq-HGSOC.

2.9 Funding

This work was funded by the National Institute of Health (BGB) (grant number), OCRA (BGB, AC) and the University of Colorado Anschutz-Boulder Nexus (AB Nexus) Seed Grant.

Chapter 3

Evaluating methods for integrating single-cell data and genetics to understand inflammatory disease complexity

The work in this chapter is as published in: **Townsend, H.A.**, Rosenberger, K.J., Vanderlinden, L.A., Inamo, J., Zhang, F. (2024). Evaluating Methods for Integrating Single-Cell Data and Genetics to Understand Inflammatory Disease Complexity. Front. Immunol. 15. doi: 10.3389/fimmu.2024.1454263. All supplemental tables can be found with the published manuscript. Figures in Appendix B were published as supplemental figures in published manuscript.

3.1 Contribution Statement

I led this work and was aided by Kaylee Rosenburger. Kaylee Rosenberger helped research different approaches to benchmark, ran scGWAS, and provided the scGWAS results. Kaylee and I wrote the initial draft which was then edited by all authors. Drs Inamo and Zhang advised on the project. I did all other work including data curation, literature reviews, and analyses.

3.2 Abstract

Background: Understanding genetic underpinnings of immune-mediated inflammatory diseases is crucial to improve treatments. Single-cell RNA sequencing (scRNA-seq) identifies cell states expanded in disease, but often overlooks genetic causality due to cost and small genotyping cohorts. Conversely, large genome-wide association studies (GWAS) are commonly accessible.

Methods: We present a 3-step robust benchmarking analysis of integrating GWAS and

scRNA-seq to identify genetically relevant cell states and genes in inflammatory diseases. First, we applied and compared the results of three recent algorithms, based on pathways (scGWAS), single-cell disease scores (scDRS), or both (scPagwas), according to accuracy/sensitivity and interpretability. While previous studies focused on coarse cell types, we used disease-specific, fine-grained single-cell atlases (183,742 and 228,211 cells) and GWAS data (Ns of 97,173 and 45,975) for rheumatoid arthritis (RA) and ulcerative colitis (UC). Second, given the lack of scRNA-seq for many diseases with GWAS, we further tested the tools' resolution limits by differentiating between similar diseases with only one fine-grained scRNA-seq atlas. Lastly, we provide a novel evaluation of noncoding SNP incorporation methods by testing which enabled the highest sensitivity/accuracy of known cell-state calls.

Results: We first found that single-cell based tools scDRS and scPagwas called superior numbers of supported cell states that were overlooked by scGWAS. While scGWAS and scPagwas were advantageous for gene exploration, scDRS effectively accounted for batch effect and captured cellular heterogeneity of disease-relevance without single-cell genotyping. For noncoding SNP integration, we found a key trade-off between statistical power and confidence with positional (e.g. MAGMA) and non-positional approaches (e.g. chromatin-interaction, eQTL). Even when directly incorporating noncoding SNPs through 5' scRNA-seq measures of regulatory elements, non disease-specific atlases gave misleading results by not containing disease-tissue specific transcriptomic patterns. Despite this criticality of tissue-specific scRNA-seq, we showed that scDRS enabled deconvolution of two similar diseases with a single fine-grained scRNA-seq atlas and separate GWAS. Indeed, we identified supported and novel genetic-phenotype linkages separating RA and ankylosing spondylitis, and UC and crohn's disease. Overall, while noting evolving single-cell technologies, our study provides key findings for integrating expanding fine-grained scRNA-seq, GWAS, and noncoding SNP resources to unravel the complexities of inflammatory diseases.

3.3 Introduction

The efficacy of treatments for immune-mediated inflammatory diseases, such as rheumatoid arthritis (RA) and ulcerative colitis (UC), varies across patients[168]. Single-cell RNA sequencing (scRNA-seq) technology enables the development of effective treatments for patients with immune-mediated inflammatory diseases by allowing the identification of specific cell states expanded in diseased tissue or blood[172]. However, most scRNA-seq analyses do not consider genetic causality, and due to its high expense, available single cell datasets are often confined to small patient cohorts. Understanding the genetic underpinnings of diseases is key for preventative care, unraveling physiological and environmental contributions to pathology, and allowing personalized treatments. Genome wide association studies (GWAS) have been the gold standard to identify disease-associated genetic loci and summary statistics for large cohorts are often publicly accessible[213]. Therefore, recent work has gone into combining the physiological insights from scRNA-seq with genetic associations from GWAS for unraveling disease causality[291, 102, 259, 199, 65, 151, 163]. Indeed, attempts to integrate bulk RNA-seq studies with GWAS have been implemented, yet still only explain about 30% of the heritability by gene expression for complex traits[277]. This pitfall is likely explained by the less fine-scale cell states available with bulk RNA-seq compared to scRNA-seq, where immune cells exhibit divergent expression profiles at nuanced cell states, and different cell phenotypes are uniquely associated with disease[101, 191, 286].

Recently, several computational tools have been developed to link disease relevant loci from GWAS to nuanced cell states revealed by scRNA-seq to identify disease-associated cell states and genes with both transcriptomic and genomic support[291, 102, 259, 199, 151, 244]. For each tool, major steps include summarizing variably expressed genes/pathways from single cell expression data, using a third-party method to link GWAS based single nucleotide polymorphisms (SNPs) to genes/pathways, and then using statistical tests to identify significant associations. However, a thorough comparison and assessment of these tools is lacking. Additionally, a critical step for all these tools, linking SNPs from GWAS to the genes they potentially impact, has been challenging

with no clear solution[134, 263, 69, 266, 272]. With more than 90% of immune-disease associated SNPs falling into noncoding regions, most of which are in cis-regulatory regions, the need to link these SNPs to physiological mechanisms cannot be overstated[39]. The most common method for linking SNPs to genes does so according to a user-selected window size outside the gene. MAGMA, one of the most common tools that does this, can take both genotype data and summary statistics as input while accounting for Linkage Disequilibrium[134]. It outputs a list of thousands of genes with the corresponding GWAS statistics reestablished at the gene level. However, many target genes of cis-regulatory regions are not the closest gene and can even be farther than 1 Mb away, contradicting the assumptions of tools like MAGMA[69]. Therefore, alternative methods focusing on eQTL, chromatin contact (e.g. Hi-C), and similarly relevant enhancer-gene linking data have been introduced[263, 97]. Additionally, newer studies have begun introducing single-cell transcriptomics methods that measure cis-regulatory elements to directly consider noncoding SNPs[163]. The influence of incorporating noncoding SNPs using non-positional compared with positional methods, specifically within the context of these algorithms, has not been formally evaluated.

Beyond SNP-gene linking complexities, transcriptomics-genomics integration algorithms have currently been assessed for capturing broad associations (e.g. metabolic cells for metabolic diseases)[291, 102, 151]. This limited analysis is primarily due to the usage of non-disease specific scRNA-seq atlases rather than disease-specific atlases with highly refined cell states identified. Disease specific, scRNA-seq atlases are quickly being developed and revolutionizing the understanding of diseased tissue heterogeneity. Yet the ability for tools tested on broader cell types to work with these more refined atlases with disease confounders has not been tested. Additionally, these tools might still be usable for diseases without atlases currently available by using atlases of similar diseases but the appropriate GWAS summary statistics.

Overall, despite the recent influx of tools integrating genetics and single-cell transcriptomics, a thorough comparison and assessment of different types of recent algorithms and major challenges of the domain is lacking. To address this, we conducted a benchmark analysis of the three most recent, open-source algorithms, scGWAS, scPagwas and scDRS, by objectively linking GWAS data

with single-cell phenotypes across four immune-mediated disease datasets[291, 102, 151, 286, 227]. We further annotated our results based on literature support of calls (detailed in Methods and Supplementary Tables 1, 2), and evaluated the computational efficiency and result interpretability. Given most immune relevant SNPs are noncoding, we then evaluated the influence of different methods incorporating these SNPs for use in the algorithms[134, 263]. As a result, we first showed that all three tools successfully identified expected significant cell types for tested diseases when using fine-grained scRNA-seq atlases, although with varying consistency and agreement. Single-cell scoring tools scDRS and scPagwas identified more significant results with literary support, although pathway-based scPagwas invokes a higher computational cost and cannot effectively consider batch effects. We also found that scDRS can be used to distinguish cell phenotypes for different diseases while using the same fine-grained scRNA-seq atlas. Finally, we provided evidence supporting the usage of positional based methods to incorporate noncoding SNPs until other methods can increase in statistical power and include more relevant atlases. Overall, our in-depth benchmarking and application on disease-tissue data demonstrated that current tools could identify associations between cell phenotypes and disease with high resolution and specificity. Our work pinpoints the capabilities and benefits of using atlases with fine-grained cell subtype annotations, while also showing that a single atlas could still be used to understand multiple diseases.

3.4 Materials and Methods

We first benchmarked the three most recent and representative algorithms in the field according to the number of literature supported clusters called significant, computational efficiency, and result interpretability (Figure 3.1A). Brief descriptions of the tools can be found in sections 2.1 and 2.4. Expected results were based on a literature search for each individual cell phenotype for expansion in a disease and/or genetic connections, the results of which can be found in Supplementary Tables 1 and 2. If a general cell state with multiple, more detailed cell states was significant, the cell states were marked as having "general" literature support while if a specific cell state was supported, it had "specific" literature support. Due to the robustness of the available atlases and studies, we

used scRNA-seq data generated from inflamed RA synovial and UC colon to determine disease-associated cell states[286, 227]. Next, we assessed the feasibility of using identical scRNA-seq atlases to distinguish between two clinically similar diseases, using RA inflamed synovial tissue for RA and ankylosing spondylitis (AS), and UC colon for UC and Crohn's disease (CD) (Figure 3.1B). Finally, we evaluated the incorporation of noncoding SNPs when using positional (MAGMA) vs non-positional based SNP-gene linking methods or cis-regulatory element focused single-cell omics like ATAC-seq or 5'-scRNA-seq (Figure 3.1C). We deploy all the code and analytical pipelines at our Github repository for reproducible research at https://github.com/fanzhanglab/SCRNA-GWAS-Benchmarking.

3.4.1 Selection of tools

We summarized the attributes of six currently available and supported packages that integrate scRNA-seq data and GWAS summary statistics to identify significant cell types and/or the GWAS-linked genes that best explain these cell types in Table 3.1. Other methods like RolyPoly, CocoNet, and sc-linker are described in Supplementary Table 3, and are either no longer maintained or not designed as user-friendly packages but instead open-source code [97, 28, 218]. Briefly, RolyPoly was one of the first tools to employ the use of polygenic modeling to identify trait-relevant cell states, CocoNet pioneered gene-network based analyses, and sc-linker leveraged enhancer-gene linkages to assign SNPs to genes. The three tools chosen for more detailed benchmarking were the most recent tools and provide unique results as either gene-gene networks or single-cell based scores. The other methods differ most by their incorporation of noncoding SNPs which is addressed separately in this work.

Table 3.1: Summary table of the currently maintained and operable packages for identifying significant cell types and/or genes based on the integration of GWAS and single-cell RNA-seq data. A similar table for methods no longer maintained (RolyPoly) or not designed as packages for complete analysis workflows (CocoNet and SC-Linker) is available in Supplementary Table 1.

Package Interface	**Inputs**	**Relevant Outputs**	**SNP-Gene Linking**	**Summary**	**Highlights**
scPagwas[151] R package	1. Seurat Object 2. GWAS summary stats	1. Cell score file 2. Cell Pathway Scores 3. Opt: Cell group score 4. Opt: Gene PCCs	Window-based	Pathway-based polygenic regression: linear regression of GWAS signals with pathway activation in cells.	Pathway-based while maintaining single-cell analysis
scGWAS[102] CL JAR, pre/post processing in R	1. Boxcox transformed gene p-values 2. Pseudobulk 3. Gene-gene network file	1. Significant gene modules in each cell type	Window-based: MAGMA	Network-based approach to identify cell types overexpressed with disease-significant genes	Pathway based for more meaningful output
scDRS[291] CLI or API	1. Anndata single cell expression data 2. Gene p-values or z-scores	1. Cell score file for a given trait 2. Opt: Cell group score and heterogeneity 3. Opt: Cell variable (e.g., gene) correlation to disease scores	Window-based: MAGMA	Monte Carlo simulation method that scores individual cells for disease association based on increased expression of sets of putative disease genes	Single-cell level allows unique post analyses
EPIC[259] R package	1. Pseudobulk gene expression 2. GWAS summary stats	1. Enrichment score of trait for each cell type 2. Relevant genes from DFBETAS	Sliding-window based LDSC	Gene-level chi-square association testing, then gene-level regression-based association testing for each cell type	Adapted for rare and common variants
ECLIPSER[199] Scripts on Github	1. GWAS summary stats 2. Gene differential expression table	1. Prioritized cell types 2. Leading edge causal genes and eQTL impact	eQTL and other functional evidence	Cell-type-specificity score for each GWAS locus, cell-type specific genes (from differential expression analysis mapped to locus)	Provides putative regulatory impact of genes
CELLECT[244] CLI	1. Specificity input from CELLEX 2. GWAS summary stats	1. Prioritized cell types 2. Opt: Gene heritability	LDSC or MAGMA	Heritability regression based method with CELLEX gene specificity scores	Allows easy usage of LDSC or MAGMA

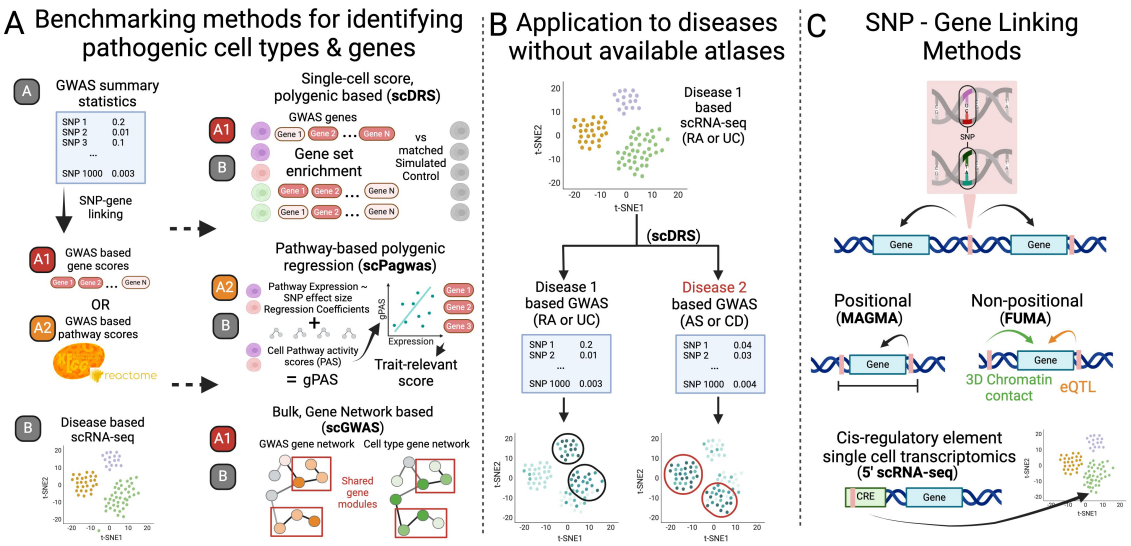

Figure 3.1: **Overview of study design**. **A.** We first benchmarked the three most recent tools built to identify cell states and genes associated to disease according to both genetics (GWAS) and transcriptomics (scRNA-seq). **B.** We next assessed if a single scRNA-seq atlas could be used with summary statistics from two diseases to reveal well separated disease associated cell states of the different diseases. **C.** Finally, we assessed the robustness and accuracy of results of these tools when using different SNP-Gene linking methods. Figure made in Biorender.

3.4.2 Data availability

The GWAS data used in this work can be found in Supplementary Table 4. Due to the most robust LD score data belonging to those with European descent, and the larger sample size of this group in both GWAS and scRNA-seq data, we focused on this subpopulation for the purpose of this benchmarking analysis. The major histocompatibility complex region was not included due to its complex genetic architecture. For GWAS summary statistics without rsids for RA, SNPs were assigned to rsids using BEDOPs and for duplicate/synonmous rsids, those with the lowest p-values were kept. The code for these steps can be found on our github under SCRNA-GWAS-Benchmarking/src/00B_Preprocess_GWAS.

For RA and AS, we analyzed a scRNA-seq data set developed by[286]. To stay consistent with GWAS data, we only included cells from individuals of European descent with RA, leaving 183,742 cells. We used their most updated cell-state and cell-type annotations determined by their

analysis of 314,011 cells with scRNA-seq, CITE-seq, experimental evidence and batch control to ensure the best validation. All expression was normalized with log(1 + UMIs for gene/totl UMIs in cell *10,000), and cells expressing fewer than 500 genes or that contained more than 20% if their total UMIs mapping to mitochondrial genes were removed. Further QC analysis is described in their paper[286]. For UC and CD, we analyzed the scRNA-seq dataset from [227] which contained 228,211 cells passing quality control by using the raw counts and metadata they provide. For batch correction in both datasets, we applied Harmony, one of the best recommended methods for correcting for technical batch effect in single-cell batch data analysis and integration[119, 247]. We used identical batch variables for correction as used in the original analysis for RA: the individual from which the cells were isolated ("sample")[287]. Combat was used for batch correction originally in Smillie et al., but is not designed for single-cell data, therefore we applied Harmony with "sample" to the UC scRNA-seq data instead[227, 296]. Both scRNA-seq data only contained individuals of non-Hispanic, European descent. For scPagwas, we created Seurat objects with the same QC-based cells but using the Seurat based normalization. Due to the high computational expense of scPagwas, we excluded certain cell states from the RA and UC datasets that were not found significant by the literature on RA and UC, including Endothelial (RA & UC), Glia, Macrophages, TA 1 & TA 2(UC), and fibroblast cell states except for F-7: NOTCH3+ sublining and F-2: CD34+ sublining (RA). The code for these steps can be found on our github repository under SCRNA-GWAS-Benchmarking/src/00A_Preprocess_scRNA.

3.4.3 SNP-gene linking

MAGMA-based SNP-gene linking was done using version v1.10 with NCBI37.3.gene.loc and NCBI38.gene.loc downloaded from the MAGMA website as the gene locations files, and European UK Biobank Phase 3 LD scores. The window sizes of 10-10kb and 50-35kb were chosen for final comparison of significant cell states as the most common window size and that used in the original scGWAS paper, respectively. When assessing the impact of this window size parameter on scDRS, sizes 0kb, 5kb, and 100kb were also chosen based on the window sizes used across the literature

(Supplementary Table 5). For this parameter stability assessment, the top-ranking genes according to MAGMA that were also found in the scRNA-seq expression data were used, with a final total of 1000 genes. Synonyms according to genecards.org and humanproteinatlas.com were also considered to verify proper comparison of genes between MAGMA and scRNA-seq. Genes from the scRNA-seq dataset still not found in the MAGMA file were added to allow their inclusion in the analysis. The genes identified by MAGMA but not found in scRNA-seq data are discussed further in the Supplementary Material, with numbers dictated in Supplementary Table 6.

The code for all these steps can be found on our github under `SCRNA-GWAS-Benchmarking/src/01_MAGMA_Gene_Alias.`

FUMA is a web-based tool that determines statistically significant disease associated genes using positional, eQTL, and 3D chromatin based mapping, but does not calculate a summary p-value like MAGMA[263]. Therefore, to explore the implications of including these forms of mapping, we used the minimum GWAS SNP P-value (minGwasP in genes.txt output file) for each gene as a proxy for a disease-association p-value to allow input for scDRS and scGWAS. FUMA identifies lead SNPs, maps to rsIDs, addresses duplicate and synonymous rsIDs, and filters out the MHC region in its analysis from the summary statistics. Default parameters were used including a MAGMA window of 10kb, with MAGMA expression data being based on GTEx v8. We also used eQTL and Chromatin Interaction Mapping, both including the options of available blood cell eQTL data. Versions include FUMA v1.5.3, MAGMA v1.08, GWAScatalog e0_r2022-11-29, and ANNOVAR 2017-07-17.

3.4.4 scGWAS, scDRS, & scPagwas

scGWAS uses a network-based approach to uncover cell types that significantly express disease-associated genes and identify gene modules representing disease-specific processes[102]. Unlike other methods where cell types are assigned a disease-significance score, scGWAS assigns significance scores to gene modules with strong representation in both scRNA-seq cell type expression and GWAS based on a proportional test (Figure 3.1A). scGWAS is imple-

mented in Java via a JAR package (ver. scGWAS_r1.jar) on the authors' GitHub repository (https://github.com/ElkonLab/scGWAS) and can be run through the command line. Based on author recommendations on their GitHub repository, configuration file parameters were kept at default values. Further, we first used the same PathwayCommons input network file as Jia et al. (5), with gene-gene relationship information used for constructing the background network. We also created a second PathwayCommons input network file following their same steps but with v14 rather than v12 (what they used originally). Briefly, housekeeping and ribosomal genes were removed as well as any genes within 50kb of one another (detailed jupyter notebook and output pathway file found on our github under SCRNA-GWAS-Benchmarking/data/Pathway). We followed the analysis pipeline described on the authors' GitHub repository for the following steps. For the screen expression input file, we processed the scRNA-seq dataset using their R-script to calculate the average log-transformed gene-based CPM per defined cell type. We processed the MAGMA output using the box-cox transformation script as the GWAS node input file. We ran scGWAS on the same scRNA-seq dataset first with general cell types and then on fine-scale defined cell states. The code for these steps can be found on our github under SCRNA-GWAS-Benchmarking/src/03_scGWAS.

scDRS assesses disease-associations at the individual cell level using a gene set enrichment analysis with genes with scored associations to the trait of interest according to a third party method[291] (Figure 3.1A). It then presents downstream analyses that use unified Monte Carlo tests to identify significant pre-annotated cell states according to a group Z score, and the genes whose expressions correlate with disease scores. It is the only tool designed to take cell-level covariates to address potential batch effects. The CLI version (Version v102 v1.0.2) of scDRS was used according to their GitHub repository (https://github.com/martinjzhang/scDRS). All default parameter values were used, and P-value files output from MAGMA served as input to scdrs munge-gs. The covariates files used in computing scDRS scores included nUMI, number of genes, and sex for both RA & UC, and age and duration for RA, and sample location, percent of mitochondrial reads, and smoking status for UC (found in our github at SCRNA-GWAS-Benchmarking/data/SC_data). We ran downstream analyses to identify significant cell groups on the same scRNA-seq dataset

using annotations of general cell types and then with fine-scale defined clusters. The code for these steps can be found on our github repository under SCRNA-GWAS-Benchmarking/src/02_scDRS.

scPagwas associates cells and cell types to traits through pathways rather than only individual genes, while maintaining associations at the individual cell level[151]. Rather than using a pre-determined GWAS based gene set list with scores like scDRS and scGWAS, scPagwas calculates genetically associated pathway activity scores (gPAS). Briefly, the gPAS is the product of a per-cell coefficient of a linear regression between SNP effect sizes and gene expression within a pathway, and the pathway activity score of the cell (first principal component of an SVD). Finally, following a similar logic of scDRS, a trait-relevance score is calculated using the Seurat cell scoring method which considers the expression of the top 1,000 genes most correlated with the summed gPAS in cells (Figure 3.1A). We followed installation instructions from the scPagwas github (https://github.com/sulab-wmu/scPagwas) for version 1.3.1, using Seurat version 5.1.0 and SeuratObject version 5.0.2. Code for these steps can be found on our github repository under SCRNA-GWAS-Benchmarking/src/04_scPagwas. To run scDRS with scPagwas genes, the 1,000 genes with the highest Pearson correlation coefficient (PCC) values output by scPagwas were used without weights (scDRS automatically assumes all weights are 1 if none are provided)[291]. The use of PCC values as weights did not lead to a significant difference, so only unweighted based results are discussed. Code to generate the scDRS input can be found in SCRNA-GWAS-Benchmarking/analysis/0A_Tool_Benchmarking/Genes/Gene_comparison.ipynb.

3.4.5 Benchmarking methods

All packages provide results indicating which cell clusters are significant for the disease, but the exact format and calculation of these results differs. scGWAS provides significance in the form of gene modules within clusters that have disease-relevance, whereas scDRS and scPagwas provide disease scores at the single cell and cluster levels. scDRS additionally provides measurements regarding the heterogeneity of these disease scores within each cluster. To compare results across the three packages, we defined significant cell clusters in scGWAS as clusters with at least one disease-

significant gene module. We then assessed whether the packages identified significant cell types similarly across a given disease. We also evaluated possible bias of scores from the health status of individuals and the sensitivity of scDRS to different numbers of top-ranking MAGMA genes (100, 300, 500, 1000, 1500, 2000). Additionally, we assessed the change in results of scGWAS to different pathway files (details in scGWAS and scDRS section above) according to both the significant gene modules and significant cell-states. Jupyter notebooks outlining these comparisons can be found at our github under SCRNA-GWAS-Benchmarking/analysis/0A_Tool_Benchmarking/Sensitivity and CT_Clusters. We also compared the genes considered most linked to the traits by the tools: scGWAS gives the significant gene modules, scDRS gives the correlation of gene expression to disease scores, and scPagwas gives the PCCs of gene expression according to a singular value decomposition method to calculate pathway activity scores in cells. We assessed the expression and correlation of significant gene modules identified by scGWAS or MAGMA top-ranking genes with scDRS and scPagwas disease scores, and compared scDRS and scPagwas correlation coefficients under SCRNA-GWAS-Benchmarking/analysis/0A_Tool_Benchmarking/Genes. Finally, the relationship of scDRS heterogeneity scores with cell-state population sizes and granularity was done with code under SCRNA-GWAS-Benchmarking/analysis/0A_Tool_Benchmarking/.

To compare genes, we analyzed the top 1,000, 500, and 100 genes ranked by MAGMA, scDRS, and scPagwas, as well as all significant gene modules identified by scGWAS. Using Gene Set Enrichment Analysis (https://www.gsea-msigdb.org/gsea/msigdb/human/compute_overlap), we examined gene sets enriched across our genes belonging to the Cell type (C8) collection, just Curated Pathways (C2-CP), or a combination of Hallmark, Curated (C2), Regulatory (C3), Biological Process (GOBP), and IMMUNESIGDB (C7-IMMUNE)[165, 236]. GSEA allows a maximum of 500 genes. We ran scGWAS with all significant gene modules collectively or individually for C8 to ensure logical results given the smaller gene numbers. We also conducted GO analysis with clusterProfiler_4.12.2 and org.Hs.eg.db_3.19.1[9, 242]. Code for this analysis can be found in SCRNA-GWAS-Benchmarking/analysis/0A_Tool_Benchmarking/Genes/Gene_comparison.ipynb.

To determine whether a single atlas could distinguish between two similar diseases, we ran

scDRS on the RA and UC cell atlases using MAGMA results from summary statistics of AS and CD GWAS, respectively. The code for analyzing scDRS results for this can be found under SCRNA-GWAS-Benchmarking/analysis/0B_Dist_path. The code for analyzing the effects of using different MAGMA window sizes and FUMA can be found under https://SCRNA-GWAS-Benchmarking/analysis/0C_Preproc.

3.5 Results

3.5.1 Single-cell disease scores allow greater sensitivity while gene-network analyses allow greater interpretability of gene targets

We built our initial benchmarking pipeline on evaluating both cell types and finer grained cell states as well as gene modules using RA and UC datasets.

3.5.1.1 Comparison of disease-significant cell types/cell states

At the scale of cell types, all tools imply significance of NK cells in RA (Appendix Figure B.1). Both scDRS and scPagwas identified T cells as significant, while scPagwas and scGWAS identified B cells as significant. scDRS alone determined Myeloid cells to be significant for RA (Appendix Figure B.1). For more specific cell-states, the three tools shared the same significance calls for 24/63 (38%) fine-grained cell states. In general, all three tools identified significant cell states within the T and B cell compartments. This overlap was particularly notable in the results from scDRS and scPagwas. scGWAS called only 20 significant cell states (45% with literary support) compared to the 46 (54% with literary support) and 43 (53% with literary support) calls from scDRS and scPagwas (Figure 3.2). scDRS alone identified MERTK+ myeloid cell states as significant[286, 123, 260]. scDRS still identified MERTK+ myeloid cell states as significant when using the same genes used by scPagwas (top 1000 correlated with gPAS cell scores) as input rather than the top 1000 MAGMA genes (Appendix Figure B.2). Additionally, scPagwas called all NK cell cell states significant for RA, while opposing subsets of NK cell states were called by scGWAS and scDRS (Figure 3.2).

There were a smaller number of significant cell types/states identified for UC. All tools identified epithelial cells as significant and T cells as not; all other cell types had mixed calls from tools (Appendix Figure B.1). For fine-grained cell states, all tools shared the same significance calls for 20/43 (47%), including M epithelial cells, Immature Enterocytes, and Secretory TA cells. Again, scPagwas called a high number of significant cell states (25, 44% with literary support) and was the only tool to identify most myeloid and fibroblast cell states as significant, including the inflammatory subtypes. scDRS and scGWAS showed similar numbers for significant cell states with seven (57% with literary support) and eight (50% with literary support), respectively (Figure 3.2). When running scDRS with the genes used by scPagwas, scDRS also identified the fibroblasts and non-mast myeloid cell states as significant (Appendix Figure B.2).

3.5.1.2 Significant genes

Significant modules identified by scGWAS are networks of genes that may represent a biological pathway and contain genes important for disease pathogenesis. scGWAS assesses these gene modules with each annotated cell type cluster. Notably, significant gene modules strongly align with functional annotations of their corresponding cell-states, as confirmed by gene set overlap analysis[165, 236](Supplementary Table 7). For example, T cell gene modules were frequently enriched with cytotoxic or T helper cell surface molecules while gene modules associated with NK cell states were enriched in genes involved in upregulating CD4 T cells and cellular responses to cytokines, chemokines, and cellular ligands. Many of these gene modules had overlapping genes and similar functions; despite having a total of 204 and 472 genes in NK and T cell cluster significant modules, there were only 63 and 87 unique genes, respectively. One gene in particular was found in nearly every significant gene module across cell states–CD2, which encodes for a surface antigen in all T cells and is involved with triggering T cells[243]. Both scDRS and scPagwas provide genes whose expressions correlate with the scDRS cell disease scores and scPagwas gPAS, respectively[291, 151]. The majority (59-85%) of the top 1,000 scoring genes in MAGMA, scPagwas, and scDRS are unique to each tool, while 75-90% scGWAS significant genes are identified by at

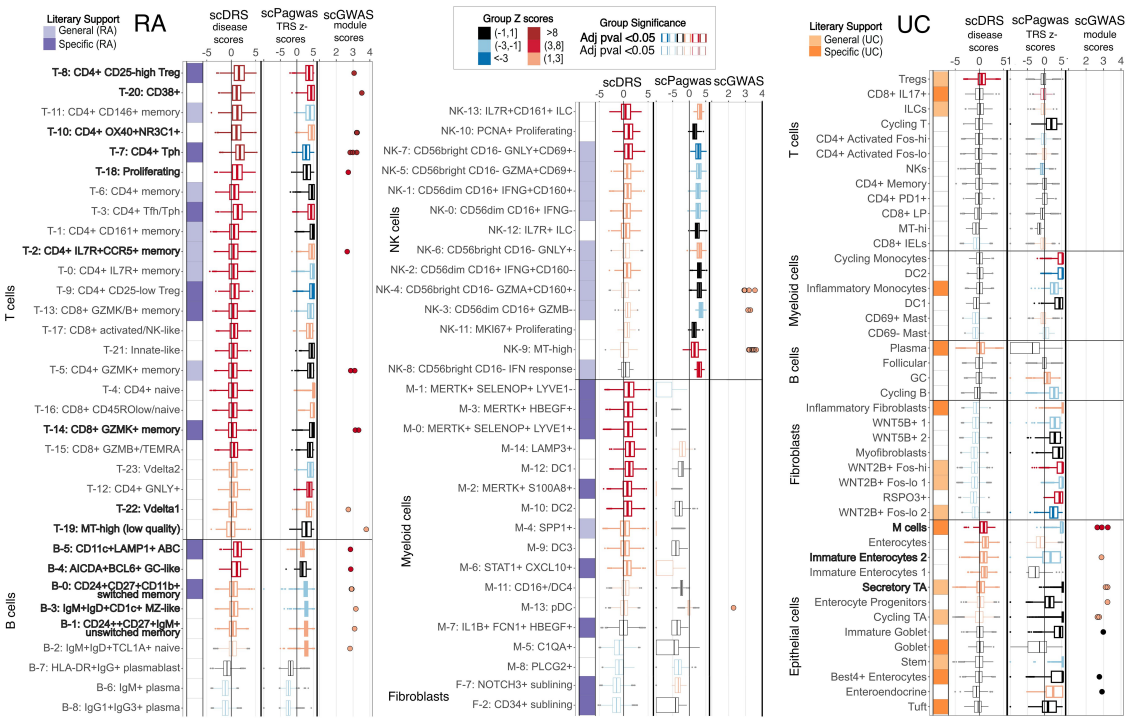

Figure 3.2: **Comparison of cell-state-specific significance results for RA and UC.**. For each cell-type and cell-state, the single-cell level scDRS Z-scores and scPagwas TRS Z-scores are displayed in boxplots colored according to the group scDRS Z-score or group scPagwas bootstrap Z-score. Non-significant cell states in scDRS or scPagwas are shown unbolded with grey outliers, while significant cell states are bolded. scGWAS called gene modules and their disease scores are also plotted with colors following the scDRS group Z-score gradient for easier comparison. Cell states considered significant by all three tools are bolded. "General literary support" means the general cell type has been shown to associate with the disease while "specific" denotes evidence in the literature linking the specific cell state. Left: RA (rheumatoid arthritis). Right: UC (ulcerative colitis).

least one other tool (Figure 3.3A). Additionally, significant genes from MAGMA and scGWAS show low median correlations to scDRS and scPagwas disease Z-scores (MAGMA: 0.02,0.05 for RA and 0.02,0.01 for UC; scGWAS: 0.06,0.09 for RA and 0.04,0.01 for UC) (Figure 3.3B). For RA, scDRS, scGWAS, and MAGMA but not scPagwas top ranked genes were enriched in myeloid cell type genesets (Supplementary Table 8). For UC, all tools except scGWAS showed myeloid cell-specific gene set enrichment, with scPagwas being the only tool to show significant enrichment for stromal terms in the top 50 pathways (Supplementary Table 9). The top 100 ranking genes for scPagwas were largely ribosomal genes regardless of the disease (43% and 68% in RA and UC, respectively)

while scDRS's top 1,000 genes contained very few if any (Figure 3.3B). Indeed, the top 20 enriched gene ontology terms for scPagwas were related to translation or general differentiation while scDRS was dominated by leukocyte-specific pathways (Appendix Figure 3). Gene sets uniquely enriched in scPagwas genes focused on translation, ribosomes, and general cell differentiation, unlike those specific to scDRS, MAGMA, and scGWAS which were immune-cell state or process focused (Supplementary Tables 8, 9). Removal of the ribosomal genes when using scPagwas genes as input to scDRS only led to one and four cell states to change in significance in RA and UC, respectively, compared to scDRS results using all scPagwas genes (Appendix Figure B.2).

3.5.1.3 Investigating result differences between pathway-based tools and scDRS

We first explored if variance in significant genes between methods might explain the different significant cell states identified by scGWAS and scDRS. We evaluated if the genes that most highly correlated with scDRS disease scores for cells in the MERTK+ cell states were found in networks in the original scGWAS pathway file and KEGG pathways. Indeed, pairs of genes that are strongly associated with scDRS disease scores were connected in the scGWAS pathway file, however, relationships between the genes beyond two were not supported and the 40 genes with the highest correlation to scDRS disease scores had only 6 pairings between them in the pathways file (Supplementary Table 10). The top 20 KEGG pathways uniquely enriched for MERTK+ cells according to scPagwas genetically associated pathway activity scores included Wnt signaling, cGMP-PKG signaling, and Inositol phosphate metabolism. We also explored the large discrepancy between NK calls across scGWAS and scDRS. As a controlled comparison, we looked at a cell cluster with strong agreement between scGWAS and scDRS: CD4+ Tph (T-7). scDRS disease scores in all cells positively correlated with the expression of the NK scGWAS module genes although T-7 scGWAS module genes had a slightly higher median correlation (0.08 vs 0.13) (Appendix Figures B.4A, B). This relative increase was maintained when the eight genes identified by scGWAS as significant for both groups were removed (median correlations 0.005 NK vs 0.02 T-7). Importantly, these correlations were comparable to that observed for all scGWAS genes and the top 100 genes ranked

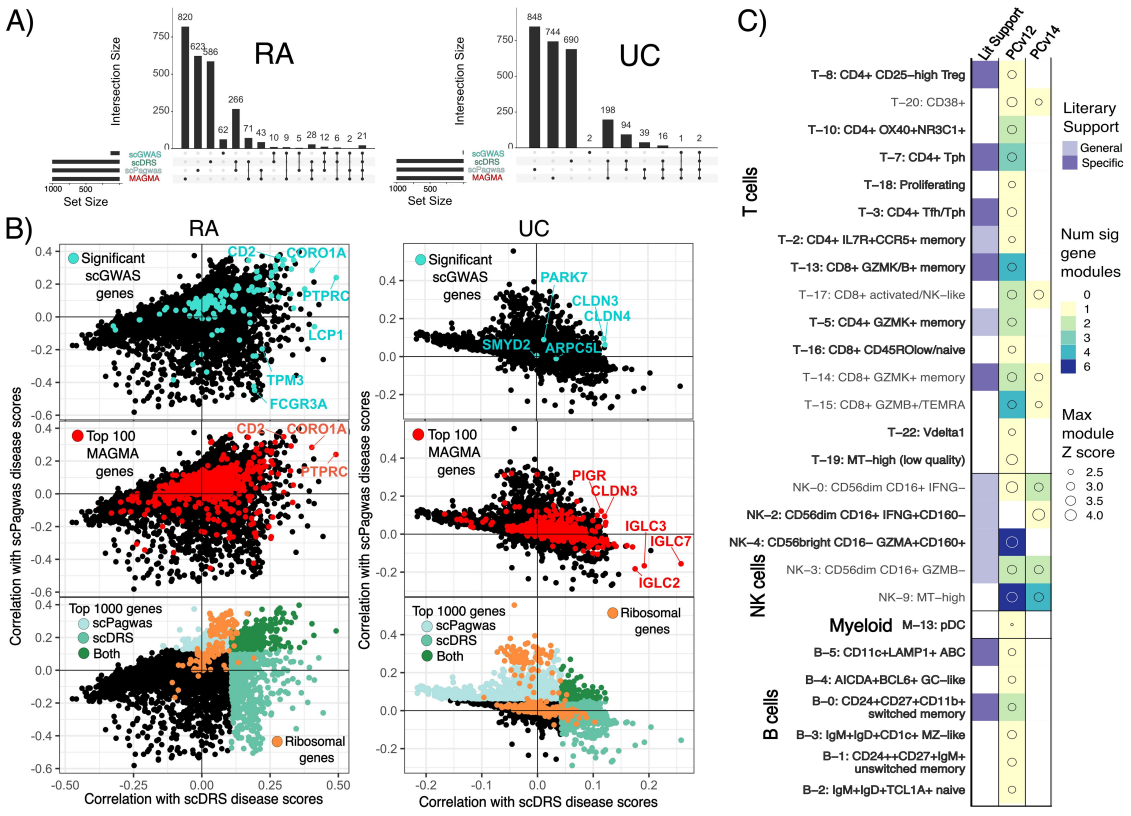

Figure 3.3: **Gene comparisons show low correlation across tool-based genes and single-cell disease scores**. **A.** UpSet plots of the top 1000 ranked genes for scDRS (highest correlation to scDRS disease scores), scPagwas (highest correlation to genetically associated pathway activity scores) and MAGMA as well as the significant scGWAS genes. RA=Rheumatoid arthritis, UC=Ulcerative colitis. **B.** Scatter plots of the correlations of all studied genes with scDRS disease scores and scPagwas gPAS with (top) scGWAS genes, (middle) MAGMA genes, or (bottom) ribosomal genes highlighted. Genes reaching the top 1000 ranked genes for scPagwas and scDRS are colored in light and dark turquoise, respectively. **C.** scGWAS results when using a pathway file based on Pathway Commons v12 or 14. Results are highlighted according to the number of significant gene modules called per RA cell state and max disease Z score across the modules for each cell state. Only cell states with a significant gene module from using either pathway file are shown. Cell states without a significant gene module called when only one of the pathway files was used are bolded.

by MAGMA with scDRS disease scores (Medians of 0.01-0.09) (Figure 3.3B). Median correlations decreased when only considering cells within the corresponding cell states (NK-cells & T-7) unlike those of the top ranking scDRS genes for each cell state (Appendix Figures B.4C, D, B.5). These findings led us to assess the impact of the pathway file used by scGWAS on results. When using gene pairings from Pathway Commons v14 instead of v12 (see Methods for details), 20 RA and 8

UC cell states changed in whether they had at least one significant gene module identified. Of these, 13 RA cell states and 1 UC cell state had been originally called significant by scDRS, scGWAS, and scPagwas (Figures 3.2, 3.3C, Appendix Figure B.6A). Extending the gene-SNP linking window from 10-10kb to 50-35kb resulted in 14 cell states no longer having a significant gene module (Appendix Figure B.6B). Despite having 319,042 more gene pairings, use of Pathway Commons v14 led to an overall decrease in significant gene modules called regardless of window size used. Even when cell states were called with both pathway files, the genes within significant gene modules were also dependent on pathway input despite all changing genes being found within both pathway input files (Appendix Figures B.7, B.8).

While scDRS single cell disease scores followed an expected normal distribution, disease scores from scPagwas or from scDRS run with scPagwas genes showed large polarization (Figure 3.2, Appendix Figures B.2, B.9). Specifically, 23% and 12% of cells in RA and UC, respectively, had scPagwas Z-scores of -10 despite the next nearest Z-score being -5. These percentages decreased to 17.5% and 3% when applying the scDRS framework to scPagwas genes, and further to 15% and 3% when ribosomal genes were removed for RA and UC, respectively. These cells were distributed across cell states, although most were found in plasma and MERTK+ cells for RA (Appendix Figures B.9, B.10).

Finally, although all tools may be impacted by covariates within the data, only scDRS allows for their inclusion for batch-effect analysis. In both RA and UC datasets, certain cell states contain significantly different proportions of cells from individuals according to health status (Appendix Figure B.11). scPagwas shows clear, significant differences in its single cell trait relevant scores, whereas scDRS exhibits minimal to no batch effects (Appendix Figures B.12, B.13). When scPagwas genes are used, biases in scDRS disease scores related to health status become more pronounced but remain less substantial than those in scPagwas disease scores (Appendix Figures B.12, B.13).

3.5.1.4 Additional features

Although all scDRS additional features are outside the scope of this work, we evaluated the usage of the tools' group-level metric to consider the heterogeneity of disease scores within a cell state[291]. This metric can hypothetically indicate if a provided cell state has inner-clusters of cells that should be further separated out based on the groupings of disease score. All large-scale cell types in RA (T cell, B cell, Myeloid, NK, Fibroblast, Endothelial) had significant heterogeneous disease scores that positively correlated with the number of cells (adjusted R2 0.29) and annotated clusters in each group (adjusted R2 0.37) (Appendix Figure 14). Eighty-seven percent (67/77) of RA fine-scale cell states had significant levels of heterogeneity in disease score with similarly low positive correlation with the number of cells (Figure 3.2, Appendix Figures B.15, B.16).

3.5.1.5 Resources

Despite these additional features and working at the single-cell level, scDRS was the most robust in memory usage and speed, although this is primarily due to the initial preprocessing step for scGWAS (Table 3.2). scPagwas took the longest by 45 hours compared to scDRS and 32 hours compared to scGWAS (Table 3.2). Notably, the number and size of cell states had a negligible effect on resource usage in scDRS and scGWAS unlike scPagwas.

3.5.2 scDRS can distinguish similar diseases from pathological cell clusters

While atlases with fine-grained annotations may allow more detailed analyses, it raises the question of whether a single atlas can still be used to study multiple diseases. This is particularly relevant for diseases without single-cell data available. Given the high sensitivity of single-cell disease scores, we used scDRS to assess the feasibility of using one atlas to identify pathological cell clusters distinguishing similar diseases. We used summary statistics from GWAS for RA and ankylosing spondylitis (AS) on the scRNA-seq data from inflamed RA synovial tissue to determine if scRNA-seq from a clinically similar disease can provide fine-grained insight on disease-relevant clusters[286, 95, 103]. We also applied the GWAS statistics from UC and crohn's disease (CD)

Table 3.2: Resource usage of each package when running for the RA cluster-level data. Memory used refers to the max amount of memory required for a single step. All tools were run with 15 CPUs

Package	**CPU used (time)**	**Wall clock time**	**Memory Used**	**Relevant Function (script)**
scDRS	00:00:05	00:00:07	488 KB	Preprocess GWAS stats (run_scdrs.sh)
scDRS	00:54:13	00:38:43	12.26 GB	Compute single cell scores(run_scdrs.sh)
scDRS	00:23:41	00:25:11	17.89 GB	Cell-type scores & Gene analysis (run_scdrs.sh)
scGWAS	04:32:03	04:33:37	208.4 GB	Preprocessing single cell data (process_sc_data_R.sh)
scGWAS	08:50:26	08:50:24	2.55 GB	Running scGWAS (run_scGWAS_2023_clusters.sh)
scPagwas	1-16:48:25	1-21:47:25	185 GB	Running scPagwas
scPagwas		1-19:00:00		Link GWAS and Pathway block annotations

on the scRNA-seq data from UC colon tissue[227, 49]. We considered both 10-10kb and 50-35kb window sizes on these analyses, focusing main figures on 50-35kb window results due to the larger number of significance calls.

3.5.2.1 RA and AS

Although both analyses used the same scRNA-seq atlas references, scDRS successfully distinguished RA from AS. We identified 46 candidate cell clusters in RA and 23 in AS, with 10 clusters shared between the two diseases. We found that while most T, myeloid, and B cell cell-states were significant for RA, very few were significantly associated with AS (Figure 3.4A). CD8+ activated/NK-like (T-17), pDC (M-13), and unswitched memory cells (B-1) were significant for AS. AS and RA showed the greatest differences across the T, NK, and myeloid cells. While essentially all T cell states showed significance for RA, only CD8+ activated NK-like (T-17) and proliferating

(T-18) T-cells showed significance for AS. Conversely, far more NK cell clusters were called significant for AS[195, 141]. Specifically, most of the CD56bright CD16- (NK4,6,8) NK cell clusters were called significant for AS. This AS and RA separation was consistent when using different MAGMA windows (Appendix Figure B.17).

3.5.2.2 UC and CD

Although fewer significant cell states were identified for UC and CD (eight and six, respectively) (Figure 3.4B), we still observed differences in pathological cell types. None of the significant cell-states were shared between UC and CD. Epithelial cells linked to UC and fibroblasts linked to CD most clearly distinguish the diseases, a finding maintained when using different MAGMA windows (Appendix Figure B.18). For example, we found that NK cells, CD4+ activated, and CD8+ lamina propria (LP) cells were enriched in CD compared to UC while only Tregs, CD8+ IL17+, and Cycling T cells were enriched in UC.

3.5.3 Positional SNP-gene linking methods provide greater statistical power than tested alternatives

Methods integrating scRNA-seq and GWAS summary statistics rely largely on the same preprocessing steps, yet a standardized guidance for these steps is lacking. Therefore, we evaluated the impact of inputs and preprocessing steps on results, focusing on scDRS due to its high sensitivity and covariate analysis.

First, we considered the robustness of results when using solely positional information to connect noncoding SNPs to genes. The primary positional method to link SNPs to genes is MAGMA which relies on a window size parameter determining the distance a SNP can be from a gene to be incorporated[134]. Because there is no standardization on MAGMA window size beyond the notion that a larger window size incorporates SNPs falling in cis-regulatory elements, we evaluated the impact of the most used window sizes on results (details in Methods)[291, 102, 65, 263, 230, 82, 264, 226, 303]. Different window sizes for RA analyses only changed the significance calls for

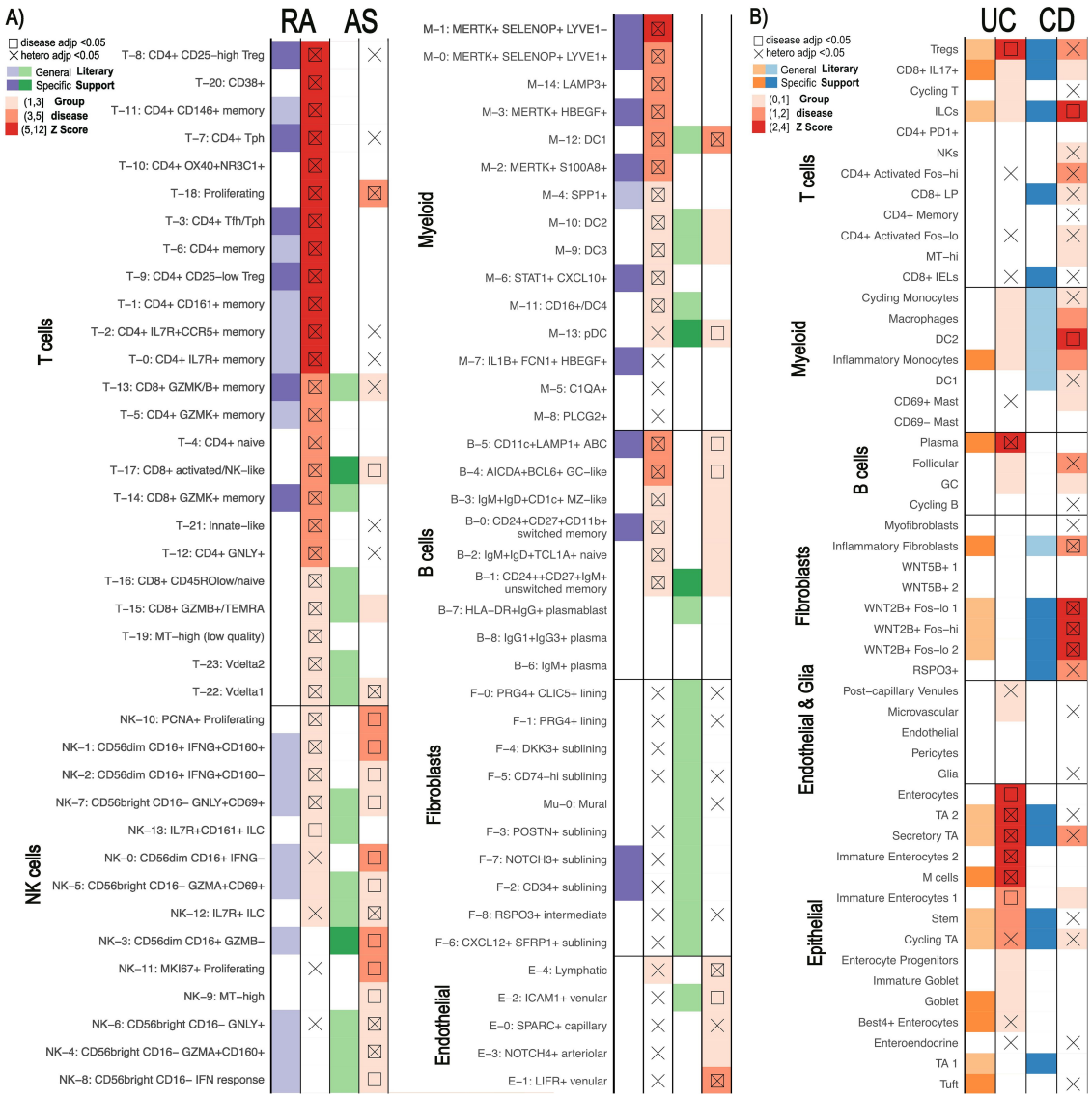

Figure 3.4: **Comparison of similar diseases with scDRS.**. Summary statistics unique to each disease were used on the same scRNA-seq data for each pair[286, 227]. scDRS defines significant clusters (annotated according to original papers) with a group disease Z-score as shown in the gradient legend. Cell clusters with literary support for either disease are labeled in purple/orange for RA/UC and green/blue for AS/CD, respectively. General literary support means that a cell type with multiple cell states is supported by the literature while specific means a specific single cell state was supported. **A.** Rheumatoid arthritis (RA) vs Ankylosing Spondylitis (AS). **B.** Ulcerative Colitis (UC) vs Crohn's Disease (CD).

16 of the 77 cell states in at least one of the window-sizes, half of which are only different in one window size (Figure 3.5). Importantly, none of these cell states had the top 20 group disease scores

in our original results (50-35kb window). There also did not appear to be a clear pattern across the window sizes in terms of the numbers of significant cell states or the cell states changing in significance. These findings were similar with our three other diseases of study, with results for CD having the greatest differences across window sizes (Appendix Figures B.17, B.18). Despite only 54% of genes being shared across the top 1000 MAGMA ranked genes in all window sizes, these shared genes consistently had most of the lowest p-values (Appendix Figure B.19). In comparison, scGWAS showed 20 cell states with change in significance just between 10-10kb and 50-35kb window sizes in RA, including four cell states originally identified as significant by all three tools: T-22, B-5, B-0, and B-1 (Figure 3.3C, Appendix Figure B.6).

Given the growing concern over positional methods inaccurately assigning SNPs to genes, we next explored the usage of non-positional based data within the framework of FUMA. Although other SNP-gene linking tools can be found in Table 3.3, we focused on FUMA as a commonly used alternative to MAGMA and because it can incorporate eQTL, chromatin contact data and positional information from MAGMA to express summary statistics at the gene-level[134, 263, 69, 266, 272]. Therefore, while FUMA uses a different summary statistics processing that doesn't allow direct comparison to our own MAGMA based analyses, we used its MAGMA pipeline to consider the impact of alternative linkage methods (details in Methods). The 1000 genes with the lowest p-values were significantly different between positional and non-positional methods, regardless of exact summary statistics used (Appendix Figure B.20). When only considering genes supported from non-positional methods, 445 genes were significant, a number consistent across usual non-positional methods (Supplementary Table 9, Table 3.3). : The smaller number of genes was maintained regardless of p-value cutoff (Supplementary Table 11). Indeed, FUMA analysis that combined positional with non-positional methods showed similar results to purely using MAGMA but with only 28 of the 52 original cell states called significant (Appendix Figure B.21). Conversely, scDRS only lost nine and five significant cell state calls when only using the top 300 and 500 ranking genes according to MAGMA, respectively (Appendix Figure B.21). Only restricting scDRS to the top 100 ranking genes allowed loss of significant results at the same magnitude (23 vs 24 by FUMA)

(Appendix Figure B.22). Still, incorporating non-positional methods added 2 significant clusters: HLA-DR+IgG+ plasmablasts (B-7) and MKI67+ Proliferating NK cells (NK-11), which were still not called significant when increasing the MAGMA window size to 100kb, a size commonly used to capture cis-regulatory element SNPs (Figure 3.5).

Table 3.3: Current methods to link SNPs to genes and the estimated number of genes output, form of significance output, and interface. All tools address linkage disequilibrium

Name (Citation)	**Method**	**Est. Gene list size**	**Score**	**Interface**
cS2G[69]	Linear combination of linking scores from main S2G strategies, exon, promoter, eQTLGen, and GTEx cis-eQTL, EpiMap, ABC, and Cicero. Restricts each strategy to gene w/ highest linking score	< 500 (depends on # lead variants)	cS2G score	Scripts provided
PoPs[266]	Similarity based filtering of MAGMA results (although paper described other input options)	<200 (depends on # lead variants)	PoPs score (for relative ranking)	CLI
nMAGMA[272]	Network-enhanced MAGMA links SNPs to genes by considering tissue specificity (Hi-C and eQTL) and functional interactions (WGNCA), then use MAGMA to get significance of genes	1000+	Z-scores and P-values	Scripts provided
FUMA[263]	SNP2GENE Module: Identifies lead SNPs, can run MAGMA or map using eQTL, position, and chromatin-interaction	MAGMA based 1000+, otherwise <700	MAGMA Z-scores/P-values or min p-value of linked SNPs	Web tool
MAGMA[134]	Maps SNPs to genes via positional window, empirical gene p-value via permutation followed by PCA regression	1000+	Z-scores and P-values	CLI

3.6 Discussion

In this study, we evaluated three software for linking genetics to single-cell phenotypes according to the enrichment of literature supported calls, robustness, and interpretability of results. Although all strategies identified disease-relevant cell states, single-cell based scDRS and scPagwas identified the greatest number supported by previous findings. B and T cell subsets were identified as significant for RA across all tools, aligning with the literature highlighting the disease relevance of lymphocytes[191, 286, 287, 260, 279, 268]. Gene set enrichment analyses indicated the significance of monocytes and macrophages across all tools for RA, consistent with the recent work discovering the cell phenotype expanded in inflamed synovial tissue. However, only scDRS called the best defined RA induced cell states, MERTK+ myeloid cells, significant[286, 123, 260]. In addition, all methods recognized autoimmune-associated B-cells (ABCs) as significant, a cell phenotype recently shown to be expanded in RA inflamed synovial tissue[286, 123, 260]. Importantly, none of the algorithms identified significant fibroblast cell types despite the expansion of NOTCH3+ and CD34+ sublining fibroblasts in RA[287, 182]. This finding supports previous hypotheses that these phenotypes arise only after the expansion of other genetically driven cell states called significant by scDRS[182]. For UC, we found few disease-significant cell states. However, all methods identified M cells – a recently discovered cell group with the highest expression of putative IBD risk genes in inflamed vs healthy tissue corroborated by two separate cohorts[227, 217]. Interestingly, no algorithm called CD8+ IL17+ T cells despite their significantly different proportions between individuals with and without UC[227, 116]. However, transcriptional changes in this group occur downstream of proportional shifts of Tregs and epithelial cells, both of which were called by scDRS[37, 193, 271].

scGWAS is more distinctly built to identify probable gene sets relevant to pathological cell states, but is significantly impacted by the pathway networks on which it bases its analyses. While removing false positives by requiring a known set of connected genes to have increased expression compared to single genes, the algorithm also assumes that the pathway file contains all possibly

relevant gene connections. Therefore, true positives can be lost such as was likely with MERTK+ cells. Additionally, many of the significantly called scGWAS gene modules overlapped, depleting information content, perhaps due to the lack of cell type specificity in the pathways. This finding underscores the importance of not necessarily using the number of significant gene modules identified as a relative metric of significance for a cell type. Although scGWAS provides gene modules more conducive for certain analyses, the original network file should be considered according to a researcher's specific focuses. In contrast, scDRS focuses on single cell based exploration by only providing genes correlated with single-cell disease scores[291]. Historically, purely correlational approaches tend to be noisy and significantly impacted by data heterogeneity[21, 83]. This fact might explain why both MAGMA and scGWAS genes showed relatively low correlation with single-cell disease scores, even within the annotated cell-state.

Although scPagwas uniquely integrates gene pathways with single-cell scoring, it currently has three limitations compared to scDRS. First, the computational expense of scPagwas makes scDRS far more feasible for large scale analyses; this could potentially be addressed by enabling multiprocessing for the current bottleneck in linking pathway blocks and GWAS, as done in the regression portion. Second, scPagwas currently lacks covariate adjustment, making it susceptible to batch effects, which may explain the highly polarized disease scores observed in scPagwas mitigated by scDRS. Finally, while both scDRS and scPagwas consider genes correlated with single-cell disease scores, scPagwas relies on these genes—rather than SNP-linked genes—for final cell-type analysis. Our results suggest that gene correlations can be heavily influenced by dataset heterogeneity and often poorly reflect SNP-based gene associations (e.g. MAGMA). This finding may help explain the overrepresentation of ribosomal genes among scPagwas genes despite their minimal impact on cell-state identification. Importantly, these results might also be based on the pathway size of scPagwas (default 5-300 genes); this range was optimized by the original authors but may require further optimizing for more heterogeneous datasets like those tested here. The scDRS simulated control set may also allow a more accurate prediction of significance given scDRS using scPagwas gene input, but not scPagwas, called MERTK+ cells significant despite the MERTK+ genetically

enriched scPagwas pathways being linked to RA[257, 52, 2, 265].

Importantly, the use of broad cell types, as mostly done in previous applications of scDRS, scPagwas and scGWAS, lacked the insight provided by fine-tuned cell state annotations. Indeed, all tools missed calling some cell types significant despite them calling significant cell states within them. The heterogeneity of disease scores as called significant by scDRS might indicate when a cell type, even when not called significant as a group, might contain cell states with significance. However, statistically significant heterogeneity does not always imply biological significance, as even small cell states with as few as 50 cells showed significant heterogeneity. Similarly, potential biases from including cells from diseased tissue in these atlases must be considered. For example, scDRS relies on normalized single-cell scores so statistical significance is partly driven by the comparison of cells. Despite these caveats, we were able to explain the lack of significance for certain cell states according to lack of genotypic support in the literature and their links to upstream cell states that had genotypic backing.

Given the increased sensitivity when using fine-grained cell states, we evaluated whether a single atlas could be used to assess multiple diseases. scDRS clearly distinguished between diseases with a single atlas, with literary support for the found differences from other single-cell based analyses. We were able to determine RA vs. AS and UC vs. CD pathogenesis based on the results of scDRS, using one scRNA-seq atlas for the respective comparisons. Cell states causally linked to AS according to a recent Mendelian randomization study were all called significant in AS: CD8+ activated/NK-like (T-17), pDC (M-13), and unswitched memory cells (B-1)[64]. Additionally, CD8+ activated NK-like (T-17) and proliferating (T-18) T-cells showed significance here and in other studies[74, 270]. NK cells were heavily implicated in AS. The unique significance of CD56dim CD16+ GZMB- cells (NK-3) in AS was supported by GZMB being expressed at much lower levels in AS patients in previous NK-focused scRNA-seq analysis and ELISA studies[195]. Similarly, the significantly called IL7R+ ILC (NK-12) cell state showed similar upregulation of genes, including IL7R, as a NK cluster upregulated in AS according to previous single cell analyses[195, 141]. Finally, most of the CD56bright CD16- (NK4,6,8) NK cell clusters were called significant for AS,

supported by the previous findings of upregulation of CD56bright NK cells in AS[195, 141]. On the other hand, epithelial cells and fibroblasts most clearly separated UC and CD respectively. Indeed, the enrichment of CD8+ LP cells, NK cells, and activated CD4+T cells has been supported by independent CD single cell analyses[96]. We were also able to distinguish fibroblasts with genetic bases for CD and UC. We called RSPO3+ fibroblasts significant when multiple CD specific SNPs have previously connected this phenotype[110]. Similarly, WNT2B+ fibroblasts were only called significant for CD, matching the previous finding that the group only shows genetic connection to CD despite it being expanded in both UC and CD[24, 63]. Publicly available scRNA-seq data is not always available or sufficient for a certain disease, so instead researchers might need to apply the existing and relevant GWAS summary statistics to the scRNA-seq data generated from a clinically similar disease. Our findings support the ability for researchers previously constrained by the lack of appropriate scRNA-seq atlases to study diseases while not sacrificing fine-scale analyses.

Finally, we also evaluated methods incorporating noncoding SNPs for identifying pathogenic cell states. Unsurprisingly, the input gene set used can have major implications on results, regardless of the tool. We determined that MAGMA-based results in scDRS are robust to window sizes while scGWAS appeared to have larger changes. This different robustness might be explained by our finding that the genes consistent across window sizes had the highest significance scores while scGWAS considers the full list of MAGMA based scores rather than the top 1000. We also considered non-positional methods to link SNPs to genes with FUMA and found the decreased power from these tools have significant impacts on results. Non-positional methods provide significantly smaller genesets due to a focus on highly confident linkages and noisy data sources (Table 3.3). Our findings show that these low gene numbers, regardless of confidence, lead to significant decline in sensitivity. Ideally, one would be able to combine strict window MAGMA results with that of a non-positional method, however the need to combine different significance scales complicates this. The p-values output by FUMA and similar methods also often do not account for the uncertainty in the predicted SNP-gene linkages. For now, if using tools reliant on a long list of genes, we suggest focusing on cell types consistent across window sizes for MAGMA and adding genes called by other

tools like FUMA. It's important to note that regardless of the window sizes used, many SNPs were still not assigned to a gene with MAGMA. For example, with a moderately large window size of 50-35kb, about 60% of SNPs for RA and UC were linked to a gene which decreased to about 40% when that window was reduced to 10-10kb. Outside of these methods, repeating analyses with multiple GWAS summary statistics and scRNA-seq cohorts is equally relevant to ensure repeatability of results.

One way to circumvent linking SNPs to genes is using cis-regulatory elements (cREs) SNPs fall in directly. Given cRE activity is highly dependent on cellular behavior and allows accurate deconvolution of cell types, this switch could also allow separation of more nuanced cellular states[152]. Additionally, tools like Cicero link cREs to their regulated genes from single cell data[184]. While classic scRNA-seq data cannot capture the activity of these elements well, 5' scRNA-seq is more sensitive to them. Moody et al.[163] successfully applied 5' sc-RNA-seq to detect the transcription of cREs and genes simultaneously and developed a metric to identify cell types enriched in trait heritability. Interestingly, they used the same summary statistics as our work for crohn's disease (CD) and ulcerative colitis (UC). Despite using gene-based methods, we captured the same fibroblast and dendritic cell enrichment for CD that they found. However, unlike their results, we did not find an overall enrichment of T/NK cells in UC compared to CD but found some specific states in these cell types oppositely enriched and supported by the literature[96]. These differences can be explained by the fact that Moody et al. relied on general lymphocyte 5'-scRNA-seq for analysis while we used scRNA-seq specifically from the colon mucosa of UC patients. The cell states we identified as seeming to conflict with findings from Moody et al. are unique to intraepithelial lymphocytes and likely would not be in their data. Overall, these results showcase the need for careful interpretation when relying on non-disease tissue specific scRNA-seq data. Exciting insight will come from evaluating the adaptation of algorithms like scDRS, scPagwas, and scGWAS to the growing cRE-based single cell data[163, 267, 145, 222].

While disease-specific and fine-scaled single-cell cRE atlases continue being developed, tools like MAGMA, scGWAS, scPagwas and scDRS provide key opportunities to identify cell states and

genes associated with disease through both transcriptomics and genomics. We've also showed that these tools can even allow single-cell level analyses for diseases without fine-scaled sc-RNA-seq atlases currently accessible if an atlas for a similar disease is available. We note that our focus on four immunological diseases, including RA, AS, UC, and CD, may not be generalizable to all other disorders. However, these analyses represent the consistency of key genetic-relevant cell phenotypes across autoimmune disorders, providing valuable guidance for future investigations to other similar diseases. Overall, the development of tools like scDRS, scGWAS, scPagwas, along with improved SNP-Gene-cell state linking methods, are essential steps for using existing data to pinpoint the search of biological targets for treatment development.

3.7 Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

3.8 Ethics statement

The studies involving humans were approved by their respective Institutional Review Boards. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

3.9 Author contributions

HT: Conceptualization, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. KR: Formal analysis, Writing – original draft, Writing – review & editing. LV: Writing – review & editing. JI: Conceptualization, Supervision, Writing – review & editing. FZ: Conceptualization, Funding acquisition, Supervision, Writing

– review & editing.

3.10 Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the Interdisciplinary Quantitative Biology (IQ Biology) PhD program at the BioFrontiers Institue, University of Colorado Boulder with the NSF NRT Integrated Data Science Fellowship (award 2022138), and the National Science Foundation NRT Integrated Data Science Fellowship (award 2022138). The PhRMA grant and the Arthritis National Research Foundation grant to FZ and the Curci Scholarship from the Shurl and Kay Curci Foundation (to HT) also enabled this work.

3.11 Acknowledgments

We appreciated the constructive feedback from the Zhang Lab members and some preliminary literature search done with Alexandra Griffin from the University of Colorado Boulder Molecular, Cellular, and Developmental Biology department. This work would also not have been possible without the IT support from CU-Anschutz Medical Campus. Finally, we appreciate the valuable insights of Dr. Kristine Kuhn from the University of Colorado Department of Medicine Division of Rheumatology, particularly regarding the clinical relevance of findings.

3.12 Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.13 Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the

reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

3.14 Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1454263/full#supplementary-material

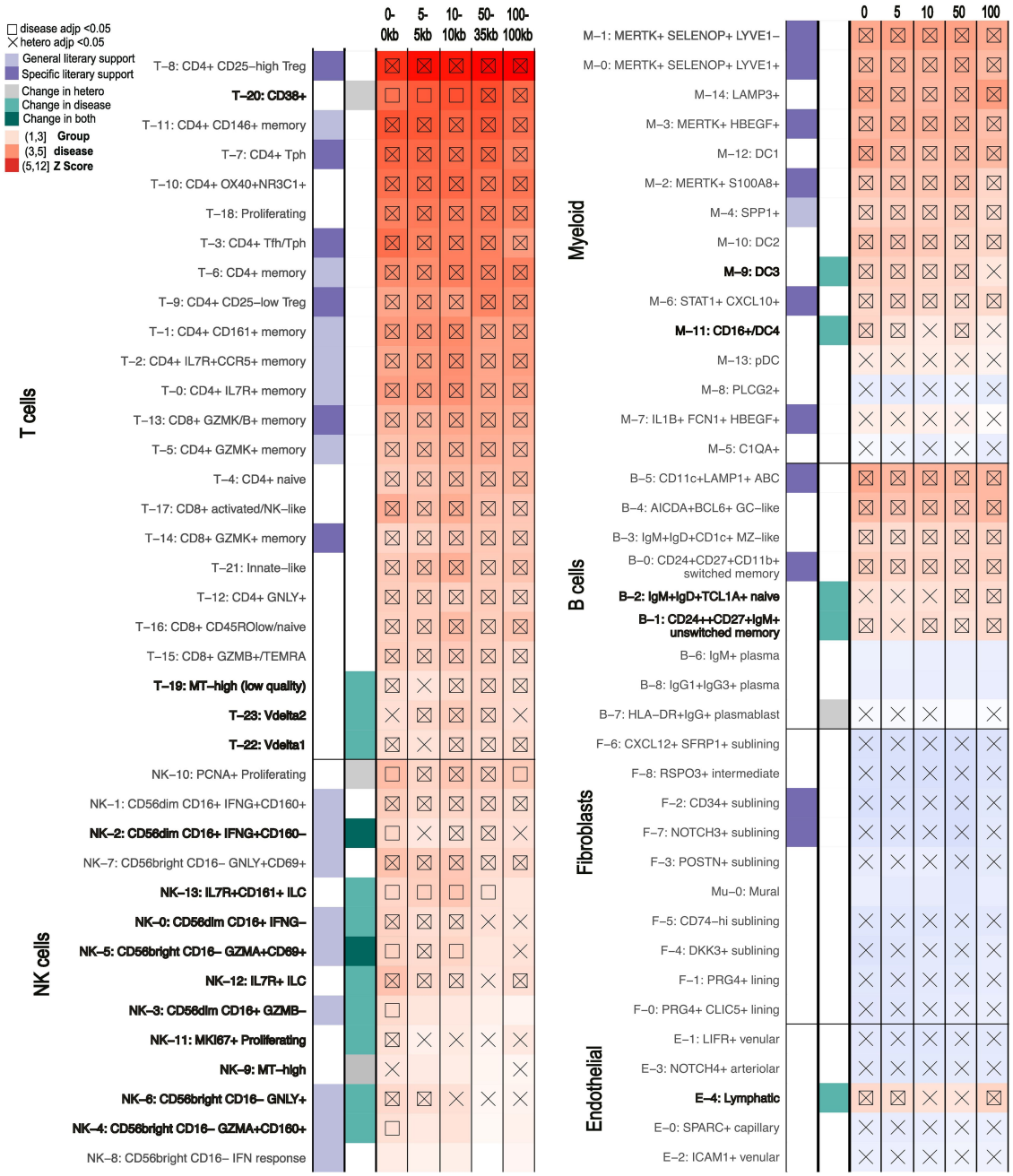

Figure 3.5: **MAGMA Window impacts on scDRS results.**. scDRS results for RA of clusters that show different levels of significance with different MAGMA windows being used to generate the GWAS inputs (0-0kb, 5-5kb, 10-10kb, 50-35kb, 100-100kb). scDRS defines significant clusters with a group disease Z-score as shown in the gradient legend (significant scores marked with square). Cell states with significant heterogeneity scores are marked by an X. General literary support means that a cell type with multiple cell states is supported by the literature while specific means a specific single cell state was supported. Cell states with changes in just scDRS disease score, heterogeneity score, or both significance calls across MAGMA windows are marked in bold and with grey or turquoise squares.

Chapter 4

Improving calls of differentially transcribed enhancers and their upstream regulators

This chapter is adapted from: **Townsend, H.A.**, Stanley, J.T., Allen, M.A., Dowell, R.D. Improving confidence of differential transcription calls in enhancers. [Under Review]. doi: 10.1101/2025.09.12.675852.

All supplemental tables can be found with the published manuscript. Supplemental Notes, Methods, and figures are found in the Appendix B but were originally published as supplemental material.

4.1 Contribution Statement

Drs. Mary Allen and Robin Dowell conceived of the original project. Dr. Dowell and I designed the algorithms for Mu Counts and the edited format of LIET. Dr. Dowell conceptualized the Leading Edge and I designed and implemented the final algorithms of it used in this work. All three of us conceptualized how to evaluate the algorithms. I implemented the algorithms, tested them, and analyzed data. Dr. Jacob Stanley confirmed accuracy of code and changes to LIET and provided general feedback. Dr. Dowell and I wrote the manuscript published on BioRxiv, which was then reviewed by all authors. I have edited that version to make the one here.

4.2 Abstract

Motivation: Most disease-associated genetic variants reside within transcribed regulatory elements (tREs). Patterns of differential transcription at tREs can be leveraged to identify upstream regulators and link enhancers to their target genes. But the low transcription levels and high variability in tREs makes identifying high confidence differentially transcribed elements challenging.

Results: We present Mu_Counts and TFEA-LE, two algorithms for robust identification of differentially transcribed tREs. The first step in accurate identification of differentially transcribed tREs is to obtain accurate RNA lengths and therefore counts over these regions. To this end we developed a method of accurate length inference (LIET-EMG) as wll as a rapid method for counting reads over tREs (Mu_Counts). Armed with newly identified and quantified tREs, TFEA-LE then integrates motif information to simultaneously identify responsive tREs and their likely upstream regulators. We show improved precision and recall over general-purpose tools (e.g. DESeq2) in detecting p53-responsive tREs. We then clarify TF-specific responses within multi-TF perturbations in lung cells. Finally we show that the TFEA-LE approach improves TF activity inference, including in complex perturbations where many TFs respond. TFEA-LE is especially effective in technically challenging datasets, whether due to highly specific or broad responses, outliers, or high GC content. Ultimately, these methods advance the systematic characterization of individual tREs, enabling their integration with regulators, target genes, and disease-associated variants for translational research.

Availability and Implementation: TFEA-LE: https://github.com/Dowell-Lab/TFEA/tree/Lead_edge. Nextflow pipeline to run Mu_Counts: https://github.com/Dowell-Lab/Bidir_Counting_Analysis. LIET (including modifications for tREs): https://github.com/Dowell-Lab/LIET/tree/LIET_EMGtoo. Source code for this work: https://github.com/Dowell-Lab/Improving_tRE_Analysis_Paper

4.3 Introduction

Enhancers are genetic sequences that regulate the transcription of genes—thereby allowing coordinated cellular responses and unique transcriptional states. Enhancer activity is historically most commonly measured by chromatin accessibility (ATAC-seq) or epigenetic markers (e.g. H3K27ac ChIP-seq). However, recent work on identifying enhancers has demonstrated that most enhancers produce lowly transcribed RNAs, hence they are more generally referred to as transcribed regulatory elements (tREs)[278]. These enhancer-associated RNAs have been found to be more reliable markers of local regulatory activity than epigenetic markers [278, 223, 6]. Despite several high-throughput measurements of enhancer activity, tissue and perturbation specificity of enhancers leads to high biological variability between samples [6, 121, 223].

Measurements of enhancer activity and transcription can vary significantly not only due to biological factors, but also technical artifacts such as depth, protocol, and analysis choices [278, 92]. This high variability complicates both the identification of tREs and detection of when they are changing [149]. Consequently, most efforts to date have focused on using tRE meta-profiles (rather than individual tREs) across conditions or samples to infer upstream regulators[200, 108, 48] or to link enhancers to their target genes[138, 223]. In these scenarios, general trends are detectable even if some changing enhancer RNAs are missed.

However, genome-wide association studies are revealing a growing need to confidently characterize individual tREs. The majority of disease-associated single-nucleotide polymorphisms (SNPs) fall within enhancers [159, 39, 284]. Given many SNPs can be inherited together (reside within linkage disequilibrium), functional data provides critical information for identifying specific causal variants[68, 209]. The relatively small lengths of tREs and their condition- and tissue-specific activities aid in fine-mapping of associated SNPs. This is particularly true when the tRE and its target gene both respond to a disease relevant perturbation[209]. These coordinated changes effectively pinpoint both the functional SNP and its relevant target, thereby implicating biological pathways, relevant cell types, genes, and/or upstream regulators. Consequently, annotating SNPs

within enhancers holds tremendous potential to aid in the identification of potential drug-targets [159, 39, 284].

Yet identifying cell type or condition-specific tREs requires precise detection of changes in individual tREs. tREs have high variability and low transcription compared to genes, making them particularly challenging for calling as differentially transcribed with high confidence[223]. Although *in-vitro* methods like massively parallel reporter assays provide a method for high-throughput evaluation of tRE activity, they do not represent *in-vivo* conditions and can lead to inaccurate indication of SNP relevance and biological mechanisms [228]. Consequently, we sought to develop a pipeline for robustly identifying and characterizing differentially transcribed tREs from high-throughput nascent run-on sequencing and similar enhancer-focused sequencing data (e.g. ATAC-seq). Our overall framework builds on prior work on transcription factor enrichment analysis (TFEA)[200], leveraging both transcription changes (via ranking metrics) and motif information to support the identification of high confidence changes in tRE expression. With this approach, we simultaneously improve identification of responsive tREs as well as the upstream regulators that activate them. Additionally, we provide the first high-throughput length estimation of tRE RNAs, both enabling more accurate differential transcription measurement and facilitating future SNP integration. Overall, this work provides key steps toward incorporating individual tRE responses into downstream studies of biological mechanisms or SNPs for translational research.

4.4 Algorithms and Methods

Detecting differential transcription of individual tREs is becoming imperative for integrating noncoding SNPs and regulatory networks into drug target discovery. However, this advance has been largely stymied due to low confidence in identifying which tREs are significantly differentially-transcribed. Given transcription best represents enhancer activity, and previous studies have already revealed poor differential calls of peak-based data[73], we focused our initial analysis using nascent RNA-sequencing.

As detailed in Supplemental Information, we show that to capture the differential activity

of tREs at the same confidence of genes, we would need unrealistically large amounts of data: either about 100 million uniquely mapped and de-duplicated reads per sample, or ten high quality replicates (each with depth of 40 million) (Appendix Figures C.1 and C.2). These sample sizes and depth are not found in any published studies; instead it is most common to have 2 replicates per sample, each with depths of about 30 million uniquely mapped and de-duplicated reads[223].

4.4.1 Truth Sets

Therefore, we sought to develop approaches to improve confidence of differentially-transcribed tREs from nascent run-on RNA-sequencing data. However, to accomplish this goal, we need a reliable truth set—a set of tREs independently known to respond in a particular condition. Although no ground truth is known, we can take advantage of the fact that Nutlin-3a is a highly specific activator of transcription factor (TF) p53 and has been characterized by both nascent run-on RNA-sequencing and chromatin immunoprecipitation (ChIP) in three cell lines: HCT116, MCF7, SJSA[5, 7]. We first defined true p53-responsive tREs by ChIP-seq peaks, and nonresponsive tREs by the lack of both a p53 motif and substantial ChIP-seq reads (details in Supplemental Methods). We also wanted to have a truth set based on transcription alone, and not biased by motifs or ChIP-peaks. Because there is no ground truth known, we instead considered the three cell types (HCT116, MCF7, SJSA) as replicates (six total replicates). Use of this combined dataset resulted in two truth sets, which include tREs called significant by any ("Combined Union") or all ("Combined Intersection") tools and parameter combinations (details in Supplemental Methods). Importantly, these truth sets cannot clarify cell type-specific vs shared calls. Using the transcription and multiomics-based truth sets, we compared the three most commonly used differential expression tools (DESeq2, Limma, and EdgeR) under 13 parameter combinations (details in Supplemental Methods) to identify tREs responding to p53 activation[149, 198, 130].

4.4.2 Algorithms and Pipelines

As high depth and replication are usually cost prohibitive, we built software tools for augmenting confidence by 1) maximizing transcriptional information gathered from nascent RNA-sequencing or peak data, and 2) integrating axillary biologically meaningful information.

4.4.2.1 Identifying tREs

The first step in enhancer analyses is to identify the (unannotated) regions of interest, in our case, sites of peaks or bidirectional transcription (Figure 4.1A Step 1). Several tools have been developed for this, including most popularly Homer and MACS2 for peak-based data and Tfit and dREG for nascent RNA-sequencing [12, 261]. The second step is to identify consensus positions of the regions, where "bedtools merge" is often used. However, because the position of RNA polymerase II initiation is critical to interpreting TF motif enrichment and similar downstream analyses, *muMerge* was created to probabilistically determine the best estimate for the midpoint of each replicate (which is assumed to be the position of RNA polymerase II initiation) (Figure 4.1A Step 2).

4.4.2.2 Characterizing tRE length and counting: LIET-EMG and Mu Counts

Next, these consensus regions are counted over. *muMerge* midpoints are key towards downstream TF activity inference, but the output region widths no longer reflect the transcribed expanse of the tRE but instead confidence windows [200, 223]. Consequently, for counting purposes, a fixed window around the center point has been used[223] (Figure 4.1A Step 3). This approach, however, does not take noise from overlapping transcripts into account, which is commonplace in nascent run-on sequencing data and intragenic tREs. Prior work benchmarked the ability of tools to identify active tREs and the position of predicted RNA polymerase initiation, but not the accuracy of transcribed lengths[278]. Therefore, in this work, we benchmark identification tools (Homer, dREG, Tfit) to assess how well they predict lengths of tRE RNAs from short-read RNA-sequencing compared to lengths based on long-read nascent RNA sequencing [56, 57, 194].

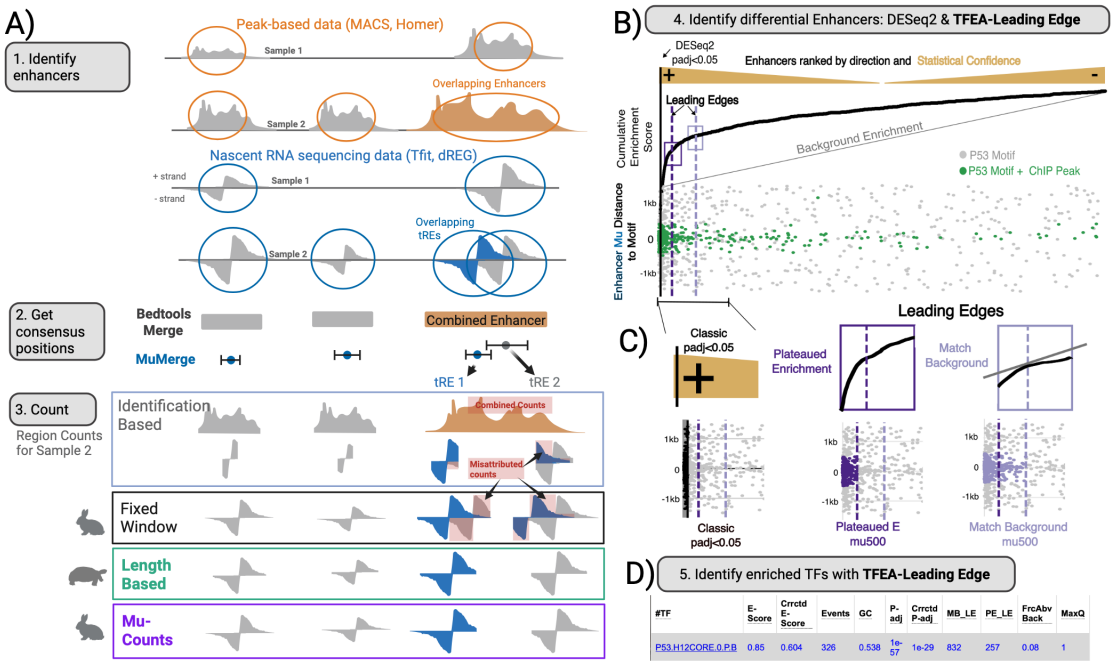

Figure 4.1: **An improved workflow for identifying differentially transcribed tREs from nascent run-on sequencing data. A.** First three steps involve 1) identifying tREs, 2) identifying consensus regions from multiple replicates, for which we show muMerge[200], and 3) counting reads over consensus regions. Shown are four methods of counting: Identification based (using Tfit or dREG), Fixed window, Length based (in this case, LIET), and Mu_Counts. Misattributed counts (red highlight) indicates counts that would be falsely considered for one tRE despite belonging to another. Fixed Window and Mu_Counts are both fast (rabbit) while LIET is slower (tortoise). **B.** F1 scores (harmonic mean of recall and precision) for p53-responsive tREs when using lengths according to various methods: tRE identification tools (Tfit and dREG, grey), HOMER (blue), LIET (green), fixed windows (black), and Mu_Counts (purple). Isolated tREs have no detectable transcription (tRE or gene) within 8kb. **C.** Scatter plots of log_2 fold change between Nutlin-3a and DMSO cells of tREs when using counts based on 600bp fixed window (top) or 6kb Mu_Counts (bottom). Truth sets according to ChIP-peaks or "Combined Union" are in green. **D.** Once regions are counted and ranked (top) by direction and significance, the leading edge is calculated on the enrichment curve (middle). The enrichment score reflects co-occurrence with motif instances (bottom, dots), with each motif instance weighted by the distance of the motif to tRE center (μ, labeled 0). tREs with p53 ChIP support are colored green. **E.** Two methods of leading edge detection are compared to the classical statistical cutoff from DESeq2 (left). Plateaued Enrichment considers when the cumulative enrichment score growth stalls significantly (dark purple). Match Background identifies when the slope of the enrichment curve matches the expected background (light purple). Final leading-edge calls can consider tREs with motifs within a certain distance of μ and within the leading edge (e.g. mu500 = 500bp, full=All). **F.** Finally, TFEA enrichment scores are used, with optional GC-bias correction (TFEA) or leading edge adjustments, to identify responding TFs.

We also add two length-focused counting approaches. First, we create a modified version of the recently published LIET approach[232] (LIET-EMG) that focuses explicitly on capturing transcription lengths of bidirectional regions (details in Supplemental Methods). Unlike Homer, LIET-EMG can incorporate the consensus midpoint positions from *muMerge* for final coordinates, providing both tRE midpoints and endpoints. The LIET algorithm is slow, so we also developed an approach called Mu_Counts that rapidly assigns reads within a fixed window (around consensus midpoints as from *muMerge*) to a specific tRE, even when another tRE or gene is nearby (Figure 4.1A Step 3 "Mu-Counts"). A figure describing the full pipeline can be found in Appendix Figure C.3.

4.4.2.3 Integrating motif information into differential expression analysis: TFEA Leading Edge

Differential expression tools like DESeq2, EdgeR, and Limma are used to capture changing enhancers with both peak-based or transcriptional data. Given these tools perform worse on features with high dispersion (such as enhancers), in addition to our previous algorithms above, we also considered how we could integrate orthogonal biological data to increase confidence. Perturbations that affect tRE transcription do so through the alteration of transcription factor function. Moreover, transcription factors most often work through DNA motifs. Therefore, we wondered if motifs within the tREs could be used to further improve our ability to detect differently transcribed tREs. Additionally, we found that regardless of differential expression tool used, ranking of tREs based on statistical confidence and direction of change was largely comparable (Appendix Figure C.4). Therefore, we developed a rank-based approach of identifying differentially transcribed tREs that could integrate orthogonal data such as DNA motifs. For this, we considered how we have already shown the power of ranking and motifs in TFEA, which uses motif co-occurrence with sites of bidirectional transcription to robustly quantify changes in TF activity between conditions[10, 200, 108].

In TFEA, tREs are first separated by direction of fold change and then ranked by the differential p-value as defined by tools like DESeq2. Iterating through the ranked tREs, TFEA then calculates a cumulative enrichment score, with the score increasing according to the proximity of

a specific TF motif to the tRE's midpoint [200]. When TF motif instances co-occur with the tRE midpoint at the extremes of the ranking, the TF is inferred as actively contributing to the differences between the conditions (Figure 4.1B). Therefore, we hypothesized that the tREs most contributing to a given TF's significant call in a perturbation would correspond well with tREs responsive to a perturbation. This concept is similar to gene set enrichment analysis identifying a "leading-edge" to pinpoint which genes are responsible for a pathway's call, and therefore relevant to the perturbation[236]. Consequently, we developed a similar approach, instead identifying the "leading-edge" of the enrichment curve within TFEA. The inflection point of this curve represents a statistically principled shift in the enrichment of TF-motif instances and changes in transcription (Figure 4.1B + C) and hence would be an alternative metric for defining regions of differential transcription that utilizes both transcription and sequence signals.

To identify the inflection point (aka leading-edge) we developed two approaches: one defines a conservative inflection point for changing tREs, and the other captures large scale trends of enrichment (Figure 4.1C). Specifically, the first looks for the point where the rate of change in the enrichment curve plateaus, hereafter called "Plateaued E" where E stands for Enrichment. The second identifies the point at which the slope of the cumulative enrichment curve is comparable to the TF motif presence in unchanged (background) tREs, hereafter called "Match Background". Details regarding these approaches can be found in Supplemental Methods. This approach extends beyond relying on a single p-value cutoff with classic statistical tools. Instead, a user can consider tREs based on multiple sources of confidence–both sequence support and statistical analyses for transcriptional change–to assess downstream biological questions.

4.4.2.4 Identifying responsive TFs

Finally, analyses usually end with identifying the TFs active within a condition, as with TFEA (4.1D Step 5). Sequence-based predictions, however, are hindered by the fact that transcription initiation sites (including tRE midpoints) tend to be GC-enriched[10, 108]. TFEA attempts to mitigate this bias with a heuristic correction term to enrichment scores before deciding significance

based on a simple regression model[200]. This was admitted to be a rudimentary solution in the original paper, and so we end this work with evaluating how to improve upon it using what is learned from the leading-edge [200]. Specifically, we develop two new metrics. First, we calculated the fraction of tREs contributing to cumulative TF enrichment scores above what is expected from background noise (hereby called "Fraction Above Background"). Second, we considered the enrichment trends across the ranked quantiles of tREs to assess if the strongest source of motif enrichment was occuring in tREs robustly changing in expression or not. Details are found in Supplemental Methods.

Overall, the resulting pipeline (Figure 4.1) still takes advantage of consensus positions from *muMerge* before combining improved feature counting (Mu_Counts) with the leading-edge approach of TFEA, the latter of which is hereafter referred to as TFEA-LE.

4.5 Results

4.5.1 Differentially transcribed tREs are poorly called by classic methods

We first clarified how poorly classic differential expression tools (DESeq2, Limma, EdgeR) perform on enhancers using our p53 truth-set. Regardless of the cell type considered, or whether truth sets were determined by ChIP-seq peaks or transcription alone, recall was poor (< 0.65) for all tools/parameter-combinations (Figure 4.2A, Appendix Figure C.5A). To test if this performance was better than random guessing, we defined an expected recall (based on chance alone, e.g. a "random set") where tREs with increased transcription (positive log fold change) were randomly selected as responsive. As expected given the stringency of the negative truth set, very few, if any, non-responsive tREs (without motif or ChIP peak), were called significant except when using the random set (Figure 4.2A, Appendix Figure C.5A). Unrealistically large p-value cutoffs were needed to reach the recall enabled from using the random set, albeit at the cost of low precision (Appendix Figures C.5B).

Despite the poor recall and variable calls between the classic tools (Deseq2, EdgeR, and

Limma), both truth sets (p53 ChIP peaks and "Combined Union"), are enriched in tREs with the highest rankings (i.e. statistical confidence) (Appendix Figure C.6). However, common significance cutoffs (e.g. alpha=0.05) of many tested parameter-tool combinations do not capture this enrichment well. This is especially true for the most popular tool according to citation numbers – DESeq2 – and across all tools in HCT116 (Appendix Figure C.6). One of the HCT116 samples has less overall transcription captured, and might therefore represent a more difficult dataset to analyze [223, 5]. Conversely, some parameter-tool combinations seemed to capture the enrichment of truth sets for MCF7 and SJSA samples well (Appendix Figure C.6). However, these approaches are the most permissive, and are therefore subject to extensive over-calling in datasets with more robust transcriptional responses[137, 23]. Our findings indicate that much of the tREs supported to be changing from multiomics data have weak statistical confidence from transcriptional data alone. We next saw if our approaches to 1) increase transcriptional data with length assessment, and 2) incorporating additional data would improve results.

4.5.2 A novel and reusable pipeline incorporates length and sequence-based support for calling tREs with poor statistical confidence

4.5.2.1 Length characterization improves differential expression analysis

To address the counting problem, we ask how accurately the region calls of tRE identification algorithms reflect transcript length. When applying these approaches to transcriptionally isolated tREs, we found that identification methods (e.g. Tfit and dREG) tended to largely underestimate length estimates while Homer and LIET-EMG lengths were closer to those suggested by long-read sequencing (C.7A). Most tREs, however, are found within genes (e.g. intragenic) or overlap other tREs [223] (4.1A). Because long-read data could not be clarified as belonging to intragenic tREs or genes, we instead simulated high noise (random reads distributed) to represent overlapping transcription of isolated tREs. While Homer falsely captured noise as extensions of transcripts, LIET-based estimates were unchanged (C.7B). The LIET-EMG model's background parameter

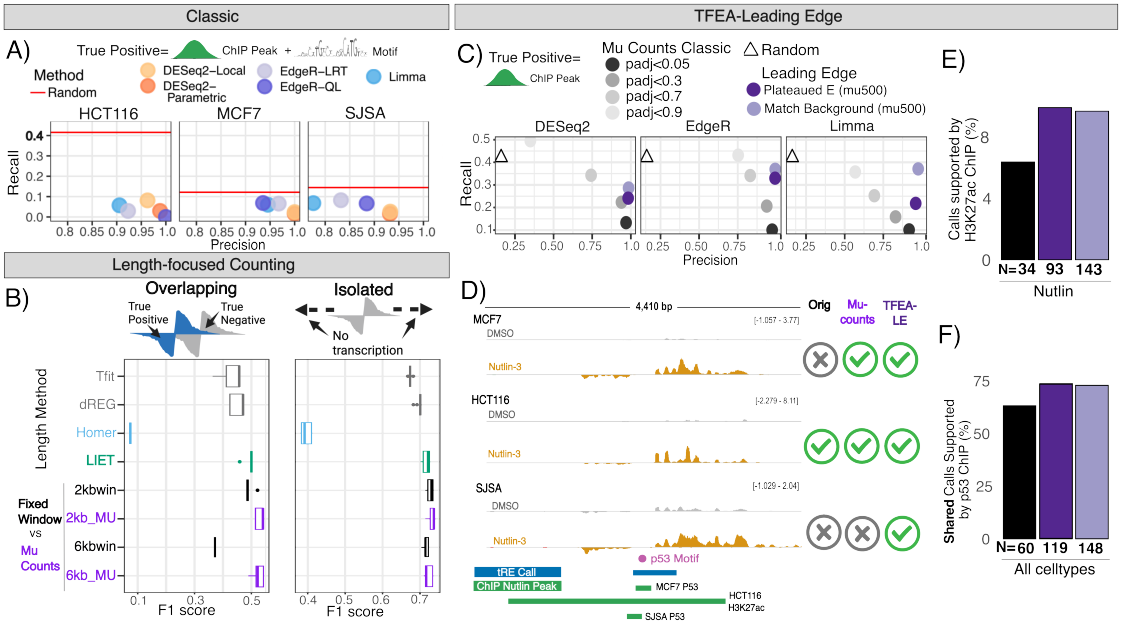

Figure 4.2: **Statistical incorporation of length and motif information significantly improve in single TF system. A.** Recall and precision of p53 responsive tREs in multiple cell lines. Red line is recall from randomly assigning tREs with positive fold change as true. Truth based on p53 ChIP peaks and false positives lack motif or ChIP peak. Data from [7]. **C.** Precision and Recall of HCT116 tREs with p53 ChIP peaks. Triangle: randomly selected tREs with positive fold change; Grey dots: adj. p-value cutoffs; Purple dots: Leading edge methods (classic padj < 0.05 + mu500). Methods correspond to DESeq2-Local-LRT (left), EdgeR-Locfit.mixed-QL (middle), and Limma-Trend-eBayes (right). **D.** An example region (chr2:113612107-113616515) for three cell types before (grey) and after Nultin-3a (orange). Data from [7]. Check-marks (far left) indicate whether the tRE was called significant in that cell type with different approaches. Org: 2kb fixed window; Mu_Counts: 2kb; TFEA-LE ranked based on 2kb Mu_Counts. **E.** Percentage of HCT116 Nutlin-responsive tRE calls from the different approaches that also have a H3K27ac peak only after Nutlin was added to the media. Black: padj$<$ 0.05 by TMM-eBayes-Trend; Dark Purple: Plateaued E with padj$<$0.05 + mu500; Light Purple: Match Background with padj$<$0.05 + mu500. **F.** Percentage of calls shared across all cell types, with p53 ChIP in Nutlin-3a also shared across cell types, according to the different approaches.

effectively captured the introduced noise and gave predicted lengths closest to the long-read data (Appendix Figure C.7B+C). Despite the accuracy of LIET-EMG, the underlying algorithm is computationally expensive and becomes obtuse when considering thousands of tREs (20,000 tREs in one sample took $>$ 24 hours) (Appendix Figure C.7D).

Therefore, we next sought if our Mu_Counts algorithm could rapidly obtain optimal counts per tRE, without requiring RNA lengths. This approach is fast, able to consider all relevant

samples in a 15 minute time frame (Appendix Figure C.7D). Both LIET and Mu_Counts not only improved differential calls of transcriptionally isolated tREs, but successfully distinguished between overlapping tREs with and without multi-omic evidence of change ("Overlapping") (Figure 4.2B). Mu_Counts and LIET's having high and comparable rankings were consistent across all cell types (Appendix Figure C.8). Importantly, Homer's low F1 scores arise in large part because it fails to call many regions (Appendix Figure C.9). Instead, LIET and Mu_Counts can focus on tRE characterization based on accurate consensus midpoints of tREs (e.g. from *muMerge*) [200].

Despite effectively maximizing transcriptional data retrieval, we still observe low statistical confidence for most tREs. While Mu_Counts shifts the density of mean counts, dispersion trends don't improve; instead many expected true positives still show transcriptional responses difficult to distinguish from noise(Appendix Figure C.10. Therefore, incorporating orthogonal biological data was still imperative for enhancers with limited data from transcription alone.

4.5.2.2 TFEA-LE identifies previously missed shared enhancer responses for p53 across cell types

We next tested TFEA-LE, first by comparing it's calls to the regions identified by those called by classic tools (adjusted p-value cutoff of 0.05, after using Mu_Counts). We find that considering tREs within either leading-edge algorithm clearly improves recall of p53 ChIP-peaks, while maintaining high precision 4.2C. These improvements, as also measured by F1-scores, are found in all cell types, regardless of non-motif-based truth set used (ChIP-peaks, Combined Intersection, Combined Union), parameter-tool combination, or length of windows used for Mu_Counts (Appendix Figure C.11). Both algorithms for obtaining the leading-edge (Plateaued E and Match Background) are also largely consistent regardless of the classic statistical tool-parameter combination used to rank tREs (Appendix Figure C.12). Figure 4.2D shows an example of a tRE that, despite having p53 and Nutlin-specific H3K27ac peaks across all cell types, would only be considered significantly changing in response to Nutlin-3a in HCT116 cells from classic statistical approaches. Using Mu_Counts allowed MCF7 to indicate a significant call as well, compared to using fixed window counts. Only

the leading-edge (TFEA-LE) allowed this tRE to be called significant across all three cell types, hence matching the multiomics support.

Likewise, the leading-edge calls are highly enriched in previously published Nutlin-induced H3K27ac peaks in MCF7 and HCT116, more so than tREs called from classic statistical approaches (SJSA H3K27ac data not available) (Figure 4.2E and Appendix Figure C.13). The leading-edge identifies a much larger number of p53 responding tREs shared across all cell types compared to the original publication using classical statistical analysis methods[7] (Appendix Figure C.14A). Despite this, the percentage of tREs called shared across all cell types that are also supported by p53 ChIP peaks [shared across cell types] is consistently higher for LE-based calls than those from only classic tools (Figure 4.2F, Appendix Figure C.14B).

4.5.3 TFEA-LE improves understanding of multi-TF transcriptional and chromatin-accessibility responses

4.5.3.1 Leading edge helps clarify GR-activating enhancers across multiple TFs

Arguably, Nutlin-3a is an exquisitely specific activator of a single transcription factor – p53. Thus, we next wondered whether the leading-edge approach works well when multiple transcription factors are responding to a perturbation. To assess this, we applied TFEA-LE to our previously published double drug study[208]. In that work, we exposed Beas-2B cells to either dexamethasone (DEX), TNF-α (TNF), or both drugs. DEX activates glucocorticoid-based receptors (GR and MCR) while TNF activates the NFκB complex, including multiple TFs (REL, RELB, TF65, NFκB1, NFκB2) (Figure 4.3A) [208]. All leading-edges for these TFs showed similar consistency across ranking methods, as with p53. The leading-edges also had greater consistency in the differential tREs called than classic statistical methods (Appendix Figure C.15). Unlike with p53, however, classic statistical tools called more tREs significant than the Plateaued E leading-edge values indicated (Figure 4.3A; Appendix Figure C.16). We predicted that the greater number of tREs called could be due to these tools not being specific towards upstream regulators, while the leading-edge is

TF-specific. These drugs also produce highly robust transcription compared to many other nascent sequencing experiments, suggesting that classic statistical tools are less disadvantaged in predicting tREs that change or might be prone to overcalling [223].

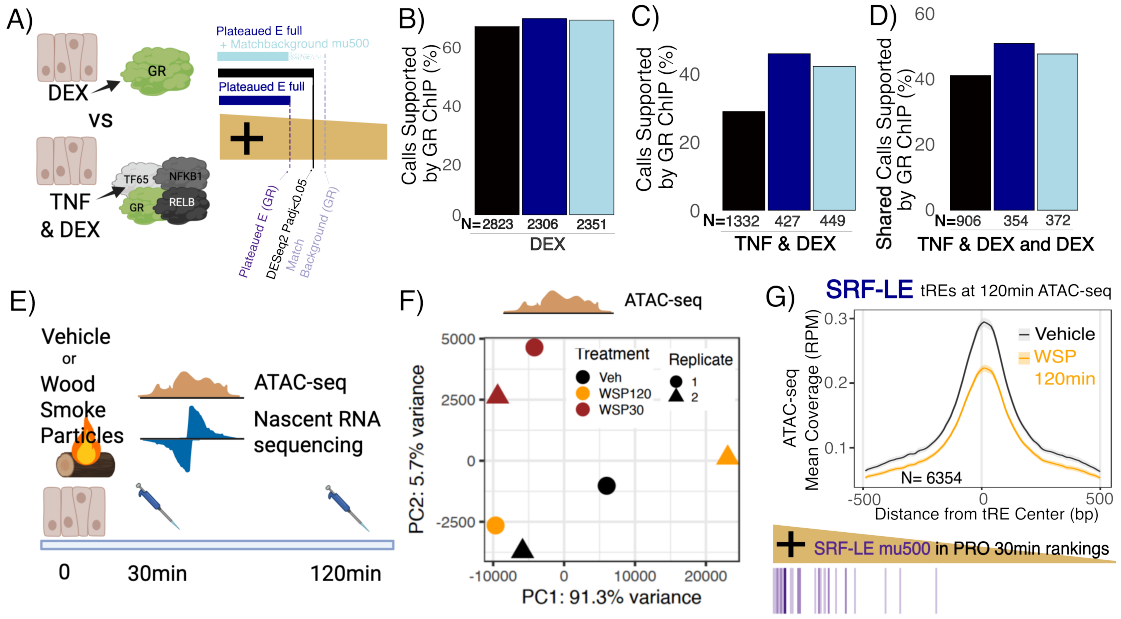

Figure 4.3: **Mu Counts and TFEA-LE significantly improve calls in multi-TF system and allow TF-enhancer predicted linkages. A.** Left: Visualization of double-drug treatment of dexamethasone (DEX) and TNF treated cells and the expected transcription factors activated. Right: Visualization of Plateaued E full and Match Background options when the classic approach calls more significant regions than the Plateaued E approach. **B.** Percentage of tRE calls in DEX-treated cells that also have a GR ChIP peak. Black: DESeq2-Ratio-LRT-Local padj < 0.05; Dark Blue: Plateaued E (full); Light Blue: Plateaued E (full) + Match Background mu500. **C.** Percentage of tRE calls in DEX and TNF treated cells that also have a GR ChIP peak. Leading-edge results are based on the GR TF motif. **D.** Percentage of tRE calls shared between cells exposed to both TNF and DEX and just DEX that also have GR ChIP peaks shared between the two conditions. In all cases, Ns listed are the total number of tREs called significant by each approach. Method selected for black (C-G) was selected for highest number of shared tREs across all samples. **E.** Visualization of wood smoke particle experiment. **F.** PCA plot of ATAC-seq BEAS-2B samples without WSP (Veh) or with WSP perturbation for 30 or 120 minutes (WSP30 and WSP120) and replicate 1 or 2. **G.** Top: Metagene showing normalized ATAC-seq reads after 120 minutes for cells perturbed with vehicle (grey) or wood smoke particles (orange) of regions within the Plateaued E leading edge for Serum Response Factor (SRF). Bottom: tREs within the Plateaued E leading edge for SRF at ATAC-seq 120 minutes (decreasing accessibility, with SRF motifs within 500bp) are colored as they are ranked by differential transcription confidence according to PRO-seq at the 30min mark (focusing only on tREs with positive log fold change).

Therefore, in this case, we hypothesized that the leading-edge would allow 1) secondary support for what adjusted p-value cutoffs should be used for classic tools in cases with robust transcription and 2) more specific identification of tREs being regulated by a single TF of interest. To test these hypotheses, we focused on considering all tREs within the leading-edge, instead of just those with strong motifs (Figure 4.3A). This approach would still consider motif enrichment trends, which could be TF specific, while allowing for secondary responses and cases where TF motifs aren't as precise. When applying this approach to the DEX condition (GR focused activation), we indeed observed either a slightly increased or comparable enrichment of calls supported by GR ChIP-seq with the leading-edge compared to classic statistical tools (Figure 4.3B DEX and Appendix Figure C.17A).

To test our second hypothesis, that the leading-edge could identify tREs most likely regulated by a specific TF, we also assessed if we could clarify GR-specific tREs amidst conditions where multiple TFs were perturbed (DEX+TNF). When considering cells perturbed with both DEX+TNF, the GR leading-edge tREs were more enriched in GR ChIP peaks than those called by classic statistical approaches; this trend is observed even after removing tREs called significant in cells perturbed with TNF alone (Figure 4.3B TNF+DEX and Appendix Figure C.17B). Next, we compared tREs responding with just DEX, and in cells perturbed with both TNF and DEX. Again, the leading-edge calls shared between these perturbations had a higher enrichment of ChIP calls also shared between the two perturbations, regardless of classic statistical approach used; leading edge calls for perturbation-unique tREs were also more enriched in ChIP-seq peaks unique to a perturbation 4.3C and Appendix Figure C.17B)).

4.5.3.2 Leading edge enables differential chromatin accessibility calls despite sample outliers

Finally, we examined whether TFEA-LE could be applied to ATAC-seq data as with its predecessor TFEA [200]. We re-examined ATAC-seq data generated from Beas-2B cells perturbed with wood smoke particles (with two time points: 30 and 120 minutes)([79]) (Figure 4.3D). In this

case, both DESeq2 and EdgeR called a maximum of one region significantly changing chromatin accessibility at the 120 minute mark, likely due to one of the two 120 minute samples showing outlier accessibility patterns (Figure 4.3E). Despite classic approaches indicating no tREs were changing, TFEA called several TF motifs as significantly enriched in the regions ranked towards decreasing accessibility at 120 minutes, including SRF. Therefore, we assessed if the leading-edge tREs of SRF would represent tREs changing in accessibility, despite generally weak agreement between samples. Indeed, normalized counts of the Plateaued E leading-edge regions for SRF show a clear depletion in accessibility at 120 minutes with wood smoke particles (Figure 4.3F). Additionally, while SRF-based leading-edge tREs showed decreased accessibility in both 120 minute replicates, random tREs of the same number with fold changes below 0.9 only showed decreased accessibility in one replicate (Appendix Figure C.18). Finally, SRF leading-edge tREs with SRF motifs have increased transcription at the 30 minute mark (which follows SRF activity at that time point) before having transcription levels again comparable to Vehicle by 120 minutes (Figure 4.3G). Similar plots of 4.3F and G for all the time points and leading-edge approaches show similar trends (Supplemental Methods). Therefore, we successfully identified tREs with changing accessibility at 120 minutes as supported by other data, even when classic statistical cutoffs of original tools could not.

4.5.3.3 The leading-edge improves identification of upstream regulators in technically challenging multiomics data

TFEA identifies upstream regulators optimally under two conditions 1) when a single or few TFs are primarily responding and 2) for TF motifs with low GC content. The more TFs that are responding, the more likely motifs of significant TFs will be found across a larger spread of ranked tREs rather than just those at the extreme poles. Similarly, motifs with high GC content are more likely to be randomly found due to the GC-enrichment of transcription-initiation sites. To mitigate this bias in TFEA, corrected Enrichment-scores are calculated as a y-offset of the observed Enrichment-scores from a linear regression fit (extreme case shown in Figure 4.4A). However, we have found that this approach fails with strong GC bias, introducing as many false-positive TFs as

it mediates against [200]. Therefore, we next sought to evaluate whether leading-edge metrics can be used to reduce false-positive TF enrichment calls.

In the Nutlin-3a and TNF/DEX data sets, false-positive and true-negative TF calls by TFEA consistently had Match Background leading-edges that suggested either half or zero tREs were contributing to the TF's enrichment score. Remarkably, these cases suggested that some significant enrichment scores were being called based on tREs with no transcriptional change between conditions (Appendix Figure C.19 top).

To robustly measure this trend, we introduce two new metrics. First, we calculated the fraction of tREs contributing to cumulative TF enrichment scores above what is expected from background noise (hereby called "Fraction Above Background"). While DEX-induced transcription factors GR and MCR have lower fractions, false-positive and true negative TF calls have fractions above .45 (Figure 4.4B). This finding is reproducible across all Nutlin-3 and DEX/TNF perturbed cells, with the fractions being lowest for highly specific TF responses (e.g. P53) (Appendix Figure C.19). Second, to more substantially characterize enrichment trends, we divided the ranked tREs into fifteen equally sized bins (Q1-Q15) and plot the maximum enrichment score per quantile (Figure 4.4C). We observed three general patterns for motifs: 1) strong motif enrichment at the extremes, as expected by TFEA; 2) weaker enrichment at the extremes, as is typical for weaker responders when multiple TFs respond; and 3) instances where the strongest motif enrichment is in the middle of the ranking, where tREs have little to no detectable change in transcription. This last group is likely false-positives. We predicted that using this positional enrichment assessment, we could both separate out weak and stronger responders, as well as remove false-positive TF calls.

Using these refined correction metrics, we turned our attention to complex perturbations with many transcription factors are responding: paired PRO-seq and ATAC-seq from Beas-2B cells perturbed with wood smoke particles (with two time points: 30 and 120 minutes). This perturbation produces a very robust transcriptional response, with multiple TFs responding. These data also have a strong GC bias in their active tREs, leading to a large linear regression coefficient in traditional TFEA (Figure 4.4A and Appendix Figure C.20). The default GC-bias correction

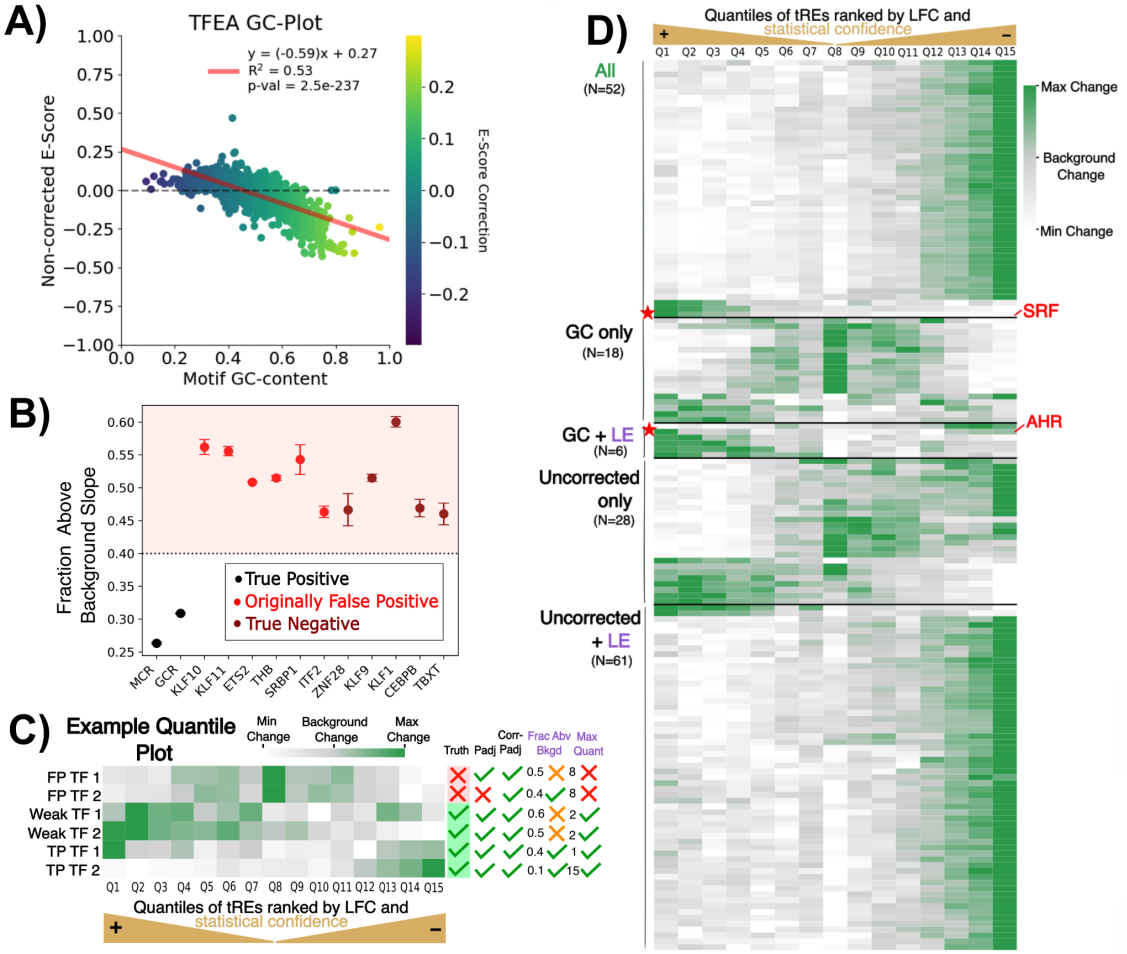

Figure 4.4: **The Leading Edge improves identification of true TF responses and responding tREs in technically challenging multi-omics data. A.** TFEA enrichment score plot for ATAC-seq at 120 min. Dots colored by GC-content, red line: regression fit. **B.** Fraction of tREs with cumulative enrichment scores above expectation from background (y-axis) is plotted for numerous TFs (x-axis). Dots are colored by correctness status based on TFEA enrichment score alone. The area about 0.4 is shaded in pink as a consistent cutoff between true positives and negatives. **C.** Cartoon example of quantiles within the ranked list, colored by their median change in enrichment. Coloring sets grey as expected background level ("background change"). Patterns represent two false positives (FP), two weak TF signals, and two strong TF signals (TP). Check marks on right indicate whether the TF is considered significant by different methods. TFEA's Padj and GC-corrected (Corr-Padj), Frac. Abv. Bkgd (Fraction tREs with cumulative enrichment increase above background expectation), and Max Quant (requires maximum value at edges of ranked list). **D.** Quantile heatmap (similar to C) based on PRO-seq from BEAS-2B cells 30 minutes after their perturbation with wood smoke[79]. TFs (rows) are sorted by whether they are called by TFEA uncorrected, TFEA GC-corrected, or TFEA-LE methods. TFs with extensive wet-lab validation are marked with red stars (SRF and AHR). Change is calculated via a binning of ranked tREs (each bin N=2994).

paradoxically results in artifically shifting tens of TFs into significance, most of which have no support in the literature for being involved in wood smoke or respiratory response. However, the GC-bias correction also enabled the calling of the TF Aryl Hydrocarbon Receptor (AhR), one of the best-known respiratory responders with confirmed activation in these same cells according to ChIP-qPCR, siRNA knockdown, and western-blots [79]. We wondered whether the improved TFEA-LE approach would both identify AhR and remove the large number of apparent false-positives.

We split wood-smoke predicted TFs according to whether they were being called significant by TFEA's GC-corrected scores, TFEA's uncorrected scores, and/or TFEA-LE based metrics. We started with strict TFEA-LE based metrics, requiring <0.45 "Fraction Above Background" and the quantile with the maximum enrichment score ("Maximum Quant") outside the middle (max q < 5 or > 10). TFs supported by all methods as responding at 30 minutes from PRO-seq showed very clear maximum changes in cumulative enrichment in tREs with the greatest confidence of statistical change (Figure 4.4D). This included the transcription factor first noted in the original publication for response: Serum Response Factor (SRF) [79]. The leading-edge approach also uniquely confirmed the significance of AhR (Figure 4.4D star). Conversely, most TFs called by GC-bias correction alone, and many called with uncorrected scores alone, had maximal enrichment scores in middle quantiles, consistent with false-positives (Figure 4.4D). These trends are generally visible at all time points and in both PRO-seq and ATAC-seq (Appendix Figure C.21). As a final support for calls, we considered which TFs were supported as relevant by both ATAC-seq and PRO-seq, in the same direction of enrichment. The percentages of TFs called with the same enrichment direction in both ATAC/PRO were highest in TFs shared across all calling approaches, and next LE-based calls at all time points (Appendix Figure C.22). A comparable but less consistent trend was observed for agreement in direction when using gene TSS bidirectionals compared to tREs (Appendix Figure C.22).

4.6 Discussion

Overall, in this work we provide a new suite of tools to improve the characterization of transcriptional regulatory networks from multiomics-data. Although especially powerful when integrated together, all tools are available as independent software packages to fit a user's needs. The high dispersion and low transcription of tREs present a difficult challenge for traditional statistical tools such as DESeq2. To identify high-confidence differential tREs, we developed two new tools: Mu_Counts and TFEA-LE. Mu_Counts maximizes the data we can consider from tREs through improved counting, while still taking advantage of consensus positions of *muMerge*. TFEA-LE then leverages motif co-localization signals to assist in the identification of statistically significant changes in transcription at lowly tREs. For this purpose, TFEA-LE improves on the original TFEA algorithm in multiple ways. First, TFEA-LE uses a leading-edge approach to identify the inflection point of co-enrichment of motif instances with the extremes of transcription changes. Finally, we leverage the TFEA-LE metrics to improve discrimination between true positive enrichment and false-positive TF calls. Collectively, these improvements enable TFEA-LE to both identify differentially transcribed tREs that are supported by multiple lines of biological information and to address complex perturbations that can have a either a highly specific or very broad impact on cells.

While prior work has benchmarked the identification of tREs[278], we extend this work to whether these methods accurately return the lengths of the transcribed region. In this extension, we also introduce a tRE-focused adaptation of the LIET model, finding that the LIET model's explicit background parameterization provides a uniquely powerful approach for distinguishing noise from signal, resulting in the best length estimates. However, LIET is computationally expensive, and tools such as Homer present an effective alternative for tREs with confirmed isolation from neighboring and overlapping transcripts. Accurate lengths of enhancer-associated transcripts are necessary for characterizing the RNA itself and using them for fine-mapping of SNPs. We and others have already shown that using tREs can be essential towards filtering and annotating SNPs[209, 68],

but were confined to about 1% of SNPs by focusing on the initiation/loading zone[31] rather than the RNA produced. We now know that some tRE RNA SNPs can lead to disease-related phenotypes [207, 171, 89, 214, 185].

The introduction of Mu_Counts was necessary to bolster counts associated with individual transcripts. The method is a fast heuristic that accounts for overlapping transcription of both genes and tREs, counting only regions where reads can be conclusively assigned. The accuracy of Mu_Counts, however, is dependent on the accurate identification of tREs and their midpoint, i.e. the position of RNA polymerase II initiation, to correctly disentangle overlapping tREs. Hence, the combination of *muMerge*[200] which seeks to preserve accuracy on the inferred midpoint of tREs across replicates, with Mu_Counts results in a measurable improvement in differential-transcription analysis of tREs through increased recall. More accurate counting can also enable more accurate linking of tREs to their target genes with correlated counts, a growing focus of research ([223, 32, 221]).

The leading-edge leverages the consistency of expression-based ranks to identify the set of tREs supported by both changes in transcription (the ranking) and co-localization of a responsive transcription factor (a supporting motif). We developed two approaches to identify the leading-edge. The Plateaued Enrichment method provided the best calls of tREs with transcription changes specific to a TF. We also show that the approach can also be applied to perturbations with more than one TF responding. Our reliance on motif instances, however, means that our assessment is limited by the accuracy and specificity of a TF's motif. For example, we were unable to decipher between the activity of AhRR and AhR due to their essentially identical motifs.

The second method of identifying the leading-edge, the Match Background approach, proved highly informative on interpreting larger-scale enrichment trends that improve false positive identification in TF enrichment. We showed that metrics based on the leading-edge can distinguish between GC-corrected significant calls with no clear enrichment and those with clear biological support from multiomics data or downstream validation. Indeed, with the leading-edge, subtle but biologically meaningful enrichment such as AhR within the woodsmoke data can now be detected

without inflating the false positive rate[79]. Some TFs can function as both an activator and a repressor, dispersing their signals to both ends of the ranked list. The quantile plots provide a quick way to identify these TFs (high change at both poles), but TFEA-LE does not currently consider simultaneous enrichment in both directions. However, the leading-edge metrics do provide an intuitive interpretation of enrichment results, allowing users to consider a number of alternative enrichment cutoffs and their tradeoffs. Therefore, the leading-edge can be essential not only for improving the call of responsive tREs, but has largely improved TFEA calling of responsive transcription factors as well.

Ultimately, this work has provided multiple, open-source tools within a novel pipeline for improved characterization of changes in expression at lowly transcribed regulatory elements. The leading-edge seems specifically advantageous when dealing with technically challenging multiomics data, in this case nascent RNA sequencing and ATAC-seq. In particular, TFEA-LE was robust to low overall transcription signal and outlier samples, cases that are problematic for classical statistical approaches yet common. Our applications simultaneously improve identification of significantly changing tREs, linking tREs to their upstream regulators, defining coordinates of tREs, and providing further confirmation of transcription factors responsible for changing transcription. Together, this work clarifies key weaknesses in current regulatory element focused analyses, particularly with nascent RNA-sequencing, and provides several methods to advance their usage.

4.7 Code Availability

All code is available on Github and zenodo at the links provided below.

(1) LIET-EMG: https://github.com/Dowell-Lab/LIET/tree/LIET_EMGtoo,

(2) Nextflow pipeline that can allow Mu_Counts or fixed-window counting: `https://github.com/Dowell-Lab/Bidir_Counting_Analysis`,

(3) Updated TFEA (with leading-edge and related metrics): https://github.com/Dowell-Lab/TFEA/tree/Lead_edge,

(4) Source code for this work: https://github.com/Dowell-Lab/Improving_tRE_Analysis_Paper

4.8 Data Availability

All sequencing data for TP53 based analysis can be found in the NCBI Gene Expression Omnibus (GEO) under accession number GSE86222 (HCT116, SJSA, MCF7). Sequencing data for GR based analysis can be found under accession numbers GSE125623 (ChIP) and GSE124916 (GRO-seq). All SRRs used with annotations can be found in Supplemental Table 1.

Chapter 5

Air pollutant multiomics improves functional annotation of SNPs associated with lung disease

5.1 Note to Readers:

We are in the process of uploading some files to GEO and finalizing the Github for this work. Supplemental Data is currently uploaded to Zenodo but will be included as supplemental tables for the final thesis. The methods section and supplemental figures can be found in Appendix D, with red parts of the methods section indicating that I am awaiting collaborators to confirm accuracy. You can request supplemental tables from me for now.

5.2 Contribution Statement

This work was made possible by collaborating with Drs. Anthony Gerber, Sarah Sasse, Arnav Gupta, and Shu-Yi Liao at National Jewish Health. Drs. Gerber and Sasse are now at the University of Kentucky while Dr. Liao is at the David Geffen School of Medicine at UCLA. Dr. Liao ran the disease-association analysis of SNPs with asthma and chronic obstructive pulmonary disorder with the All of Us Research cohort. Dr. Sarah Sasse performed the initial PRO-seq for UPM. Dr. Sasse provided BAMS (intermediate files) from two currently unpublished ATAC-seq experiments in nasal epithelial cells. Dr. Arnav Gupta performed all experimental validation (qRT-PCR, ELISA), and did the HYA-gene regression analyses in COPDGene. I performed all other computational analyses. I wrote the initial draft, with Methods sections for the association analysis provided by Dr. Liao and for the ELISA by Dr. Gupta. The version used for this thesis was then

reviewed by Dr. Dowell.

5.3 Introduction

Both asthma and COPD are prevalent lung conditions affecting a combined 15% of U.S. adults[33]. Treatment development has been largely stymied by the fact that treatment response and pathogenesis for asthma and COPD are largely dictated by interactions between both genetic background and environment[117]. Air pollutants, specifically, such as those found from dust, wildfires, or urban areas, explain half of COPD risk and have often been incorporated into clinical models[225, 54, 143, 157]. Air pollution exposure is also increasing worldwide, with a projected 22% increase in the number of people exposed to hazardous pollution by 2030[188]. Air pollutants and lung disease are hence a growing medical concern, and yet there remains a very limited understanding of the biological mechanisms by which pollutants increase risk of disease.

Genome-wide association studies identify genetic variants associated with disease yet have had limited impact on treatment development for COPD and asthma. Due to the sheer number of correlated variants across the genome, interpretation of these studies is often limited to only single nucleotide polymorphisms (SNPs) that pass extremely stringent statistical significance (adjusted p-value $< 5x10^{-8}$)[213]. Rare SNPs, or those found across regulatory regions that function together, often have low individual effect sizes despite functional relevance [234]. Additionally, the SNPs prioritized downstream are limited due to feasibility. SNPs inherited together (due to linkage disequilibrium) are commonly analyzed with using the SNP with the highest association with the disease (lead SNP)[213]. However, these lead SNPs are commonly not the SNPs whose functionality is relevant to the disease. Limiting downstream analysis to these lead, significant SNPs increases feasibility of functional annotation and experimental validation, but also removes the majority of genetic variance explaining a disease. In the case of COPD, only considering variants passing strict statistical significance cutoffs leaves 95% of heritability still unexplained[41, 298].

Interpretation of disease-associated variants for treatment stability is additionally stymied by the fact that most disease-associated variants, including those for COPD and asthma, are found

in noncoding regions[54, 112, 225]. Linking noncoding SNPs to druggable targets, like the genes whose functionality SNPs impact, is a challenging problem. These noncoding SNPs are often functionally annotated by identifying transcription factor binding motif instances or sites that might be functionally impacted by a sequence change[213]. However, the functional relevance of these changes is dependent on whether the impacted transcription factors are active for the disease, which is typically unknown. Active transcription factors can be deciphered from experimental data in disease-related cell types and conditions, but such data are often not available[108]. Overall, genome-wide association studies often end with a long list of associated variants with unknown functional activity and pertinence to experimental validation.

Our lab and others have recently established experimental and computational pipelines that address these challenges. Our processes primarily take advantage of nascent RNA run-on sequencing because it captures the direct transcriptional activity of genes and transcribed regulatory elements (tREs), like enhancers, while also enabling inference of transcription factor activity. In our previous work, we pinpointed key noncoding variants associated with asthma and their functions, focusing on SNPs found within tREs responding in lung cells perturbed with inflammatory signals [209]. We also recently developed several algorithms that improve our ability to accurately identify when noncoding SNPs correspond to tREs, their target genes, and upstream regulators[223, 246].

In this work, we expand this framework, applying it to COPD and asthma based on the response of lung cells to air pollutants. To this end, we consider published and new nascent sequencing data of lung cells perturbed with an expanse of air pollutants: wood smoke particles (WSP), urban particulate matter (UPM), and Afghan dust particles (ADP)[79, 78]. This work presents the first analysis of UPM nascent transcriptional response in primary lung cells over time. Therefore, we compare transcriptional regulatory networks (transcription factors, genes, enhancers) based on both time-scale dynamics and general response, specifically between WSP and UPM. Subsequently, we perform a focused association study with both COPD and asthma of genetic variants falling within air-pollutant-responding genomic regions. By combining our novel and previously published transcriptomics, we effectively rank and predict the functionality of variants. From these

analyses, we prioritize and predict biological mechanisms for genetic variants associated with lung disease and several other inflammatory diseases. Beyond our biological insights, readers can apply our computational pipeline to other diseases and data for improved understanding of noncoding variants in disease physiology.

5.4 Results

5.4.1 Genes show dynamic transcriptional responses to UPM comparable to other particulate matter responses

We first asked how responses to UPM transforms across time, and if genes showing certain timings show shared functions (via Gene Ontology enrichment analyses). By comparing to other particulate matter responses, we can also clarify if the dynamics are general to particulate matter responses or specific to compounds in UPM. To determine the earliest transcriptional responses to UPM, we conducted nascent RNA sequencing on primary samples of small airway epithelial cells (smAECs) using precision run-on sequencing (PRO-Seq). PRO-Seq measures both gene and regulatory element (e.g. enhancer) transcription at high temporal resolution based on RNA polymerase II activity. To confine our analyses to direct transcriptional responses, we used both a 30-min and 120-min exposure time to UPM and analyzed the results relative to treatment with vehicle (PBS). Transcriptional responses were defined into three large-scale groups: Early response (rise or fall at 30 minutes before returning to vehicle baseline at 120 minutes), Early-Plateau response (rise or fall at 30 minutes that is sustained at 120 minutes), Late response (rise or fall only after 120 minutes) (Supplemental Data). We found that the general transcriptional patterns of genes within these categories were comparable, with the largest number of genes responding within 30 minutes (Figure 5.1A, Appendix Figure D.1).

.

Some of the strongest early responses were marked by upregulated genes attributed to cellular adhesion and cilia pathways, and downregulated genes involved in hyaluronan biosynthesis and

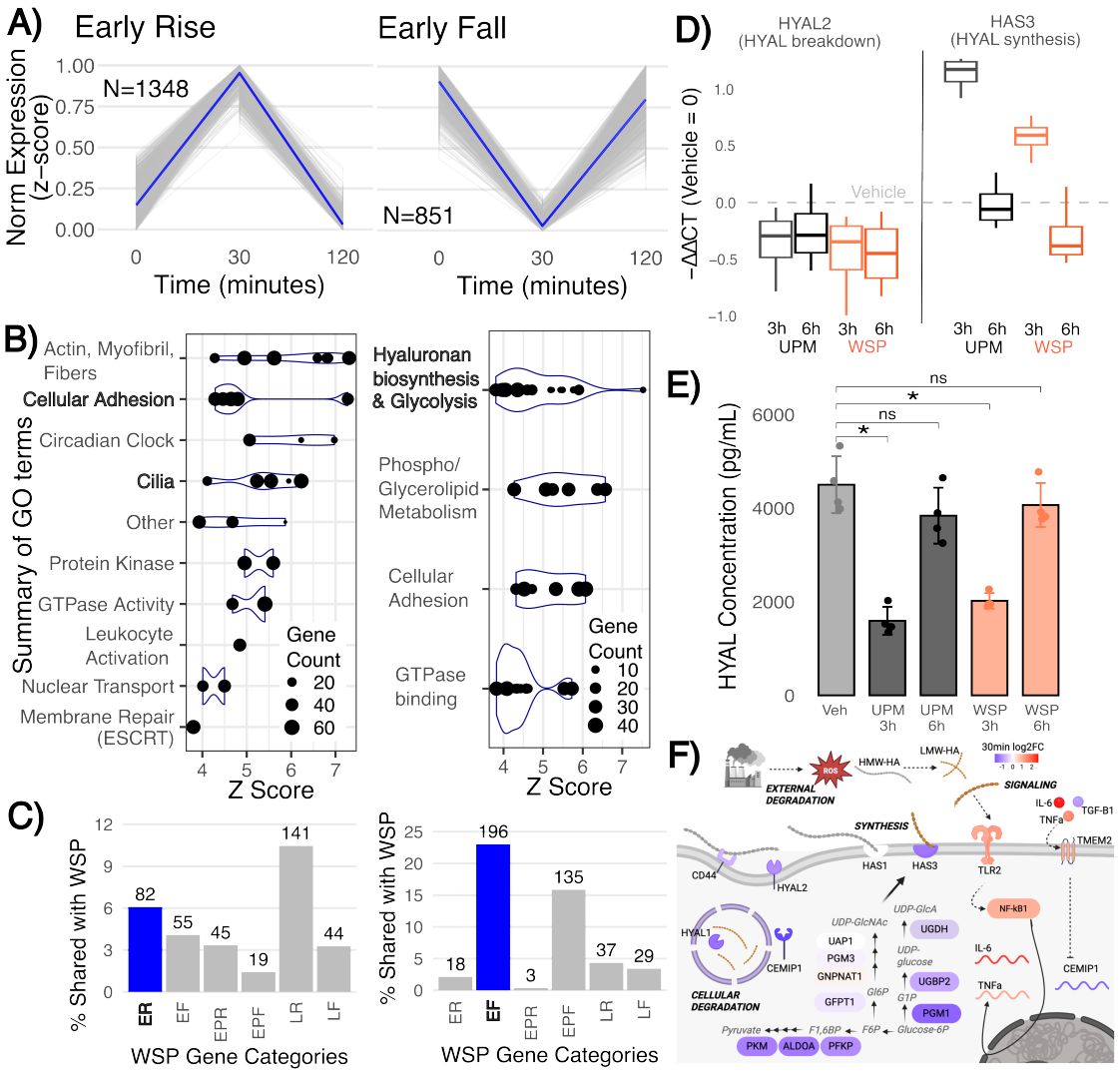

Figure 5.1: **UPM produces a timeline of responses comparable to other particle-based perturbations in lung cells. A.** Line plots of the mean (blue, light blue 95% confidence interval) and individual (grey) normalized transcriptional changes of statistically significant (adj-p-value ¡ 1x10-10) genes in early response (details in Methods). Norm Expression (z-score) is normalized counts min-max scaled. 0 minutes is Vehicle response. **B.** Summary of significant Gene Ontology (GO) terms of the matching genes with the size of the dot corresponding to the number of significant genes matching the GO term. **C.** Percentage of significant UPM genes that were found across the timeline categories of WSP responsive genes (adjusted-p-value ¡ 1x10-10). ER=Early Rise, EF=Early Fall, EPR=Early Rise-Plateau, EPF=Early Fall-Plateau, LR=Late Rise, LF=Late Fall. The bar corresponding to the same time point as UPM for WSP is highlighted in blue. Numbers above bars are number of UPM genes shared. **D.** qRT-PCR or **E.** ELISA (HYA levels in media) results from smAECs perturbed with UPM or WSP for 3 and 6 hours. **F.** Log2FC expression at UPM 30 min for genes involved in HYAL degradation, synthesis, and related signaling. H/LMW-HA=high/low molecular weight hyaluronic acid. ROS=reactive oxygen species.

glycolysis (Figure 5.1B, Supplemental Data). Hyaluronans (HYA), built from glycolysis products, are abundant in the lung extracellular matrix and coat alveolar macrophages[107]. While the robust early response could reflect a large-scale stress response, early-plateaued or late-responding genes appeared more characteristic of targeted particulate matter degradation. Glucuronidation-related pathways were strongly enriched by genes either increasing transcription early and maintaining high transcription at 120 minutes, or only increasing transcription after 120 minutes (Appendix Figure D.1B). Similarly, other metabolic response pathways for Vitamins, Iso/Flavo/Terpinoids, alcohols, and steroids dominated the late rise response (Appendix Figure D.1B Right).

When comparing gene response patterns in UPM to those found in WSP, we saw an average of 30% of UPM responsive genes also being found significantly responsive in WSP, with about 50% of early fall UPM genes (which includes HYAL metabolism) also responding in WSP. Although the overall direction of gene expression was largely maintained, the precise timing of the response was less so. For example, most UPM early-rise genes only rose at the 120-minute time mark for WSP (Figure 5.1C), while more than 15% of the early fall UPM genes had maintained lower expression levels compared to vehicle at 120 minutes in WSP. Enzyme *CYP1B1* had induced transcription at 30 minutes for WSP and UPM, but while it stayed highly activated at 120 minutes for UPM, the gene was back to baseline with WSP by 120 minutes, and in the case of small airway epithelial cells perturbed by ADP, down-regulated by 20 hours (Appendix Figure D.1C).

5.4.2 Hyaluronan metabolism and signaling response to both WSP and UPM is highly dynamic

HYA metabolism and signaling relevance to lung function has shown increasing evidence of relevance in the literature, so we saw if our data could further clarify its timing and cellular response with pollutants[86, 107, 212]. Specifically, smoke can cause degradation of high-molecular weight (HMW) hyaluronic acid into low-molecular weight (LMW) hyaluronic acid fragments, which then serve as pro-inflammatory signals[86]. Genes showing downregulation by thirty minutes in UPM include genes whose proteins are directly involved in both the synthesis (HAS2, HAS3) (HAS1

not expressed) and breakdown (CEMIP, HYAL1, HYAL2, HYAL3) of hyaluronic acid (Appendix Figure D.2). We observe comparable nascent RNA early response in Beas-2B cells perturbed with WSP at the same time points, except that HAS3 and PFKP show initial upregulation before being downregulated compared to baseline at 120 minutes (Appendix Figure D.2).

Given the transience of hyaluronic acid-related response in nascent transcription levels, we evaluated if these responses were also observed at steady state RNA levels. RNA levels of *HYAL2* and its paralogs *HYAL1/3* all showed small downregulation by 3 and 6 hr post WSP or UPM perturbations according to qRT-PCR in smAECs, matching downward trends shown in PRO-seq (Figure 5.1D, Appendix Figure D.3, Supplemental Table 1). Conversely, synthesis genes (*HAS2* and *HAS3*) showed upregulated steady state RNA levels by 3 hr. RNA-seq of Beas-2B cells perturbed with WSP indicated that at 2 hours, RNA levels of all *HYAL* and *HAS* genes followed the trends observed with qRT-PCR, with limited change observed in glycolysis-related genes (Appendix Figure D.2). Finally, we confirmed that levels of hyaluronic acid in media decreased at 3 hours for smAECs perturbed with either UPM or WSP, but returned to vehicle levels by 6 hours based on ELISA (Figure 5.1E, Supplemental Table 2).

We then evaluated whether predicted signaling pathways involving HYA followed expected trends. Pro-inflammatory signals from HYA fragments have been shown as largely mediated via toll-like receptors TLR4 and TLR2, which then activate NFKB for increased expression of cytokines(Campo et al., 2012); this same pathway shows increased expression at 30 minutes in PRO-seq and 2h in RNA-seq, specifically expression of cytokines IL-6 and TNF-a for UPM (Figure 5.1F, Appendix Figure D.2). Similarly, TMEM2 (textitCEMIP2) mediates the decreased expression of *CEMIP1* by cytokines like IL1-B and TG1-B and is shown to be upregulated at 30 minutes in our data (Figure 5.1F) [211, 212]. Finally, we found that HYA metabolism can be directly linked to lung-based disease measurements. Specifically, a 1 unit increase in HYAL1 protein levels predicted increased FEV1 (ml/yr) by 0.0024 (high-precision p-value 0.0016) in a linear regression model of the COPDGene cohort; this is after accounting for smoking status, gender, age, BMI, and total cumulative cigarette exposure.

5.4.3 Robust transcriptional network responses are comparable across particulate matter perturbations and cell types

Nascent RNA sequencing allows us to not only consider the nascent transcription of genes, but also of transcribed regulatory elements (tREs). The response of tREs and motif enrichment can then allow us to predict the transcription factors responding to UPM and the tREs they are regulating with Transcription Factor Enrichment Analysis (TFEA-LE)[200, 246]. Therefore, we identified and characterized the responding tREs and transcription factors in our three lung cell experiments (smAEC-UPM, BEA2B-WSP, smAEC-ADP) (Supplemental Data). Since transcription factors can also be impacted by a change in chromatin accessibility, we also considered ATAC-seq in nasal airway epithelial cells perturbed by UPM or WSP at the same time points for PRO-seq.

Similar to genes, tREs show robust early responses in UPM and WSP (early rise and fall) (Appendix Figure D.4A left). However, unlike with genes, a similar number of tREs also show maintained decreased expression after 30 or only at 120 minutes (Appendix Figure D.4A middle and right). This finding suggests that a similar number of tREs are being deactivated as being activated in response. Similar to genes, tREs responding in both WSP and UPM show very similar directionality with more variability in exact timing (most rising in UPM also rise in WSP, etc.) (Appendix Figure D.4B). The group of tREs with the largest variability between WSP and UPM were the "late fall" category, with equal proportions of UPM late-fall tREs showing both late-fall and early-fall response in WSP.

While responses of general stress-focused transcription factor were shared, some transcription factors showed different predicted activities depending on perturbation. Both WSP and UPM responsive tREs are strongly enriched in the GC-rich motifs of several transcription factor families known to respond to respiratory stress (e.g. KLF, SP families, Nuclear respiratory factor), primarily at 30 minutes (Figure 5.2A). We also observe strong regulation of the genes encoding for these transcription factors at the 30-minute mark in both perturbations (Appendix Figure D.5). However, while Serum Response Factor (SRF) was the primary predicted transcription factor re-

sponding to WSP, tREs responding to UPM showed no clear enrichment for the motif (Figure 5.2A). Similarly, the gene encoding for SRF saw significant upregulation at 30 minutes for WSP, but with no clear transcriptional change for UPM (Figure 5.2B). The tREs predicted as regulated by SRF in WSP response (details in Methods) also showed upregulation at 30 minutes for WSP but have no transcriptional response to UPM at any timepoint (Appendix Figure D.6). Conversely, NFKB transcription factors (REL, NFKB1, NFKB2) had significantly positive enrichment at 30 minutes for UPM but not WSP, and genes encoding for NFKB TFs increased expression at 30 minutes under UPM conditions but not WSP (Figure 5.2A, Appendix Figure D.5). Not all transcription factors change chromatin accessibility with their activation or repression, but we do see that the KLF TFs, SP TFs, NRF1, AhRR, and AhR show the same trends (as seen in PRO seq) in ATAC-seq at the 30-minute mark of nasal airway epithelial cells perturbed with WSP or UPM (Appendix Figure D.7).

Both WSP and UPM early response tREs were enriched for the transcription factor Aryl-hydrocarbon Receptor (AhR) and its antagonizer AhRR, particularly at 30 minutes (Figure 5.2A). However, while AhR still had a positive enrichment score at 120 minutes for UPM, it had a negative enrichment score for WSP (Appendix Figure D.8). These results match the transcription levels of *AhR* and *AhRR*. In WSP, *AhR* steadily increases transcription while *AhRR* decreases by 120 minutes; in UPM, *AhR* returns to baseline expression at 120 minutes while AhRR continues to increase expression at 120 minutes (Figure 5.2B). Previously confirmed target genes of AhR also show similar trends of upregulation at 30 minutes with both UPM and WSP (Appendix Figure D.9)[79]. We then compared the tREs predicted as responding to AhRR/AhR between perturbations, focusing on AhRR as it has a slightly more specific motif less likely to lead to false motif hits (details in Methods).

Following the transcriptional trends of *AhRR* and TFEA-LE results, many of the predicted AhRR-responding tREs at 30 minutes for UPM still maintain high upregulation compared to vehicle at 120 minutes (Figure 2C, grey lines UPM 120min). Conversely, WSP-responsive AhRR tREs become mostly downregulated compared to vehicle at 120 minutes (Figure 5.2C, pink lines

WSP 120 min). Despite the different perturbations and cell types, both WSP and UPM shared 18 tREs at 30 minutes predicted to be regulated by AhRR. And when highlighting how the 30-minute responding AhRR tREs are ranked in expression changes in opposite perturbations, we see the tREs follow the same response as the perturbation they map to (Figure 5.2C, comparing WSP 30 min to UPM 30 min within colors). In other words, many of the same tREs appear to be responding but with timings dependent on the experiment.

Finally, the response of smAECs treated with ADP after 20 hours was predicted to be largely based on a single transcription factor: P53. Interestingly, TFEA-LE noted P53 (and its motif neighbor P63) as having positive, but weak enrichment in UPM and WSP. The P53 binding motif was confined to weakly upregulated tREs at both 120 and 30 minutes for UPM and WSP rather than the strongly downregulated tREs for afghan dust particles after 20 hours for ADP (Figure 5.2D). We observed the same weak upregulation in accessibility of tREs with P53 motifs in nasal epithelial cells perturbed with WSP and UPM (Supplemental Data).

.

5.4.4 Particulate matter responsive tREs pinpoint SNPs associated with COPD, asthma, and other lung-relevant diseases

Our characterized transcriptional responses of pollutant-responding lung systems then improved the interpretation and annotation of genetic variants associated with COPD and asthma. We specifically asked whether focusing association analyses on SNPs within particulate-matter responsive lung tREs can pinpoint SNPs associated with COPD/asthma, or related diseases, even when working with a small and heterogeneous association dataset. In this case, we perform an association study of SNPs in pollutant-responsive tREs in the All of Us cohorts for COPD and asthma[176](Figure 5.3A Steps 1 and 2). We consider all available ethnicities to increase the generalizability of our study at the expense of statistical power. Therefore, we considered functional annotations along with less stringent association cutoffs since noncoding SNPs, regardless of functionality in disease, are known to have weaker effect sizes.

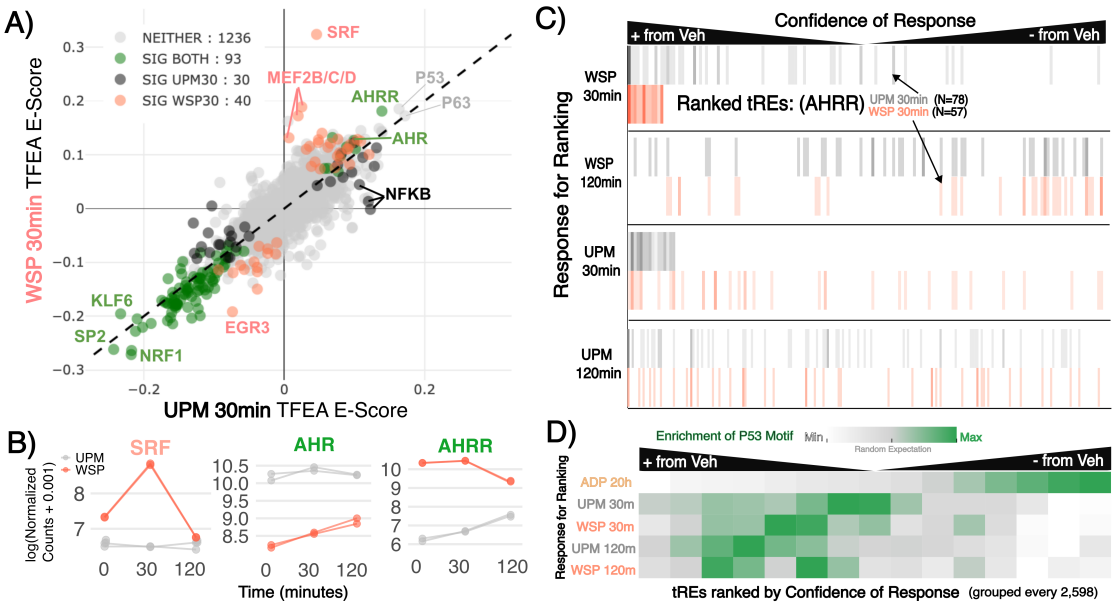

Figure 5.2: **UPM produces a specific response of transcriptional networks (transcription factors and corresponding tREs). A.** Scatterplot of TFEA GC-corrected enrichment scores for WSP and UPM 30 minutes. Transcription factors are highlighted as being a call in neither UPM/WSP, only one or the other, or both (where significance requires adjusted p-value < 0.01 and Fraction above Background < 0.5 – see Methods). **B.** Log normalized counts of genes encoding transcription factors SRF, AHR, and AHRR in UPM and WSP treated cells across time. **C.** tREs called responding via AhRR to WSP and UPM at 30 minutes are colored pink and grey, respectively. The x axes are rankings of tREs according to confidence of up(+) or down(-) regulation compared to Vehicle across 4 different responses: WSP 30min, WSP 120min, UPM 30min, UPM 120min. Left (or Right) most x-coordinates are tREs with the lowest adjusted p-values for a positive (or negative) log2fold change for the listed response. **D.** Quantiles of tREs are ranked as done in C for 5 different responses: ADP 20, WSP 120min, UPM 30min, UPM 120min. Relative enrichment of the P53 motif (weighted score from TFEA) is colored within each quantile (scaled within each row) so that green shows the maximum enrichment.

A total of 15,196 tREs responded to particulate matter, and we captured full RNA coordinates for 15,136 (99.6%) with the LIET-EMG approach[246] (Figure 5.3A Step 1). 12,576 (83%) of these tREs were uniquely responsive to a particulate, with most of the remaining tREs being shared between WSP and UPM (Appendix Figure D.12A). Since bidirectional transcription in PRO-seq enables precise identification of the transcriptional start site (TSS) of genes[278], we also considered Gene TSS bidirectionals of genes with significant changes across particulate responses. We again saw that most calls were unique to a given particulate response (6,935 – 77%) (Appendix Figure

D.12B). After merging any RNAs with overlapping genomic coordinates, a total of 20,764,608 bp were considered with 28% coming from either Gene TSS bidirectionals or tREs whose RNAs overlapped the 1kb gene TSS region, 36% coming from tREs within the exons/introns of genes, and the final 36% coming from tREs outside of annotated genes.

We then predicted the regulatory networks of a given tRE (and overlapping SNP), including the likely target gene and, if applicable, the transcription factor who's binding a SNP could disrupt. To this end, we considered several factors from the particulate matter data along with publicly available databases, including our previous nascent run-on sequencing database (DBNascent) (full details in Methods, Figure 5.3A Step 3)[164, 223]. We then ranked SNPs with the strongest support for downstream validation based on association with COPD/asthma, previously found association with traits in ClinVar/GWAS Catalog[35, 127], confidence of regulatory network support, and overall response to particulate matter (Figure 5.3A Step 4; details in Methods).

The All of Us cohort had 6,164 SNPs mapping within our coordinate search space, with 51% of these SNPs (N=3,144) associated with COPD and asthma in the same direction (Figure 5.3B). Of the total SNPs, 10% (N=681) had p-values below 0.05 for association with asthma or COPD. Compared to the original full set of SNPs, a similar proportion of significant COPD- or asthma-associated SNPs associated in the same direction with both diseases (57%) and all SNPs containing TF binding site according to ENCODE (6.5% to 58.4%) or having genome-wide significance with external studies were maintained (0.7% to 6.5%) (Figure 5.3B). Of the 44 SNPs correlated to external traits, only 3 (7%) were found within exons, with 23 (52%) and 18 (41%) SNPs being intronic and intergenic, respectively. SNPs and their functional annotations (predicted target TFs/genes, phenotypes, etc.) can all be found in Supplemental Tables 3 and 4.

The traits these SNPs were externally associated with could all be directly linked back to lung function and inflammatory response except for Peutz-Jeghers syndrome and non-specific traits like neuroimaging and protein measurements (Figure 5.3C, Supplemental Table 5)[4, 13, 15, 25, 30, 34, 58, 67, 72, 85, 91, 111, 115, 142, 144, 179, 180, 219, 235, 237, 238, 241, 256, 288, 301, 283]. Two SNPs associated with direct lung-related measurements – smoking initiation and a COPD

indicator (FEV/FVC ratio). This second SNP (rs12894780), showing a significant odds ratio of 1.06 for COPD (p-value=$5.9x10^{-3}$) in our study, was previously associated with FEV/FVC ratio across two different genome-wide association studies and ancestry cohorts (p-values 0.00237 and $3.8x10^{-20}$) [45, 140]. It is an intronic variant found within *ITPK* and is 67bp from the predicted transcription initiation site (μ) of a tRE responding to WSP that also has evidence as a cis-regulatory element in lung based on ENCODE[164] (Appendix Figure D.10). *ITPK* indeed shows upregulation at all time points and air pollutant perturbations compared to vehicle, although it is only statistically significant in WSP (Appendix Figure D.11). Two other SNPs were associated with diseases impacting lung development: Ellis-van Creveld syndrome and Hutchinson-Gilford syndrome[13, 256]. The largest category (N=12) of non-lung phenotypes corresponded with either immune-cell measurements (N=7) or red blood cell measurements (N=5). The second largest group of phenotypes focused on markers of muscle/liver/kidney function (N=8).

Overall, the enrichment of SNPs showing independently significant correlation with lung and inflammatory based traits support that our filtering approach improved identification of lung disease relevant variants.

.

5.4.5 Functional relevance of coordinate space is not tied to timing of response

PRO-seq allows cleaner evaluation of noncoding SNP linkages, specifically in linking TF and target genes, but whether SNPs within transient transcriptional patterns are also relevant has not been established. Given most of the tREs and genes responses in PRO-seq were transient, we evaluated whether these responses 1) had SNP enrichment comparable to consistent responses (e.g. plateau) and 2) impacted steady-state RNA levels. For the first question, we considered transient responses based on both transcription (PRO-seq) and chromatin accessibility (ATAC-seq). We saw no systematic difference in enrichment of SNPs associated with asthma or COPD across time categories, even when only considering transient (early) responses in both WSP and UPM (Figure 5.4A). We then compared steady-state RNA changes from previously published RNA-seq (2h and

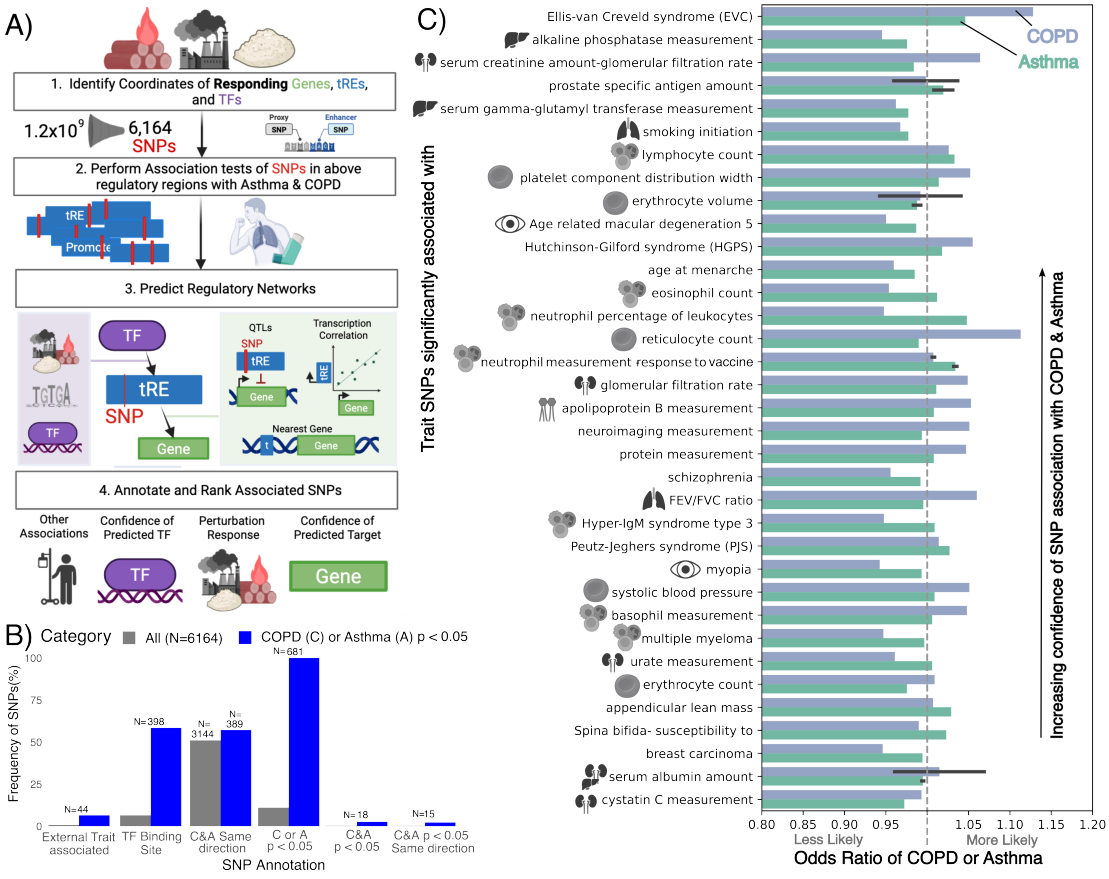

Figure 5.3: **Perturbation-focused regulatory region SNPs associated with COPD and asthma are externally associated with other traits related to lung-function.** **A.** Visualization of pipeline. 1) Identify the genes, tREs, and transcription factors responding to air pollutant perturbations and the coordinates of gene and tRE bidirectionals. 2) Evaluate the association of SNPs within these coordinates and with asthma and COPD according to All of Us cohorts. 3) Predict transcription factor-tRE linkages by considering the transcription factors responding to perturbations, motifs within SNPs, and binding sites of transcription factors. Predict gene-tRE linkages based on the nearest gene, quantitative trait loci for the SNPs, and correlation of transcription across lung cells. 4) Rank candidate SNPs based on associations, regulatory networks, and response of tREs to the perturbations. **B.**Percentage of all region SNPs (grey) vs those associated with COPD (C) or asthma (A) (blue) that fall into different functional annotations (details in Methods). **C.** shows the traits associated with the SNPs associated with COPD or asthma in this work based on GWAS Catalog and ClinVar. Bar plots indicate odds-ratios for COPD and asthma where error bars indicate there are two SNPs associating to the trait.

4h) to the nascent transcription changes in PRO-seq (30m and 120m) in Beas-2B cells perturbed with WSP[77]. Regardless of timing in PRO-seq (early, plateau, late), about 50% of responding genes also showed change in steady-state RNA levels (Appendix Figure D.13). Early response

genes then continuously showed the highest percentage shared with RNA-seq for significant change in the same direction (Appendix Figure D.13). Overall, transient transcriptional responses still showed important functional insight for SNP filtering in addition to enabling functional annotation of SNPs.

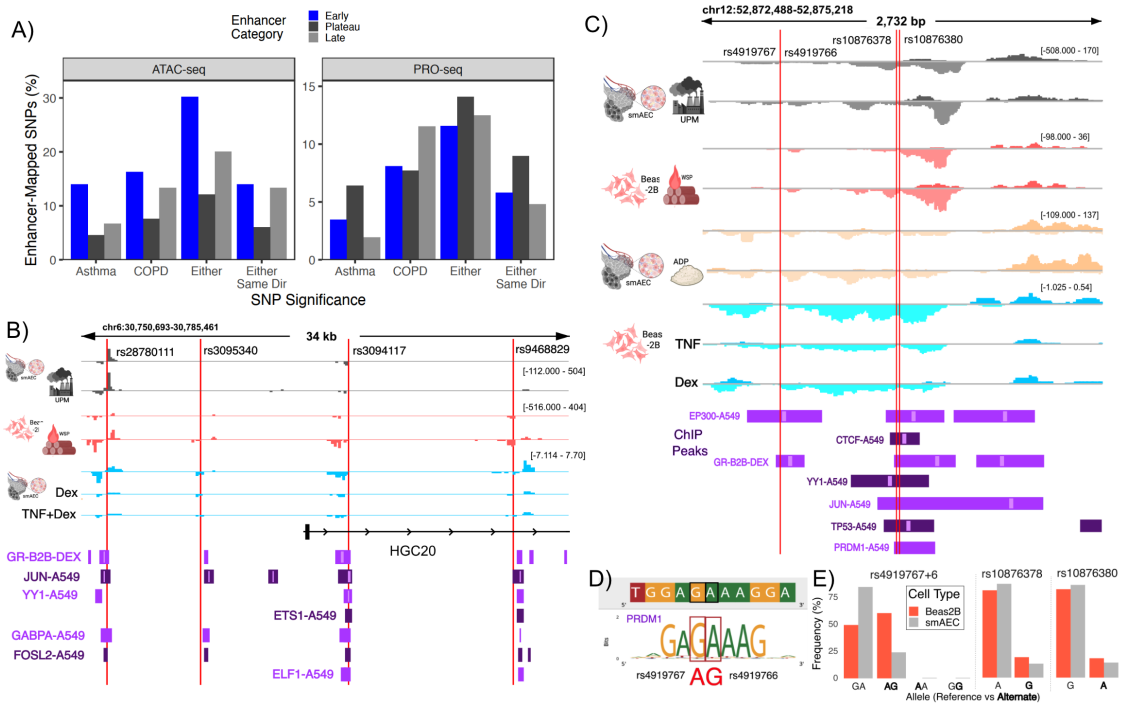

Figure 5.4: **Focused multiomics data clarifies grouped SNP relevance across cohorts into biological features.** **A.** Percentage of SNPs in enhancer timing categories (agreed upon between WSP and UPM) based on ATAC-seq and PRO-seq that are associated with asthma or COPD. **B.** and **C.** IGV tracks for two SNP-enriched enhancer groups, of PRO -seq of smAEC perturbed with UPM and Beas-2B cells perturbed with WSP, and GRO-seq of Beas-2B cells perturbed with dexamethasone (Dex), TNF, or both. 30 minute time points (compared to Vehicle above) are shown. Transcription factor or EP300/CTCF ChIP-peaks of either Beas-2B perturbed with Dex (B2B-DEX) or A549 cells. The numerical read distributions are not directly comparable outside of experiments due to different normalization approaches (scale noted on the right top of each experiment). SNPs are labeled by their rsids. **D.** Annotated opentargets screenshot of SNPs rs4919767 and rs4919766 (chr12:52873025 and 6, respectively), and indicate change of GA to AG within a PRDM1 binding motif. **E.** Frequency of alleles of chr12 enhancer group SNPs (C) across cell types used in this study.

.

5.4.6 Focused multiomics links mixed SNP significance across cohorts to shared target-gene and enhancer functionalities

Lung cell multiomics data can confer key insight into the mechanisms by which noncoding SNPs might have association with COPD/asthma. This can include clarifying the possible role of SNPs associated with other diseases, like those noted above. We specifically asked whether, if by focusing on biological features (e.g. enhancers) rather than individual SNPs, we could identify enhancer groups that showed consistent significant association with Asthma/COPD across cohorts.

Although some immune-associated SNPs were intronic to genes that logically linked to inflammatory signals (e.g. *PFKP*, *IL2RA*, *BAG4*), many were intergenic and could be linked to target genes based on our data. One tRE with a SNP associated with reticulocyte count (rs54211) showed high correlation of transcription with the gene encoding platelet-derived growth factor (PCC=0.91, p-value$< 1x10^{-20}$, expressed in 87% samples), both of which also showed strong early response in WSP. Instead, when relying on quantitative trait loci alone, GTEx has indicated this SNP was associated with isoform transcription of its nearest gene, which has no known correspondence with reticulocyte function: *RPL3* [22]. Similarly, SNP rs4989187 associated with eosinophil counts was found within an intergenic tRE predicted to target *CYP26A1* based on our data (PCC=0.95, p-value$< 1x10^{-20}$, expressed in 49% samples). A knockout of this same gene has been previously shown to increase eosinophilic globules[229]. Finally, one of the SNPs associated with serum albumin, rs10145747, was in an intergenic tRE between *SERPINA* genes, with our data suggesting the tRE transcription was best correlated with transcription of *FAM181A* and *SERPINA6*. In each case, direct transcriptional data in a disease-relevant cell type and perturbations have allowed noncoding-SNPs to be linked to a target gene of logical function.

When evaluating if biological features have consistent results across cohorts, we often found the same group of enhancers contained SNPs associated with asthma and COPD, albeit with SNPs showing varying significance across cohorts. For example, we considered the All of Us Research cohort and GERA cohorts based on our previous work associating SNPs within tREs responsive

to TNF and dexamethasone in Beas-2B cells with asthma from the GERA cohort[209]. Only one of the original 36 significant SNPs was also considered in the All of Us cohort and responsive to WSP and UPM: rs3094117 (chr6:30769709). Interestingly, while this SNP did not have significant association with either COPD or asthma in the All of Us cohort (p-values 0.62 and 0.75), three SNPs found within tREs just proximal to the SNP tRE showed the same transcription patterns and significant correlation with asthma: rs28780111, rs3095340, and rs9468829. All four tREs showed upregulation with at least one particulate matter type, and downregulation in response to dexamethasone at 30 minutes (Figure 5.4B, Appendix Figure D.14). Despite two of the four tREs falling within introns of *HGC20*, all four tREs coordinates had accessibility and transcriptional peaks specific to their regions far above any reads dispersed across HGC20 in all particulate matter datasets (Appendix Figure D.14). The SNPs were close to or within A549 or Beas-2B ChIP peaks of transcription factors that were also identified as responsive to particulate matter or dexamethasone based on TFEA-LE: GR, JUN, YY1, ETS1, GABPA, FOSL2, ELF1 (Figure 5.4B)[164]. One of the SNPs, rs3095340, was also found significantly associated with FEV1/FVC and inflammatory bowel syndrome in a genome-wide pleiotropic analysis of Europeans (N=400,102 COPD cases and 709,884 controls, p-value $6x10^{-11}$)[104]. In the smaller All of Us cohort (N=5,843 cases and 158,404 controls), this SNP had the same odds ratio direction (1.03 for COPD) but did not reach statistical significance (p-value 0.32). Therefore, by considering enhancer groups and genes, we pinpoint a group of features that might perform a shared functionality relevant to lung disease.

Finally, of the top ten associated SNPs with COPD and asthma in our study, four were found within the same enhancer group (overlapping tREs) on chromosome 12 that showed robust early response in both UPM and WSP (Figure 5.4C). There appear to be three positions of transcription initiation across these tREs according to PRO-seq and GRO-seq data in BEAS-2B and smAEC cells as well as EP300 ChIP peaks from A549 cells. All three tREs show accessibility peaks in nasal epithelial cell lines while falling within one vast H3K27ac peak in A549 (Appendix Figure D.15)[164]. The odds ratios for all four SNPs were almost identical and indicated decreased likelihood of COPD or asthma with the presence of the SNP: 0.945 for COPD and 0.98 for asthma. One of the four SNPs

(rs10876380) had also been independently shown to be associated with alkaline phosphatase levels (Beta=0.0088, p-value 4.17x10-13), a phenotype which has also been associated with poor lung function [203]. For all four SNPs, the alternative allele is more common across all ancestries (range from 59-81% of population), suggesting a possible fitness advantage as supported by decreased odds of lung-diseases(Appendix Figure D.16)[22, 38].

Interestingly, our data suggests that purely database-based predictions of impacted target-gene(s) and transcription factors for this enhancer group might not be relevant to pollutant-based response. SNPs rs4919767 and rs4919766 correspond to adjacent positions in chr12 (52873025 and 52873026), and disrupt two sites corresponding to a PRDM1 motif (Figure 5.4D). Indeed, these two SNPs appear to be almost exclusively found together in reads from the cell lines we considered (Beas2B and smAECs), with few, if any, cases of the alleles being disrupted separately (Figure 5.4E). SNPs rs10876378 and rs10876380 were similarly shown to disrupt key bases for the binding motifs of SOX5 and MEIS1/MEIS2, respectively (Appendix Figure D.17). Neither Beas2B nor smAEC cells appeared to show these alleles in our data (Figure 5.4E). Despite predicted impacts on motifs, all transcription factors whose motif instances the SNPs were suggested to disrupt did not contain significantly enriched motifs in responding tREs across any particulate responses, nor appropriate binding sites based on lung-cell CHIP-seq (Figure 5.4C, Supplemental Data). Although, raw signal suggests that there may be a significant PRDM1 ChIP-seq signal at the appropriate SNP sites (Appendix Figure D.18). We next turned our focus to the target-gene. All four SNPs fall within an eQTL region associated with PFDN5 in the upper lobe of the left lung[164]. This gene, however, the opposite transcription trend of the tREs, showing nonsignificant (adjusted p-value > 0.05) downregulation at 30 minutes in both perturbations (Appendix Figure D.11). It is possible that these tREs repress the gene, there is simply a different gene target, or that these SNPs are operating within a specific subset of cells in the lung not well-captured by Beas-2B or smAEC cell studies.

Overall, PRO-seq data proved essential to link several SNPs to a shared biological feature that can then undergo downstream experimental testing to evaluate the exact functionalities these

SNPs confer.

5.5 Discussion

We first highlight dynamic transcriptional responses of lung cells to air pollutants. The strong initial transient response might represent a core initial stress response found in lung cells before stabilization to stress-specific responses, and merits additional comparison across non-particulate perturbations. By 120 minutes, we saw a clear upregulation of genes previously noted as involved in modulating response to foreign particles such as smoke. WNT signaling is involved in lung cell regeneration/repair and has been previously shown to regulate inflammation from cigarette smoke[75, 192]. Vitamin D metabolism and similar detoxification-focused genes like *CYP1A1* and *UGT* genes have also been shown to help modulate air particulate responses[27, 158]. Finally, despite differences in cell type and air pollutant, a significant fraction of both transcription factors, tREs, and genes are responding regardless of time point – suggesting some universal pollutant responses worth exploring in the future.

Our work specifically shows a complex but rapid response to hyaluronan molecules that may be relevant not only to cigarette-smoke response, but also to environmental pollutants. Other studies have shown the importance of HYA response in COPD, both with HYA plasma levels and *HYAL2* expression [252]. With reactive oxygen species from smoke particles inducing HMW breakdown, it is possible that the rapid response we observed reflected a need for HYA metabolism being shifted towards other stress-related responses, with synthesis genes rebounding to increased steady-state RNA levels by 2-3 hours. Indeed, the mixed timing response of *HAS3* in Beas-2B cells compared to smAECs might be reflective of cell type differences or stress resilience, as total lung tissue vs lung macrophages have shown opposite results of *HAS3* and *HAS1* expression response to cigarette-smoke[19]. Finally, the rapid HYA response in our work provides biological explanation behind a previously discovered treatment for smoking-based inflammation. Aerosolized HMW can have anti-inflammatory effects and reduce lung injury if taken with smoking, with the positive impact lowering with delayed HMW intake [29].

Our predicted responses of transcription factors both confirm and further clarify transcription factor particulate-based studies in the literature. Previous work showed that cigarette smoke increases expression of SP- and KLF-based transcription factor genes, as observed in our data, and also show that knockdown of KLF6 reverses mitochondrial dysfunction and apoptosis[51, 253]. Nuclear respiration receptors NRF1 and NRF2 have both been shown to compete in response to cigarette smoke, following our data, with NRF2 specifically protecting lung cells from wood smoke particle toxicity[125, 248]. Indeed, our ATAC-seq indicated that while NRF2 regions had decreased accessibility in wood smoke, they had increased accessibility in UPM at 30 minutes, thus downregulation of NRF2 target genes *SRXN1* and *NQO1* in WSP but not UPM. Finally, the finding that p53 target tREs are only slightly upregulated at WSP and UPM early responses but strongly downregulated by 20 hours after ADP response might indicate key timings for oxidative stress and DNA damage and p53 response. Indeed, repeated exposure of particulate matter can also lead to hypermethylation and therefore downregulation of P53[299].

We also confirm that rapid, transient transcriptional responses can still be biologically relevant for SNP filtering and functional annotation, and can vary across conditions. Indeed, such transient responses may still confer changes in steady-state, or might prime enhancers or genes for easier accessibility if perturbed again[250]. We considered a broader range of cell types and perturbations to ensure less bias in SNP filtering from enhancers, but that limited the direct comparability of our results across experiments. It would be interesting to compare such transient responses between cells from patients of increasing disease severity, and see how RNA decay rates help balance pollutant impact[245]. Regardless of the disease-relevance of these dynamics, considering short time points and time-series data allowed more specific SNP functional annotation (e.g. linkage to target genes and transcription factor binding) than provided by general public databases.

Finally, our results show that taking advantage of precise enhancer coordinates and transcriptional network characterization across multiple conditions enabled more refined association analyses of SNPs with COPD and asthma. First, we improved clarification of secondary vs primary associations with SNPs by considering whether the SNPs were found in lung-enhancers or

genes. For example, SNPs associated with markers of kidney function like creatinine or glomerular filtrate rate have often been associated with COPD or asthma as secondary impacts from increased work of skeletal muscle for breathing or increased inflammatory processes[4, 25, 115, 179, 241]. Second, multiomics data allowed us to refine the predicted transcription factor and target genes explaining phenotypic connections. For example, we were able to predict target-genes of multiple SNPs associated with external associations along with COPD and asthma. Finally, we showed how assessing associations based on enhancers rather than SNPs alone allows us to still capture disease-relevant functions even when individual SNPs show mixed significance across cohorts. In this study, we were limited to small cohort sizes for association studies, with mixed ancestries increasing generalizability of our results but also limiting statistical power from heterogeneity. Therefore, we focused on cases where multiple individual SNPs showing significance in at least one cohort all pointed to the same enhancer group. Noncoding SNPs often have lower individual effect sizes (and therefore significance), regardless of causality, due to the higher level of redundancy in enhancer behavior compared to coding regions[234]. Therefore, SNP-focused analyses would miss that several cohorts confirm the relevance of the same enhancer, and therefore potentially important disease mechanisms. Instead, we were able to identify three promising enhancer groups with validation across cohorts based on shared SNPs as well as response in disease-relevant perturbations worthy of downstream experimental exploration. Overall, this work provides the first pan-pollutant based nascent transcriptional characterization along with a framework and dataset of promising lung-disease SNPs/enhancers for downstream testing.

5.6 Data availability

DEX-TNF Beas2B GRO-seq (GSE124916), WSP Beas2B RNA-seq (GSE283113), WSP Beas2B PRO-seq (GSE167372), UPM smAEC PRO-seq (uploading to GEO), ADP smAEC PRO-seq (GSE201150). ENCODE experiments and files, all in A549, include ENCSR977FEF (PRDM1 Experiment), ENCFF096PTH (EP300), ENCFF229ULU (CTCF), ENCFF835PPK (YY1), ENCFF618CNL (JUN), ENCFF229ULU (TP53), ENCFF271WHK (ETS1), ENCFF343BIK

(GABPA), ENCFF232TVP (FOSL2), ENCFF169QCM (ELF1). UPM and WSP NEC ATAC-seq de-deuplicated bams were received by Dr. Sarah Sasse (personal communication).

Supplemental Data can be found at Zenodo: https://doi.org/10.5281/zenodo.18929828.

Chapter 6

Conclusions and Future Directions

6.1 Summary of Contributions

Transcriptomics and genomics provide complementary insights into disease biology: genetic variation offers a clearer framework for causal inference, while transcriptomics contextualizes gene function within cell types and environmental conditions. Despite their complementary strengths, integrating these data types has remained challenging due to technical heterogeneity across assays, ambiguity in which cell types drive the most relevant pathology, and limited ability to resolve the molecular mechanisms of noncoding genetic variants. This thesis provides computational frameworks and benchmarking that address these challenges, enabling integrative analyses that connect genetic risk loci to regulatory networks.

I first address (Chapter 2) the challenge of integrating heterogeneous transcriptomic datasets, which is essential for assembling sufficiently large and diverse cohorts to model complex diseases. Not only did I provide guidelines for optimized preprocessing and assay-integration, but I also introduced several new potential biomarkers of ovarian cancer. This framework establishes the statistical and technical basis required for downstream integration with genetic data.

In my third chapter, I pinpoint some key challenges behind integrating genetic variants with disease-relevant cell types and genes. Briefly, I benchmarked methods that integrate GWAS summary statistics with scRNA-seq data to identify the cell types that best represent causal disease pathology, as well as the SNP-gene linking methods on which these approaches rely heavily. I show that the use of scRNA-seq atlases that only consider healthy tissue can lead to misleading

results, further demonstrating the importance of considering the proper cell type for physiological interpretation of SNPs. Similarly, I show the high variability of SNP-gene linking approaches for non-coding SNPs and how this variability can severely change what biological mechanisms and drug targets are predicted. Given this need, I focus the remainder of my thesis on using specific transcriptomics data to improve our ability to link noncoding SNPs to druggable targets/mechanisms likely relevant to the disease.

In Chapter 4, I develop several algorithms to optimize 1) linking noncoding SNPs to transcriptional units (e.g. enhancers/genes), and 2) mapping active transcriptional networks under a given condition. Briefly, I introduced LIET-EMG, based largely on the original foundation model from my colleague Dr. Jacob Stanley, to probabilistically assign SNPs to a given enhancer [232]. Next, I provided a new counting algorithm (Mu-counts) and pipeline that addresses overlapping transcription while maximizing signal from lowly expressed enhancers. Finally, I introduce TFEA-LE to improve our abilities to assign TFs to the enhancers they regulate in a given treatment. These efforts improved how SNPs might be assigned to druggable targets such as enhancers, genes, and transcription factors.

Finally, my thesis culminates with showcasing how my algorithms enable identification of novel biological mechanisms of lung disease by integrating genomics with transcriptomics (Chapter 5). My improvements first allowed me to clarify shared biological mechanisms of response across environmental perturbations in different lung cell types previously unestablished. I then combined publicly available databases with our own nascent sequencing data to filter and assign COPD and Asthma associated SNPs to predicted functional mechanisms and biological pathways. From these efforts, my collaborators were able to experimentally validate key biological pathways or super enhancer regulation relevant to lung disease.

Collectively, this thesis advances the frameworks by which integrating genomics with transcriptomics can clarify disease mechanisms and support translational discovery in complex human diseases.

6.2 Future Work

This work also provides several clear avenues for future research that can continue to strengthen the open-source abilities to identify disease mechanisms. I provide three brief example areas below.

6.2.1 Combining single-cell with bulk approaches for network prediction

Nascent RNA sequencing is largely optimal for transcriptional regulatory network prediction due to its ability to accurately quantify regulatory element activity from transcription. However, the chemistry behind nascent RNA sequencing combined with the low nascent RNA content in cells has limited the approach to bulk sequencing. Attempts for single-cell nascent approaches have shown extremely sparse data that would make enhancer-based studies infeasible [154]. Instead, as noted in Chapter 2, 5′-scRNA-seq might provide a more feasible alternative, showing a decent capture of enhancer transcription compared to classic scRNA-seq approaches [163]. Future efforts could compare regulatory networks predicted based on scATAC-seq, 5′-scRNA-seq, and bulk PRO-seq. We could evaluate how much loss we see of enhancer identification across approaches, and compare network prediction from deconvoluted PRO-seq and single cell alternatives. Most of my work focused on using perturbation data for network prediction so it would also be interesting to integrate Bayesian (or other statistical) priors of networks predicted in PRO-seq into methods for predicting networks from single-cell data.

6.2.2 Improved large-scale enhancer characterization

I introduce a novel way of fine-mapping SNPs to functional units, particularly enhancers, using a probabilistic coordinate space of enhancer RNAs using the LIET model. This can currently be done at a per-experiment level, as done in Chapter 4, but we can extend this to a universal dataset based on the predicted enhancer lengths across the full 3,000+ nascent sequencing samples in DBNascent [223]. This work could also consider whether lengths of enhancers seem to systematically

change based on perturbation. Previous work had used these same sequencing data to predict enhancer-gene linkages across cell types [223]. A similar effort can be made but with my optimal counting methodologies outlined in Chapter 4. The accuracy of enhancer-SNP-gene linkages can then be validated with eQTL and perturb-seq data.

6.2.3 Novel database and resource for functional prioritization of disease pathways

Finally, my above efforts could be combined into a single, shared resource to optimize target and disease pathway discovery. Unsurprisingly, some of the most successful insights into physiologically come from combining multiple sources of evidence, but accessing and modeling these data can become overwhelming. First, we could develop a relational database based on SNPs, enhancers, genes and their linkages. For example, a SNP-focused table could contain predicted disease or gene-linkages from genome wide association studies as well as the enhancers it is predicted to fall within. Enhancers would have their linked genes under given cell types/conditions, whether or not they have been assigned to a given TF based on the leading edge approach, and conditions/cell types under which they are active. Finally, genes would have similar activity metrics across conditions/-cell types. This database could be integrated into a web-server similar to the FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies) web-server evaluated in Chapter 3, that allows users to rank functional units of study (e.g. SNPs, enhancer, transcriptional networks) for downstream validation based on criteria like disease, condition, confidence of measurements, etc [263, 264]. The value of this general capability has already been shown in Chapter 5, but is not yet widely available and easily reproducible. Taken together, a full resource could become essential for exploring and understanding the relevant transcriptional regulatory networks for disease to optimize downstream drug target prediction and functional validation.

Bibliography

[1] Nicholas Adzibolosu, Ayesha B. Alvero, Rouba Ali-Fehmi, Radhika Gogoi, Logan Corey, Roslyn Tedja, Hussein Chehade, Vir Gogoi, Robert Morris, Matthew Anderson, Julie Vitko, Clarissa Lam, Douglas B. Craig, Sorin Draghici, Thomas Rutherford, and Gil Mor. Immunological modifications following chemotherapy are associated with delayed recurrence of ovarian cancer. Frontiers in Immunology, 14, June 2023. Publisher: Frontiers.

[2] Hyoungjoon Ahn, Jong Seong Roh, Seulgi Lee, Jiyoon Beon, Beomgu Lee, Dong Hyun Sohn, and Seyun Kim. Myeloid IPMK promotes the resolution of serum transfer-induced arthritis in mice. Animal Cells and Systems, 25(4):219, July 2021.

[3] Katherine M Aird, Hua Li, Frances Xin, Panagiotis A Konstantinopoulos, and Rugang Zhang. Identification of ribonucleotide reductase M2 as a potential target for pro-senescence therapy in epithelial ovarian cancer. Cell Cycle, 13(2):199–207, January 2014. Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/cc.26953.

[4] W. Michael Alberts, James H. Williams, and Joe W. Ramsdell. Clinical Implications of Serum Creatine Kinase Levels in Acute Asthma. Western Journal of Medicine, 144(3):321–323, March 1986.

[5] Mary Ann Allen, Zdenek Andrysik, Veronica L Dengler, Hestia S Mellert, Anna Guarnieri, Justin A Freeman, Kelly D Sullivan, Matthew D Galbraith, Xin Luo, W Lee Kraus, Robin D Dowell, and Joaquin M Espinosa. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. eLife, 3:e02200, May 2014. Publisher: eLife Sciences Publications, Ltd.

[6] Robin Andersson. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. BioEssays, 37(3):314–323, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201400162.

[7] Zdenek Andrysik, Matthew D. Galbraith, Anna L. Guarnieri, Sara Zaccara, Kelly D. Sullivan, Ahwan Pandey, Morgan MacBeth, Alberto Inga, and Joaquín M. Espinosa. Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. Genome Research, 27(10):1645–1657, October 2017. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[8] Rebecca C. Arend, Angelina I. Londoño, Allison M. Montgomery, Haller J. Smith, Zachary C. Dobbin, Ashwini A. Katre, Alba Martinez, Eddy S. Yang, Ronald D. Alvarez, Warner K. Huh, Kerri S. Bevis, J. Michael Straughn, Jr, Jacob M. Estes, Lea Novak, David K. Crossman, Sara J. Cooper, Charles N. Landen, and Charles A. Leath, III. Molecular Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma. Molecular Cancer Research, 16(5):813–824, April 2018.

[9] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1):25–29, May 2000. Publisher: Nature Publishing Group.

[10] Joseph G. Azofeifa, Mary A. Allen, Josephina R. Hendrix, Timothy Read, Jonathan D. Rubin, and Robin D. Dowell. Enhancer RNA profiling predicts transcription factor activity. Genome Research, 28(3):334–344, March 2018. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[11] Joseph G. Azofeifa, Mary A. Allen, Manuel E. Lladser, and Robin D. Dowell. An Annotation Agnostic Algorithm for Detecting Nascent RNA Transcripts in GRO-Seq. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(5):1070–1081, September 2017.

[12] Joseph G Azofeifa and Robin D Dowell. A generative model for the behavior of RNA polymerase. Bioinformatics, 33(2):227–234, January 2017.

[13] Geneviève Baujat and Martine Le Merrer. Ellis-Van Creveld syndrome. Orphanet Journal of Rare Diseases, 2(1):27, June 2007.

[14] Felipe Beckedorff, Ezra Blumenthal, Lucas Ferreira daSilva, Yuki Aoi, Pradeep Reddy Cingaram, Jingyin Yue, Anda Zhang, Sadat Dokaneheifard, Monica Guiselle Valencia, Gabriel Gaidosh, Ali Shilatifard, and Ramin Shiekhattar. The human integrator complex facilitates transcriptional elongation by endonucleolytic cleavage of nascent transcripts. Cell Reports, 32(3), 2024/10/02 2020.

[15] S. Bekaert, N. Rocks, C. Vanwinge, A. Noel, and D. Cataldo. Asthma-related inflammation promotes lung metastasis of breast cancer cells through CCL11–CCR3 pathway. Respiratory Research, 22(1):61, February 2021.

[16] Brooks A Benard, Chinmay K Lalgudi, Ilayda Ilerten, Ruo Han Wang, and Andrew J Gentles. PRECOG update: an augmented resource of clinical outcome associations with gene expression for adult, pediatric, and immunotherapy cohorts. Nucleic Acids Research, 54(D1):D1579–D1589, January 2026.

[17] Tomas Bonome, Douglas A. Levine, Joanna Shih, Mike Randonovich, Cindy A. Pise-Masison, Faina Bogomolniy, Laurent Ozbun, John Brady, J. Carl Barrett, Jeff Boyd, and Michael J. Birrer. A Gene Signature Predicting for Survival in Suboptimally Debulked Patients with Ovarian Cancer. Cancer Research, 68(13):5478–5486, July 2008.

[18] Daniel A. Bose, Greg Donahue, Danny Reinberg, Ramin Shiekhattar, Roberto Bonasio, and Shelley L. Berger. RNA Binding to CBP Stimulates Histone Acetylation and Transcription. Cell, 168(1):135–149.e22, January 2017.

[19] Ken R. Bracke, Mieke A. Dentener, Eleni Papakonstantinou, Juanita H. J. Vernooy, Tine Demoor, Nele S. Pauwels, Jack Cleutjens, Robert Jan van Suylen, Guy F. Joos, Guy G. Brusselle, and Emiel F. M. Wouters. Enhanced Deposition of Low-Molecular-Weight Hyaluronan in Lungs of Cigarette Smoke–Exposed Mice. American Journal of Respiratory Cell and Molecular Biology, 42(6):753–761, June 2010. Publisher: American Thoracic Society - AJRCMB.

[20] Cassandra L. Brenner. Role of ID Proteins in Supporting Ovarian Cancer Cells. Master's thesis, San Diego State University, United States – California, 2025. ISBN: 9798314882115.

[21] Armin Bunde. The Different Types of Noise and How They Effect Data Analysis. Chemie Ingenieur Technik, 95(11):1758–1767, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cite.202300031.

[22] Annalisa Buniello, Daniel Suveges, Carlos Cruz-Castillo, Manuel Bernal Llinares, Helena Cornu, Irene Lopez, Kirill Tsukanov, Juan María Roldán-Romero, Chintan Mehta, Luca Fumis, Graham McNeill, James D Hayhurst, Ricardo Esteban Martinez Osorio, Ehsan Barkhordari, Javier Ferrer, Miguel Carmona, Prashant Uniyal, Maria J Falaguera, Polina Rusina, Ines Smit, Jeremy Schwartzentruber, Tobi Alegbe, Vivien W Ho, Daniel Considine, Xiangyu Ge, Szymon Szyszkowski, Yakov Tsepilov, Maya Ghoussaini, Ian Dunham, David G Hulcoop, Ellen M McDonagh, and David Ochoa. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. Nucleic Acids Research, 53(D1):D1467–D1475, January 2025.

[23] Alessia Buratin, Stefania Bortoluzzi, and Enrico Gaffo. Systematic benchmarking of statistical methods to assess differential expression of circular RNAs. Briefings in Bioinformatics, 24(1):bbac612, January 2023.

[24] John P. Burke, Jurgen J. Mulsow, Conor O'Keane, Neil G. Docherty, R. William G. Watson, and P. Ronan O'Connell. Fibrogenesis in Crohn's disease. The American Journal of Gastroenterology, 102(2):439–448, February 2007.

[25] N. K. Burki and L. Diamond. Serum Creatine Phosphokinase Activity in Asthma. American Review of Respiratory Disease, 116(2):327–331, August 1977. Publisher: American Thoracic Society - AJRCCM.

[26] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology, 8(12):e1002822, December 2012. Publisher: Public Library of Science.

[27] Junhyoung Byun, Boa Song, Kyungwoo Lee, Byoungjae Kim, Hae Won Hwang, Myung-Ryul Ok, Hojeong Jeon, Kijeong Lee, Seung-Kuk Baek, Sang-Heon Kim, Seung Ja Oh, and Tae Hoon Kim. Identification of urban particulate matter-induced disruption of human respiratory mucosa integrity using whole transcriptome analysis and organ-on-a chip. Journal of Biological Engineering, 13(1):88, November 2019.

[28] Diego Calderon, Anand Bhaskar, David A. Knowles, David Golan, Towfique Raj, Audrey Q. Fu, and Jonathan K. Pritchard. Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. American Journal of Human Genetics, 101(5):686–699, November 2017.

[29] Jerome O. Cantor, Joseph M. Cerreta, Marcos Ochoa, Shuren Ma, Ming Liu, and Gerard M. Turino. Therapeutic Effects of Hyaluronan on Smoke-induced Elastic Fiber Injury: Does Delayed Treatment Affect Efficacy? Lung, 189(1):51–56, February 2011.

[30] A Capelli, M Lusuardi, C G Cerutti, and C F Donner. Lung alkaline phosphatase as a marker of fibrosis in chronic interstitial disorders. American Journal of Respiratory and Critical Care Medicine, 155(1):249–253, January 1997. Publisher: American Thoracic Society - AJRCCM.

[31] Joseph F. Cardiello, Gilson J. Sanchez, Mary A. Allen, and Robin D. Dowell. Lessons from eRNAs: understanding transcriptional regulation through the lens of nascent RNAs. Transcription, 11(1):3–18, January 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21541264.2019.1704128.

[32] Nancy V N Carullo, Robert A Phillips III, Rhiana C Simon, Salomon A Roman Soto, Jenna E Hinds, Aaron J Salisbury, Jasmin S Revanna, Kendra D Bunner, Lara Ianov, Faraz A Sultan, Katherine E Savell, Charles A Gersbach, and Jeremy J Day. Enhancer RNAs predict enhancer–gene regulatory links and are critical for enhancer function in neuronal systems. Nucleic Acids Research, 48(17):9550–9570, September 2020.

[33] Centers for Disease Control and Prevention. National Center for Health Statistics. Behavioral Risk Factor Surveillance System, 2023. Calculations by the American Lung Association Research and Program Services Division using SPSS software., 2023.

[34] Ivana Cepelak, Slavica Dodig, Dominik Romic, Nedeljka Ruljancic, Sanja Popovic-Grle, and Ana Malic. Enzyme Catalytic Activities in Chronic Obstructive Pulmonary Disease. Archives of Medical Research, 37(5):624–629, July 2006.

[35] Maria Cerezo, Elliot Sollis, Yue Ji, Elizabeth Lewis, Ala Abid, Karatuğ Ozan Bircan, Peggy Hall, James Hayhurst, Sajo John, Abayomi Mosaku, Santhi Ramachandran, Amy Foreman, Arwa Ibrahim, James McLaughlin, Zoë Pendlington, Ray Stefancsik, Samuel A Lambert, Aoife McMahon, Joannella Morales, Thomas Keane, Michael Inouye, Helen Parkinson, and Laura W Harris. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Research, 53(D1):D998–D1005, January 2025.

[36] Rinal Chavda, Mayuri Inchanalkar, Jitendra Gawde, and Manoj B. Mahimkar. Comparison of real-time PCR and nCounter NanoString techniques to validate copy number alterations in oral cancer. Scientific Reports, 15(1):23015, July 2025. Publisher: Nature Publishing Group.

[37] Chaithanya Chelakkot, Jaewang Ghim, and Sung Ho Ryu. Mechanisms regulating intestinal barrier integrity and its pathological implications. Experimental & Molecular Medicine, 50(8):1–9, August 2018. Publisher: Nature Publishing Group.

[38] Lexin Chen, Can Li, Hangang Chen, Yangli Xie, Nan Su, Fengtao Luo, Junlan Huang, Ruobin Zhang, Lin Chen, Bo Chen, and Jing Yang. Cross-sectional studies of the causal link between asthma and osteoporosis: insights from Mendelian randomization and bioinformatics analysis. Osteoporosis International, 35(6):1007–1017, June 2024.

[39] Xiao-Feng Chen, Ming-Rui Guo, Yuan-Yuan Duan, Feng Jiang, Hao Wu, Shan-Shan Dong, Xiao-Rong Zhou, Hlaing Nwe Thynn, Cong-Cong Liu, Lin Zhang, Yan Guo, and Tie-Lin Yang. Multiomics dissection of molecular regulatory mechanisms underlying autoimmune-associated noncoding SNPs. JCI Insight, 5(17):e136477, 2020.

[40] Jarosław Chilimoniuk, Anna Erol, Stefan Rödiger, and Michał Burdukiewicz. Challenges and opportunities in processing NanoString nCounter data. Computational and Structural Biotechnology Journal, 23:1951–1958, December 2024.

[41] Michael H Cho, Brian D Hobbs, and Edwin K Silverman. Genetics of chronic obstructive pulmonary disease: understanding the pathobiology and heterogeneity of a complex disorder. The Lancet Respiratory Medicine, 10(5):485–496, May 2022.

[42] M. Ryan Corces, Alexandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, Adam J. Rubin, Kathleen S. Montine, Beijing Wu, Arwa Kathiria, Seung Woo Cho, Maxwell R. Mumbach, Ava C. Carter, Maya Kasowski, Lisa A. Orloff, Viviana I. Risca, Anshul Kundaje, Paul A. Khavari, Thomas J. Montine, William J. Greenleaf, and Howard Y. Chang. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nature Methods, 14(10):959–962, October 2017. Publisher: Nature Publishing Group.

[43] Leighton J. Core, André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics, 46(12):1311–1320, December 2014. Publisher: Nature Publishing Group.

[44] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science, 322(5909):1845–1848, December 2008. Publisher: American Association for the Advancement of Science.

[45] Justin Cosentino, Babak Behsaz, Babak Alipanahi, Zachary R. McCaw, Davin Hill, Tae-Hwi Schwantes-An, Dongbing Lai, Andrew Carroll, Brian D. Hobbs, Michael H. Cho, Cory Y. McLean, and Farhad Hormozdiari. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. Nature Genetics, 55(5):787–795, May 2023. Publisher: Nature Publishing Group.

[46] Anne P. G. Crijns, Rudolf S. N. Fehrmann, Steven de Jong, Frans Gerbens, Gert Jan Meersma, Harry G. Klip, Harry Hollema, Robert M. W. Hofstra, Gerard J. te Meerman, Elisabeth G. E. de Vries, and Ate G. J. van der Zee. Survival-Related Profile, Pathways, and Transcription Factors in Ovarian Cancer. PLOS Medicine, 6(2):e1000024, February 2009. Publisher: Public Library of Science.

[47] Adrian V. Dalca and Michael Brudno. Genome variation discovery with high-throughput sequencing data. Briefings in Bioinformatics, 11(1):3–14, January 2010.

[48] Charles G. Danko, Stephanie L. Hyland, Leighton J. Core, Andre L. Martins, Colin T. Waters, Hyung Won Lee, Vivian G. Cheung, W. Lee Kraus, John T. Lis, and Adam Siepel. Identification of active transcriptional regulatory elements from GRO-seq data. Nature Methods, 12(5):433–438, May 2015. Publisher: Nature Publishing Group.

[49] Katrina M. de Lange, Loukas Moutsianas, James C. Lee, Christopher A. Lamb, Yang Luo, Nicholas A. Kennedy, Luke Jostins, Daniel L. Rice, Javier Gutierrez-Achury, Sun-Gou Ji, Graham Heap, Elaine R. Nimmo, Cathryn Edwards, Paul Henderson, Craig Mowat, Jeremy Sanderson, Jack Satsangi, Alison Simmons, David C. Wilson, Mark Tremelling, Ailsa Hart, Christopher G. Mathew, William G. Newman, Miles Parkes, Charlie W. Lees, Holm Uhlig, Chris Hawkey, Natalie J. Prescott, Tariq Ahmad, John C. Mansfield, Carl A. Anderson, and Jeffrey C. Barrett. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nature Genetics, 49(2):256–261, February 2017.

[50] Joshua M. Dempster, Isabella Boyle, Francisca Vazquez, David E. Root, Jesse S. Boehm, William C. Hahn, Aviad Tsherniak, and James M. McFarland. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biology, 22(1):343, December 2021.

[51] Y. Peter Di, Jinming Zhao, and Richart Harper. Cigarette Smoke Induces MUC5AC Protein Expression through the Activation of Sp1*. Journal of Biological Chemistry, 287(33):27948–27958, August 2012.

[52] Qian Ding, Wei Hu, Ran Wang, Qinyan Yang, Menglin Zhu, Meng Li, Jianghong Cai, Peter Rose, Jianchun Mao, and Yi Zhun Zhu. Signaling pathways in rheumatoid arthritis: implications for targeted therapy. Signal Transduction and Targeted Therapy, 8(1):1–24, February 2023. Publisher: Nature Publishing Group.

[53] Tengteng Ding, Yuanbin Zhang, Zhixuan Ren, Ying Cong, Jingyi Long, Manli Peng, Oluwasijibomi Damola Faleti, Yinggui Yang, Xin Li, and Xiaoming Lyu. EBV-Associated Hub Genes as Potential Biomarkers for Predicting the Prognosis of Nasopharyngeal Carcinoma. Viruses, 15(9):1915, September 2023.

[54] Dany Doiron, Kees de Hoogh, Nicole Probst-Hensch, Isabel Fortier, Yutong Cai, Sara De Matteis, and Anna L. Hansell. Air pollution, lung function and COPD: results from the population-based UK Biobank study. European Respiratory Journal, 54(1), July 2019. Publisher: European Respiratory Society Section: Original Articles.

[55] Boryana Doyle, Geoffrey Fudenberg, Maxim Imakaev, and Leonid A. Mirny. Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. PLOS Computational Biology, 10(10):e1003867, October 2014. Publisher: Public Library of Science.

[56] Heather L. Drexler, Karine Choquet, and L. Stirling Churchman. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. Molecular cell, 77(5):985, March 2020. Publisher: NIH Public Access.

[57] Heather L. Drexler, Karine Choquet, Hope E. Merens, Paul S. Tang, Jared T. Simpson, and L. Stirling Churchman. Revealing nascent RNA processing dynamics with nano-COP. Nature Protocols, 16(3):1343–1375, March 2021. Number: 3 Publisher: Nature Publishing Group.

[58] Wencong Du, Haoyu Guan, Xinglin Wan, Zheng Zhu, Hao Yu, Pengfei Luo, Lulu Chen, Jian Su, Yan Lu, Dong Hang, Ran Tao, Ming Wu, Jinyi Zhou, and Xikang Fan. Circulating liver function markers and the risk of COPD in the UK Biobank. Frontiers in Endocrinology, 14:1121900, March 2023.

[59] Aurelien Dugourd, Christoph Kuppe, Marco Sciacovelli, Enio Gjerga, Attila Gabor, Kristina B. Emdal, Vitor Vieira, Dorte B. Bekker-Jensen, Jennifer Kranz, Eric.M.J. Bindels, Ana S.H. Costa, Abel Sousa, Pedro Beltrao, Miguel Rocha, Jesper V. Olsen, Christian Frezza, Rafael Kramann, and Julio Saez-Rodriguez. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Molecular Systems Biology, 17(1):e9730, January 2021. Publisher: John Wiley & Sons, Ltd.

[60] Noah Dukler, Gregory T. Booth, Yi-Fei Huang, Nathaniel Tippens, Colin T. Waters, Charles G. Danko, John T. Lis, and Adam Siepel. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Research, 27(11):1816–1829, November 2017. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[61] Leslie Dunipace, Zsuzsa Ákos, and Angelike Stathopoulos. Coacting enhancers can have complementary functions within gene regulatory networks and promote canalization. PLOS Genetics, 15(12):e1008525, December 2019. Publisher: Public Library of Science.

[62] Jaikumar Duraiswamy, Riccardo Turrini, Aspram Minasyan, David Barras, Isaac Crespo, Alizée J. Grimm, Julia Casado, Raphael Genolet, Fabrizio Benedetti, Alexandre Wicky, Kalliopi Ioannidou, Wilson Castro, Christopher Neal, Amandine Moriot, Stéphanie Renaud-Tissot, Victor Anstett, Noémie Fahr, Janos L. Tanyi, Monika A. Eiva, Connor A. Jacobson, Kathleen T. Montone, Marie Christine Wulff Westergaard, Inge Marie Svane, Lana E. Kandalaft, Mauro Delorenzi, Peter K. Sorger, Anniina Färkkilä, Olivier Michielin, Vincent Zoete, Santiago J. Carmona, Periklis G. Foukas, Daniel J. Powell, Sylvie Rusakiewicz, Marie-Agnès Doucey, Denarda Dangaj Laniti, and George Coukos. Myeloid antigen-presenting cell niches sustain antitumor T cells and license PD-1 blockade via CD28 costimulation. Cancer Cell, 39(12):1623–1642.e20, December 2021. Publisher: Elsevier.

[63] Silvia D'Alessio, Federica Ungaro, Daniele Noviello, Sara Lovisa, Laurent Peyrin-Biroulet, and Silvio Danese. Revisiting fibrosis in inflammatory bowel disease: the gut thickens. Nature Reviews Gastroenterology & Hepatology, 19(3):169–184, March 2022. Publisher: Nature Publishing Group.

[64] Yuchang Fei, Huan Yu, Yulun Wu, and Shanshan Gong. The causal relationship between immune cells and ankylosing spondylitis: a bidirectional Mendelian randomization study. Arthritis Research & Therapy, 26(1):24, January 2024.

[65] Hilary K. Finucane, Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, Giulio Genovese, Arpiar Saunders, Evan Macosko, Samuela Pollack, John R. B. Perry, Jason D. Buenrostro, Bradley E. Bernstein, Soumya Raychaudhuri, Steven McCarroll, Benjamin M. Neale, and Alkes L. Price. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature Genetics, 50(4):621–629, April 2018. Number: 4 Publisher: Nature Publishing Group.

[66] Kaitlin C. Fogg, Will R. Olson, Jamison N. Miller, Aisha Khan, Carine Renner, Isaac Hale, Paul S. Weisman, and Pamela K. Kreeger. Alternatively activated macrophage-derived secre-

tome stimulates ovarian cancer spheroid spreading through a JAK2/STAT3 pathway. Cancer Letters, 458:92–101, August 2019.

[67] Michael Fricker, Peter G. Gibson, Heather Powell, Jodie L. Simpson, Ian A. Yang, John W. Upham, Paul N. Reynolds, Sandra Hodge, Alan L. James, Christine Jenkins, Matthew J. Peters, Guy B. Marks, Melissa Baraket, and Katherine J. Baines. A sputum 6-gene signature predicts future exacerbations of poorly controlled asthma. Journal of Allergy and Clinical Immunology, 144(1):51–60.e11, July 2019.

[68] Fabienne Gally, Sarah K. Sasse, Jonathan S. Kurche, Margaret A. Gruca, Jonathan H. Cardwell, Tsukasa Okamoto, Hong W. Chu, Xiaomeng Hou, Olivier B. Poirion, Justin Buchanan, Sebastian Preissl, Bing Ren, Sean P. Colgan, Robin D. Dowell, Ivana V. Yang, David A. Schwartz, and Anthony N. Gerber. The MUC5B-associated variant rs35705950 resides within an enhancer subject to lineage- and disease-dependent epigenetic remodeling. JCI Insight, 6(2):e144294, 2021.

[69] Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal K. Dey, Joseph Nasser, Karthik A. Jagadeesh, Daniel J. Weiner, Huwenbo Shi, Charles P. Fulco, Luke J. O'Connor, Bogdan Pasaniuc, Jesse M. Engreitz, and Alkes L. Price. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nature Genetics, 54(6):827–836, June 2022. Number: 6 Publisher: Nature Publishing Group.

[70] Gary K. Geiss, Roger E. Bumgarner, Brian Birditt, Timothy Dahl, Naeem Dowidar, Dwayne L. Dunaway, H. Perry Fell, Sean Ferree, Renee D. George, Tammy Grogan, Jeffrey J. James, Malini Maysuria, Jeffrey D. Mitton, Paola Oliveri, Jennifer L. Osborn, Tao Peng, Amber L. Ratcliffe, Philippa J. Webster, Eric H. Davidson, Leroy Hood, and Krassen Dimitrov. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nature Biotechnology, 26(3):317–325, March 2008. Publisher: Nature Publishing Group.

[71] Andrew J. Gentles, Aaron M. Newman, Chih Long Liu, Scott V. Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S. Nair, Yue Xu, Amanda Khuong, Chuong D. Hoang, Maximilian Diehn, Robert B. West, Sylvia K. Plevritis, and Ash A. Alizadeh. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nature Medicine, 21(8):938–945, August 2015. Publisher: Nature Publishing Group.

[72] Dipender Gill, Nuala A. Sheehan, Matthias Wielscher, Nick Shrine, Andre F. S. Amaral, John R. Thompson, Raquel Granell, Bénédicte Leynaert, Francisco Gómez Real, Ian P. Hall, Martin D. Tobin, Juha Auvinen, Susan M. Ring, Marjo-Riitta Jarvelin, Louise V. Wain, John Henderson, Deborah Jarvis, and Cosetta Minelli. Age at menarche and lung function: a Mendelian randomization study. European Journal of Epidemiology, 32(8):701–710, August 2017.

[73] Paul Gontarz, Shuhua Fu, Xiaoyun Xing, Shaopeng Liu, Benpeng Miao, Viktoriia Bazylianska, Akhil Sharma, Pamela Madden, Kitra Cates, Andrew Yoo, Anna Moszczynska, Ting Wang, and Bo Zhang. Comparison of differential accessibility analysis strategies for ATAC-seq data. Scientific Reports, 10(1):10150, June 2020. Publisher: Nature Publishing Group.

[74] Eric Gracey, Yuchen Yao, Zoya Qaiyum, Melissa Lim, Michael Tang, and Robert D. Inman. Altered Cytotoxicity Profile of CD8+ T Cells in Ankylosing Spondylitis. Arthritis & Rheumatology, 72(3):428–434, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/art.41129.

[75] Lingli Guo, Tao Wang, Yanqiu Wu, Zhicheng Yuan, Jiajia Dong, Xiao'ou Li, Jing An, Zenglin Liao, Xue Zhang, Dan Xu, and Fu-Qiang Wen. WNT/β-catenin signaling regulates cigarette smoke-induced airway inflammation via the PPARδ/p38 pathway. Laboratory Investigation, 96(2):218–229, February 2016.

[76] Yang Guo and Zhiqiang Xiao. Constructing the dynamic transcriptional regulatory networks to identify phenotype-specific transcription regulators. Briefings in Bioinformatics, 25(6):bbae542, September 2024.

[77] Arnav Gupta, Amber Dahlin, Alejandra Macario, Fabienne Gally, Michael R. Weaver, Samuel Guarino, Louisa Kahn, Lynn Sanford, Margaret A. Gruca, Michael H. Cho, Robin D Dowell, Scott T. Weiss, Sarah K Sasse, and Anthony N Gerber. Functional Variant Discovery Identifies a Novel Genetic Link Between SPRY2, Wood Smoke, and Asthma. American Journal of Respiratory Cell and Molecular Biology, pages rcmb.2024–0587OC, September 2025.

[78] Arnav Gupta, Sarah K. Sasse, Reena Berman, Margaret A. Gruca, Robin D. Dowell, Hong Wei Chu, Gregory P. Downey, and Anthony N. Gerber. Integrated genomics approaches identify transcriptional mediators and epigenetic responses to Afghan desert particulate matter in small airway epithelial cells. Physiological Genomics, 54(10):389–401, October 2022. Publisher: American Physiological Society.

[79] Arnav Gupta, Sarah K. Sasse, Margaret A. Gruca, Lynn Sanford, Robin D. Dowell, and Anthony N. Gerber. Deconvolution of multiplexed transcriptional responses to wood smoke particles defines rapid aryl hydrocarbon receptor signaling dynamics. Journal of Biological Chemistry, 297(4), October 2021. Publisher: Elsevier.

[80] Balázs Győrffy. Discovery and ranking of the most robust prognostic biomarkers in serous ovarian cancer. GeroScience, 45(3):1889–1898, June 2023.

[81] Cecil Han, Yunhua Liu, Guohui Wan, Hyun Jin Choi, Luqing Zhao, Cristina Ivan, Xiaoming He, Anil K. Sood, Xinna Zhang, and Xiongbin Lu. The RNA-Binding Protein DDX1 Promotes Primary MicroRNA Maturation and Inhibits Ovarian Tumor Progression. Cell Reports, 8(5):1447–1460, September 2014. Publisher: Elsevier.

[82] Praveen Hariharan and Josée Dupuis. Mapping gene and gene pathways associated with coronary artery disease: a CARDIoGRAM exome and multi-ancestry UK biobank analysis. Scientific Reports, 11(1):16461, August 2021. Number: 1 Publisher: Nature Publishing Group.

[83] Uwe Hassler and Thorsten Thadewald. Nonsensical and biased correlation due to pooling heterogeneous samples. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3):367–379, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9884.00365.

[84] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38(4):576–589, May 2010.

[85] Jung Won Heo, Hwa Young Lee, Solji Han, Hye Seon Kang, Soon Seog Kwon, and Sook Young Lee. The association between serum apolipoprotein B and fractional exhaled nitric oxide in bronchial asthma patients. Journal of Thoracic Disease, 13(7):4195–4206, July 2021.

[86] Antony Hoarau, Myriam Polette, and Christelle Coraux. Lung Hyaluronasome: Involvement of Low Molecular Weight Ha (Lmw-Ha) in Innate Immunity. Biomolecules, 12(5):658, May 2022. Publisher: Multidisciplinary Digital Publishing Institute.

[87] Inga Hoffmann, Mihnea P. Dragomir, Nanna Monjé, Carlotta Keunecke, Catarina Alisa Kunze, Simon Schallenberg, Sofya Marchenko, Wolfgang D. Schmitt, Hagen Kulbe, Jalid Sehouli, Ioana Elena Braicu, Paul Jank, Carsten Denkert, Silvia Darb-Esfahani, David Horst, Bruno V. Sinn, Christine Sers, Philip Bischoff, and Eliane T. Taube. Increased expression of IDO1 is associated with improved survival and increased number of TILs in patients with high-grade serous ovarian cancer. Neoplasia, 44:100934, October 2023.

[88] Nathalie Hoffmann. Tumor-associated Natural Killer cells in ovarian cancer ascites: Molecular and functional characterization. PhD thesis, Philipps University of Marburg, September 2020. Publisher: Philipps-Universität Marburg.

[89] Tim Y. Hou and W. Lee Kraus. Analysis of estrogen-regulated enhancer RNAs identifies a functional motif required for enhancer assembly and gene expression. Cell Reports, 39(11), June 2022. Publisher: Elsevier.

[90] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C. Hicks. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biology, 21(1):218, August 2020.

[91] Sheng Hu, Qiang Guo, Silin Wang, Wenxiong Zhang, Jiayue Ye, Lang Su, Sheng Zou, Deyuan Zhang, Yang Zhang, Dongliang Yu, Jianjun Xu, and Yiping Wei. Supplementation of serum albumin is associated with improved pulmonary function: NHANES 2013–2014. Frontiers in Physiology, 13, October 2022. Publisher: Frontiers.

[92] Samuel Hunter, Rutendo F. Sigauke, Jacob T. Stanley, Mary A. Allen, and Robin D. Dowell. Protocol variations in run-on transcription dataset preparation produce detectable signatures in sequencing libraries. BMC Genomics, 23(1):187, March 2022.

[93] Xiao Huo, Hengzi Sun, Shuangwu Liu, Bing Liang, Huimin Bai, Shuzhen Wang, and Shuhong Li. Identification of a Prognostic Signature for Ovarian Cancer Based on the Microenvironment Genes. Frontiers in Genetics, 12, May 2021. Publisher: Frontiers.

[94] BROAD Institute. DepMap 25Q3, 2025.

[95] Kazuyoshi Ishigaki, Saori Sakaue, Chikashi Terao, Yang Luo, Kyuto Sonehara, Kensuke Yamaguchi, Tiffany Amariuta, Chun Lai Too, Vincent A. Laufer, Ian C. Scott, Sebastien Viatte, Meiko Takahashi, Koichiro Ohmura, Akira Murasawa, Motomu Hashimoto, Hiromu Ito, Mohammed Hammoudeh, Samar Al Emadi, Basel K. Masri, Hussein Halabi, Humeira Badsha, Imad W. Uthman, Xin Wu, Li Lin, Ting Li, Darren Plant, Anne Barton, Gisela Orozco, Suzanne M. M. Verstappen, John Bowes, Alexander J. MacGregor, Suguru Honda, Masaru Koido, Kohei Tomizuka, Yoichiro Kamatani, Hiroaki Tanaka, Eiichi Tanaka, Akari Suzuki, Yuichi Maeda, Kenichi Yamamoto, Satoru Miyawaki, Gang Xie, Jinyi Zhang, Christopher I. Amos, Edward Keystone, Gertjan Wolbink, Irene van der Horst-Bruinsma, Jing Cui,

Katherine P. Liao, Robert J. Carroll, Hye-Soon Lee, So-Young Bang, Katherine A. Siminovitch, Niek de Vries, Lars Alfredsson, Solbritt Rantapää-Dahlqvist, Elizabeth W. Karlson, Sang-Cheol Bae, Robert P. Kimberly, Jeffrey C. Edberg, Xavier Mariette, Tom Huizinga, Philippe Dieudé, Matthias Schneider, Martin Kerick, Joshua C. Denny, Koichi Matsuda, Keitaro Matsuo, Tsuneyo Mimori, Fumihiko Matsuda, Keishi Fujio, Yoshiya Tanaka, Atsushi Kumanogoh, Matthew Traylor, Cathryn M. Lewis, Stephen Eyre, Huji Xu, Richa Saxena, Thurayya Arayssi, Yuta Kochi, Katsunori Ikari, Masayoshi Harigai, Peter K. Gregersen, Kazuhiko Yamamoto, S. Louis Bridges, Leonid Padyukov, Javier Martin, Lars Klareskog, Yukinori Okada, and Soumya Raychaudhuri. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. Nature Genetics, 54(11):1640–1651, November 2022. Number: 11 Publisher: Nature Publishing Group.

[96] Natalia Jaeger, Ramya Gamini, Marina Cella, Jorge L. Schettini, Mattia Bugatti, Shanrong Zhao, Charles V. Rosadini, Ekaterina Esaulova, Blanda Di Luccia, Baylee Kinnett, William Vermi, Maxim N. Artyomov, Thomas A. Wynn, Ramnik J. Xavier, Scott A. Jelinsky, and Marco Colonna. Single-cell analyses of Crohn's disease tissues reveal intestinal intraepithelial T cells heterogeneity and altered subset distributions. Nature Communications, 12(1):1921, March 2021. Number: 1 Publisher: Nature Publishing Group.

[97] Karthik A. Jagadeesh, Kushal K. Dey, Daniel T. Montoro, Rahul Mohan, Steven Gazal, Jesse M. Engreitz, Ramnik J. Xavier, Alkes L. Price, and Aviv Regev. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nature Genetics, 54(10):1479–1492, October 2022. Number: 10 Publisher: Nature Publishing Group.

[98] Mariam Jamal-Hanjani, Sergio A. Quezada, James Larkin, and Charles Swanton. Translational Implications of Tumor Heterogeneity. Clinical cancer research : an official journal of the American Association for Cancer Research, 21(6):1258–1266, March 2015.

[99] Nicole E. James, Morgan Woodman, Payton De La Cruz, Katrin Eurich, Melih Arda Ozsoy, Christoph Schorl, Linda C. Hanley, and Jennifer R. Ribeiro. Adaptive transcriptomic and immune infiltrate responses in the tumor immune microenvironment following neoadjuvant chemotherapy in high grade serous ovarian cancer reveal novel prognostic associations and activation of pro-tumorigenic pathways. Frontiers in Immunology, 13, September 2022. Publisher: Frontiers.

[100] Melissa Javellana, Mark A. Eckert, Janna Heide, Katarzyna Zawieracz, Melanie Weigert, Sarah Ashley, Elizabeth Stock, David Chapel, Lei Huang, S. Diane Yamada, Ahmed Ashour Ahmed, Ricardo R. Lastra, Mengjie Chen, and Ernst Lengyel. Neoadjuvant Chemotherapy Induces Genomic and Transcriptomic Changes in Ovarian Cancer. Cancer Research, 82(1):169–176, January 2022.

[101] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M. Garske, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature Communications, 11(1):1971, April 2020. Number: 1 Publisher: Nature Publishing Group.

[102] Peilin Jia, Ruifeng Hu, Fangfang Yan, Yulin Dai, and Zhongming Zhao. scGWAS: landscape of trait-cell type associations by integrating single-cell transcriptomics-wide and genome-wide association studies. Genome Biology, 23(1):220, October 2022.

[103] Longda Jiang, Zhili Zheng, Hailing Fang, and Jian Yang. A generalized linear mixed model association tool for biobank-scale data. Nature Genetics, 53(11):1616–1621, November 2021. Number: 11 Publisher: Nature Publishing Group.

[104] Minghui Jiang, Xingjie Hao, Yi Jiang, Si Li, Chaolong Wang, and Shanshan Cheng. Genetic and observational associations of lung function with gastrointestinal tract diseases: pleiotropic and mendelian randomization analysis. Respiratory Research, 24(1):315, December 2023.

[105] Yinan Jiang, Shuting Huang, Lan Zhang, Yun Zhou, Wei Zhang, Ting Wan, Haifeng Gu, Yi Ouyang, Xiaojing Zheng, Pingping Liu, Baoyue Pan, Huiling Xiang, Mingxiu Ju, Rongzhen Luo, Weihua Jia, Shenjiao Huang, Jundong Li, and Min Zheng. Targeting the Cdc2-like kinase 2 for overcoming platinum resistance in ovarian cancer. MedComm, 5(4):e537, 2024. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mco2.537.

[106] Alejandro Jiménez-Sánchez, Paulina Cybulska, Katherine LaVigne Mager, Simon Koplev, Oliver Cast, Dominique-Laurent Couturier, Danish Memon, Pier Selenica, Ines Nikolovski, Yousef Mazaheri, Yonina Bykov, Felipe C. Geyer, Geoff Macintyre, Lena Morrill Gavarró, Ruben M. Drews, Michael B. Gill, Anastasios D. Papanastasiou, Ramon E. Sosa, Robert A. Soslow, Tyler Walther, Ronglai Shen, Dennis S. Chi, Kay J. Park, Travis Hollmann, Jorge S. Reis-Filho, Florian Markowetz, Pedro Beltrao, Hebert Alberto Vargas, Dmitriy Zamarin, James D. Brenton, Alexandra Snyder, Britta Weigelt, Evis Sala, and Martin L. Miller. Unraveling tumor–immune heterogeneity in advanced ovarian cancer uncovers immunogenic effect of chemotherapy. Nature Genetics, 52(6):582–593, June 2020. Publisher: Nature Publishing Group.

[107] Pauline Johnson, Arif A. Arif, Sally S. M. Lee-Sayer, and Yifei Dong. Hyaluronan and Its Interactions With Immune Cells in the Healthy and Inflamed Lung. Frontiers in Immunology, 9, November 2018. Publisher: Frontiers.

[108] Taylor Jones, Rutendo F. Sigauke, Lynn Sanford, Dylan J. Taatjes, Mary A. Allen, and Robin D. Dowell. TF Profiler: a transcription factor inference method that broadly measures transcription factor activity and identifies mechanistically distinct networks. Genome Biology, 26(1):92, April 2025.

[109] Kimberly R. Jordan, Matthew J. Sikora, Jill E. Slansky, Angela Minic, Jennifer K. Richer, Marisa R. Moroney, Junxiao Hu, Rebecca J. Wolsky, Zachary L. Watson, Tomomi M. Yamamoto, James C. Costello, Aaron Clauset, Kian Behbakht, T. Rajendra Kumar, and Benjamin G. Bitler. The Capacity of the Ovarian Cancer Tumor Microenvironment to Integrate Inflammation Signaling Conveys a Shorter Disease-free Interval. Clinical Cancer Research, 26(23):6362–6373, December 2020.

[110] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N Ananthakrishnan, Vibeke Andersen, Jane M Andrews, Leonard

Baidoo, Tobias Balschun, Peter A Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C Lawrance, Charlie W Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y Ponsioen, Uros Potocnik, Natalie J Prescott, Miguel Regueiro, Jerome I Rotter, Richard K Russell, Jeremy D Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A Simms, Jurgita Sventoraityte, Stephan R Targan, Kent D Taylor, Mark Tremelling, Hein W Verspaget, Martine De Vos, Cisca Wijmenga, David C Wilson, Juliane Winkelmann, Ramnik J Xavier, Sebastian Zeissig, Bin Zhang, Clarence K Zhang, Hongyu Zhao, Mark S Silverberg, Vito Annese, Hakon Hakonarson, Steven R Brant, Graham Radford-Smith, Christopher G Mathew, John D Rioux, Eric E Schadt, Mark J Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C Barrett, and Judy H Cho. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature, 491(7422):119–124, November 2012.

[111] Jenny Juschten, Sarah A. Ingelse, Lieuwe D. J. Bos, Armand R. J. Girbes, Nicole P. Juffermans, Tom van der Poll, Marcus J. Schultz, Pieter Roel Tuinman, F. M. de Beer, L. D. Bos, T. A. Claushuis, G. J. Glas, J. Horn, A. J. Hoogendijk, R. T. van Hooijdonk, M. A. Huson, M. D. de Jong, N. P. Juffermans, W. K. Lagrand, T. van der Poll, B. Scicluna, L. R. Schouten, M. J. Schultz, K. F. van der Sluijs, M. Straat, L. A. van Vught, L. Wieske, M. A. Wiewel, E. Witteveen, and for the BASIC study investigators. Alkaline phosphatase in pulmonary inflammation—a translational study in ventilated critically ill patients and rats. Intensive Care Medicine Experimental, 8(1):46, December 2020.

[112] Manale El Kharbili, Sarah K. Sasse, Lynn Sanford, Sean Jacobson, Katja Aviszus, Arnav Gupta, Claire Guo, Susan M. Majka, Robin D. Dowell, Anthony N. Gerber, Russell P. Bowler, and Fabienne Gally. Noncoding SNPs decrease expression of FABP5 during COPD exacerbations. The Journal of Clinical Investigation, 134(3), February 2024. Publisher: American Society for Clinical Investigation.

[113] Nayoung Kim, Huiram Kang, Areum Jo, Seung-Ah Yoo, and Hae-Ock Lee. Perspectives on single-nucleus RNA sequencing in different cell types and tissues. Journal of Pathology and Translational Medicine, 57(1):52–59, January 2023. Publisher: The Korean Society of Pathologists/The Korean Society for Cytopathology.

[114] Tae-Kyung Kim and Ramin (second) Shiekhattar. Architectural and Functional Commonalities between Enhancers and Promoters. Cell, 162(5):948–959, August 2015. Publisher: Cell Press.

[115] Young Soo Kim, Hee Yeon Kim, Hyo-Suk Ahn, Tae Seo Sohn, Jae Yen Song, Young Bok Lee, Dong-Hee Lee, Jae-Im Lee, Tae-Kyu Lee, Seong Cheol Jeong, Mihee Hong, Hiun Suk Chae, Kyungdo Han, and Chang Dong Yeo. Glomerular filtration rate affects interpretation of pulmonary function test in a Korean general population: results from the Korea National Health and Nutrition Examination Survey 2010 to 2012. The Korean Journal of Internal Medicine, 31(6):1101–1109, March 2016. Publisher: The Korean Association of Internal Medicine.

[116] Sebastian Kjærgaard, Thorbjørn S.R. Jensen, Ulrike R. Feddersen, Niels Bindslev, Kaare V. Grunddal, Steen S. Poulsen, Hanne B. Rasmussen, Esben Budtz-Jørgensen, and Mark Berner-Hansen. Decreased number of colonic tuft cells in quiescent ulcerative colitis patients. European Journal of Gastroenterology & Hepatology, 33(6):817–824, June 2021.

[117] Steven R. Kleeberger and David Peden. Gene-Environment Interactions in Asthma and Other Respiratory Diseases. Annual Review of Medicine, 56(1):383–400, February 2005.

[118] Yifat Koren Carmi, Hazem Khamaisi, Rina Adawi, Eden Noyman, Jacob Gopas, and Jamal Mahajna. Secreted Soluble Factors from Tumor-Activated Mesenchymal Stromal Cells Confer Platinum Chemoresistance to Ovarian Cancer Cells. International Journal of Molecular Sciences, 24(9):7730, January 2023. Publisher: Multidisciplinary Digital Publishing Institute.

[119] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods, 16(12):1289–1296, December 2019. Publisher: Nature Publishing Group.

[120] Katla Kristjánsdóttir, Alexis Dziubek, Hyun Min Kang, and Hojoong Kwak. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. Nature Communications, 11(1):5963, November 2020. Number: 1 Publisher: Nature Publishing Group.

[121] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. Nature, 518(7539):317–330, February 2015. Publisher: Nature Publishing Group.

[122] Charles Kunos, Tomas Radivoyevitch, Fadi W. Abdul-Karim, James Fanning, Ovadia Abulafia, Albert J. Bonebrake, and Lydia Usha. Ribonucleotide reductase inhibition restores platinum-sensitivity in platinum-resistant ovarian cancer: a Gynecologic Oncology Group Study. Journal of Translational Medicine, 10(1):79, April 2012.

[123] David Kuo, Jennifer Ding, Ian S. Cohn, Fan Zhang, Kevin Wei, Deepak A. Rao, Cristina Rozo, Upneet K. Sokhi, Sara Shanaj, David J. Oliver, Adriana P. Echeverria, Edward F. DiCarlo, Michael B. Brenner, Vivian P. Bykerk, Susan M. Goodman, Soumya Raychaudhuri, Gunnar Rätsch, Lionel B. Ivashkiv, and Laura T. Donlin. HBEGF+ macrophages in rheumatoid arthritis induce fibroblast invasiveness. Science Translational Medicine, 11(491):eaau8587, May 2019. Publisher: American Association for the Advancement of Science.

[124] Kevser Kübra Kırboğa, Ecir Uğur Küçüksille, Mithun Rudrapal, Emre Aktaş, Raghu Ram Achar, Gouri Deshpande, Victor Stupin, and Ekaterina Silina. Tumor Microenvironment in Ovarian Cancer through Spatial Transcriptomics and Identification of Key Gene Expression Profiles, April 2025. Pages: 2025.04.25.650590 Section: New Results.

[125] Sarah E. Lacher, Tessa Schumann, Ryan Peters, Christopher Migliaccio, Andrij Holian, and Matthew Slattery. NRF2 protects lung epithelial cells from wood smoke particle toxicity. Advances in Redox Research, 13:100115, December 2024.

[126] Alexandra Lahtinen, Kari Lavikka, Anni Virtanen, Yilin Li, Sanaz Jamalzadeh, Aikaterini Skorda, Anna Røssberg Lauridsen, Kaiyang Zhang, Giovanni Marchi, Veli-Matti Isoviita, Valeria Ariotta, Oskari Lehtonen, Taru A. Muranen, Kaisa Huhtinen, Olli Carpén, Sakari Hietanen, Wojciech Senkowski, Tuula Kallunki, Antti Häkkinen, Johanna Hynninen, Jaana Oikkonen, and Sampsa Hautaniemi. Evolutionary states and trajectories characterized by distinct pathways stratify patients with ovarian high grade serous carcinoma. Cancer Cell, 41(6):1103–1117.e12, June 2023. Publisher: Elsevier.

[127] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1):D1062–D1067, January 2018.

[128] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1):559, December 2008.

[129] Inga-Maria Launonen, Iga Niemiec, María Hincapié-Otero, Erdogan Pekcan Erkan, Ada Junquera, Daria Afenteva, Matias M. Falco, Zhihan Liang, Matilda Salko, Foteini Chamchougia, Angela Szabo, Fernando Perez-Villatoro, Yilin Li, Giulia Micoli, Ashwini Nagaraj, Ulla-Maija Haltia, Essi Kahelin, Jaana Oikkonen, Johanna Hynninen, Anni Virtanen, Ajit J. Nirmal, Tuulia Vallius, Sampsa Hautaniemi, Peter K. Sorger, Anna Vähärautio, and Anniina Färkkilä. Chemotherapy induces myeloid-driven spatially confined T-cell exhaustion in ovarian cancer. Cancer Cell, 42(12):2045–2063.e10, December 2024.

[130] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology, 15(2):R29, February 2014.

[131] Hyewon Lee, Jung Yoon Ho, In Sun Hwang, and Youn Jin Choi. Indoleamine 2,3-dioxygenase 1 inhibition reverses cancer-associated fibroblast-mediated immunosuppression in high-grade serous ovarian cancer. FEBS Open Bio, n/a(n/a), October 2025. _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/2211-5463.70126.

[132] Seungha Alisa Lee, Katla Kristjánsdóttir, and Hojoong Kwak. eRNA co-expression network uncovers TF dependency and convergent cooperativity. Scientific Reports, 13(1):19085, November 2023. Publisher: Nature Publishing Group.

[133] Yong-Jae Lee, Eun-Ji Nam, Sunghoon Kim, Young-Tae Kim, Pamela Itkin-Ansari, and Sang-Wun Kim. Expression Profiles of ID and E2A in Ovarian Cancer and Suppression of Ovarian Cancer by the E2A Isoform E47. Cancers, 14(12):2903, January 2022. Publisher: Multidisciplinary Digital Publishing Institute.

[134] Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Computational Biology, 11(4):e1004219, April 2015. Publisher: Public Library of Science.

[135] Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, Soohwan Oh, Hong-Sook Kim, Christopher K. Glass, and Michael G. Rosenfeld. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature, 498(7455):516–520, June 2013. Number: 7455 Publisher: Nature Publishing Group.

[136] Xiao-fei Li, Hai-yan Sun, Tian Hua, Hai-bo Zhang, Yun-jie Tian, Yan Li, and Shan Kang. Promoter Methylation of the MGRN1 Gene Predicts Prognosis and Response to Chemotherapy of High-Grade Serous Ovarian Cancer Patients. Frontiers in Oncology, 11, June 2021. Publisher: Frontiers.

[137] Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li, and Jingyi Jessica Li. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biology, 23(1):79, March 2022.

[138] Katja Lidschreiber, Lisa A. Jung, Henrik von der Emde, Kashyap Dave, Jussi Taipale, Patrick Cramer, and Michael Lidschreiber. Transcriptionally active enhancers in human cancer cells. Molecular Systems Biology, 17(1):MSB20209873, January 2021.

[139] Bo Lin, Shunji Wang, Youdan Yao, Yuehong Shen, and Hongyu Yang. Comprehensive co-expression analysis reveals TMC8 as a prognostic immune-associated gene in head and neck squamous cancer. Oncology Letters, 22(1):1–15, July 2021. Publisher: Spandidos Publications.

[140] Wei-De Lin, Wen-Ling Liao, Wei-Cheng Chen, Ting-Yuan Liu, Yu-Chia Chen, and Fuu-Jen Tsai. Genome-wide association study identifies novel susceptible loci and evaluation of polygenic risk score for chronic obstructive pulmonary disease in a Taiwanese population. BMC Genomics, 25(1):607, June 2024.

[141] Jing Liu, Yulong Tang, Yan Huang, Jian Gao, Shuai Jiang, Qingmei Liu, Yanyun Ma, Xiaolin Qian, Feng Qian, John D. Reveille, Dongyi He, Hejian Zou, Li Jin, Qi Zhu, Weilin Pu, and Jiucun Wang. Single-cell analysis reveals innate immunity dynamics in ankylosing spondylitis. Clinical and Translational Medicine, 11(3):e369, March 2021.

[142] Lei Liu, Cong Li, Honghua Yu, and Xiaohong Yang. A critical review on air pollutant exposure and age-related macular degeneration. Science of The Total Environment, 840:156717, September 2022.

[143] Qi Liu, Jiali Weng, Chenfei Li, Yi Feng, Meiqin Xie, Xiaohui Wang, Qing Chang, Mengnan Li, Kian Fan Chung, Ian M Adcock, Yan Huang, Hai Zhang, and Feng Li. Attenuation of PM2.5-induced alveolar epithelial cells and lung injury through regulation of mitochondrial fission and fusion. Particle and Fibre Toxicology, 20:28, July 2023.

[144] Qiao Liu, Biao Wu, Ruijie Xie, Yuling Luo, Du Zheng, Guang Liu, and Huihai Zhang. Association between serum albumin and pulmonary function in adolescents: analyses of NHANES 2007–2012. BMC Pulmonary Medicine, 24(1):554, November 2024.

[145] Marina Lizio, Jayson Harshbarger, Hisashi Shimoji, Jessica Severin, Takeya Kasukawa, Serkan Sahin, Imad Abugessaisa, Shiro Fukuda, Fumi Hori, Sachi Ishikawa-Kato, Christopher J. Mungall, Erik Arner, J. Kenneth Baillie, Nicolas Bertin, Hidemasa Bono, Michiel de Hoon, Alexander D. Diehl, Emmanuel Dimont, Tom C. Freeman, Kaori Fujieda, Winston Hide, Rajaram Kaliyaperumal, Toshiaki Katayama, Timo Lassmann, Terrence F. Meehan, Koro Nishikata, Hiromasa Ono, Michael Rehli, Albin Sandelin, Erik A. Schultes, Peter AC 't Hoen, Zuotian Tatum, Mark Thompson, Tetsuro Toyoda, Derek W. Wright, Carsten O. Daub, Masayoshi Itoh, Piero Carninci, Yoshihide Hayashizaki, Alistair RR Forrest, Hideya Kawaji, and the FANTOM consortium. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology, 16(1):22, January 2015.

[146] David Llères, Andres Cardozo Gizzi, and Marcelo Nollmann. Redefining enhancer action: Insights from structural, genomic, and single-molecule perspectives. Current Opinion in Cell Biology, 95:102527, August 2025.

[147] I. Lodewijk, A. Bernardini, C. Suárez-Cabrera, E. Bernal, R. Sánchez, J. L. Garcia, K. Rojas, L. Morales, S. Wang, X. Han, M. Dueñas, J. M. Paramio, and L. Manso. Genomic landscape and immune-related gene expression profiling of epithelial ovarian cancer after neoadjuvant chemotherapy. npj Precision Oncology, 6(1):7, January 2022. Publisher: Nature Publishing Group.

[148] Rebecca Ting Jiin Loo, Lukas Pavelka, Graziella Mangone, Fouad Khoury, Marie Vidailhet, Jean-Christophe Corvol, and Enrico Glaab. Multi-cohort machine learning identifies predictors of cognitive impairment in Parkinson's disease. npj Digital Medicine, 8(1):482, July 2025. Publisher: Nature Publishing Group.

[149] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12):550, December 2014.

[150] Yipin Lv, Yongliang Zhao, Xianhua Wang, Na Chen, Fangyuan Mao, Yongsheng Teng, Tingting Wang, Liusheng Peng, Jinyu Zhang, Ping Cheng, Yugang Liu, Hui Kong, Weisan Chen, Chuanjie Hao, Bin Han, Qiang Ma, Quanming Zou, Jun Chen, and Yuan Zhuang. Increased intratumoral mast cells foster immune suppression and gastric cancer progression through TNF-α-PD-L1 pathway. Journal for ImmunoTherapy of Cancer, 7(1):54, February 2019.

[151] Yunlong Ma, Chunyu Deng, Yijun Zhou, Yaru Zhang, Fei Qiu, Dingping Jiang, Gongwei Zheng, Jingjing Li, Jianwei Shuai, Yan Zhang, Jian Yang, and Jianzhong Su. Polygenic regression uncovers trait-relevant cellular contexts through pathway activation transformation of single-cell RNA sequencing data. Cell Genomics, 3(9):100383, September 2023. Publisher: Elsevier.

[152] Zachary Maas, Rutendo Sigauke, and Robin Dowell. Deconvolution of Nascent Sequencing Data Using Transcriptional Regulatory Elements, October 2023. Pages: 2023.10.11.561942 Section: New Results.

[153] Zachary L Maas and Robin D Dowell. Internal and external normalization of nascent RNA sequencing run-on experiments. BMC Bioinformatics, 25(1):19, Jan 2024.

[154] Dig B. Mahat, Nathaniel D. Tippens, Jorge D. Martin-Rufino, Sean K. Waterton, Jiayu Fu, Sarah E. Blatt, and Phillip A. Sharp. Single-cell nascent RNA sequencing unveils coordinated global transcription. Nature, 631(8019):216–223, July 2024. Publisher: Nature Publishing Group.

[155] Dig Bijay Mahat, Hojoong Kwak, Gregory T. Booth, Iris H. Jonkers, Charles G. Danko, Ravi K. Patel, Colin T. Waters, Katie Munson, Leighton J. Core, and John T. Lis. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nature Protocols, 11(8):1455–1476, August 2016. Publisher: Nature Publishing Group.

[156] Pacharla Manasa, Chirukandath Sidhanth, Syama Krishnapriya, Sekar Vasudevan, and Trivadi S. Ganesan. Oncogenes in high grade serous adenocarcinoma of the ovary. Genes & Cancer, 11(3-4):122–136, November 2020.

[157] Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health, 8:14, February 2020.

[158] Carolien Mathyssen, Jef Serré, Annelore Sacreas, Stephanie Everaerts, Karen Maes, Stijn Verleden, Lieve Verlinden, Annemieke Verstuyf, Charles Pilette, Ghislaine Gayan-Ramirez, Bart Vanaudenaerde, and Wim Janssens. Vitamin D Modulates the Response of Bronchial Epithelial Cells Exposed to Cigarette Smoke Extract. Nutrients, 11(9):2138, September 2019. Publisher: Multidisciplinary Digital Publishing Institute.

[159] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science, 337(6099):1190–1195, September 2012. Publisher: American Association for the Advancement of Science.

[160] James M. McFarland, Zandra V. Ho, Guillaume Kugener, Joshua M. Dempster, Phillip G. Montgomery, Jordan G. Bryan, John M. Krill-Burger, Thomas M. Green, Francisca Vazquez, Jesse S. Boehm, Todd R. Golub, William C. Hahn, David E. Root, and Aviad Tsherniak. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nature Communications, 9(1):4610, November 2018. Publisher: Nature Publishing Group.

[161] Carlos A. Melo, Jarno Drost, Patrick J. Wijchers, Harmen van de Werken, Elzo de Wit, Joachim A. F. Oude Vrielink, Ran Elkon, Sónia A. Melo, Nicolas Léveillé, Raghu Kalluri, Wouter de Laat, and Reuven Agami. eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. Molecular Cell, 49(3):524–535, February 2013.

[162] Samuel C. Mok, Tomas Bonome, Vinod Vathipadiekal, Aaron Bell, Michael E. Johnson, Kwong-Kwok Wong, Dong-Choon Park, Ke Hao, Daniel K. P. Yip, Howard Donninger, Laurent Ozbun, Goli Samimi, John Brady, Mike Randonovich, Cindy A. Pise-Masison, J. Carl Barrett, Wing H. Wong, William R. Welch, Ross S. Berkowitz, and Michael J. Birrer. A Gene Signature Predictive for Outcome in Advanced Ovarian Cancer Identifies a Survival Factor: Microfibril-Associated Glycoprotein 2. Cancer Cell, 16(6):521–532, December 2009. Publisher: Elsevier.

[163] Jonathan Moody, Tsukasa Kouno, Miki Kojima, Ikuko Koya, Julio Leon, Akari Suzuki, Akira Hasegawa, Taishin Akiyama, Nobuko Akiyama, Masayuki Amagai, Jen-Chien Chang, Ayano Fukushima-Nomura, Mika Handa, Kazunori Hino, Mizuki Hino, Tomoko Hirata, Yuuki Imai, Kazunori Inoue, Hiroshi Kawasaki, Toshihiro Kimura, Tomofumi Kinoshita, Ken-ichiro Kubo, Yasuto Kunii, Fernando López-Redondo, Riichiro Manabe, Tomohiro Miyai, Satoru Morimoto, Atsuko Nagaoka, Jun Nakajima, Shohei Noma, Yasushi Okazaki, Kokoro Ozaki, Noritaka Saeki, Hiroshi Sakai, Kuniaki Seyama, Youtaro Shibayama, Tomohisa Sujino, Michihira Tagami, Hayato Takahashi, Masaki Takao, Masaru Takeshita, Tsuyoshi Takiuchi, Chikashi Terao, Chi Wai Yip, Satoshi Yoshinaga, Hideyuki Okano, Kazuhiko Yahamoto, Takeya Kasukawa, Yoshinari Ando, Piero Carninci, Jay W. Shin, and Chung-Chau Hon. A single-cell atlas of transcribed cis-regulatory elements in the human genome, November 2023. Pages: 2023.11.13.566791 Section: New Results.

[164] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583(7818):699–710, July 2020. Number: 7818 Publisher: Nature Publishing Group.

[165] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 34(3):267–273, July 2003. Number: 3 Publisher: Nature Publishing Group.

[166] Samuel Morabito, Fairlie Reese, Negin Rahimzadeh, Emily Miyoshi, and Vivek Swarup. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Reports Methods, 3(6), June 2023. Publisher: Elsevier.

[167] Margaret Morash, Hannah Mitchell, Himisha Beltran, Olivier Elemento, and Jyotishman Pathak. The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. Journal of Personalized Medicine, 8(3), September 2018. Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher.

[168] C. Morgan, M. Lunt, H. Brightwell, P. Bradburn, W. Fallow, M. Lay, A. Silman, and I. N. Bruce. Contribution of patient related differences to multidrug resistance in rheumatoid arthritis. Annals of the Rheumatic Diseases, 62(1):15–19, January 2003. Publisher: BMJ Publishing Group Ltd Section: Extended report.

[169] Alexander H. Morrison, Mark S. Diamond, Ceire A. Hay, Katelyn T. Byrne, and Robert H. Vonderheide. Sufficiency of CD40 activation and immune checkpoint blockade for T cell priming and tumor immunity. Proceedings of the National Academy of Sciences, 117(14):8022–8031, April 2020. Publisher: Proceedings of the National Academy of Sciences.

[170] Hakhamanesh Mostafavi, Jeffrey P. Spence, Sahin Naqvi, and Jonathan K. Pritchard. Systematic differences in discovery of genetic effects on gene expression and complex traits. Nature Genetics, 55(11):1866–1875, November 2023. Publisher: Nature Publishing Group.

[171] Kambiz Mousavi, Hossein Zare, Stefania Dell'Orso, Lars Grontved, Gustavo Gutierrez-Cruz, Assia Derfoul, Gordon L. Hager, and Vittorio Sartorelli. eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. Molecular Cell, 51(5):606–617, September 2013. Publisher: Elsevier.

[172] Method of the Year 2019: Single-cell multimodal omics, January 2020. Number: 1 Publisher: Nature Publishing Group.

[173] Yu-Mei Ning, Kun Lin, Xiao-Ping Liu, Yang Ding, Xiang Jiang, Zhang Zhang, Yu-Ting Xuan, Li Dong, Lan Liu, Fan Wang, Qiu Zhao, Hai-Zhou Wang, and Jun Fang. NAPSB as a predictive marker for prognosis and therapy associated with an immuno-hot tumor microenvironment in hepatocellular carcinoma. BMC Gastroenterology, 22(1):392, August 2022.

[174] Na Niu, Weiwei Shen, Yanping Zhong, Robert C. Bast, Amir Jazaeri, Anil K. Sood, and Jinsong Liu. Expression of B7–H4 and IDO1 is associated with drug resistance and poor prognosis in high-grade serous ovarian carcinomas. Human Pathology, 113:20–27, July 2021.

[175] David Ochoa, Mohd Karim, Maya Ghoussaini, David G. Hulcoop, Ellen M. McDonagh, and Ian Dunham. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nature Reviews Drug Discovery, 21(8):551–551, July 2022. Bandiera_abtest: a Cg_type: Biobusiness Briefs Publisher: Nature Publishing Group.

[176] The All of Us Research Program Investigators. The "All of Us" Research Program. New England Journal of Medicine, 381(7):668–676, August 2019. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMsr1809937.

[177] Ozgur Oksuz, Jonathan E. Henninger, Robert Warneford-Thomson, Ming M. Zheng, Hailey Erb, Adrienne Vancura, Kalon J. Overholt, Susana Wilson Hawken, Salman F. Banani, Richard Lauman, Lauren N. Reich, Anne L. Robertson, Nancy M. Hannett, Tong I. Lee, Leonard I. Zon, Roberto Bonasio, and Richard A. Young. Transcription factors interact with RNA to regulate genes. Molecular Cell, 83(14):2449–2463.e13, July 2023.

[178] Alyssa Oplt. Determining the Role of DDR2 in Acquired Chemo-Resistance in Ovarian Cancer. Ph.D., Washington University in St. Louis, United States – Missouri, 2024. ISBN: 9798382310411.

[179] Asishana A. Osho, Anthony W. Castleberry, Laurie D. Snyder, Asvin M. Ganapathi, Paul J. Speicher, Sameer A. Hirji, Mark Stafford-Smith, Mani A. Daneshmand, R. Duane Davis, and Matthew G. Hartwig. Determining eligibility for lung transplantation: A nationwide assessment of the cutoff glomerular filtration rate. The Journal of Heart and Lung Transplantation, 34(4):571–579, April 2015. Publisher: Elsevier.

[180] Cengiz Ozge, Murat Bozlu, Eylem Sercan Ozgur, Mesut Tek, Ahmet Tunckiran, Necati Muslu, and Ahmet Ilvan. The impact of hypoxemia on serum total and free prostate-specific antigen levels in patients with chronic obstructive pulmonary disease. Medical Oncology, 32(5):156, April 2015.

[181] Anna Perkiö, Barun Pradhan, Fatih Genc, Anna Pirttikoski, Sanna Pikkusaari, Erdogan Pekcan Erkan, Matias Marin Falco, Kaisa Huhtinen, Sara Narva, Johanna Hynninen, Liisa Kauppi, and Anna Vähärautio. Locus-specific LINE-1 expression in clinical ovarian cancer specimens at the single-cell level. Scientific Reports, 14(1):4322, February 2024. Publisher: Nature Publishing Group.

[182] Robert Phillips. NK cells induce a pro-inflammatory phenotype in RA synovial fibroblasts. Nature Reviews Rheumatology, 17(11):645–645, November 2021. Publisher: Nature Publishing Group.

[183] Elina A. Pietilä, Jordi Gonzalez-Molina, Lidia Moyano-Galceran, Sanaz Jamalzadeh, Kaiyang Zhang, Laura Lehtinen, S. Pauliina Turunen, Tomás A. Martins, Okan Gultekin, Tarja Lamminen, Katja Kaipio, Ulrika Joneborg, Johanna Hynninen, Sakari Hietanen, Seija Grénman, Rainer Lehtonen, Sampsa Hautaniemi, Olli Carpén, Joseph W. Carlson, and Kaisa Lehti. Co-evolution of matrisome and adaptive adhesion dynamics drives ovarian cancer chemoresistance. Nature Communications, 12(1):3904, June 2021. Publisher: Nature Publishing Group.

[184] Hannah A. Pliner, Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, Andrew C. Adey, Frank J. Steemers, Jay Shendure, and Cole Trapnell. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular Cell, 71(5):858–871.e8, September 2018.

[185] Lilach Pnueli, Sergei Rudnizky, Yahav Yosefzon, and Philippa Melamed. RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin α-subunit gene. Proceedings of the National Academy of Sciences of the United States of America, 112(14):4369–4374, April 2015.

[186] Tessa M. Popay and Jesse R. Dixon. Coming full circle: On the origin and evolution of the looping model for enhancer–promoter communication. Journal of Biological Chemistry, 298(8), August 2022. Publisher: Elsevier.

[187] Eleonora Porcu, Marie C. Sadler, Kaido Lepik, Chiara Auwerx, Andrew R. Wood, Antoine Weihs, Maroun S. Bou Sleiman, Diogo M. Ribeiro, Stefania Bandinelli, Toshiko Tanaka, Matthias Nauck, Uwe Völker, Olivier Delaneau, Andres Metspalu, Alexander Teumer, Timothy Frayling, Federico A. Santoni, Alexandre Reymond, and Zoltán Kutalik. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nature Communications, 12(1):5647, September 2021. Publisher: Nature Publishing Group.

[188] Emmanouil Proestakis, Kyriakoula Papachristopoulou, Thanasis Georgiou, Sofia Eirini Chatoutsidou, Mihalis Lazaridis, Antonis Gkikas, Ilias Fountoulakis, Ioanna Tsikoudi, Manolis P. Petrakis, and Vassilis Amiridis. Atmospheric dust and air quality over large-cities and megacities of the world. Atmospheric Chemistry and Physics, 25(21):14777–14823, November 2025. Publisher: Copernicus GmbH.

[189] Nicholas Pudjihartono, Tayaza Fadason, Andreas W. Kempa-Liehr, and Justin M. O'Sullivan. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2, June 2022. Publisher: Frontiers.

[190] Homa Rahnamoun, Jihoon Lee, Zhengxi Sun, Hanbin Lu, Kristen M. Ramsey, Elizabeth A. Komives, and Shannon M. Lauberth. RNAs interact with BRD4 to promote enhanced chromatin engagement and transcription activation. Nature Structural & Molecular Biology, 25(8):687–697, August 2018. Number: 8 Publisher: Nature Publishing Group.

[191] Deepak A. Rao, Michael F. Gurish, Jennifer L. Marshall, Kamil Slowikowski, Chamith Y. Fonseka, Yanyan Liu, Laura T. Donlin, Lauren A. Henderson, Kevin Wei, Fumitaka Mizoguchi, Nikola C. Teslovich, Michael E. Weinblatt, Elena M. Massarotti, Jonathan S. Coblyn, Simon M. Helfgott, Yvonne C. Lee, Derrick J. Todd, Vivian P. Bykerk, Susan M. Goodman, Alessandra B. Pernis, Lionel B. Ivashkiv, Elizabeth W. Karlson, Peter A. Nigrovic, Andrew Filer, Christopher D. Buckley, James A. Lederer, Soumya Raychaudhuri, and Michael B. Brenner. Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. Nature, 542(7639):110–114, February 2017. Number: 7639 Publisher: Nature Publishing Group.

[192] Ahmed A. Raslan and Jeong Kyo Yoon. WNT Signaling in Lung Repair and Regeneration. Molecules and Cells, 43(9):774–783, September 2020.

[193] Eva Rath and Dirk Haller. Intestinal epithelial cell metabolism at the interface of microbial dysbiosis and tissue injury. Mucosal Immunology, 15(4):595–604, April 2022. Publisher: Nature Publishing Group.

[194] Kirsten A. Reimer, Claudia A. Mimoso, Karen Adelman, and Karla M. Neugebauer. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. Molecular Cell, 81(5):998–1012.e7, March 2021.

[195] Conglin Ren, Mingshuang Li, Yang Zheng, Bingbing Cai, Weibin Du, Helou Zhang, Fengqing Wu, Mengsha Tong, Fu Lin, Jinfu Wang, and Renfu Quan. Single-cell RNA-seq reveals

altered NK cell subsets and reduced levels of cytotoxic molecules in patients with ankylosing spondylitis. Journal of Cellular and Molecular Medicine, 26(4):1071–1082, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcmm.17159.

[196] Mostafa Rezapour, Patrick M. McNutt, David A. Ornelles, Stephen J. Walker, Sean V. Murphy, Anthony Atala, and Metin Nafi Gurcan. Cross-modal predictive modeling of multi-omic data in 3D airway organ tissue equivalents during viral infection. Frontiers in Genetics, 16, September 2025. Publisher: Frontiers.

[197] Mostafa Rezapour, Stephen J. Walker, David A. Ornelles, Muhammad Khalid Khan Niazi, Patrick M. McNutt, Anthony Atala, and Metin Nafi Gurcan. A comparative analysis of RNA-Seq and NanoString technologies in deciphering viral infection response in upper airway lung organoids. Frontiers in Genetics, 15, June 2024. Publisher: Frontiers.

[198] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, January 2010.

[199] John M. Rouhana, Jiali Wang, Gokcen Eraslan, Shankara Anand, Andrew R. Hamel, Brian Cole, Aviv Regev, François Aguet, Kristin G. Ardlie, and Ayellet V. Segrè. ECLIPSER: identifying causal cell types and genes for complex traits through single cell enrichment of e/sQTL-mapped genes in GWAS loci, November 2021. Pages: 2021.11.24.469720 Section: New Results.

[200] Jonathan D. Rubin, Jacob T. Stanley, Rutendo F. Sigauke, Cecilia B. Levandowski, Zachary L. Maas, Jessica Westfall, Dylan J. Taatjes, and Robin D. Dowell. Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment. Communications Biology, 4(1):1–15, June 2021. Number: 1 Publisher: Nature Publishing Group.

[201] Polina V. Rusina, Maria J. Falaguera, Juan Maria R. Romero, Ellen M. McDonagh, Ian Dunham, and David Ochoa. Genetic support for FDA-approved drugs over the past decade. Nature Reviews Drug Discovery, 22(11):864–864, October 2023. Bandiera_abtest: a Cg_type: Biobusiness Briefs Publisher: Nature Publishing Group.

[202] Paul Sabharwal, Jillian H. Hurst, Rohit Tejwani, Kevin T. Hobbs, Jonathan C. Routh, and Benjamin A. Goldstein. Combining adult with pediatric patient data to develop a clinical decision support tool intended for children: leveraging machine learning to model heterogeneity. BMC Medical Informatics and Decision Making, 22(1):84, March 2022.

[203] Saori Sakaue, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, Takahiro Konuma, Kenichi Yamamoto, Masato Akiyama, Kazuyoshi Ishigaki, Akari Suzuki, Ken Suzuki, Wataru Obara, Ken Yamaji, Kazuhisa Takahashi, Satoshi Asai, Yasuo Takahashi, Takao Suzuki, Nobuaki Shinozaki, Hiroki Yamaguchi, Shiro Minami, Shigeo Murayama, Kozo Yoshimori, Satoshi Nagayama, Daisuke Obata, Masahiko Higashiyama, Akihide Masumoto, Yukihiro Koretsune, Kaoru Ito, Chikashi Terao, Toshimasa Yamauchi, Issei Komuro, Takashi Kadowaki, Gen Tamiya, Masayuki Yamamoto, Yusuke Nakamura, Michiaki Kubo, Yoshinori Murakami, Kazuhiko Yamamoto, Yoichiro Kamatani, Aarno Palotie, Manuel A. Rivas, Mark J. Daly, Koichi Matsuda, and Yukinori Okada. A

cross-population atlas of genetic associations for 220 human phenotypes. Nature Genetics, 53(10):1415–1424, October 2021. Publisher: Nature Publishing Group.

[204] Eleanor Sanderson, M. Maria Glymour, Michael V. Holmes, Hyunseung Kang, Jean Morrison, Marcus R. Munafò, Tom Palmer, C. Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. Nature reviews. Methods primers, 2:6, February 2022.

[205] Yasmarie Santana-Rivera, Robert J Rabelo-Fernández, Blanca I Quiñones-Díaz, Nilmary Grafals-Ruíz, Ginette Santiago-Sánchez, Eunice L Lozada-Delgado, Ileabett M Echevarría-Vargas, Juan Apiz, Daniel Soto, Andrea Rosado, Loyda Meléndez, Fatima Valiyeva, and Pablo E Vivas-Mejía. Reduced expression of enolase-1 correlates with high intracellular glucose levels and increased senescence in cisplatin-resistant ovarian cancer cells. American Journal of Translational Research, 12(4):1275–1292, April 2020.

[206] Siddik Sarkar, Sarbar Ali Saha, Abhishek Swarnakar, Arnab Chakrabarty, Avipsa Dey, Poulomi Sarkar, Sarthak Banerjee, and Pralay Mitra. The molecular prognostic score, a classifier for risk stratification of high-grade serous ovarian cancer. Journal of Ovarian Research, 17(1):159, August 2024.

[207] Vittorio Sartorelli and Shannon M. Lauberth. Enhancer RNAs are an important regulatory layer of the epigenome. Nature Structural & Molecular Biology, 27(6):521–528, June 2020. Number: 6 Publisher: Nature Publishing Group.

[208] Sarah K. Sasse, Amber Dahlin, Lynn Sanford, Margaret A. Gruca, Arnav Gupta, Fabienne Gally, Ann Chen Wu, Carlos Iribarren, Robin D. Dowell, Scott T. Weiss, and Anthony N. Gerber. Glucocorticoid-regulated bidirectional enhancer RNA transcription pinpoints functional genetic variants linked to asthma, September 2023. Pages: 2022.11.10.22281906.

[209] Sarah K. Sasse, Amber Dahlin, Lynn Sanford, Margaret A. Gruca, Arnav Gupta, Fabienne Gally, Ann Chen Wu, Carlos Iribarren, Robin D. Dowell, Scott T. Weiss, and Anthony N. Gerber. Enhancer RNA transcription pinpoints functional genetic variants linked to asthma. Nature Communications, 16(1):2750, March 2025. Publisher: Nature Publishing Group.

[210] Sarah K. Sasse, Margaret Gruca, Mary A. Allen, Vineela Kadiyala, Tengyao Song, Fabienne Gally, Arnav Gupta, Miles A. Pufall, Robin D. Dowell, and Anthony N. Gerber. Nascent transcript analysis of glucocorticoid crosstalk with TNF defines primary and cooperative inflammatory repression. Genome Research, 29(11):1753–1765, November 2019. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[211] Shinya Sato, Megumi Miyazaki, Shinji Fukuda, Yukiko Mizutani, Yoichi Mizukami, Shigeki Higashiyama, and Shintaro Inoue. Human TMEM2 is not a catalytic hyaluronidase, but a regulator of hyaluronan metabolism *via* HYBID (KIAA1199/CEMIP) and HAS2 expression. Journal of Biological Chemistry, 299(6):104826, June 2023.

[212] Shinya Sato, Yukiko Mizutani, Yuta Yoshino, Manami Masuda, Megumi Miyazaki, Hideaki Hara, and Shintaro Inoue. Pro-inflammatory cytokines suppress HYBID (hyaluronan (HA) -binding protein involved in HA depolymerization/KIAA1199/CEMIP) -mediated

HA metabolism in human skin fibroblasts. Biochemical and Biophysical Research Communications, 539:77–82, February 2021.

[213] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics, 19(8):491–504, August 2018. Number: 8 Publisher: Nature Publishing Group.

[214] Katie Schaukowitch, Jae-Yeol Joo, Xihui Liu, Jonathan K. Watts, Carlos Martinez, and Tae-Kyung Kim. Enhancer RNA Facilitates NELF Release from Immediate Early Genes. Molecular Cell, 56(1):29–42, October 2014.

[215] Stefan Schoenfelder and Peter Fraser. Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics, 20(8):437–455, August 2019. Publisher: Nature Publishing Group.

[216] Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. TT-seq maps the human transient transcriptome. Science, 352(6290):1225–1228, June 2016. Publisher: American Association for the Advancement of Science.

[217] Joao M. Serigado, Jennifer Foulke-Abel, William C. Hines, Joshua A Hanson, Julie In, and Olga Kovbasnjuk. Ulcerative Colitis: Novel Epithelial Insights Provided by Single Cell RNA Sequencing. Frontiers in Medicine, 9:868508, April 2022.

[218] Lulu Shang, Jennifer A. Smith, and Xiang Zhou. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. PLoS Genetics, 16(4):e1008734, April 2020.

[219] Yi-Chen Shen, Ning-Yi Hsia, Wan-Hua Wu, Cheng-Li Lin, Te-Chun Shen, and Wei-Chien Huang. Age-related macular degeneration and premorbid allergic diseases: a population-based case–control study. Scientific Reports, 11(1):16537, August 2021. Publisher: Nature Publishing Group.

[220] Huajuan Shi, Ying Zhou, Erteng Jia, Min Pan, Yunfei Bai, and Qinyu Ge. Bias in RNA-seq Library Preparation: Current Challenges and Solutions. BioMed Research International, 2021(1):6647597, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/6647597.

[221] Norio Shinkai, Ken Asada, Hidenori Machino, Ken Takasawa, Satoshi Takahashi, Nobuji Kouno, Masaaki Komatsu, Ryuji Hamamoto, and Syuzo Kaneko. SEgene identifies links between super enhancers and gene expression across cell types. npj Systems Biology and Applications, 11(1):49, May 2025. Publisher: Nature Publishing Group.

[222] Rutendo F. Sigauke, Lynn Sanford, Zachary L. Maas, Taylor Jones, Jacob T. Stanley, Hope A. Townsend, Mary A. Allen, and Robin D. Dowell. Atlas of nascent RNA transcripts reveals enhancer to gene linkages, December 2023. Pages: 2023.12.07.570626 Section: Confirmatory Results.

[223] Rutendo F. Sigauke, Lynn Sanford, Zachary L. Maas, Taylor Jones, Jacob T. Stanley, Hope A. Townsend, Mary A. Allen, and Robin D. Dowell. Atlas of nascent RNA transcripts reveals tissue-specific enhancer to gene linkages. BMC Genomics, 26(1):406, April 2025.

[224] Scott Silvey, Amy Olex, Shaojun Tang, and Jinze Liu. Sample size requirements for machine learning classification of binary outcomes in bulk RNA-Seq data. BMC Bioinformatics, January 2026.

[225] Don D. Sin, Dany Doiron, Alvar Agusti, Antonio Anzueto, Peter J. Barnes, Bartolome R. Celli, Gerard J. Criner, David Halpin, MeiLan K. Han, Fernando J. Martinez, Maria Montes de Oca, Alberto Papi, Ian Pavord, Nicolas Roche, Dave Singh, Robert Stockley, M. Victorina Lopez Varlera, Jadwiga Wedzicha, Claus Vogelmeier, and Jean Bourbeau. Air pollution and COPD: GOLD 2023 committee report. European Respiratory Journal, 61(5), May 2023. Publisher: European Respiratory Society Section: Perspective.

[226] Nathan G. Skene, Julien Bryois, Trygve E. Bakken, Gerome Breen, James J. Crowley, Héléna A. Gaspar, Paola Giusti-Rodriguez, Rebecca D. Hodge, Jeremy A. Miller, Ana B. Muñoz-Manchado, Michael C. O'Donovan, Michael J. Owen, Antonio F. Pardiñas, Jesper Ryge, James T. R. Walters, Sten Linnarsson, Ed S. Lein, Patrick F. Sullivan, and Jens Hjerling-Leffler. Genetic identification of brain cell types underlying schizophrenia. Nature Genetics, 50(6):825–833, June 2018. Number: 6 Publisher: Nature Publishing Group.

[227] Christopher S. Smillie, Moshe Biton, Jose Ordovas-Montanes, Keri M. Sullivan, Grace Burgin, Daniel B. Graham, Rebecca H. Herbst, Noga Rogel, Michal Slyper, Julia Waldman, Malika Sud, Elizabeth Andrews, Gabriella Velonias, Adam L. Haber, Karthik Jagadeesh, Sanja Vickovic, Junmei Yao, Christine Stevens, Danielle Dionne, Lan T. Nguyen, Alexandra-Chloé Villani, Matan Hofree, Elizabeth A. Creasey, Hailiang Huang, Orit Rozenblatt-Rosen, John J. Garber, Hamed Khalili, A. Nicole Desch, Mark J. Daly, Ashwin N. Ananthakrishnan, Alex K. Shalek, Ramnik J. Xavier, and Aviv Regev. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell, 178(3):714–730.e22, July 2019. Publisher: Elsevier.

[228] Gabrielle D. Smith, Wan Hern Ching, Paola Cornejo-Páramo, and Emily S. Wong. Decoding enhancer complexity with machine learning and high-throughput discovery. Genome Biology, 24(1):116, May 2023.

[229] Jessica M. Snyder, Guo Zhong, Cathryn Hogarth, Weize Huang, Traci Topping, Jeffrey LaFrance, Laura Palau, Lindsay C. Czuba, Michael Griswold, Gabriel Ghiaur, and Nina Isoherranen. Knockout of Cyp26a1 and Cyp26b1 during postnatal life causes reduced lifespan, dermatitis, splenomegaly, and systemic inflammation in mice. The FASEB Journal, 34(12):15788–15804, 2020. _eprint: https://faseb.onlinelibrary.wiley.com/doi/pdf/10.1096/fj.202001734R.

[230] Lucia Sobrin, Gayatri Susarla, Lynn Stanwyck, John M. Rouhana, Ashley Li, Samuela Pollack, Robert P. Igo Jr, Richard A. Jensen, Xiaohui Li, Maggie C. Y. Ng, Albert V. Smith, Jane Z. Kuo, Kent D. Taylor, Barry I. Freedman, Donald W. Bowden, Alan Penman, Ching J. Chen, Jamie E. Craig, Sharon G. Adler, Emily Y. Chew, Mary Frances Cotch, Brian Yaspan, Paul Mitchell, Jie Jin Wang, Barbara E. K. Klein, Tien Y. Wong, Jerome I. Rotter, Kathyrn P. Burdon, Sudha K. Iyengar, and Ayellet V. Segrè. Gene Set Enrichment Analsyes Identify Pathways Involved in Genetic Risk for Diabetic Retinopathy. American Journal of Ophthalmology, 233:111–123, January 2022. Publisher: Elsevier.

[231] Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J. E. Matson, Quentin Barraud, Ariel J.

Levine, Gioele La Manno, Michael A. Skinnider, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. Nature Communications, 12(1):5692, September 2021. Publisher: Nature Publishing Group.

[232] Jacob T Stanley, Georgia E F Barone, Hope A Townsend, Rutendo F Sigauke, Mary A Allen, and Robin D Dowell. LIET model: capturing the kinetics of RNA polymerase from loading to termination. Nucleic Acids Research, 53(7):gkaf246, April 2025.

[233] Halley R. Steiner, Nickolaus C. Lammer, Robert T. Batey, and Deborah S. Wuttke. An Extended DNA Binding Domain of the Estrogen Receptor Alpha Directly Interacts with RNAs in Vitro. Biochemistry, 61(22):2490–2494, November 2022. Publisher: American Chemical Society.

[234] Sven Stringer, Naomi R. Wray, René S. Kahn, and Eske M. Derks. Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. PLOS ONE, 6(11):e27964, November 2011. Publisher: Public Library of Science.

[235] Yu-Li Su, Ching-Lan Chou, Kun-Ming Rau, and Charles Tzu-Chi Lee. Asthma and Risk of Prostate Cancer. Medicine, 94(36):e1371, September 2015.

[236] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, October 2005. Publisher: Proceedings of the National Academy of Sciences.

[237] Shuichi Suetani, Faraz Honarparvar, Dan Siskind, Guy Hindley, Nicola Veronese, Davy Vancampfort, Lauren Allen, Marco Solmi, John Lally, Fiona Gaughran, Brendon Stubbs, and Toby Pillinger. Increased rates of respiratory disease in schizophrenia: A systematic review and meta-analysis including 619,214 individuals with schizophrenia and 52,159,551 controls. Schizophrenia Research, 237:131–140, November 2021.

[238] Desheng Sun, Hongyan Liu, Yao Ouyang, Xiansheng Liu, and Yongjian Xu. Serum Levels of Gamma-Glutamyltransferase During Stable and Acute Exacerbations of Chronic Obstructive Pulmonary Disease. Medical Science Monitor, 26:0–0, October 2020. Publisher: International Scientific Information, Inc.

[239] Kazuhiro Suzuki, Akira Yokoi, Kosuke Yoshida, Hironori Suzuki, Masami Kitagawa, Eri Asano-Inami, Seiko Matsuo, Masato Yoshihara, Satoshi Tamauchi, Nobuhisa Yoshikawa, Kaoru Niimi, Tamotsu Sudo, Satoshi Yamaguchi, Yusuke Yamamoto, and Hiroaki Kajiyama. Overcoming platinum-resistant ovarian cancer targeting the activated JAK-STAT pathways via extracellular vesicles. Communications Biology, 8(1):1305, August 2025. Publisher: Nature Publishing Group.

[240] Tuan Zea Tan, He Yang, Jieru Ye, Jeffrey Low, Mahesh Choolani, David Shao Peng Tan, Jean-Paul Thiery, and Ruby Yun-Ju Huang. CSIOVDB: a microarray gene expression database of epithelial ovarian cancer subtype. Oncotarget, 6(41):43843–43852, November 2015. Publisher: Impact Journals.

[241] Selda Telo, Mutlu Kuluöztürk, Figen Deveci, Gamze Kırkıl, Önsel Öner, and Dilara Kaman. Serum Cystatin C Levels in COPD: Potential Diagnostic Value and Relation between Respiratory Functions. Journal of Medical Biochemistry, 37(4):434–440, December 2018.

[242] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. Genetics, 224(1):iyad031, May 2023.

[243] Immune cell - CD2 - The Human Protein Atlas.

[244] Pascal N Timshel, Jonatan J Thompson, and Tune H Pers. Genetic mapping of etiologic brain cell types for obesity. eLife, 9:e55851, September 2020. Publisher: eLife Sciences Publications, Ltd.

[245] Izabela Todorovski, Mary-Jane Tsang, Breon Feran, Zheng Fan, Sreeja Gadipally, David Yoannidis, Isabella Y Kong, Stefan Bjelosevic, Sarahi Rivera, Olivia Voulgaris, Magnus Zethoven, Edwin D Hawkins, Kaylene J Simpson, Gisela Mir Arnau, Anthony T Papenfuss, Ricky W Johnstone, and Stephin J Vervoort. RNA kinetics influence the response to transcriptional perturbation in leukaemia cell lines. NAR Cancer, 6(4):zcae039, October 2024.

[246] Hope A. Townsend, Jacob T. Stanley, Mary A. Allen, and Robin D. Dowell. Improving confidence of differential transcription calls in enhancers, September 2025. ISSN: 2692-8205 Pages: 2025.09.12.675852 Section: New Results.

[247] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology, 21(1):12, January 2020.

[248] Christy B. M. Tulen, Ying Wang, Daan Beentjes, Phyllis J. J. Jessen, Dennis K. Ninaber, Niki L. Reynaert, Frederik-Jan van Schooten, Antoon Opperhuizen, Pieter S. Hiemstra, and Alexander H. V. Remels. Dysregulated mitochondrial metabolism upon cigarette smoke exposure in various human bronchial epithelial cell models. Disease Models & Mechanisms, 15(3):dmm049247, March 2022.

[249] Alexey Uvarovskii, Isabel S. Naarmann-de Vries, and Christoph Dieterich. On the optimal design of metabolic RNA labeling experiments. PLoS Computational Biology, 15(8):e1007252, August 2019.

[250] Anniina Vihervaara, Dig Bijay Mahat, Michael J. Guertin, Tinyi Chu, Charles G. Danko, John T. Lis, and Lea Sistonen. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nature Communications, 8(1):255, August 2017. Publisher: Nature Publishing Group.

[251] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics, 101(1):5–22, July 2017. Publisher: Elsevier.

[252] Kiki Waeijen-Smit, Niki L. Reynaert, Rosanne J. H. C. G. Beijers, Sarah Houben-Wilke, Sami O. Simons, Martijn A. Spruit, and Frits M. E. Franssen. Alterations in plasma hyaluronic acid in patients with clinically stable COPD versus (non)smoking controls. Scientific Reports, 11(1):15883, August 2021. Publisher: Nature Publishing Group.

[253] Menghong Wan, Chen Wang, Jiamin Cui, Qing Xia, and Lei Zhang. Silencing KLF6 Alleviates Cigarette Smoke Extract-Induced Mitochondrial Dysfunction in Bronchial Epithelial Cells by SIRT4 Upregulation. International Journal of Chronic Obstructive Pulmonary Disease, 19:815–828, March 2024.

[254] Allen Wang, Feng Yue, Yan Li, Ruiyu Xie, Thomas Harper, Nisha A Patel, Kayla Muth, Jeffrey Palmer, Yunjiang Qiu, Jinzhao Wang, Dieter K Lam, Jeffrey C Raum, Doris A Stoffers, Bing Ren, and Maike Sander. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell Stem Cell, 16(4):386–99, Apr 2015.

[255] Fengxu Wang, Xuehai Wang, Lei Liu, Siyuan Deng, Wenqian Ji, Yang Liu, Xiangdong Wang, Rui Wang, Xinyuan Zhao, and Erli Gao. Comprehensive analysis of PTPN gene family revealing PTPN7 as a novel biomarker for immuno-hot tumors in breast cancer. Frontiers in Genetics, 13, September 2022. Publisher: Frontiers.

[256] Jingjing Wang, Yuelin Guan, Yue Wang, Junyi Tan, Zhongkai Cao, Yuhan Ding, Langping Gao, Haidong Fu, Xiangjun Chen, Jianyu Lin, Ning Shen, Xudong Fu, Fangqin Wang, Jianhua Mao, and Lidan Hu. Disease pathogenicity in Hutchinson–Gilford progeria syndrome mice: insights from lung-associated alterations. Molecular Medicine, 31(1):114, March 2025.

[257] Ke-xin Wang, Yao Gao, Cheng Lu, Yao Li, Bo-ya Zhou, Xue-mei Qin, Guan-hua Du, Li Gao, Dao-gang Guan, and Ai-ping Lu. Uncovering the Complexity Mechanism of Different Formulas Treatment for Rheumatoid Arthritis Based on a Novel Network Pharmacology Model. Frontiers in Pharmacology, 11, July 2020. Publisher: Frontiers.

[258] Ruike Wang, Xia Du, and Yaqin Zhi. Screening of Critical Genes Involved in Metastasis and Prognosis of High-Grade Serous Ovarian Cancer by Gene Expression Profile Data. Journal of Computational Biology, 27(7):1104–1114, July 2020. Publisher: Mary Ann Liebert, Inc., publishers.

[259] Rujin Wang, Dan-Yu Lin, and Yuchao Jiang. EPIC: Inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing. PLOS Genetics, 18(6):e1010251, June 2022. Publisher: Public Library of Science.

[260] Yan Wang, Katy A. Lloyd, Ioannis Melas, Diana Zhou, Radha Thyagarajan, Joakim Lindqvist, Monika Hansson, Anna Svärd, Linda Mathsson-Alm, Alf Kastbom, Karin Lundberg, Lars Klareskog, Anca I. Catrina, Stephen Rapecki, Vivianne Malmström, and Caroline Grönwall. Rheumatoid arthritis patients display B-cell dysregulation already in the naïve repertoire consistent with defects in B-cell tolerance. Scientific Reports, 9(1):19995, December 2019. Publisher: Nature Publishing Group.

[261] Zhong Wang, Tinyi Chu, Lauren A. Choate, and Charles G. Danko. Identification of regulatory elements from nascent transcription using dREG. Genome Research, 29(2):293–303, February 2019. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[262] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1):57–63, January 2009. Publisher: Nature Publishing Group.

[263] Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with FUMA. Nature Communications, 8(1):1826, November 2017. Number: 1 Publisher: Nature Publishing Group.

[264] Kyoko Watanabe, Maša Umićević Mirkov, Christiaan A. de Leeuw, Martijn P. van den Heuvel, and Danielle Posthuma. Genetic mapping of cell type specificity for complex traits. Nature Communications, 10(1):3222, July 2019. Number: 1 Publisher: Nature Publishing Group.

[265] Yinshen Wee, Chieh-Hsiang Yang, Shau-Kwaun Chen, Yu-Chun Yen, and Ching-Shuen Wang. Inositol hexaphosphate modulates the behavior of macrophages through alteration of gene expression involved in pathways of pro- and anti-inflammatory responses, and resolution of inflammation pathways. Food Science & Nutrition, 9(6):3240–3249, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/fsn3.2286.

[266] Elle M. Weeks, Jacob C. Ulirsch, Nathan Y. Cheng, Brian L. Trippe, Rebecca S. Fine, Jenkai Miao, Tejal A. Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P. Fulco, Katherine C. Tashman, Francois Aguet, Taibo Li, Jose Ordovas-Montanes, Christopher S. Smillie, Moshe Biton, Alex K. Shalek, Ashwin N. Ananthakrishnan, Ramnik J. Xavier, Aviv Regev, Rajat M. Gupta, Kasper Lage, Kristin G. Ardlie, Joel N. Hirschhorn, Eric S. Lander, Jesse M. Engreitz, and Hilary K. Finucane. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nature Genetics, 55(8):1267–1276, August 2023. Number: 8 Publisher: Nature Publishing Group.

[267] Sean Whalen, Rebecca M. Truty, and Katherine S. Pollard. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature Genetics, 48(5):488–496, May 2016. Number: 5 Publisher: Nature Publishing Group.

[268] Fengping Wu, Jinfang Gao, Jie Kang, Xuexue Wang, Qing Niu, Jiaxi Liu, and Liyun Zhang. B Cells in Rheumatoid Arthritis: Pathogenic Mechanisms and Treatment Prospects. Frontiers in Immunology, 12:750753, September 2021.

[269] Yong Wu, Lingfang Xia, Ping Zhao, Yu Deng, Qinhao Guo, Jun Zhu, Xiaojun Chen, Xingzhu Ju, and Xiaohua Wu. Immune profiling reveals prognostic genes in high-grade serous ovarian cancer. Aging, 12(12):11398–11415, June 2020.

[270] Huixuan Xu, Haiyan Yu, Lixiong Liu, Hongwei Wu, Cantong Zhang, Wanxia Cai, Xiaoping Hong, Dongzhou Liu, Donge Tang, and Yong Dai. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Peripheral Mononuclear Cells in Patients With Ankylosing Spondylitis. Frontiers in Immunology, 12:760381, November 2021.

[271] Akiko Yamada, Rieko Arakaki, Masako Saito, Takaaki Tsunematsu, Yasusei Kudo, and Naozumi Ishimaru. Role of regulatory T cell in the pathogenesis of inflammatory bowel disease. World Journal of Gastroenterology, 22(7):2195–2205, February 2016.

[272] Anyi Yang, Jingqi Chen, and Xing-Ming Zhao. nMAGMA: a network-enhanced method for inferring risk genes from GWAS summary statistics and its application to schizophrenia. Briefings in Bioinformatics, 22(4):bbaa298, July 2021.

[273] Jin H. Yang and Anders S. Hansen. Enhancer selectivity in space and time: from enhancer–promoter interactions to promoter activation. Nature Reviews Molecular Cell Biology, 25(7):574–591, July 2024. Publisher: Nature Publishing Group.

[274] Luhan Yang, Hongping Zhang, Junfeng Wang, Jing Ge, Rushan Hao, Junxu Yu, and Bingrong Zheng. Study on the effects and mechanism of RRM2 on three gynecological malignancies. Cellular Signalling, 129:111674, May 2025.

[275] Mei Yang, Ji Hoon Lee, Zhao Zhang, Richard De La Rosa, Mingjun Bi, Yuliang Tan, Yiji Liao, Juyeong Hong, Baowen Du, Yanming Wu, Jessica Scheirer, Tao Hong, Wei Li, Teng Fei, Chen-Lin Hsieh, Zhijie Liu, Wenbo Li, Michael G. Rosenfeld, and Kexin Xu. Enhancer RNAs Mediate Estrogen-Induced Decommissioning of Selective Enhancers by Recruiting ERα and Its Cofactor. Cell Reports, 31(12):107803, June 2020.

[276] Pengyi Yang, Hao Huang, and Chunlei Liu. Feature selection revisited in the single-cell era. Genome Biology, 22(1):321, December 2021.

[277] Douglas W. Yao, Luke J. O'Connor, Alkes L. Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nature Genetics, 52(6):626–633, June 2020. Number: 6 Publisher: Nature Publishing Group.

[278] Li Yao, Jin Liang, Abdullah Ozer, Alden King-Yung Leung, John T. Lis, and Haiyuan Yu. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nature Biotechnology, 40(7):1056–1065, July 2022. Number: 7 Publisher: Nature Publishing Group.

[279] Hooi-Yeen Yap, Sabrina Zi-Yi Tee, Magdelyn Mei-Theng Wong, Sook-Khuan Chow, Suat-Cheng Peh, and Sin-Yeang Teow. Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. Cells, 7(10):161, October 2018.

[280] John Yeh, Beom Su Kim, Jennifer Peresie, and Carly Page. Declines in Levels of Hyperpolarization-Activated Cation (HCN) Channels in the Rat Ovary After Cisplatin Exposure. Reproductive Sciences, 16(10):986–994, October 2009. Publisher: SAGE Publications Inc.

[281] Kosuke Yoshihara, Tatsuhiko Tsunoda, Daichi Shigemizu, Hiroyuki Fujiwara, Masayuki Hatae, Hisaya Fujiwara, Hideaki Masuzaki, Hidetaka Katabuchi, Yosuke Kawakami, Aikou Okamoto, Takayoshi Nogawa, Noriomi Matsumura, Yasuhiro Udagawa, Tsuyoshi Saito, Hiroaki Itamochi, Masashi Takano, Etsuko Miyagi, Tamotsu Sudo, Kimio Ushijima, Haruko Iwase, Hiroyuki Seki, Yasuhisa Terao, Takayuki Enomoto, Mikio Mikami, Kohei Akazawa, Hitoshi Tsuda, Takuya Moriya, Atsushi Tajima, Ituro Inoue, Kenichi Tanaka, and for The Japanese Serous Ovarian Cancer Study Group. High-Risk Ovarian Cancer Based on 126-Gene Expression Signature Is Uniquely Characterized by Downregulation of Antigen Presentation Pathway. Clinical Cancer Research, 18(5):1374–1385, March 2012.

[282] Zhi Yu, Tim H. H. Coorens, Md Mesbah Uddin, Kristin G. Ardlie, Niall Lennon, and Pradeep Natarajan. Genetic variation across and within individuals. Nature Reviews Genetics, 25(8):548–562, August 2024. Publisher: Nature Publishing Group.

[283] Tianyi Yuan and Haidong Zou. Effects of air pollution on myopia: an update on clinical evidence and biological mechanisms. Environmental Science and Pollution Research International, 29(47):70674–70685, 2022.

[284] Judith Barbara Zaugg, Pelin Sahlén, Robin Andersson, Meritxell Alberich-Jorda, Wouter de Laat, Bart Deplancke, Jorge Ferrer, Susanne Mandrup, Gioacchino Natoli, Dariusz Plewczynski, Alvaro Rada-Iglesias, and Salvatore Spicuglia. Current challenges in understanding the role of enhancers in disease. Nature Structural & Molecular Biology, 29(12):1148–1158, December 2022. Number: 12 Publisher: Nature Publishing Group.

[285] Yaqiong Zhan, Lushun Jiang, Xuehang Jin, Shuaibing Ying, Zhe Wu, Li Wang, Wei Yu, Jiepeng Tong, Li Zhang, Yan Lou, and Yunqing Qiu. Inhibiting RRM2 to enhance the anticancer activity of chemotherapy. Biomedicine & Pharmacotherapy, 133:110996, January 2021.

[286] Fan Zhang, Anna Helena Jonsson, Aparna Nathan, Nghia Millard, Michelle Curtis, Qian Xiao, Maria Gutierrez-Arcelus, William Apruzzese, Gerald F. M. Watts, Dana Weisenfeld,

Saba Nayar, Javier Rangel-Moreno, Nida Meednu, Kathryne E. Marks, Ian Mantel, Joyce B. Kang, Laurie Rumker, Joseph Mears, Kamil Slowikowski, Kathryn Weinand, Dana E. Orange, Laura Geraldino-Pardilla, Kevin D. Deane, Darren Tabechian, Arnoldas Ceponis, Gary S. Firestein, Mark Maybury, Ilfita Sahbudin, Ami Ben-Artzi, Arthur M. Mandelin, Alessandra Nerviani, Myles J. Lewis, Felice Rivellese, Costantino Pitzalis, Laura B. Hughes, Diane Horowitz, Edward DiCarlo, Ellen M. Gravallese, Brendan F. Boyce, Larry W. Moreland, Susan M. Goodman, Harris Perlman, V. Michael Holers, Katherine P. Liao, Andrew Filer, Vivian P. Bykerk, Kevin Wei, Deepak A. Rao, Laura T. Donlin, Jennifer H. Anolik, Michael B. Brenner, and Soumya Raychaudhuri. Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. Nature, 623(7987):616–624, November 2023. Number: 7987 Publisher: Nature Publishing Group.

[287] Fan Zhang, Kevin Wei, Kamil Slowikowski, Chamith Y. Fonseka, Deepak A. Rao, Stephen Kelly, Susan M. Goodman, Darren Tabechian, Laura B. Hughes, Karen Salomon-Escoto, Gerald F. M. Watts, A. Helena Jonsson, Javier Rangel-Moreno, Nida Meednu, Cristina Rozo, William Apruzzese, Thomas M. Eisenhaure, David J. Lieb, David L. Boyle, Arthur M. Mandelin, Brendan F. Boyce, Edward DiCarlo, Ellen M. Gravallese, Peter K. Gregersen, Larry Moreland, Gary S. Firestein, Nir Hacohen, Chad Nusbaum, James A. Lederer, Harris Perlman, Costantino Pitzalis, Andrew Filer, V. Michael Holers, Vivian P. Bykerk, Laura T. Donlin, Jennifer H. Anolik, Michael B. Brenner, and Soumya Raychaudhuri. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nature Immunology, 20(7):928–942, July 2019. Publisher: Nature Publishing Group.

[288] Guanran Zhang, Yanlin Qu, Zhenyu Wu, Wenjia Liu, Huihuan Luo, Renjie Chen, Huixun Jia, and Xiaodong Sun. Association between low lung function and the increased risk of age-related macular degeneration: A population-based prospective cohort study. Journal of Global Health, 14:04102, June 2024.

[289] Han Zhang, Yijun Wu, Hao Li, Liping Sun, and Xiangkai Meng. Model constructions of chemosensitivity and prognosis of high grade serous ovarian cancer based on evaluation of immune microenvironment and immune response. Cancer Cell International, 21(1):593, November 2021.

[290] Kaiyang Zhang, Erdogan Pekcan Erkan, Sanaz Jamalzadeh, Jun Dai, Noora Andersson, Katja Kaipio, Tarja Lamminen, Naziha Mansuri, Kaisa Huhtinen, Olli Carpén, Sakari Hietanen, Jaana Oikkonen, Johanna Hynninen, Anni Virtanen, Antti Häkkinen, Sampsa Hautaniemi, and Anna Vähärautio. Longitudinal single-cell RNA-seq analysis reveals stress-promoted chemoresistance in metastatic ovarian cancer. Science Advances, 8(8):eabm1831, February 2022. Publisher: American Association for the Advancement of Science.

[291] Martin Jinye Zhang, Kangcheng Hou, Kushal K. Dey, Saori Sakaue, Karthik A. Jagadeesh, Kathryn Weinand, Aris Taychameekiatchai, Poorvi Rao, Angela Oliveira Pisco, James Zou, Bruce Wang, Michael Gandal, Soumya Raychaudhuri, Bogdan Pasaniuc, and Alkes L. Price. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. Nature Genetics, 54(10):1572–1580, October 2022. Number: 10 Publisher: Nature Publishing Group.

[292] Tiantian Zhang, Zhuqiang Zhang, Qiang Dong, Jun Xiong, and Bing Zhu. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. Genome Biology, 21(1):45, February 2020.

[293] Wei Zhang, Raphael Petegrosso, Jae-Woong Chang, Jiao Sun, Jeongsik Yong, Jeremy Chien, and Rui Kuang. A large-scale comparative study of isoform expressions measured on four platforms. BMC Genomics, 21(1):272, March 2020.

[294] Yuanfu Zhang, Shu Sun, Yue Qi, Yifan Dai, Yangyang Hao, Mengyu Xin, Rongji Xu, Hongyan Chen, Xiaoting Wu, Qian Liu, Congcong Kong, Guangmei Zhang, Peng Wang, and Qiuyan Guo. Characterization of tumour microenvironment reprogramming reveals invasion in epithelial ovarian carcinoma. Journal of Ovarian Research, 16(1):200, October 2023.

[295] Yueqi Zhang, Xinhui Liu, Kairui Sun, Yue Luo, Jack Yang, Aimin Li, Matti Kiupel, Stefanie Fenske, Martin Biel, Qing-Sheng Mi, Hongbing Wang, and Hua Xiao. Hyperpolarization-activated cyclic nucleotide-gated cation channel 3 promotes HCC development in a female-biased manner. Cell Reports, 42(10), October 2023. Publisher: Elsevier.

[296] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics, 2(3):lqaa078, September 2020.

[297] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLOS ONE, 9(1):e78644, January 2014. Publisher: Public Library of Science.

[298] Jin J. Zhou, Michael H. Cho, Peter J. Castaldi, Craig P. Hersh, Edwin K. Silverman, and Nan M. Laird. Heritability of Chronic Obstructive Pulmonary Disease and Related Phenotypes in Smokers. American Journal of Respiratory and Critical Care Medicine, 188(8):941–947, October 2013. Publisher: American Thoracic Society - AJRCCM.

[299] Wei Zhou, Dongdong Tian, Jun He, Yimei Wang, Lijun Zhang, Lan Cui, Li Jia, Li Zhang, Lizhong Li, Yulei Shu, Shouzhong Yu, Jun Zhao, Xiaoyan Yuan, and Shuangqing Peng. Repeated PM2.5 exposure inhibits BEAS-2B cell P53 expression through ROS-Akt-DNMT3B pathway-mediated promoter hypermethylation. Oncotarget, 7(15):20691–20703, March 2016. Publisher: Impact Journals.

[300] Weijian Zhou, Gaoshaer Yeerkenbieke, Yumei Zhang, Mingwang Zhou, and Jin Li. Guanylate binding protein 4 shapes an inflamed tumor microenvironment and identifies immuno-hot tumors. Journal of Cancer Research and Clinical Oncology, 150(2):90, February 2024.

[301] Jianguo Zhu, Jukun Song, Zezhen Liu, Jin Han, Heng Luo, Yunlin Liu, Zhenyu Jia, Yuanbo Dong, Wei Zhang, Funeng Jiang, Chinlee Wu, Zaolin Sun, and Weide Zhong. Association between allergic conditions and risk of prostate cancer: A Prisma-Compliant Systematic Review and Meta-Analysis. Scientific Reports, 6(1):35682, October 2016. Publisher: Nature Publishing Group.

[302] Tao Zhu, Chunhe Li, and Xiakun Chu. Transcriptional condensates encode a "golden mean" to optimize enhancer–promoter communication across genomic distances. Proceedings of the National Academy of Sciences, 122(43):e2513371122, October 2025. Publisher: Proceedings of the National Academy of Sciences.

[303] Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nature Communications, 9(1):4361, October 2018. Number: 1 Publisher: Nature Publishing Group.

[304] Yun Zhu, Lin Sun, Zhao Chen, John W. Whitaker, Tao Wang, and Wei Wang. Predicting enhancer transcription and activity from chromatin modifications. Nucleic Acids Research, 41(22):10032–10043, December 2013.

[305] Chang Zou, Jiacheng Shen, Fangfang Xu, Yingjun Ye, Yuanyuan Wu, and Shaohua Xu. Immunoreactive Microenvironment Modulator GBP5 Suppresses Ovarian Cancer Progression by Inducing Canonical Pyroptosis. Journal of Cancer, 15(11):3510–3530, May 2024. Publisher: Ivyspring International Publisher.

[306] Daniela Šimčíková, Dominik Gardáš, Tomáš Pelikán, Lukáš Moráň, Martin Hruda, Kateřina Hložková, Tiziana Pivetta, Michal Hendrych, Júlia Starková, Lukáš Rob, Petr Vaňhara, and Petr Heneberg. Metabolism of primary high-grade serous ovarian carcinoma (HGSOC) cells under limited glutamine or glucose availability. Cancer & Metabolism, 12(1):27, September 2024.

Appendix A

Appendix A. Supplemental Material to Addressing transcriptomic assay heterogeneity for predictive modeling in cancer

A.1 Supplemental Results

A.1.1 Harmonizing Counts between RNA-seq and Nanostring

A.1.1.1 Genes Considered in Harmonization

The NanoString platform uses probes considering a specific sequence subset of a gene's exons. This means that a NanoString count can apply to multiple isoforms of the gene. NanoString provides the RefSeq isoform to which a given probe is meant to capture so we first tried to link said isoforms to the most current RefSeq annotations. Of the 770 genes in the NanoString panel, 31 had exact NanoString naming matches to RefSeq, while 723 had isoform names matching except for the version number (e.g. A2M NM_000014.4 in Nanostring is NM_000014.6 in the current RefSeq). The remaining 16 isoforms are not currently annotated in the most recent RefSeq version and therefore we instead linked the NanoString probe based on where it mapped in the genome (details in Methods). Three of these probes aligned to genes no longer existing in RefSeq (*CD45RA*, *CD45RB*, and *CD45RO*) but instead had to be manually linked to different isoforms of the same gene: textitPTPRC. Every one of the 770 probes mapped to the genome clearly only once, supporting NanoString's claim that NanoString is specific and accurate.

We attempted two methods for harmonizing counts: isoform-focused and exon-focused. NanoString annotation files provide the Refseq or ENSEMBL isoform of the gene used. There-

fore, we first retrieved coordinates of gene isoforms matching to the names (Figure A.5B). Probe sequences are about 100bp long and therefore best correspond to an exon rather than a complete isoform (i.e. collection of exons). Therefore, we also mapped probe sequences to individual exons for counting that might better match with NanoString counts. 765 probes successfully mapped to exons (details in Methods).

A.1.1.2 Normalization

Both NanoString and RNA-seq then rely on normalization to mitigate bias from a sample simply having higher general transcription levels, or sequencing/measuring biases of machines. NanoString uses relatively standardized normalization approaches, relying on positive and negative control probes along with pre-specified housekeeping genes to calculate normalization factors. Conversely, RNA-seq normalization approaches are widespread and often rely entirely on external factors (the length of regions being counted and the number of sequences/reads mapping to the regions being considered) rather than specifically calibrated internal controls. Due to the high variability of normalization approaches for RNA-seq and the fact that some data is only available with already normalized counts, we considered three normalization methods: Median-Ratio from DESeq2 (MR), Reads Per Kilobase of transcript per Million mapped reads (RPKM), and Transcripts per Million (TPM) (details in Methods, Supplemental Figure A.5C) [100].

A.1.1.3 Isoform-based Counts allow improved harmonization

Some biological explanation of isoform-counts being better regardless of limits of detection can be found by the fact that genes showing better harmonization with isoform-based counts tend to have longer probe-matching exons; therefore an exon is less comparable to the length of the probe (Supplemental Figure A.10). We observed no clear feature differences for exon-improved counts nor between genes with improved harmonization from length-based vs non-length-based normalization (Supplemental Figure A.11). Ultimately, isoform-based counts with length-based normalization, allowed more similar low-dimensional spaces of samples in NanoString and RNA-seq, regardless

of removing genes below limits of detection (Supplemental Figure A.12). Therefore, isoform-based counts seem to better harmonize with NanoString, likely by maintaining enough counts to overcome sampling and length-based biases.

When considering highest-expressed isoforms vs NanoString-based isoforms, all cases with changed log2FC (outliers labeled in Figure 2.2E) were explained by expression levels changing between values already below our limits of detection and zero (details in Supplemental Results). Removing these cases led to R^2 values of 0.99 and below 1% of sample-gene combinations with different isoforms having changes in log2FC exceeding 1 and below 0.05% changing in sign. Of these, 1% problem cases, 60%-88% were explained by at least one of teh transcription cases again being below the limit of detection.

A.1.2 Poor correlation for select samples in Matched Dataset

We wanted to make sure that poor or high correlation between NanoString and RNAseq of a gene was not being driven by a single outlier. Therefore, we recalculated spearman correlation coefficients after removing the top (maximum 2) samples with greatest distance from the line of best fit. Some genes significantly increased in correlation when only removing one sample (from spearman coefficients of 0.5 to greater than 0.85), regardless of harmonization approach. When looking into these genes, we saw that the outlier sample in each case was well within the range of expression (therefore not simply caused due to having expression levels not considered by other samples) and was almost always one of two samples: P_09 and P_18. We then checked if these sample outliers were explained by poor QC. Both samples, however, had uniquely mapped read numbers comparable to that of the other samples (31.6M for P_09 and 30.3M for P_18) and did not show any clear differences in mapping, sequencing, or trimming compared to the other samples. These disrepancies could be caused by multiple factors. First, mutations in the patient might allow specific genes to be poorly mapped to the probe. Additionally, it is possible that exact samples of the FFPE tissue blocks used for sequencing and NanoString were different enough to cause these disrepancies. Because we wanted to have strict consideration for whether a gene had poor or strong

correlation, without bias from one specific sample, we removed these two samples for final coefficient calculations. For final results, we used the number of genes with correlation trends (good, poor, mid, mixture) across methodologies before filtering. After removing the samples, the number of genes with good correlation across all increased (345 to 390) and poor correlation decreased (116 to 78). We saw limited changes to distributions of spearman correlations (average increase of 0.04 and 0.002 increase in median spearman correlation for across and within patients, respectively). When determining characteristics of genes with improved correlation, we did not include the outliers to ensure we had clean comparisons. Importantly, we performed all RNAseq-vs-NanoString analyses with and without the samples, with no changes to trends. The genes with the largest impacts from outliers include: textitBLM/IRF2/RELA, and *G6PD/HLA-F* changed correlations of $0.5 - 0.6$ to ≥ 0.85 when removing P_09 or P_18, respectively; *FUT4*, textitUBA7, textitIFI35, textitCLEC4E saw vast improvements when removing P_09 and P_18.

Full comparisons of with and without the outlier samples can be found at Evaluate_Harmonization.ipynb.

A.1.3 Relevance of Harmonization in single-assay predictive modeling

We wanted to assess if our harmonization insights have improved upon the interpretation of the top 20 predictable genes based on several different modeling approaches (AUROCs < 0.6) for the same NanoString data[cite lucy paper]. When training a model using only these 20 genes, we achieved comparable performance to our original model (T/V AUROC=0.84, hold-out=0.78), but wit as almost entirely reliant on two genes: *CD274* and *MTOR* (Supplemental Figure A.19B and C). *CD274* (protein PD-L1) was also among the top 25 predictive frequencies in both our RNA-seq and NanoString analyses. All genes with independent predictive power (measured by AUROC of Tr/V and hold-out above 0.55) showed expression levels above the limits of detection (*CD274*, *PDCD1*, *MTOR*, *IL11RA*, and *RAD51*), with 7/20 genes showing median expression levels below the limit in most experiments (Supplemental Figure A.19A and D). Therefore, our harmonization approaches clearly allow improved insight both for feature filtering and model interpretation.

A.1.4 Evaluation of Predictive Models for PFS

A.1.4.1 Feature Relevance

To help clarify if some features captured redundant signals, we assessed how well each model performed when removing each individual gene, a leave one feature out strategy. For the combined (NanoString+RNA-seq) model, removal of most genes independently led to small (< 0.05) changes to AUROCs of the hold-out patients, with the most severe changes coming from removing *EGR1* (decrease > 0.1), *CTSS* (increase 0.1), *JAK2*, and *ENO1* (both decrease around 0.05) (Supplemental Figure A.17 Top). When used as the only gene in the model, most of the model genes had AUROCs only slightly above 0.5, with *CD40* having the strongest predictive power (hold-out AUROC 0.81) and *JAK2*, *ENO1*, *RRM2*, and *IDO1* following as still having AUROCs clearly above 0.5 for both the full validation/training set and hold-out set (Supplemental Figure A.17 Bottom). For the RNA-seq model, removal of *ESYT3* and *CLK2* showed no change in model performance, indicating clearly redundant signals from these and other genes in the model. However, removing either *CDKL2* or *GBP4* caused sharp declines (¿0.15) in AUROCs in the hold-out set, while *DDX11*, *HCN3*, *GPR173*, and *AMOTL2* showed only a slight decline (¿0.05) (Supplemental Figure A.22). These findings further support that these six genes contain non-redundant information in the model. Conversely, removing *SHROOM1* caused a decrease in training/validation AUROC but an increase in the hold-out AUROC, suggesting that the gene's patterns might be more population-specific or the coefficient is prone to overfitting (Supplemental Figure A.22). Indeed, we were unable to find clear literature support for the *SHROOM1*'s relevance in ovarian cancer.

A.1.4.2 Model confidence reflects biology

We then confirmed that the model confidence (logistic probability) matched that expected given our data. The combined model clearly favored one of the classifications (probabilities < 0.4 or > 0.6) for 100 of the 140 patients and low confidence classifications (near the decision boundary, i.e. ± 0.1 of 0.5) were enriched in samples with PFS closer to the cutoff of 12 months (Supplemental

Figure A.15). As with the combined model, the vast majority of cases (84%) were high confidence (predicted probabilities from the model were outside 0.4-0.6) and lower confidence predictions (probabilities of 0.5 ± 0.1) were enriched in patients with PFS values close to the cutoff (12 mths) (Supplemental Figure A.14).

A.1.4.3 Models perform better than random chance after feature filtering

Since we have a feature filtering approach before modeling, we wanted to estimate 1) if we might be missing key genes with predictive power in the model, and 2) what a fair "random guess" baseline might be to compare for our model given we already limit features. To do this, we trained 1000 models on randomly selected sets of two genes and recorded the AUROCs on both the full training/validation set and hold-out set. For the NanoString+RNA-seq models, Median AUROCs from these random combinations were well below what our final model reached: 0.6 for training/validation and 0.5 for hold-out (Supplemental Figure A.16A). There were a total of 17/1000 gene sets where the training/validation set had AUROC above 0.6 (max of 0.71) and the hold-out set above 0.8, with two of the cases including one of our final model genes: *CD40*. For the RNA-seq only model of PFS, we again found that the AUROCs values were significantly higher than AUROCs acheived across 1000 models using random sets of 2 genes (median AUC 0.45 Tr/V and 0.5 Hold out) (Supplemental Figure A.16B). Only ten of the 1000 models had AUROCs above 0.8 (max 0.92)in one of the datasets (Tr/V or Hold-out) and above 0.7 (max 0.75) in the other, with *HCN3* being in one of the models. There was some stochasticity to this process, but the general results were observed regardless of seed used.

A.1.5 Guanylate-binding genes

While *GBP4* showed the most consistent and strongest trends and the highest AUROCs when considering different GBP genes in the full model, other guanylate-binding genes still show comparable if not greater predictive power individually (Supplemental Figure A.23). *GBP3/4* showed consistent trends across all tested assays but *GBP3* was not included in the NanoString

panel (Supplemental Figure A.23C). Similarly, when considering the 4 HGSOC non-longitudinal micorarray datasets, *GBP4/5* showed consistently significant results across 2 cohorts while *GBP1/2* showed in only one, and *GBP3/7* in none (Supplemental Table 5).

A.2 Supplemental Methods

A.2.1 Longitudinal Cohort Information

Datasets were named based on the first or last authors of the papers from which the dataset was collected and as noted in the Data Availability Statement. All source and clinical information for patients are available in Supplemental Table 1.

A.2.2 Nanostring Analysis

Nanostring counts were achieved by using nSolver Analysis Software version 4.0.70 on MacOS. Lanes were flagged when 0.5fM positive control $\leq$ 2 standard deviations above the mean from negative controls. Background thresholding was done according to the geometric mean of negative control counts. Positive control normalization was done using the geometric mean with lanes flagged if normalization factors were outside the 0.3-3 range. Reference/housekeeping normalization was done using the geometric mean with lanes flagged if normalization factors were outside the 0.1-10 range. No samples/lanes were flagged. RLF files from nCounter PanCancer IO 360 panel were used with platform and annotation information available on GEO using GEO accession GPL27956.

A.2.3 RNA-seq Analysis

Reads were mapped to the hg38 genome where NCBI RefSeq annotations were used (hg38 release GCF 000001405.40-RS 2023 03) for all analyses except comparison with scRNA-seq in which case GENCODE v25 annotation was used to match what was used for the scRNA-seq. Bams were created by using the RNAseq-Flow nextflow pipeline at https://github.com/Dowell-Lab/RNAseq-Flow under commit 97c703b using the options –genome_id 'hg38', -profile slurm, and –forwardStranded. Versions hisat2/2.1.0, fastqc/0.11.8, and bbmap/38.05 were used. Bbduk was used to trim fastqs which were then checked via fastqc. Hisat2 was used to map to hg38. Counting was done using featureCounts using subread v1.6.2. If using DESeq2 median-ratio normalization, RNA-seq counts were normalized using size factors from according to 191 housekeeping genes (full

list found at 05_Get_Final_Counts.ipynb). Annotated genes not captured by Nanostring probes were counted as all exons corresponding to the gene (no single isoform), or the isoform most highly expressed in the experiment.

For the matched RNA-seq and NanoString dataset, RNA-seq counts were filtered so that genes not also considered in Nanostring were only kept if the median TPM value across all 24 samples was above 1 (leaving 17,976 genes).

A.2.4 Harmonizing Counts

The code needed to replicate these analyses are found in https://github.com/Hope2925/NanoString-RNAseq-HGSOC/RNA-Nano-Matched. We first tried to link the NanoString probes to the isoform most similarly representing them by 1) using the isoform annotated by the NanoString company for the probe, or 2) considering the longest isoform that uses the exon(s) to which the probe maps (Isoform Counting). We also counted just over the exon(s) to which the probe mapped for a gene (Exon Counting), hypothesizing that this may allow for cleaner harmonization in the event of changing isoform usage (Supplemental Figure A.5B). We next considered three different normalization approaches for RNA-seq: Median-Ratio from DESeq2 (MR), Reads Per Kilobase of transcript per Million mapped reads (RPKM), and Transcripts per Million (TPM) (details in Methods, Supplemental Figure A.5C) [100]. Further details are below.

A.2.4.1 Probe Mapping

To account for probes extending across multiple exons, the probe sequences for the 770 Nanostring genes were mapped to hg38 using HISAT2. These sequences were also blasted (https://blast.ncbi.nlm.nih.gov/ against the database Refseq genome RS_2023_03 using highly similar sequence algorithm (megablast).

A.2.4.2 Exon matching

HISAT2 mappings of probe sequences were overlapped with hg38 exons (RefSeq hg38 release GCF 000001405.40-RS 2023 03). If a probe sequence was completely within an exon, those exon co-

ordinates were used. Otherwise, all exons with the probe overlapping were considered. Probes that mapped to multiple genes (6 total probes) due to overlapping exons were linked to the coordinates of the exon(s) for the same gene as Nanostring. A subset of PTPRC exons simultaneously mapped to the probes for CD45RB, CD45RO, CD45RA, and PTPRC. Only the PTPRC probe was kept to avoid the same exons being considered across multiple probes, and because the PTPRC probe covered all the exons. A GTF file was produced to assign multiple exon coordinates to a single probe for counting. The full code and results for this can be found in subfolder Preprocessing

A.2.4.3 Matching Isoforms between RNA-seq and Nanostring

To identify isoforms best matching to NanoString probes, we first saw if the RefSeq isoforms NanoString offered as annotation of the probes could be linked to the most recent RefSeq hg38 annotations that we mapped RNA-seq reads to. CD45RA, CD45RB, CD45RO, and PTPRC in the Nanostring samples map to different exons of PTPRC, with no clear isoform matching up (CD45RA and CD45B and CD45RO no longer exist in the most recent annotation). For sake of completion, Nanostring probes for PTPRC, CD45RA, CD45RB, and CD45RO were assigned to isoforms that used the exon in which the probe mapped most exclusively: NM_080921.4, NM_002838.5, XM_047426398.1, and XM_006711472.5 of PTPRC, respectively. The full list of isoforms linked to probes based on HISAT and BLAST results can be found at 05_Get_Final_Counts.ipynb. All probes, if successfully mapping to the genome, only mapped once. LILRA3 has the Nanostring isoform of NM_001172654.2 which maps to an alternative reference assembly due to large differences to primary reference chromosome sequences (chr19 ALT_REF_LOCI_9), (both according to the isoform name and BLAST). Therefore the gene was not easily integrated into RNA-seq analysis and showed 0 counts across all RNA-seq datasets.

A.2.4.4 Determining similarity between RNA-seq and NanoString

Spearman correlation coefficients were calculated to capture how NanoString expression compared to RNA-seq expression. We considered correlation of gene expression across patients (1 coef-

ficient calculated for each gene) or within patients (1 coefficient calculated for each patient). These calculations were done for NanoString expression compared to each of the six approaches described in the RNA-seq Analysis section individually for RNA-seq based isoforms and NanoString-based isoforms (12 total comparisons). A gene was considered to have poor correlation if it had a spearman correlation coefficient across patients < 0.65 and high correlation if ≥ 0.8. Notes on two outlier samples can be found in Supplementary Info, but had little impact on final trends.

A gene was considered to be better captured by Isoform- or Exon-based counts if the spearman correlation changed by at least 0.1 across at least two of the three normalization approaches. A gene was considered to be better captured by a normalization approach if there was a change of at least 0.1 across both Exon- and Isoform-based counts.

For evaluation of correlation and similarity of expression after considering low expression and range genes, we did the following. Genes that had ranges below the limits of detection were removed. Expression levels below the limit of detection were replaced with NA (therefore not included in analyses) and genes with NA values in 10 or more patients (out of 24) were removed. This led to about 635-700 genes kept when using strict cutoffs and 702-730 genes kept when using loose limits of detection.

RV coefficients were used to assess how similar the relationships were between samples in low-dimensional space without directly integrating RNA-seq with NanoString. We calculated them using FactoMineR::coeffRV and only considering genes with no samples having expression levels below the limits of detection. We then compared how samples mapped in low dimensional space (using PCA R function prcomp) when combined, with different approaches of scaling. In all cases, we used z-score scaling. We only changed scaling for the NanoString and RNA-seq samples, seperately, before performing gene-based scaling in the full (NanoString and RNA-seq) data-set as is the usual action for PCA. We either did not scale, scaled per-patient, or scaled per-gene separately before combination. The analysis for scaling comparisons are found in Evaluate_Harmonization_Post.ipynb.

We also investigated whether normalization approaches led to biases, but very few genes saw systematic improvements to correlation due to normalization: 9 genes better with length-

based methods (RPKM and TPM) compared to MR, 5 genes better with MR than length-based methods, 5 genes better in RPKM compared to both MR and TPM.

A.2.4.5 Estimating limits of detection for RNA-seq and NanoString

First, to bin genes generally according to low expression or range, we split genes according to 10 quantiles of expression and range (highest-lowest expression) for each of the six harmonization approaches. If considering all six harmonization approaches at once, each measurement would serve as a gene-harmonization approach pair rather than simply a gene. To estimate limits of detection, we first plotted NanoString expression and RNA-seq expression (separately for each of the six harmonization approaches), with each gene colored by the spearman correlation coefficient. We observed the worse correlation appeared close to low expression levels for both NanoString and RNAseq, mostly below where the largest mass of genes clearly began to cluster around a line-of-best fit. We first estimated loose cutoffs, where the NanoString (or RNA-seq) expression level reached the first point of the large cluster of genes. Strict cutoffs were determined by similar logic except where the first point of the large cluster of genes that also contain Spearman correlation coefficients clearly above 0.65. Without technical replicates, these limits were estimated largely by visual clustering and are not expected to be used as official limits of detection. The final estimated limits of detection for harmonization can be found in Supplemental Table 2. The code for this analysis is found in Evaluate_Harmonization.ipynb.

A.2.5 Predictive Modeling of PFS and NanoString/RNAseq

The code and results for this section, and an outline of the number of samples in each prediction category and cohorts are found under RNA-PrePost/Modeling/.

A.2.5.1 Feature Selection

Due to our relatively small sample sizes, including the full feature list of a minimum shared 770 or all expressed 18,000+ genes would likely cause overfitting. Therefore, we performed an initial

feature selection to confine the feature space considered by a model to a maximum of 25 genes, as is often done in high-dimensional datasets [189, 276]. First, we performed analyses with and without filtering of genes with median log2FCs ≤ 0.5 to optionally focus on genes with clearer changes in expression. When considering the full RefSeq annotated genes, we removed any genes where the median expression was below 0.4 RPKM (right below strict limit of detection) in both classifications: patients with PFS $\leq 12mths$ and $> 12mths$. This would ensure we are not removing genes that have large changes in expression (e.g. from no expression to high expression) but instead only genes with expression levels difficult to ascertain consistently. This led to gene numbers declining from 18,035 genes to 12,974, and keeping 652 NanoString genes.

For choosing the genes considered in model validation from the bootstrapping approach, the 10-40 genes with the highest predictive frequencies excluding that of a chosen threshold after 390 bootstrapped samples (cumulative proportions remained stable after around 200 samples). The threshold was chosen based on a clear visual clustering of genes that was consistently obvious as shown in these jupyter notebooks RNA-PrePost/Modeling/02_NanoRNAseq_Btsp.ipynb and RNA-PrePost/Modeling/03_RNAseqFull_BtspCons.ipynb.

Scaling of features showed little to no difference in results. Therefore, we reported coefficients without scaling for more immediately interpretable results.

A.2.5.2 Assessing Model performance

Since we performed some initial feature selection, we wanted to ensure that the random baseline from which to compare model performance was indeed still about 0.5. Therefore, we randomly selected two genes from the genes that had log2FC above 0.5 1000 different times and evaluated model performance using each of these subsets. We recorded AUROCs on the full training/validation set and hold-out set to evaluate the AUROCs for each of these when using random subsets of genes.

We also calculated sensitivity and precision. We did this either using the whole dataset, or after splitting patients based on whether or not the probability from the model was greater than

0.1 distance from 0.5. Briefly, low-confidence patients had probabilities from the model of 0.4-0.6 (model is almost randomly guessing) while high-confidence patients had probabilities either below 0.4 or above 0.6.

We also assessed how much each feature of the final model was contributing to the model (e.g. if there were features capturing redundant signal). To do so, we reperformed training, and recalculated AUROC on the training/validation and hold-out sets when removing the single gene from the model. Similarly, we evaluated the predictive power of each feature individually by reperforming training and evaluation using only the single gene in the model.

A.2.6 Non-longitudinal Microarray

Code and results for this analysis can be found at RNA-scRNA-Microarray/SC_Micro/03_Format_Microarray.

Expression data was downloaded from GEO numbers GSE32062 (Yoshishara)[281], GSE26712 (Bonome)[17], GSE13876 (Crijns)[46], and GSE18521 (Mok)[162]. Metadata was downloaded from Precog (https://precog.stanford.edu/) [16, 71].

Univariate Cox regression was performed on each of the four Microarray cohorts and gene-probes, using both a categorical variable (low vs high expression) or treating the expression as a numerical variable. For categorical Cox regression and Kaplan-Meier curves, we split the patients within each cohort based on the median expression. In both cases we used 'coxph()' for the regression and 'survfit()' for the Kaplain-Meier curves from survminer v0.5.1, survival v3.8-3, survMisc v0.5.6, and KMsurv v0.1-6. All model genes assessed for survival regression were considered by probes in at least two of the microarray datasets.

A.2.7 scRNA-seq

Code and results for this analysis can be found at RNA-scRNA-Microarray/SC_Micro/. scRNA-seq data was pseudobulked (at the count level) for samples as well as sample/cell cluster levels based on the previous annotations from Zhang et al[290] using SingleCellExperiment v1.26.0 and muscat v1.18.0 in R. There were 11 patients with SC data, 5 which had paired bulk

data as well. There are 13 total samples originating from the same patient and tissue with both sequencing from scRNA-seq and bulk RNA-seq. Pseudobulked counts were then normalized using counts per million (CPM) or DESeq2 v1.44.0 (MR) to ensure results were not dirven by normalization approaches. Due to the high sparsity of scRNA-seq data, log2FC was calculated as log2((Post_Norm+1)/(Pre_Norm+1)). A total of 18,921 genes had non-zero counts in both scRNA-seq and bulk RNA-seq paired samples.

A.2.8 Gene-Network Analyses

A.2.8.1 hdWGCNA

Before running hdWGCNA v0.04.09, a Seurat object v5.3.0 and Seurat v5.4.0 for the 11 patients was normalized (NormalizeData()), the top 3000 variable features were found (FindVariableFeatures), and the expression was scaled (ScaleData()) with variables "nCount_RNA" and "percent.mt" regressed out using option vars.to.regress. PCA was then run using RunPCA(seurat, npcs = 50). To maintain reasonable N, we did not include Endothelial cells (only 79 cells), combined all epithelial cancer cells (EOC) into one group (new N of 8,806), combined dendritic cells (pDC, DC-1, DC-2) into "Myeloid_APC" (new N of 1,297), and combined NK cells (N=1,744) with ILC cells (N=288) into "Innate_Lymphoid" or "NK_ILC." Harmony v1.2.4 was then run using the first 30 PCs of PCA and group.by.vars of "sample" and "patient_id". Metacells were then built within each treatment phase (Pre or Post NACT) and cell type (with the following changes above). Briefly, genes expressed in at least 4% of cells were considered and then MetacellsByGroups() was run with groupings being based on the cell types and patients with minimums of 50 cells needed. Several cell types lost patient representation due to not having enough cells, with the exact lost patients noted in the jupyter notebook below. Topological Matrices (TOMs) were calculated based on TestSoftPowers() using a signed network and ConstructNetwork() with default parameters. Briefly, the TOM similarity scores measure the relative interconnectedness of two nodes (genes) by comparing their shared neighbors and direct expression correlations across samples, thereby

being more robust to noise than correlation alone[128]. Code for this analysis can be found at RNA-scRNA-Microarray/Network/04_hdWGCNA_Networks.ipynb.

A.2.8.2 WGCNA

Variance stabilization transformation from DESeq2 v1.44.0 was used to get normalized counts for all available bulk raw RNA-seq data for pre or post-NACT conditions (N=80 or N=83). Code for this step can be found in RNA-scRNA-Microarray/Network/04_hdWGCNA_Networks.ipynb These data were then used to run pre/post NACT specific network analyses with WGCNA (v1.73) to create signed TOMs, with genes being split into two TOMs to avoid memory surges. This split is determined internally by WGCNA to keep similarly networked genes together. Code for this step can be found at RNA-scRNA-Microarray/Network/05_WGCNA_Networks.R

Two different approaches were used to identify *GBP4* gene networks specific to a certain timing (shared, Pre, Post). Both approaches led to the same findings described in results. First, genes that had TOM similarities above 0.1 in only one timing were considered timing-specific, and those above 0.1 in both were considered shared. Since Pre networks tended to have overall higher connectivities, we then classified genes to more specifically consider this potential bias based on both scores and differences in scores. Genes shared between Pre and Post NACT networks had TOM similarity scores above 0.11 in both Pre and Post, and differences ¡ 0.02. Genes only found in Pre networks had score differences above 0.08 or the gene was not considered in the Post TOM despite having a connectivity with *GBP4* above 0.12 in Pre. Genes only found in Pre networks had score differences above 0.03 or the gene was not considered in the Pre TOM despite having a connectivity with *GBP4* above 0.11 in Post. Gene enrichment analysis was done with clusterProfiler::enrichGO where the universe of genes considered was the 23,557 genes reaching the required expression levels for the TOM in either pre- or post-NACT conditions. Code for this analysis can be found at RNA-scRNA-Microarray/Network/07_GBP4_Networks.ipynb

A.2.8.3 Comparing

For each gene, the spearman correlation of TOM similarity scores were calculated between bulk and cell-type networks at Pre and Post conditions. Only unfiltered genes in both cases were considered. Code for this analysis can be found at . RNA-scRNA-Microarray/Network/06_Network_Compare.R

A.3 Supplemental Figures

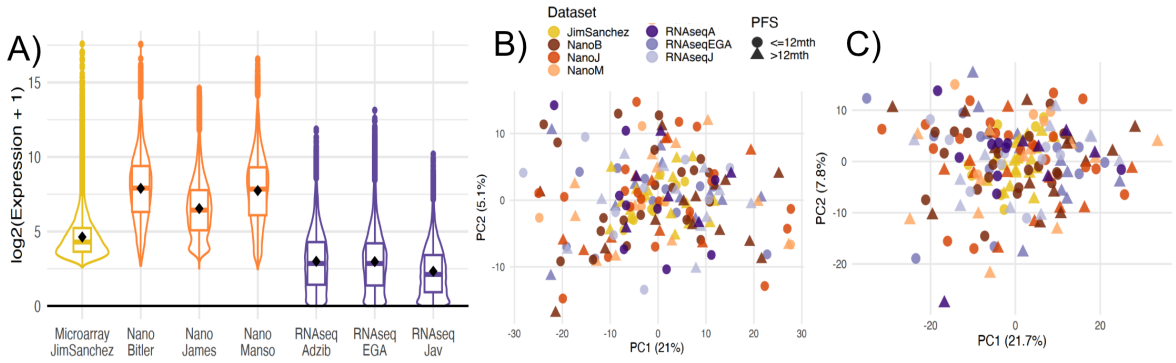

Figure A.1: **Microarray shows a limited dynamic range with gene-scaling required for somewhat comparable levels** A. Violin and box plots of normalized expression levels (log2 with pseudocount) of 731 genes shared across all assays and cohorts (RNA-seq normalization was RPKM). Black diamonds indicate means. B. PCA from pre-NACT expression alone of all cohorts/assays when scaling per gene before combining assays. C. PCA from Post-Pre NACT log2FC of all cohorts/assays when scaling per gene before combining assays.

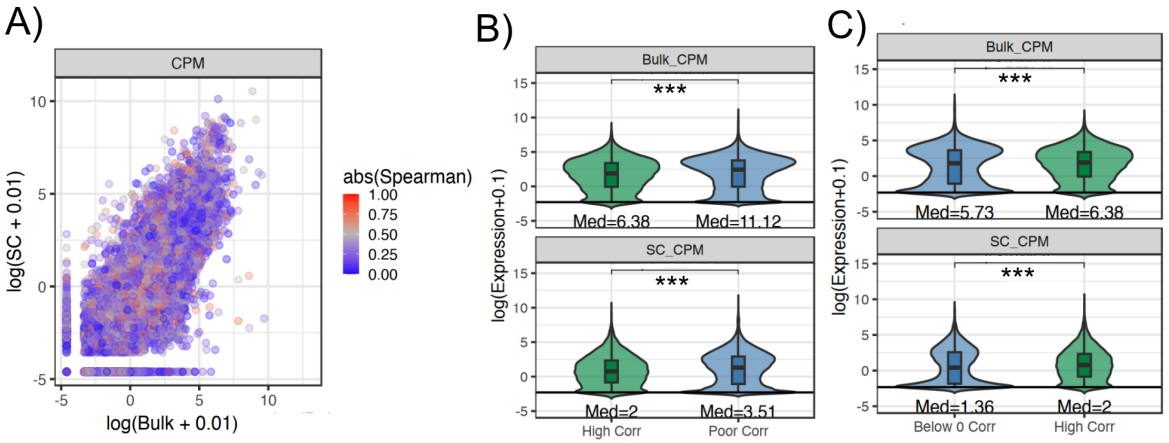

Figure A.2: **Poor correlation between matched pseudobulked scRNA-seq and bulk RNA-seq is not explained by low expression genes** A. Genes with low spearman correlation across matched pseudobulked scRNA-seq and bulk RNA-seq samples (N=13) (blue) are well spread across transcription levels predicted by both scRNA-seq and Bulk RNA-seq (x and y-axes). B. Genes with poor correlation () have higher expression levels on average than those with high correlation (N=). C. Genes with below 0 spearman correlation (N=) have loower transcription than those with high correlation (N). *** indicates p-value from Mann-Whitney test below 0.001.

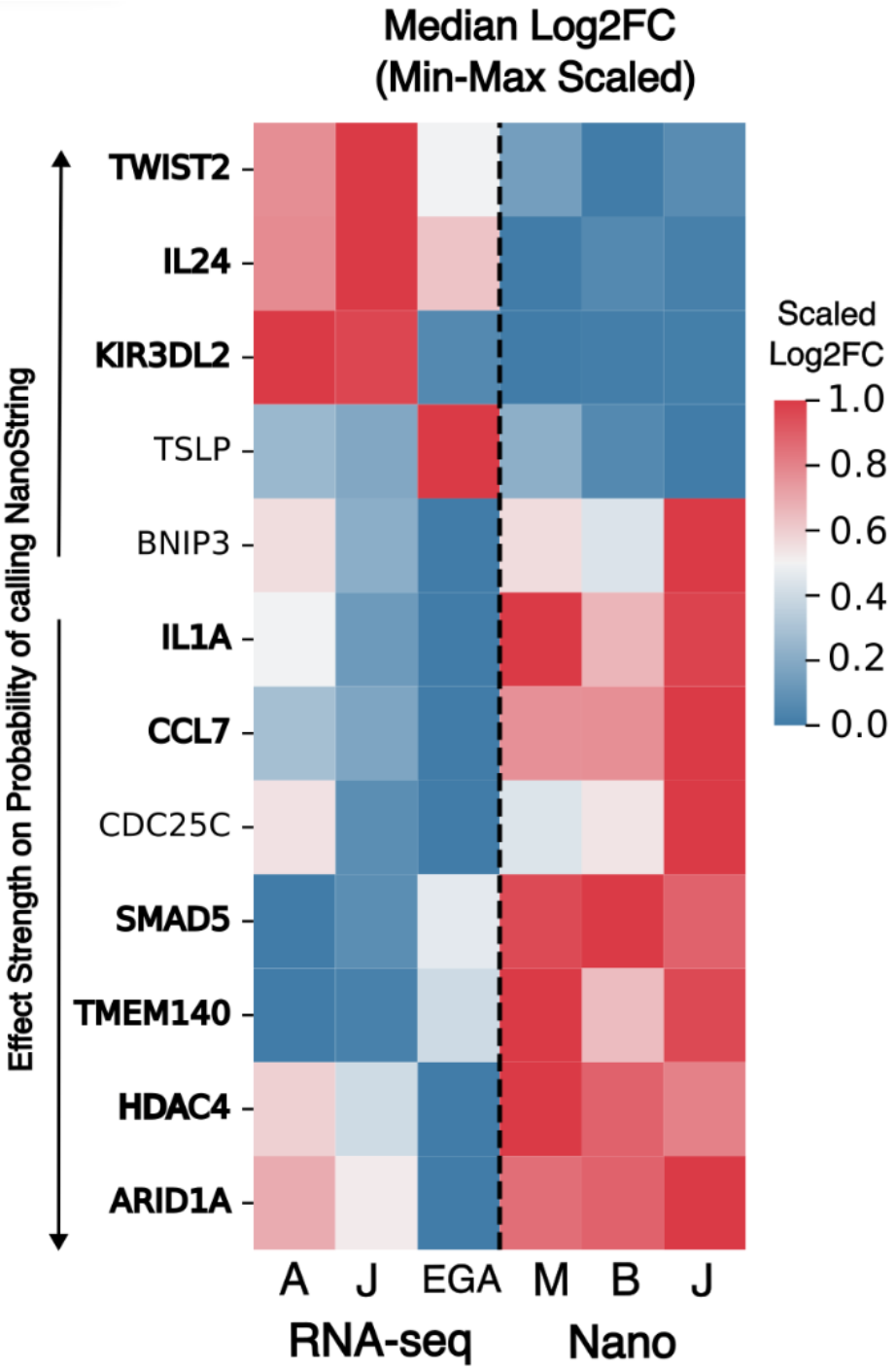

Figure A.3: **Genes predictive of NanoString vs RNA-seq show consistent trends regardless of experiments** Heatmap of final genes used as features in the model predicting NanoString from RNA-seq samples based on log2FC. Genes are ranked based on direction and effect strength (coefficient size) in the model calling NanoString. Bolded genes have the same trends across all experiments (NanoString vs RNAseq). Median Log2FC for each experiment are row-scaled with min-max scaling to more clearly visualize trends.

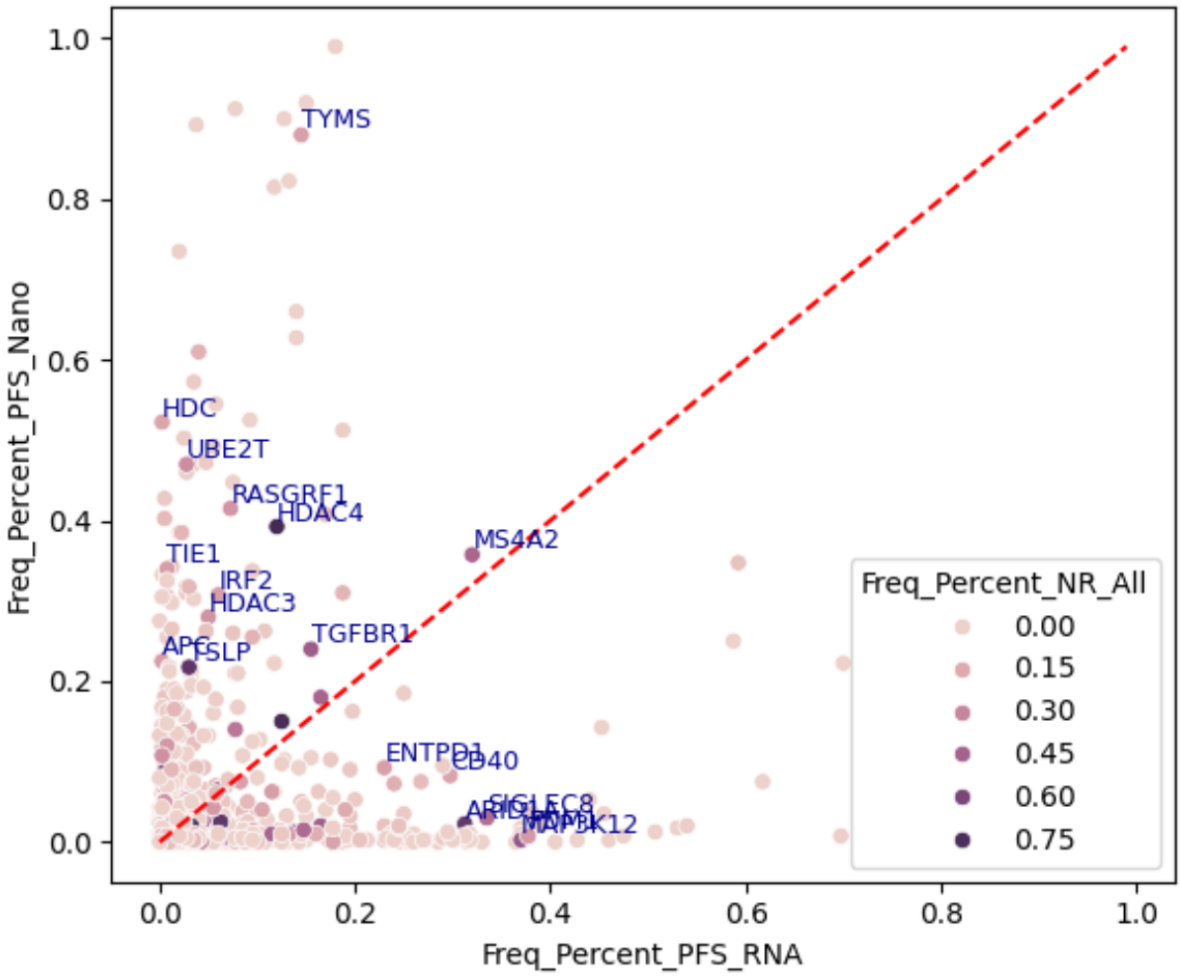

Figure A.4: **Genes predictive of NanoString vs RNA-seq often shown high predictive power in either RNA-seq or NanoString** Scatter plot of the cumulative frequency of a gene being within the top 100 predictive features across boostrapped samples (e.g. 1 means 100% of bootstrapped samples had the gene within the top 100 predictive features) for NanoString (y-axis) or RNA-seq (x-axis). Genes are then colored according to their cumulative frequency for bootstrapped samples in predicting NanoString vs RNA-seq samples (Freq_Percent_NR_All).

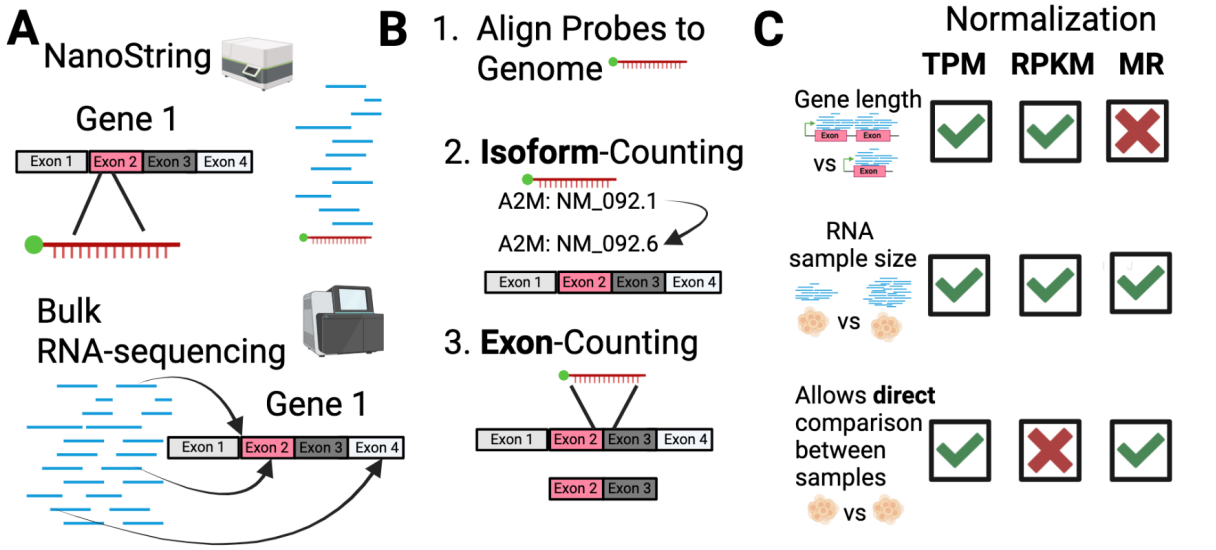

Figure A.5: **Several preprocessing approaches were tested to harmonize counts across NanoString and RNA-seq A.** Visualization of what NanoString vs RNA-seq might capture based on a gene and its exons. NanoString will be confined to probe location while RNA-seq can map anywhere. **B.** We mapped probes to the genome and then used that to match a NanoString probe to either a given isoform (Isoform-Counting) or Exon(s) (Exon-Counting). **C.** Three RNA-seq normalization approaches were considered: Transcripts per kilobase million (TPM), Reads per kilobase million (RPKM), and DESeq2's median of ratios (MR). Unlike TPM and RPKM, MR does not account for gene length. TPM and MR attempt to allow direct comprison between samples. All approaches consider library size. More details can be found in Methods.

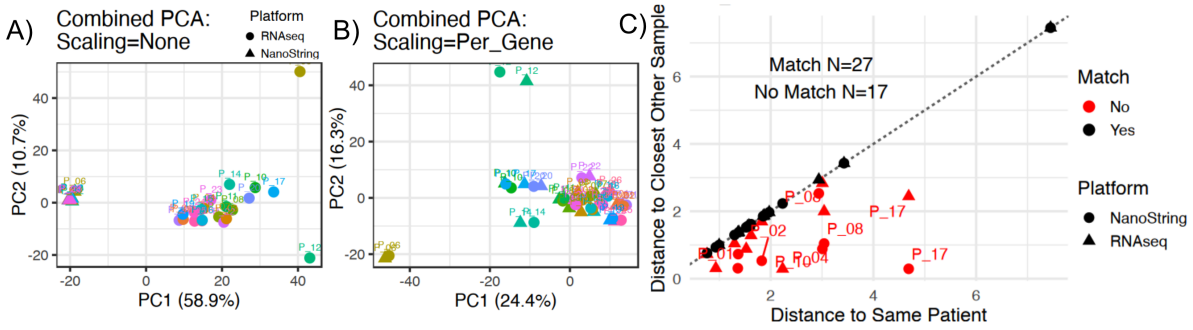

Figure A.6: **RNA-seq and NanoString matched-samples are similar in low-dimensional space after gene-based scaling A.** PCA of RNA-seq and NanoString matched samples without gene-based scaling. Patients are color coded and NanoString vs RNA-seq define most of the variance (PC1). NanoString and RNA-seq originally cleanly separate, with RNA-seq showing larger inner-variability. **B.** When performing gene-based scaling (z-score scaling for each gene across patients), patients are instead close together. **C.** When graphing the euclidean distance of a sample (NanoString or RNA-seq) to it's nearest other sample of the opposite assay (y-axis), most samples were closest to the sample of the same patient (27) (on the line so y=x).

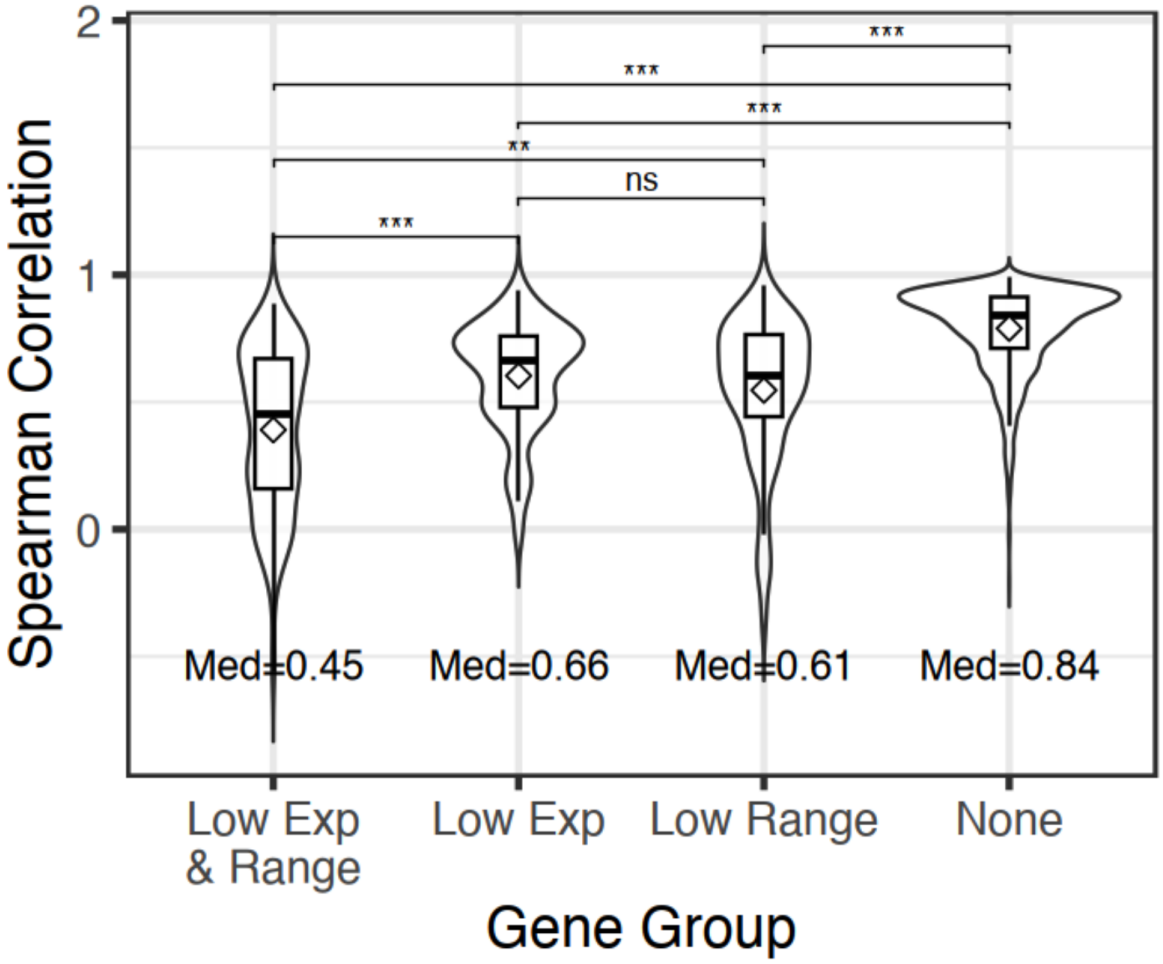

Figure A.7: **Genes with low expression or low ranges of expression explain most of the poor correlation between NanoString and RNA-seq A.** Violin/Box plots of spearman correlation coefficients of genes between matched NanoString and RNA-seq split across four categories: those with both low range of expression and expression (Low Exp & Range), just low expression (Low Exp), just low range of expression (Low Range), or neither (None). Means are indicated by a diamond. Non-parameteric Dunn's test was used for post-hoc pairwise comparisons and bonferroni adjusted p-values are indicated by ns > 0.05, ** < 0.01, *** < 0.001.

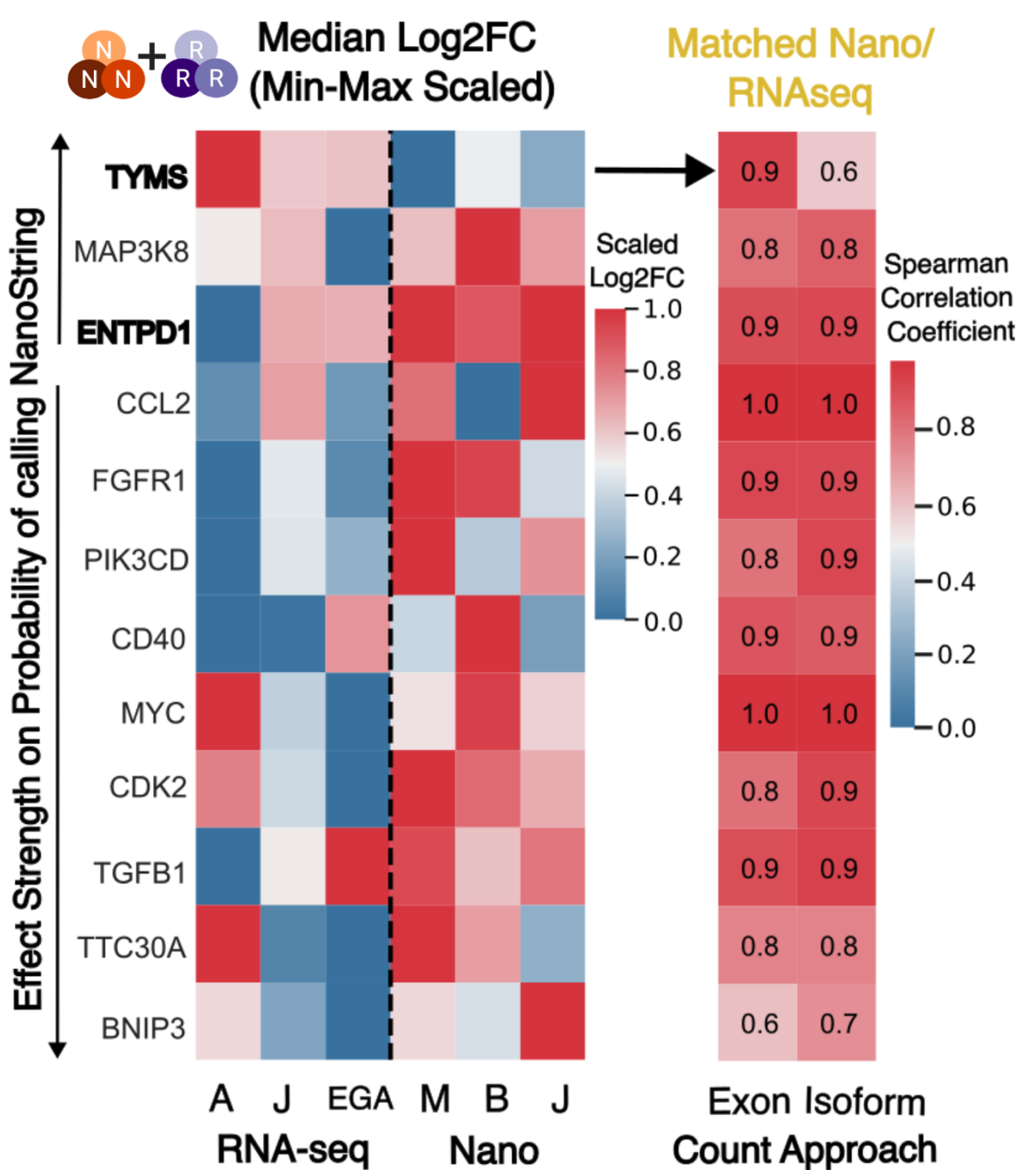

Figure A.8: **Non low expression features marking differences between RNA-seq and NanoString log2FC are not consistent or improved by exon counts**. Heatmap of final genes used as features in the model predicting NanoString from RNA-seq samples based on log2FC. Genes are ranked based on direction and effect strength (coefficient size) in the model calling NanoString. Bolded genes have the same trends across all experiments (NanoString vs RNAseq). Median Log2FC for each experiment are row-scaled with min-max scaling to more clearly visualize trends. Right shows heatmap of spearman correlation coefficients for the gene when comparing matched RNA-seq and NanoString used either Exon-based or Isoform-based counting.

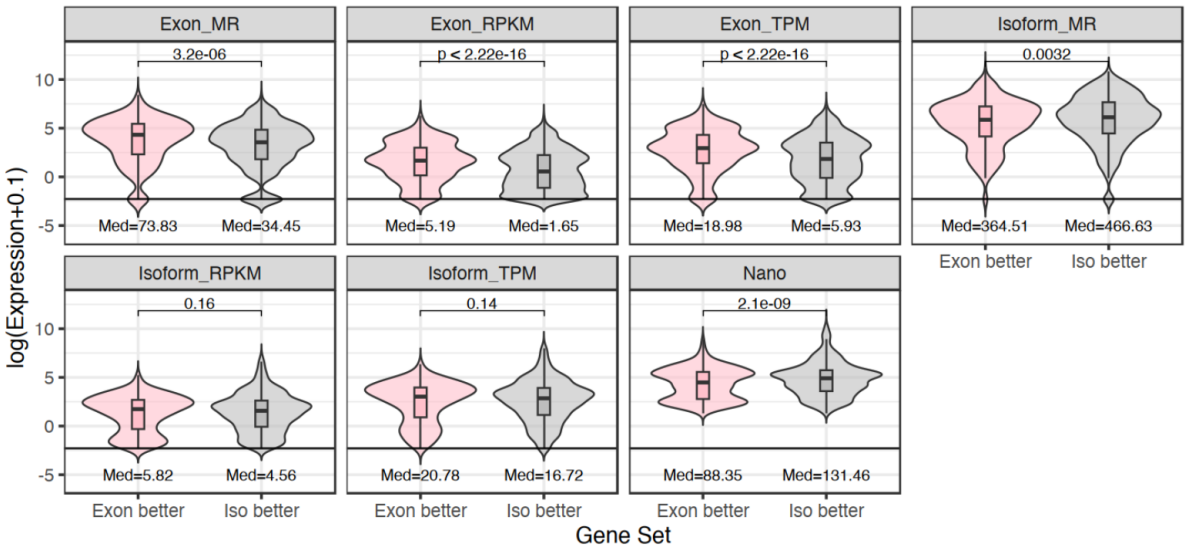

Figure A.9: **Genes where exon counts have better harmonization have higher expression in exons than genes where isoform counts have better harmonization.** Expression levels (y-axis) for genes where exon-based counts show improved correlation between NanoString and RNA-seq than isoform counts (pink) compared to genes where isoform-based counts are better (grey). Exon-better genes have significantly higher expression in exons than isoform-better genes. There is no clear difference in expression levels for isoform-based counts. Pairwise comparisons were made using a Student's t-test with p-values shown.

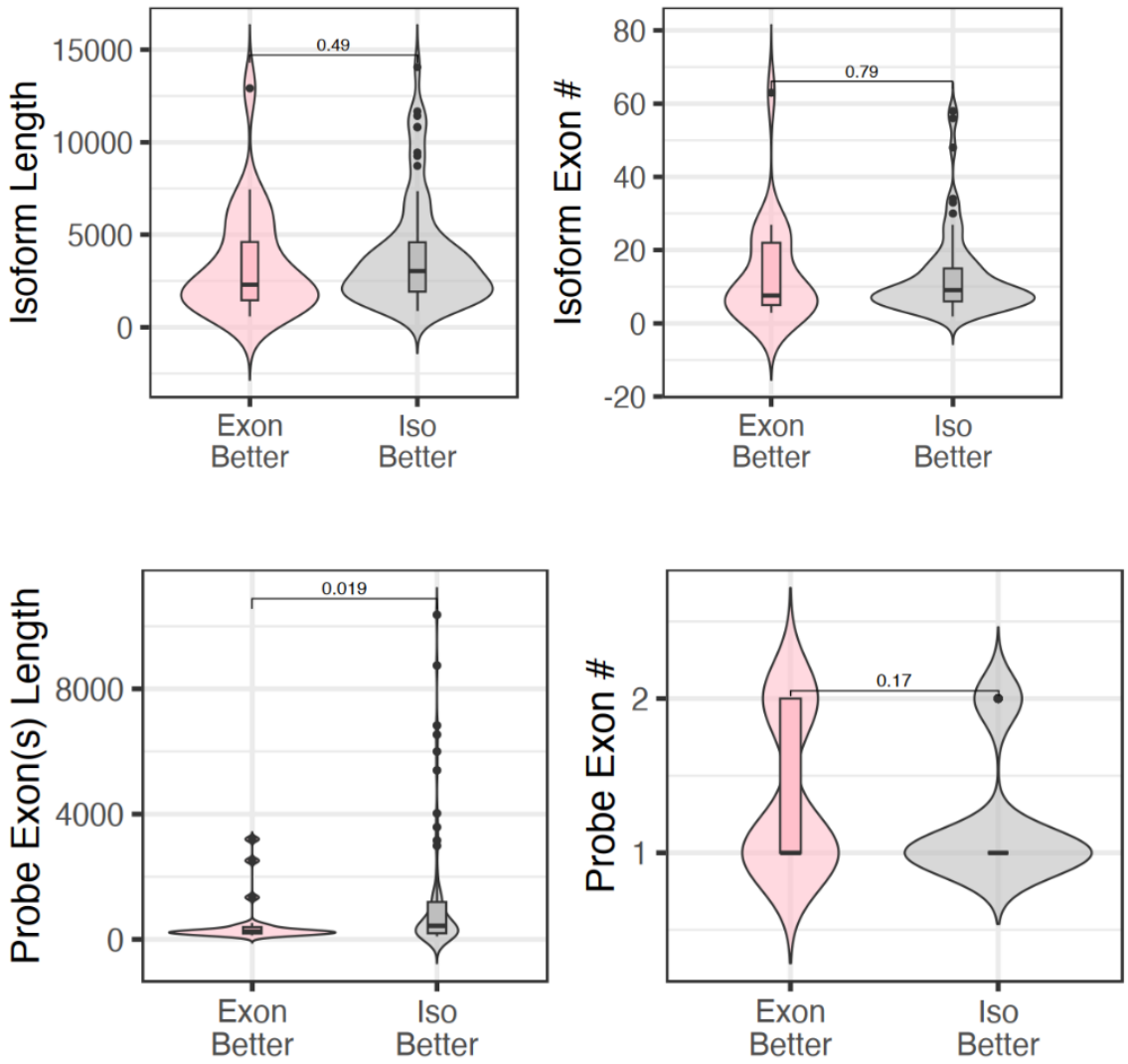

Figure A.10: **Genes with better harmonization from isoform counts tend to have longer exons.** Genes with better harmonization between NanoString and RNA-seq from Exon-based counts (Exon Better - pink) or Isoform-based counts (Iso Better - grey) are compared for their isoform length, number of exons for the isoform (Isoform Exon #), lengths of the exons used for exon-based counting (Probe Exon(s) Length) and the number of exons to which the probe aligned (Probe Exon #). Pairwise comparisons were made using a Student's t-test with p-values shown.

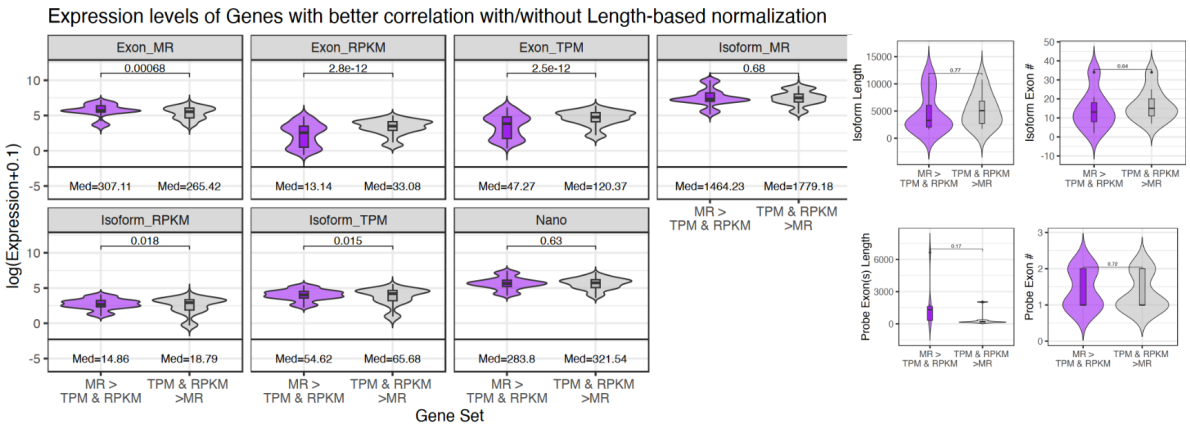

Figure A.11: **There is no clear biological explanation for genes with higher harmonization via a normalization approach** Genes are split into whether non-length based normalization (MR) allowed better harmonization between RNA-seq and NanoString (purple) or if length-based normalization (TPM and RPKM) was better (grey). There is no consistent difference in Expression levels (left) or other features (right) (length of isoforms or exons, number of exons to the isoform or probes) between the two categories. Pairwise comparisons were made using a Student's t-test with p-values shown.

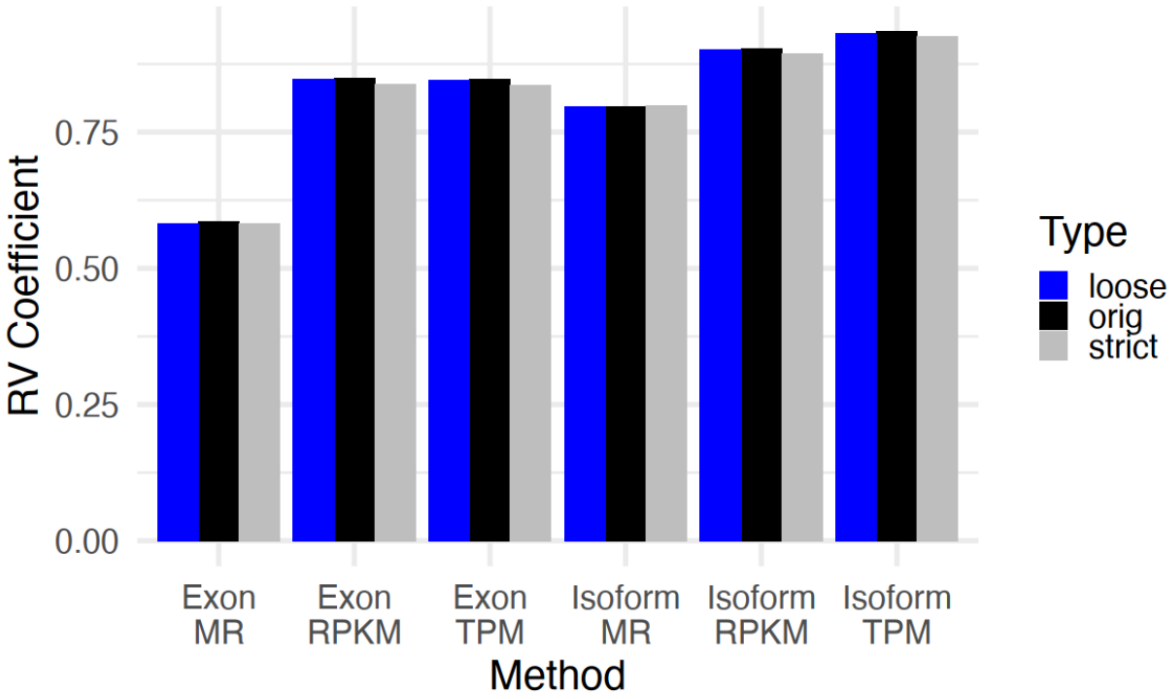

Figure A.12: **Isoform and length-based normalization allow better harmonization in low-dimensional space between RNA-seq and NanoString.** RV coefficients (y-axis) measure the similarity of NanoString and RNA-seq in low-dimensional space when using the different harmonization methods (x-axis), where a higher coefficient indicates more similarity. Original refers to use of all 770 genes while loose and strict refer to when genes not passing the loose and strict limits of detection are removed.

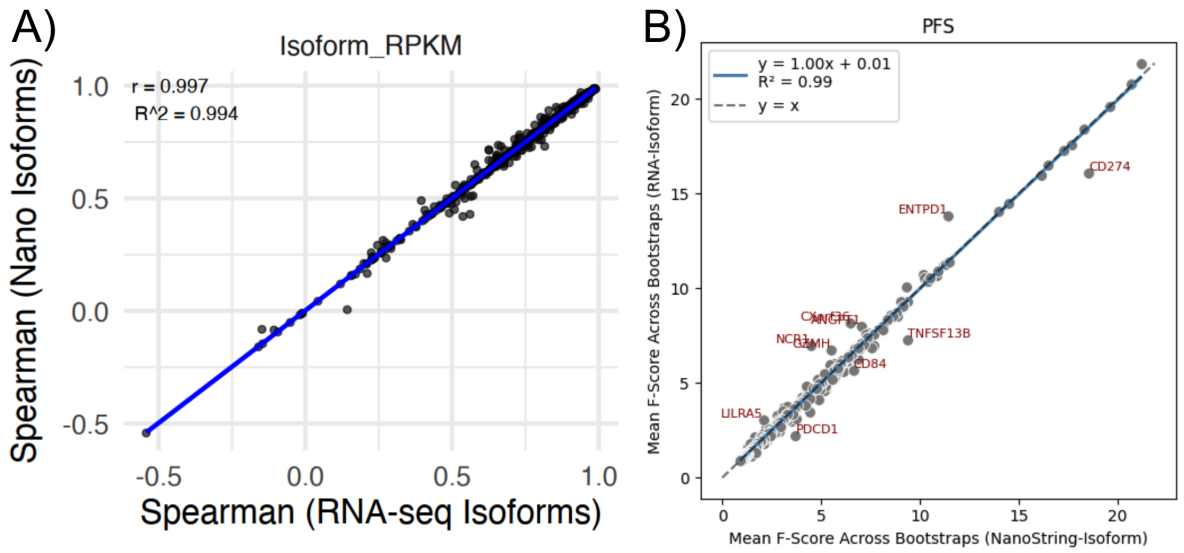

Figure A.13: **Use of RNAseq-based isoforms has limited change on harmonization and predictive comparison between NanoString and RNA-seq A.** RNA-seq isoforms were chosen as the isoform most strongly expressed in TPM compared to other isoforms for a given gene. About 60% of isoforms changed from the original NanoString isoforms. Spearman correlations of the genes that changed isoforms between NanoString and RNA-seq are plotted when using NanoString-based isoforms (y-axis) and RNA-seq based isoforms (x-axis) cross ll three normalization approaches. The line of best fit along with r and R^2 values are indicated. **B.** Mean F-score of genes across bootstraps when using RNA-seq isoforms (y-axis) or NanoString isoforms (x-axis) are plotted with genes farther than 2 standard deviations from the fitted line highlighted.

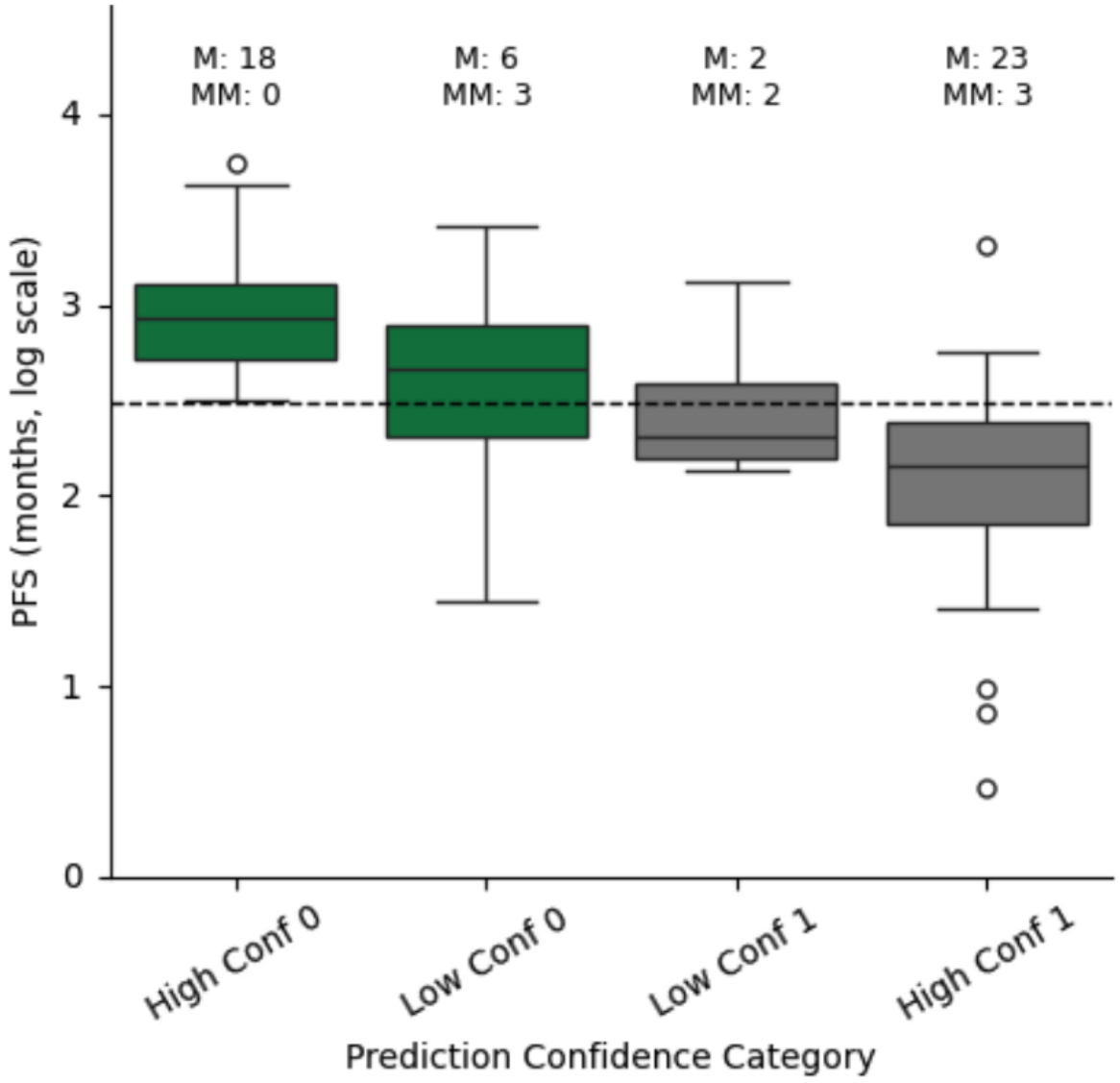

Figure A.14: **Probabilities generated from the RNA-seq only model of PFS seem to reflect confidence of calls.** Patients were split into four categories based on whether the model called them PFS$\leq 12 mths$ (1) with high confidence (probability above 0.6, High Conf 1) or not (Low Conf 1), or PFS$> 12 mths$ (0) with high confidence (probability below 0.4, High Conf 0) or not (Low Conf 0). The true PFS are plotted on the y-axis with a line indicating the 12mth mark. Low confidence patients had PFS values closer to the cutoff of 12 and a higher ratio of mismatched call to reality (MM) to matches (M).

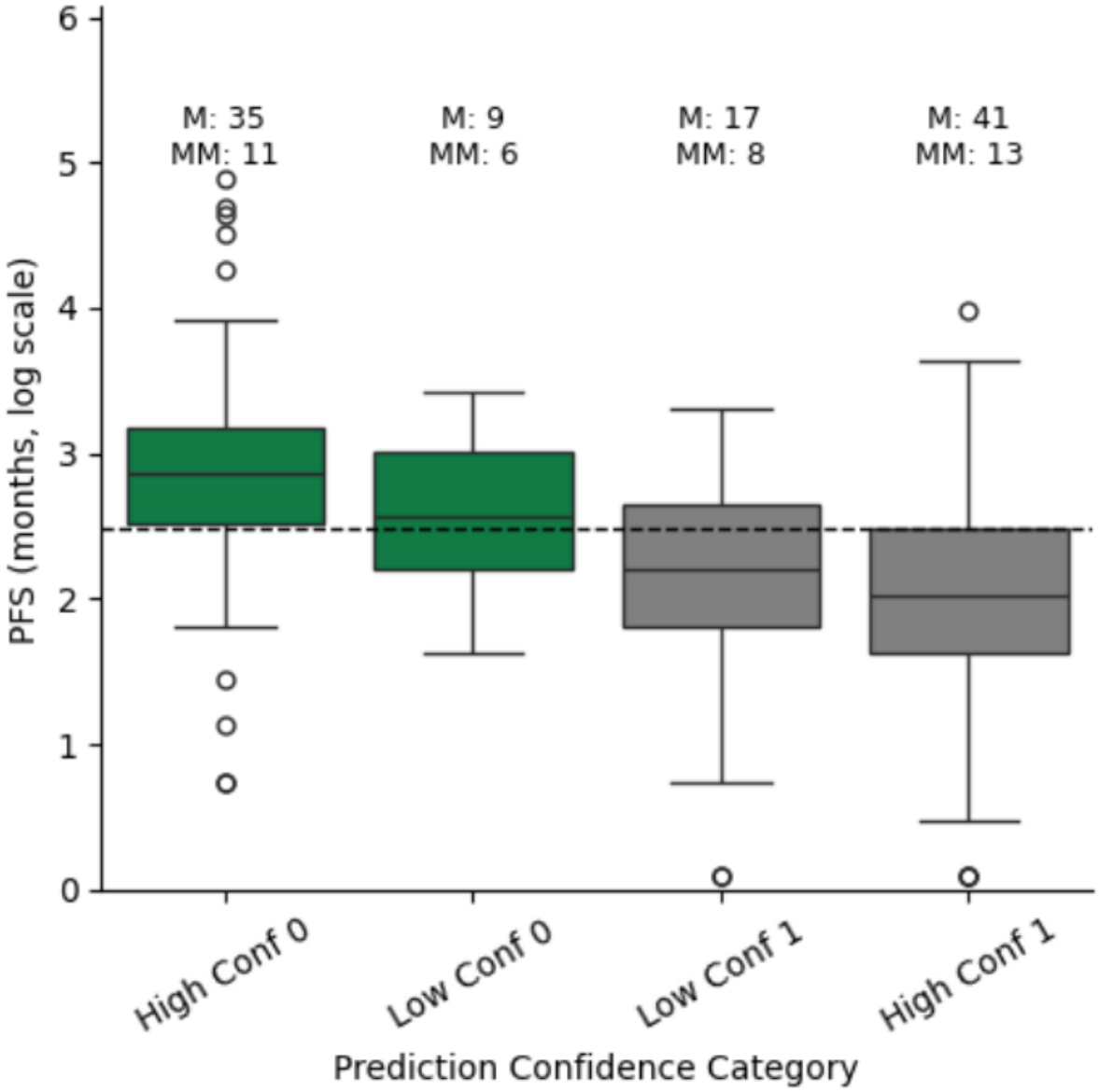

Figure A.15: **Probabilities generated from the NanoString-RNAseq combined model of PFS seem to reflect confidence of calls.** Patients were split into four categories based on whether the model called them PFS$\leq 12 mths$ (1) with high confidence (probability above 0.6, High Conf 1) or not (Low Conf 1), or PFS$> 12 mths$ (0) with high confidence (probability below 0.4, High Conf 0) or not (Low Conf 0). The true PFS are plotted on the y-axis with a line indicating the 12mth mark. Low confidence patients had PFS values closer to the cutoff of 12 and a higher ratio of mismatched call to reality (MM) to matches (M).

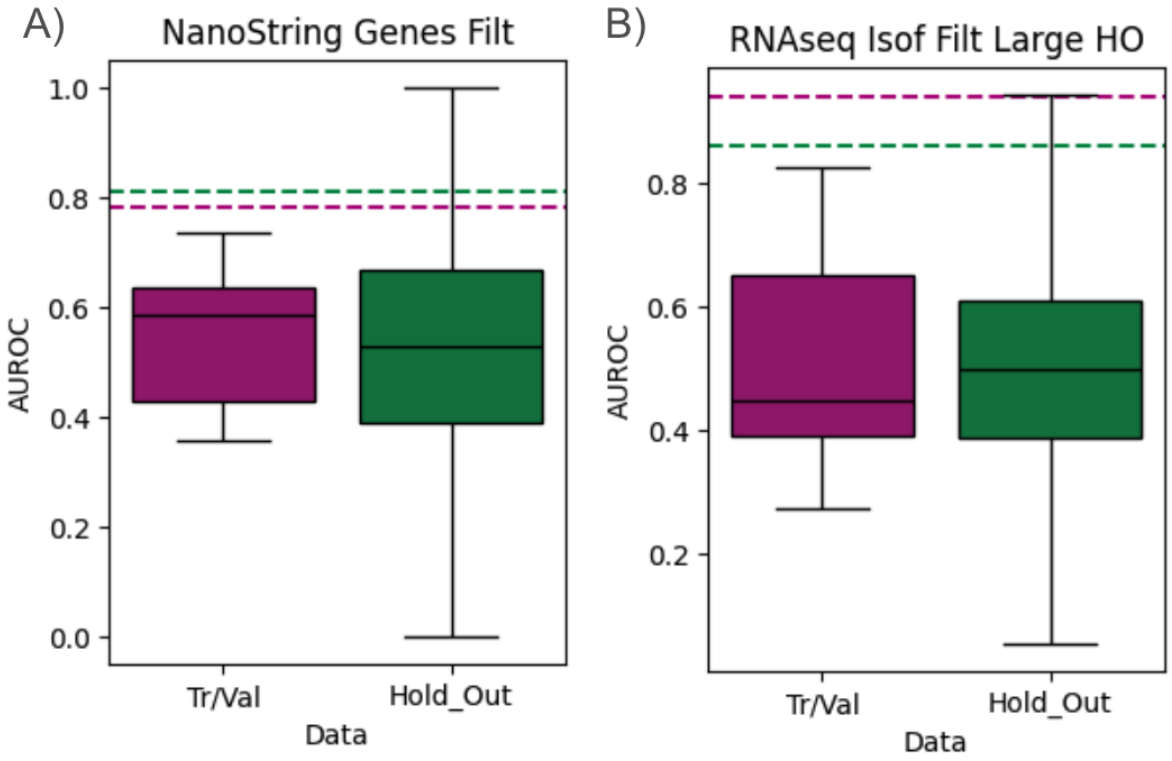

Figure A.16: **Models using random sets of genes cannot recapitulate the same AUROCs of our final models.** Results from considering only the NanoString panel gene and combined dataset are in A while the full RNA-seq genes and RNA-seq dataset is in B. 1000 models were generated, each using random sets of 2 genes with median log2FC above 0.5 with AUROCs recorded for both the full training/validation set (All - purple) and hold out test set (Hold out - green). AUROCs generated from the final model from feature selection are shown as lines. For the combined datasets, all cases where the Hold out AUROC above 0.8 had All AUROC below 0.7 except IL2RA and CENPF (0.71 All and 0.82 AUROC). For the RNA-seq dataset, only 10 combinations had Hold out $AUROC > 0.8$ and All $AUROC > 0.7$ with the highest combination being 0.92 HO / 0.75 All when using genes GNB4 and CTSW.

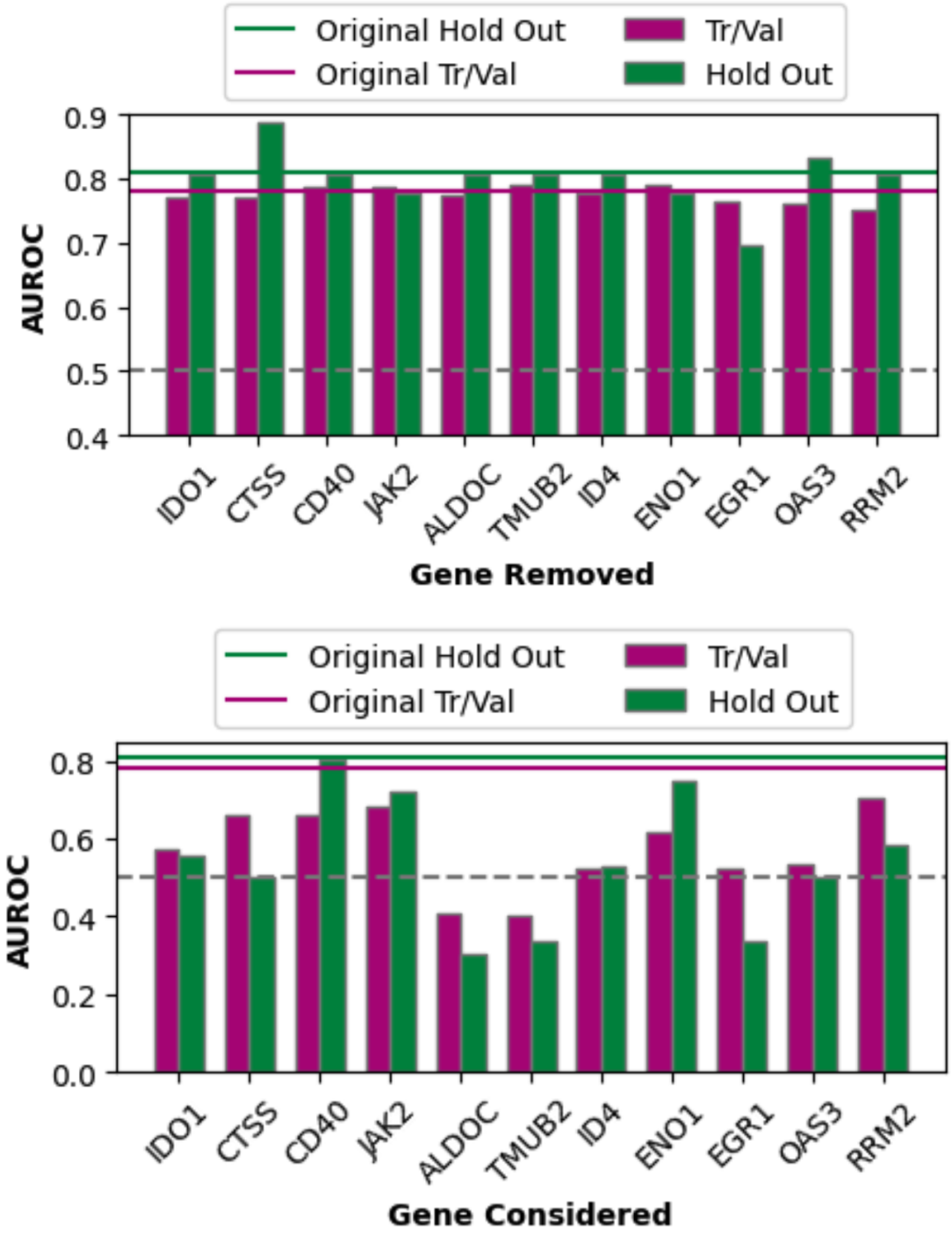

Figure A.17: **There is some redundancy in features in the NanoString-RNAseq PFS model and varying individual predictive power across genes.** AUROCs of models including either all but one of the genes (top) or only a single gene (bottom) for the full training/validation set (All) or the Hold out set. A grey dotted line marks 0.5 AUROC (what is expected from random guessing) and solid lines indicate the AUROCs using the complete original model.

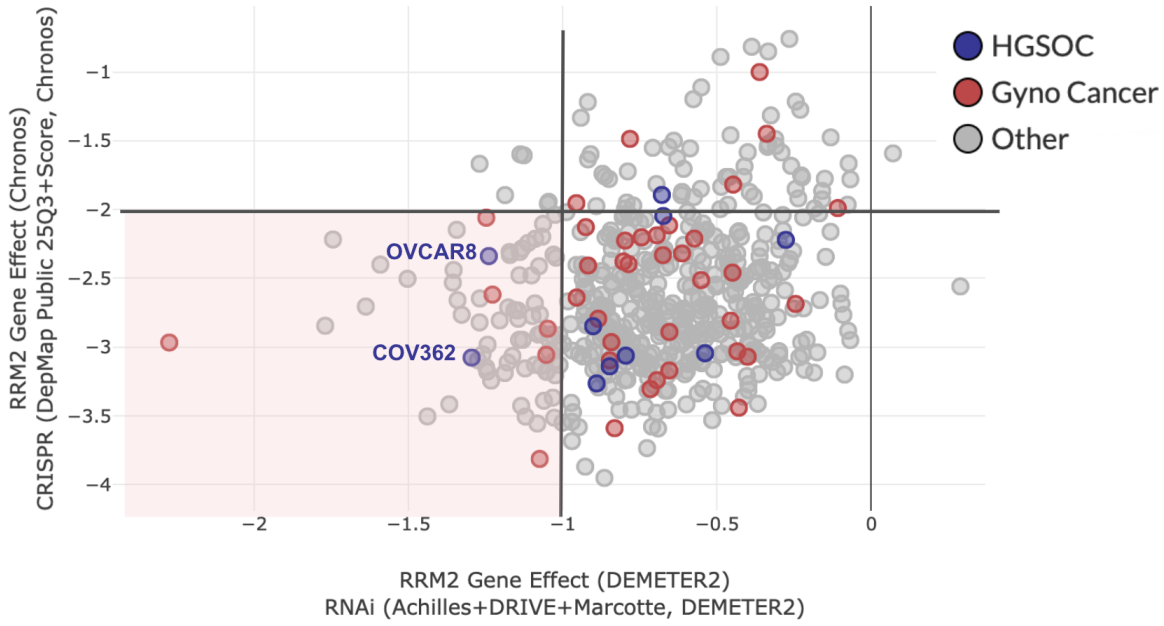

Figure A.18: **RRM2 is a largely essential gene for ovarian cancer cell lines.** Scatterplot of gene effects of RRM2 from DepMap across cancer cell lines from CRISPR (y-axis) or RNA-interference (x-axis) where more negative values indicate more essential genes. DepMap suggested cutoffs of essentiality include -2 for CRISPR (Chronos algorithm) and -1 for RNAi (DEMETER2 algorithm) which are highlighted. Cancer cell lines from HGSOC (purple) and gynecological cancers (red) are highlighted.

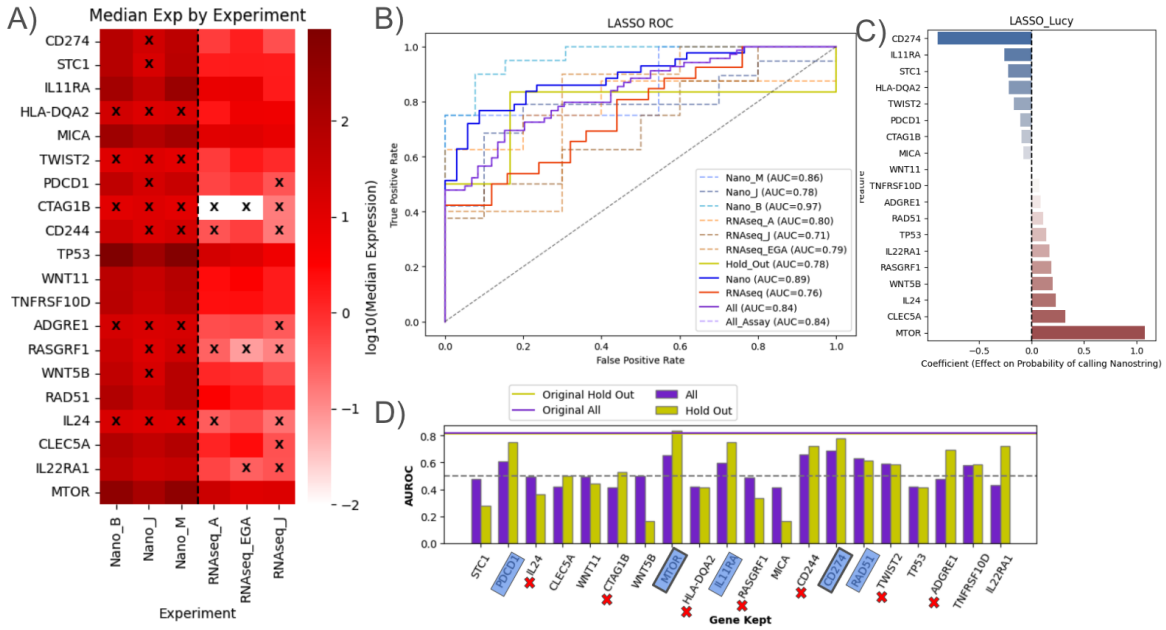

Figure A.19: **Our harmonization approach informs about the top 20 ranked genes according to several different models and the same NanoString data** A) log10(Median expression) of genes across the mini-cohorts (experiments) where an X indicates the value is below the predicted limit of detection, B) AUROC plot for LASSO model using genes in C, C) Coefficient values for genes from the model, D) AUROC values for hold-out and training (All) set when each gene is removed. Xs are placed by the genes with median expression levels below the limits of detection and blue highlights indicate where AUROCs of both training and hold out are above 0.6.

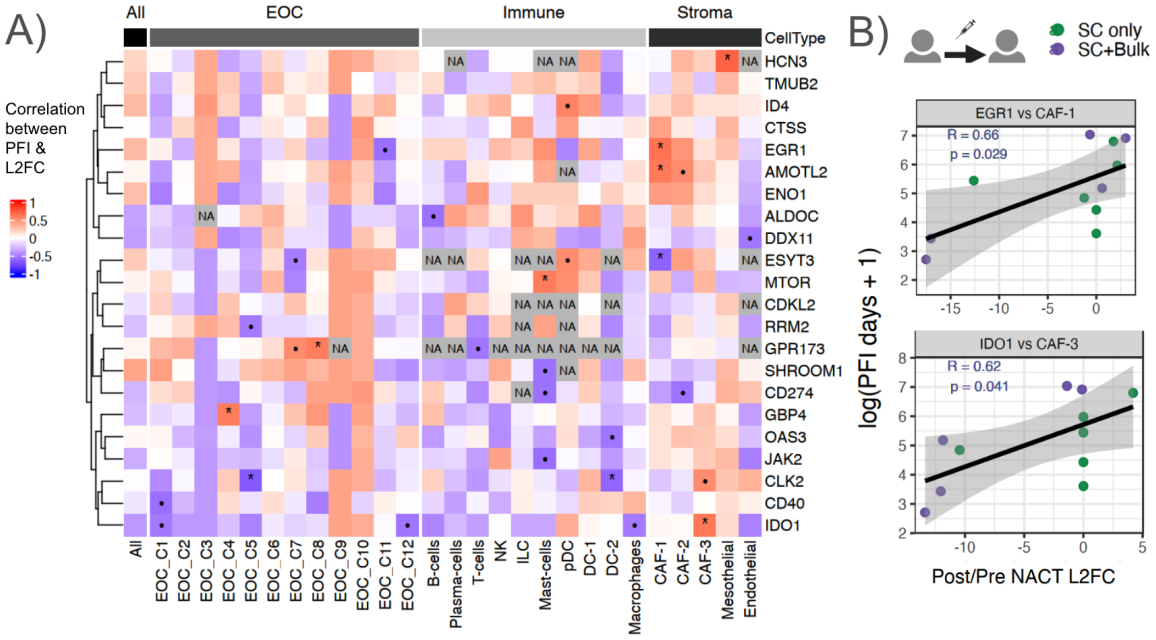

Figure A.20: **Despite low correlation between pseudobulk scRNA-seq and bulk RNA-seq, several model genes show significant correlation with PFI in a cell-state dependent manner A.** Heatmap of Pearson correlation coefficients between Post/Pre NACT log2FC of genes in the pseudobulked cell states and log(PFI+1) of 11 patients. An * indicates a Pearson correlation p-value below 0.05 and • below 0.1. *B.* Scatter plot of log2FC of gene in the noted cell type (x-axis) and log(PFI+1) of patients (y-axis) with the R-squared values noted according to best-fit line and dots colored by whether they have samples in Bulk (SC+Bulk) or just single-cell.

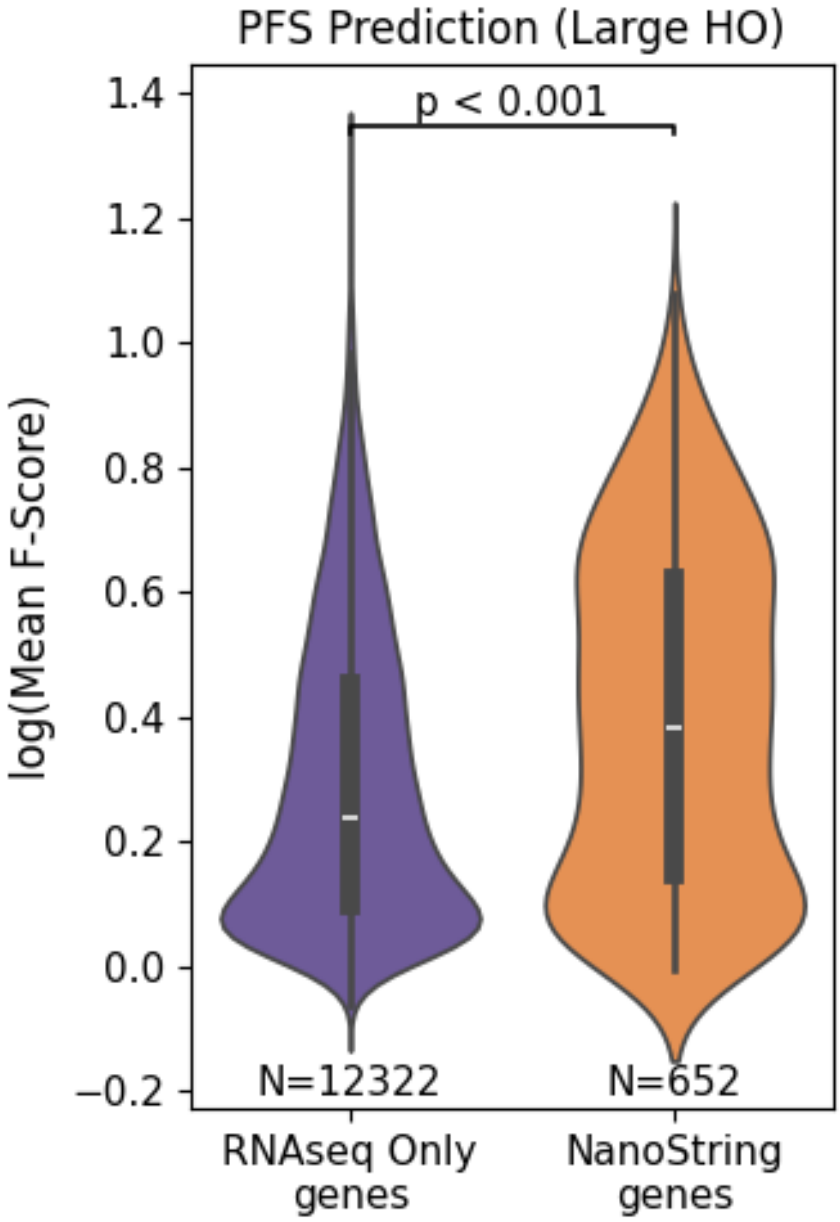

Figure A.21: **NanoString genes had high predictive power compared to most RNA-seq genes.** Mean F-score of either genes only considered in RNA-seq (red) or in both NanoString panel and RNA-seq (blue) across bootstraps when using RNA-seq isoforms. Genes below count thresholds are not included. *GBP4* and *CD274* are NanoString panel genes that were consistently found in the top 25 genes with the highest predictive frequency.

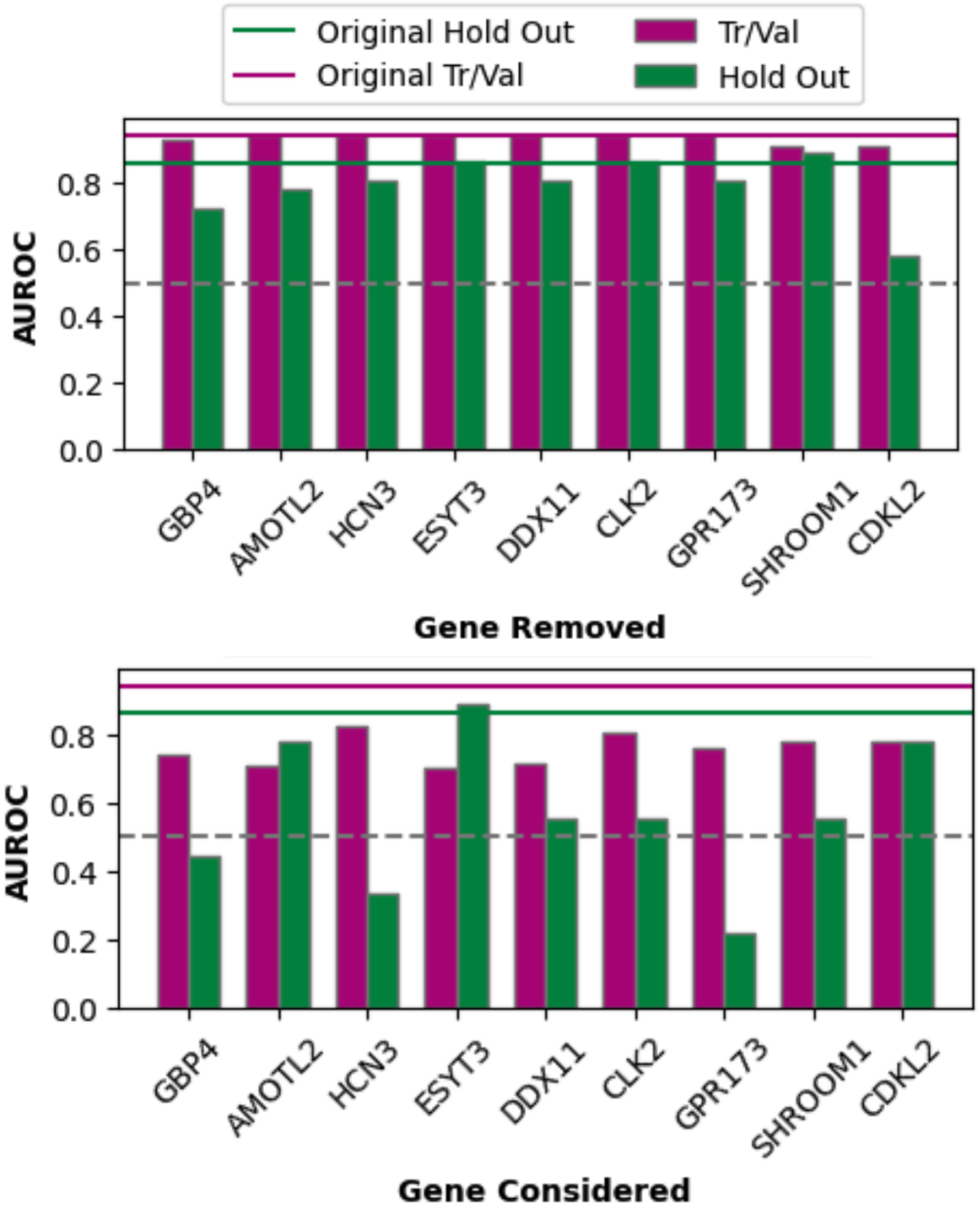

Figure A.22: **There is some redudancy in features in the RNAseq PFS model and varying individual predictive power across genes.** AUROCs of models including either all but one of the genes (top) or only a single gene (bottom) for the full training/validation set (All) or the Hold out set. A grey dotted line marks 0.5 AUROC (what is expected from random guessing) and solid lines indicate the AUROCs using the complete original model. Removal of ESYT3, GPR173, or CLK2 had no impact on model performance despite having varying predictive power.

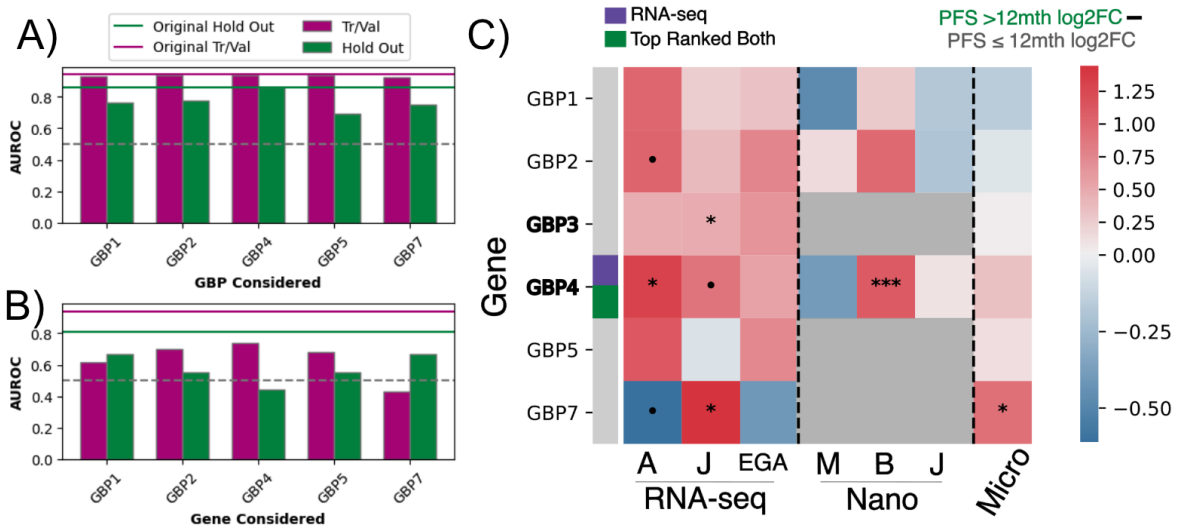

Figure A.23: **Any one of the GBP genes might explain the GBP relevance A.** AUROC for the RNA-seq based model to predict PFS category on RNA-seq data when considering the different GBP genes in the dataset in the full model. **B.** AUROC when only using one of the GBP genes on RNA-seq data to predict PFS category. **C.** Heatmap comparable to Figure 4A of difference in median log2FC between patients with PFS ¿12mth and $\leq$ 12mth. $\bullet$ = p-value < 0.1, * < 0.05, *** < 0.001 for t-test (non-parameteric Mann-Whitney shows comparable results).

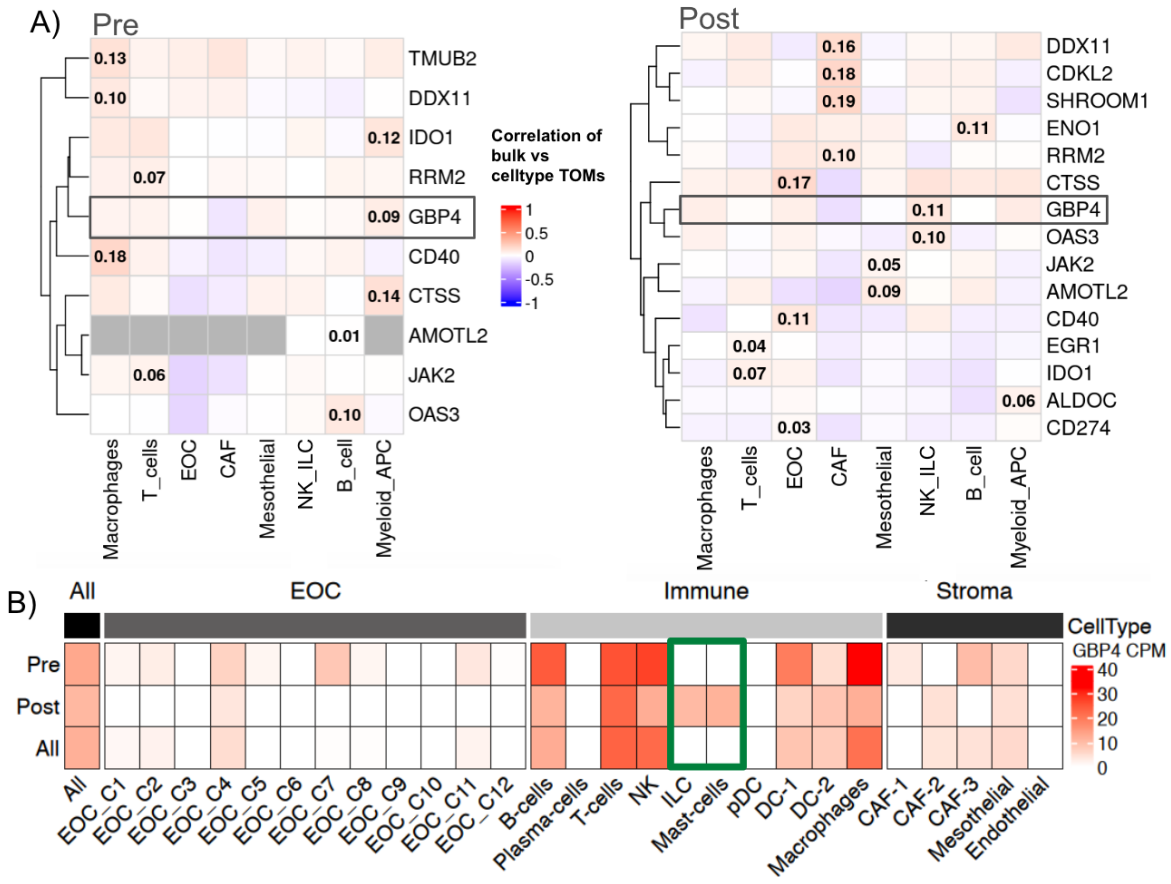

Figure A.24: **Comparing GBP4 bulk expression within the context scRNA-seq A.** Spearman correlation of TOM similarity scores (edges) between model genes and other genes in Pre-NACT and Post-NACT conditions between bulk RNA-seq WGCNA and scRNA-seq hdWGCNA for cell-types. Grey or missing genes indicates that a gene did not pass filters for hdWGCNA for a given or any cell type, respsectively. Low correlation scores are expected given the consideration of 1000s of genes for each. **B.** Counts per million (CPM) within each pseudobulked cell-type from scRNA-seq Pre and Post NACT, or combined (All). Green boxes are around cell types that show an increase in GBP4 expression after NACT, as shows the highest correlation with improved PFS in the bulk model.

Appendix B

Appendix B. Supplemental Material to Evaluating methods for integrating single-cell data and genetics to understand inflammatory disease complexity

B.1 Supplemental Figures

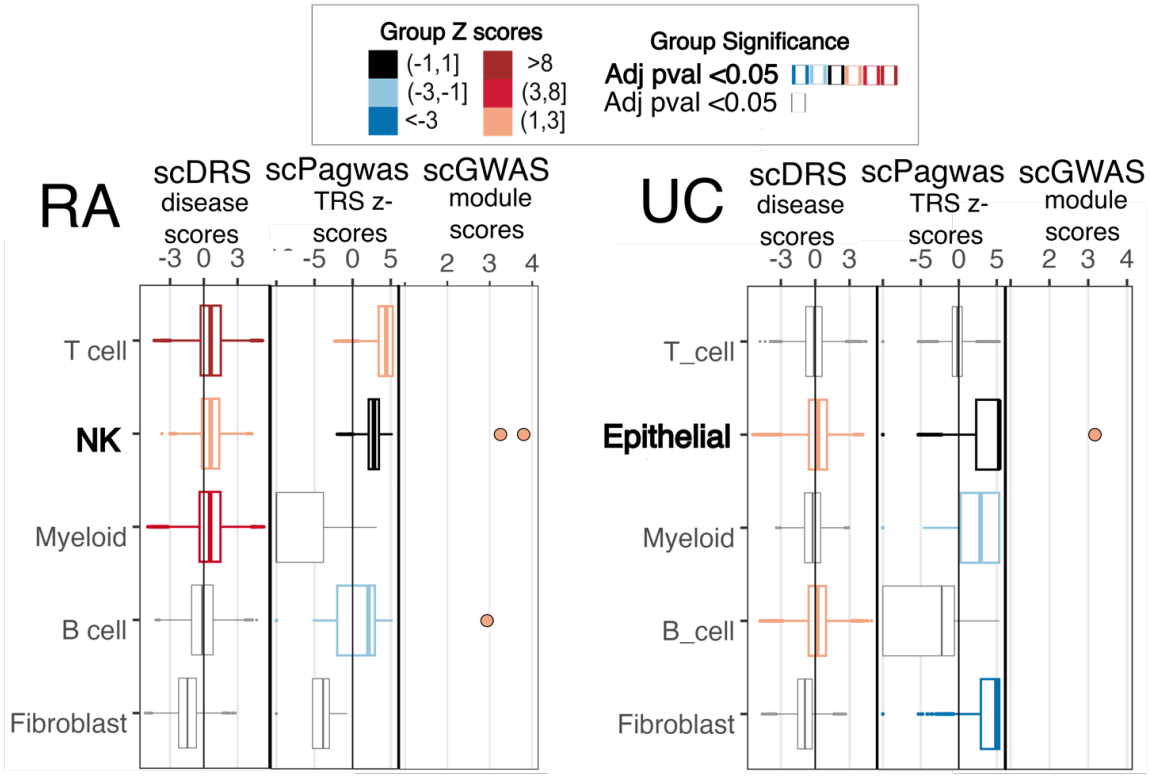

Figure B.1: For each cell-type, the single-cell scDRS Z-scores and scPagwas TRS Z-scores are displayed in boxplots colored according to the group scDRS Z-score or group scPagwas bootstrap Z-score. Non-significant cell types are shown as non-bolded and grey, while significant cell states are bolded. scGWAS called gene modules and their disease scres are plotted with colors according to the scDRS group Z-score gradient for easier comparison. Cell types considered significant by all three tools are bolded. Left: RA (rheumatoid arthritis). Right: UC (ulcerative colitis).

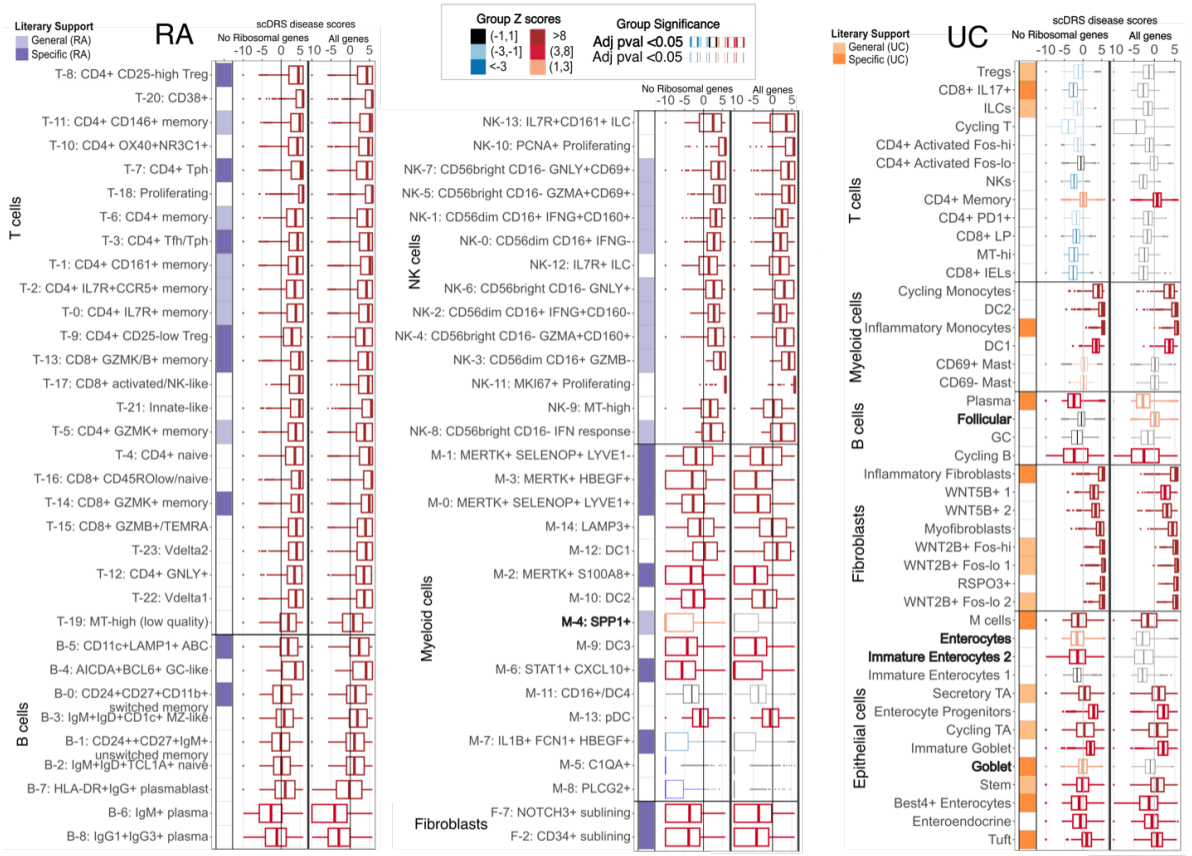

Figure B.2: scDRS results when using 1000 genes with the highest pearson correlation coefficients with scPagwas single-cell genetically associated disease scores, with or without ribosomal genes included. Significant calls have bolded box plots. Single cell scDRS z-scores are plotted as box-plots that are then outlined in the color of the cell group scDRS score. Cell states with different significance calls depending on ribosome gene inclusion are bolded.

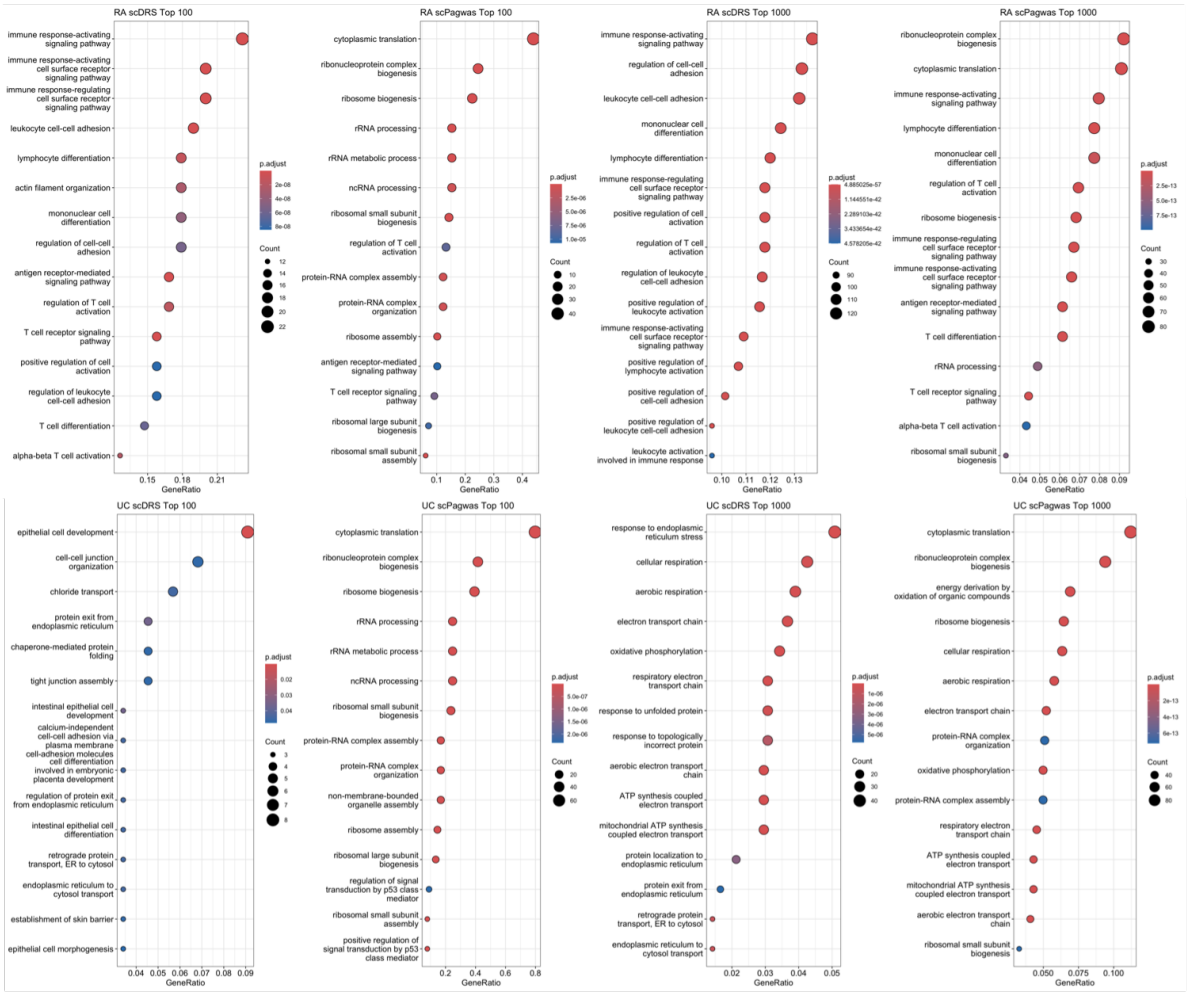

Figure B.3: Top 15 Gene Ontology results for the top 100 and 1000 ranked genes according to correlation with scDRS disease scores and scPagwas genetically associated pathway activity scores. Top is RA and Bottom is UC.

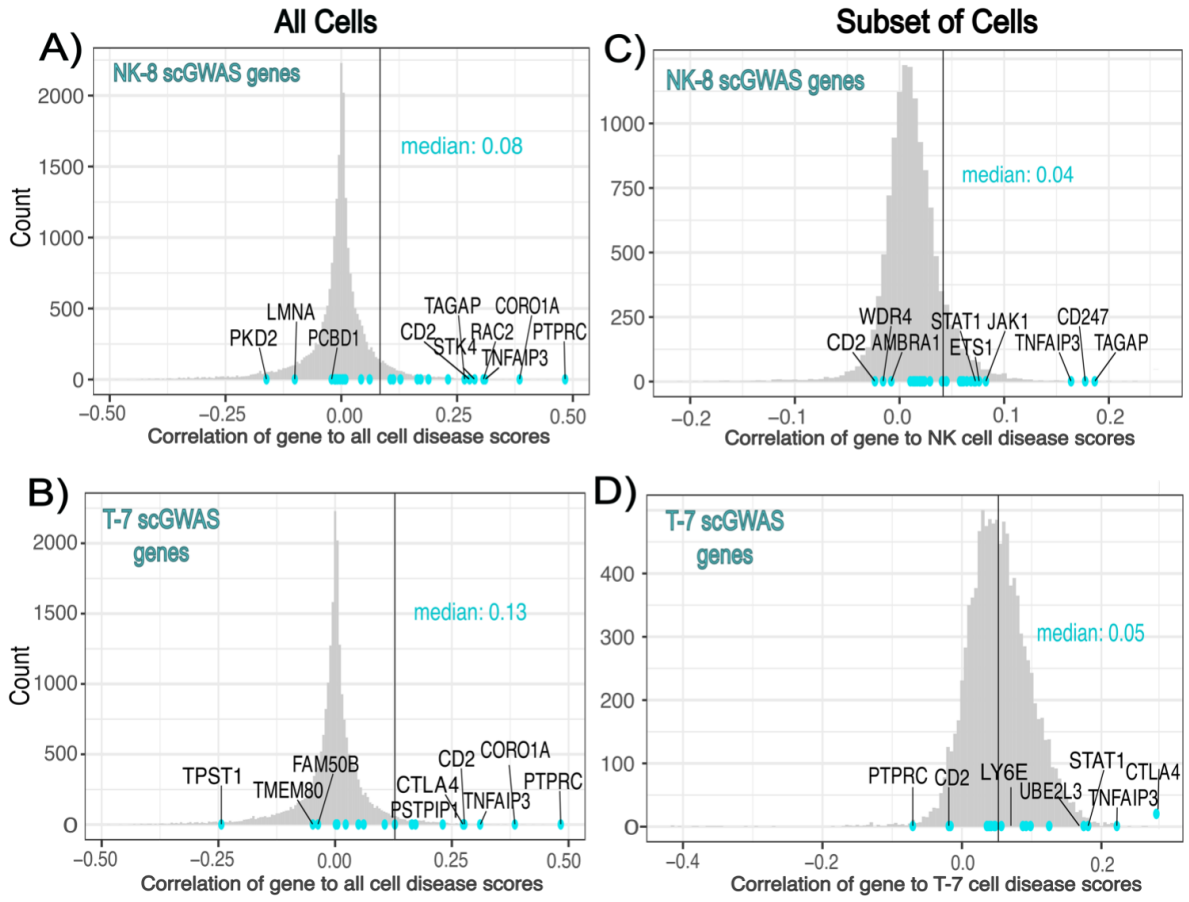

Figure B.4: Histograms of the correlations of all studied genes with scDRS disease scores and scPagwas gPA scores (grey) in all cells (A,B), NK cells (C), or T-7 cells (D) with the appropriately labeled scGWAS module genes highlighted (turquoise). Median correlation score of scGWAS genes is written and shown as a vertical line. scGWAS genes with highest and lowest scores are labeled.

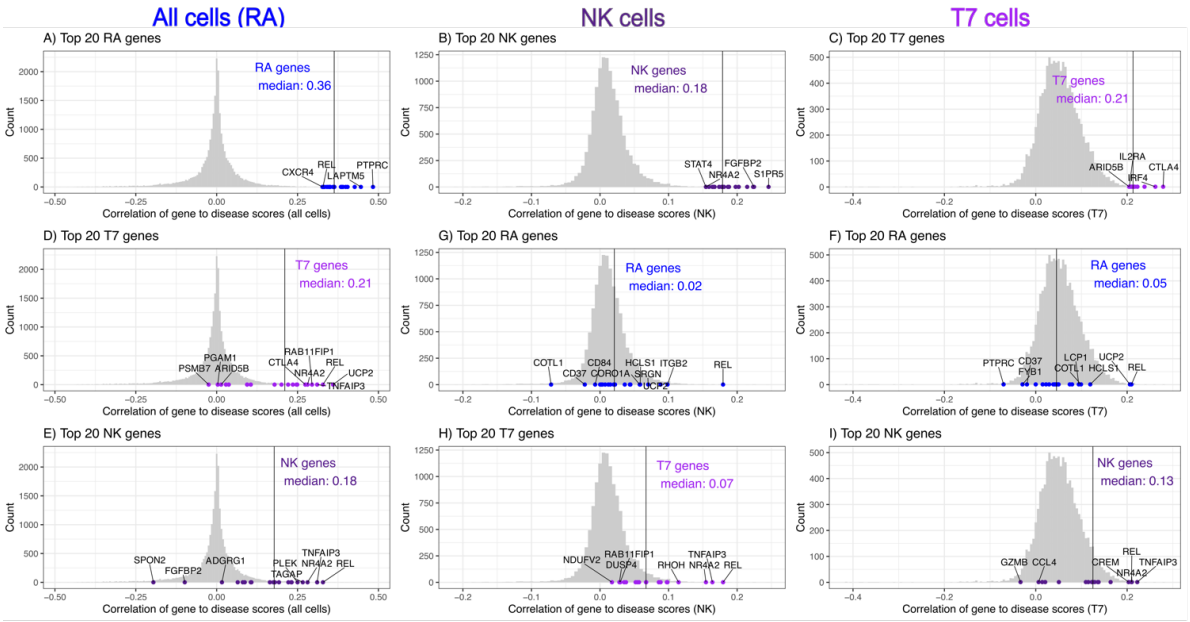

Figure B.5: Histograms of the scDRS correlation scores of the top 20 genes correlated with disease scores in all cells (RA), NK cells, or T-7 cells, within the different cell type options. The genes with the highest and lowest scDRS correlations within the top 20 list are annotated. The median correlation of the top 20 genes are listed. All data is from the RA analysis.

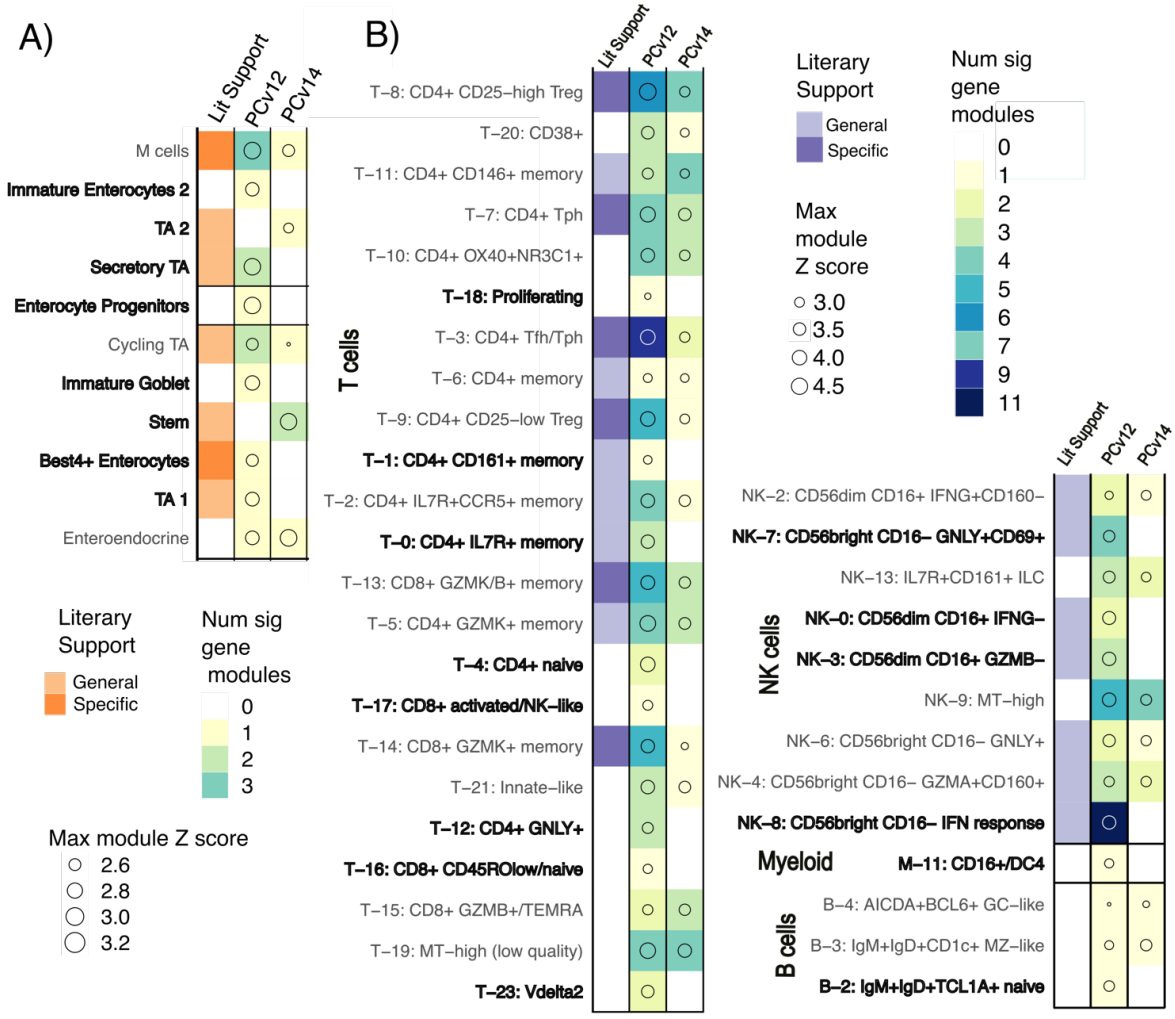

Figure B.6: scGWAS results when using a pathway file based on Pathway Commons v12 or 14 for Ulcerative Colitis with 10kb-10kb MAGMA windows **A.** and Rheumatoid Arthritis with 50-35kb MAGMA windows **B.** Results are highlighted according to the number of significant gene modules called per RA cell state and max disease Z score across the modules for each cell state. Only cell states with a significant gene module from using either pathway file are shown. Cell states without a significant gene module called when only one of the pathway files was used are bolded. Max module Z score refers to the maximum z score value of all significant gene modules called for a cell state.

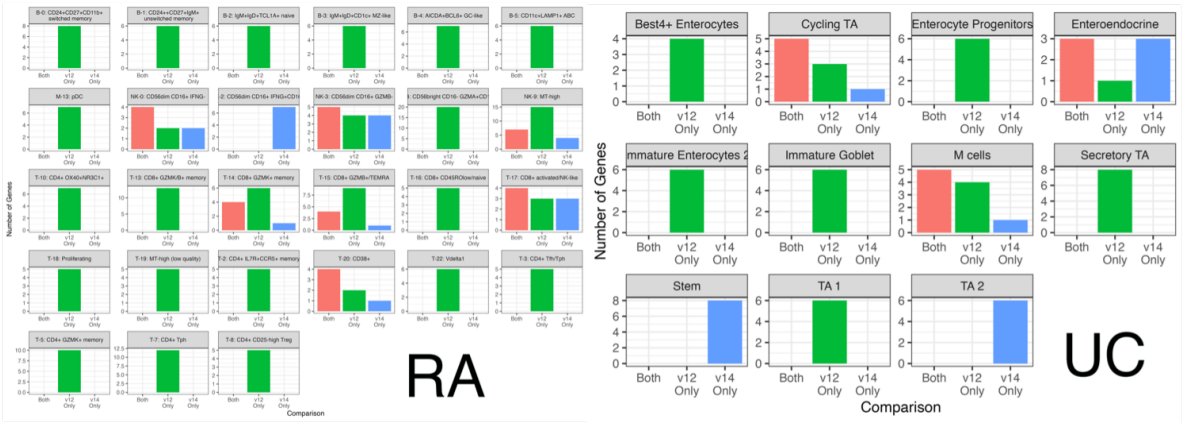

Figure B.7: The number of genes of significant modules of RA and UC cell states called when using input from Pathway Commons v12 (v12 Only), v14 (v14 Only), or either (Both) for 10-10kb MAGMA windows. Only cell states with significant gene modules are shown.

Figure B.8: The number of genes of significant modules of RA cell states called when using input from Pathway Commons v12 (v12 Only), v14 (v14 Only), or either (Both) for 50-35kb MAGMA windows. Only RA cell states with significant gene modules are shown.

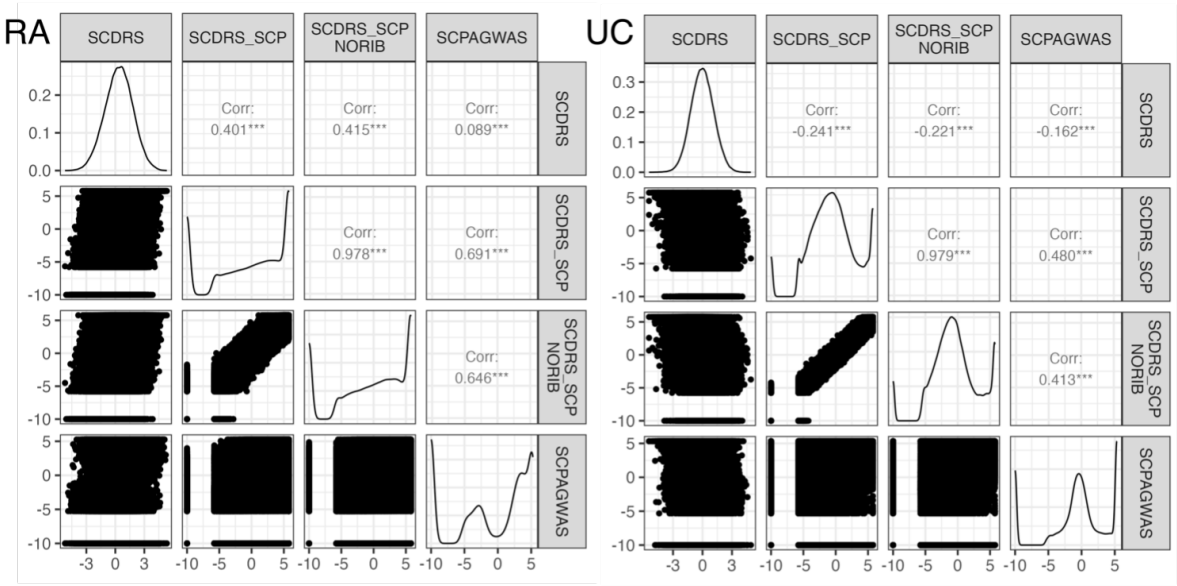

Figure B.9: Distribution and correlation of scDRS z-scores with MAGMA input (SCDRS), scPagwas gene input (SCDRS_SCP), scPagwas gene input without ribosomal genes (SCDRS_SCP NORIB), and scPagwas z-scores for trait relevant scores (SCPAGWAS). Diagonals show the distributions of the scores and correlation coefficients were calculated with the Spearman method using base R function cor. Left: Rheumatoid arthritis (RA), Right: Ulcerative colitis (UC).

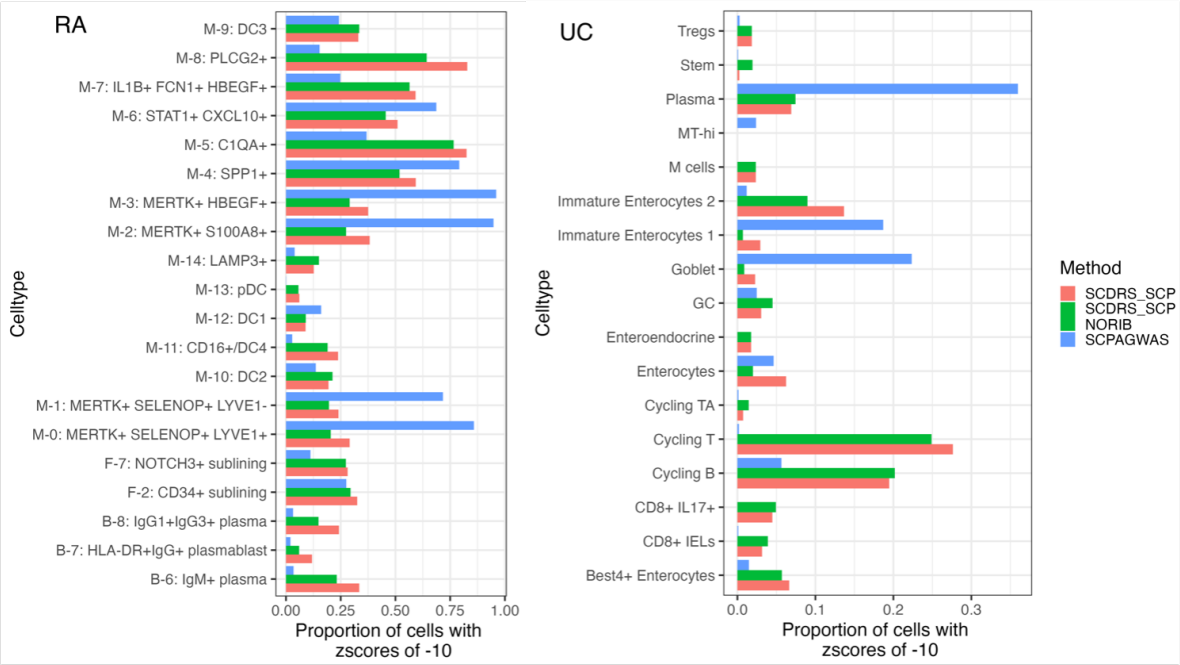

Figure B.10: The proportion of cells within each cell type that have disease zscores of -10 according to scDRS z-scores using scPagwas gene input (SCDRS_SCP) or scPagwas gene input without ribosomal genes (SCDRS_SCP NORIB), and scPagwas z-scores for trait relevant scores (SCPAGWAS). Left: Rheumatoid arthritis (RA), Right: Ulcerative colitis (UC).

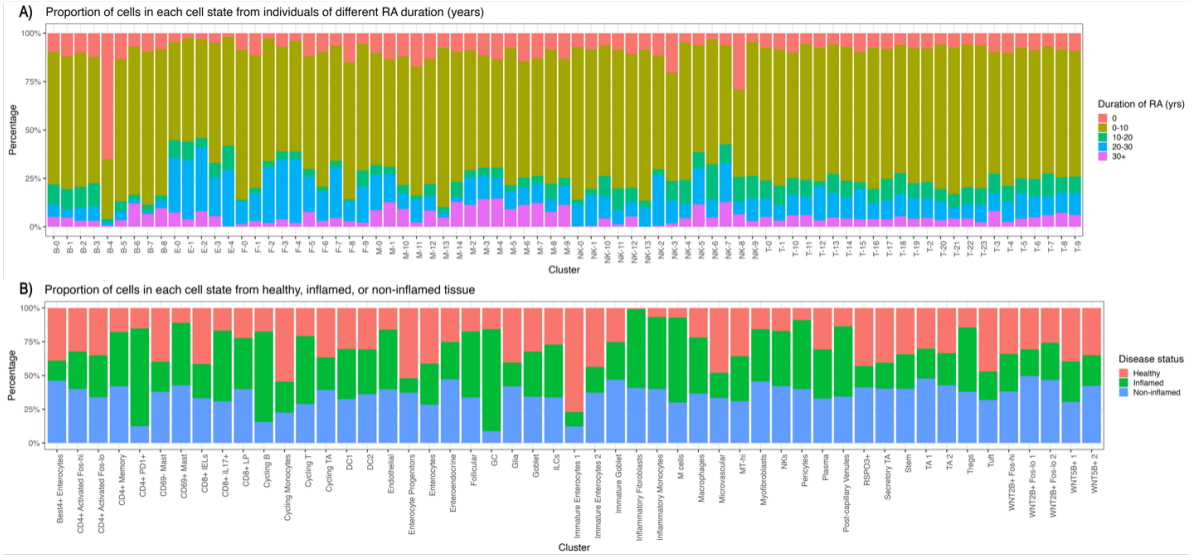

Figure B.11: **Proportion of disease status across cell-states.** The proportion of cells belong in either: A) RA duration for RA and B) disease tissue status for UC were graphed for each of the relevant cell states.

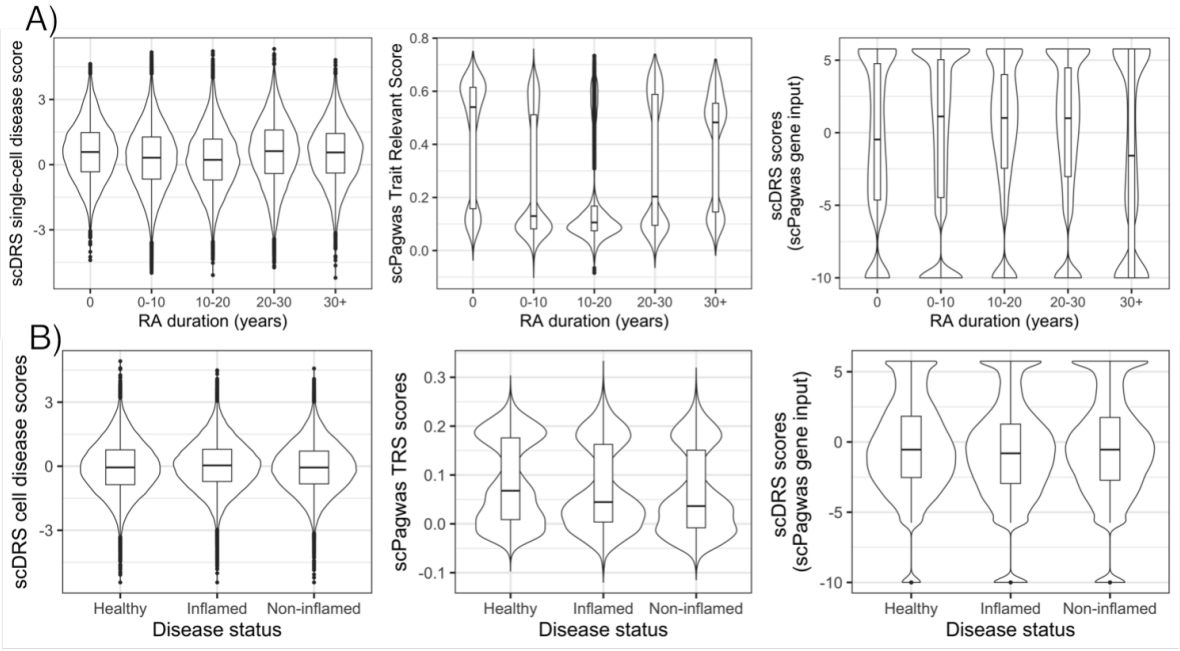

Figure B.12: **Single-cell disease scores connection to disease status.** Single- cell disease scores (from scDRS, scDRS with scPagwas input, and scPagwas) were graphed for cells belonging in difference disease annotations: A) RA duration for RA and B) disease tissue status for UC. ANOVA was followed by Tukey multiple comparison analysis (95% confident intervals).

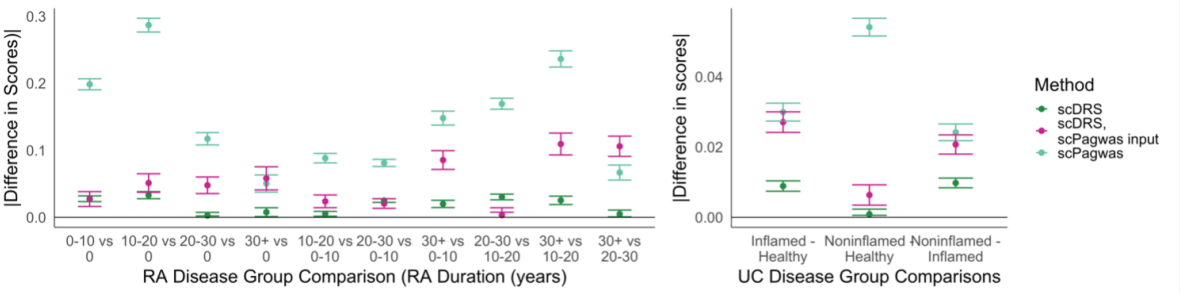

Figure B.13: **scDRS improves covariate bias**. Confidence intervals around mean differences (absolute value) from Tukey-based ANOVA post-hoc comparison analysis of cell disease scores (from scDRS, scDRS with scPagwas input, and scPagwas) according to the disease status of cells (RA duration in years for RA (left) and Inflammation status for UC (right)). Confidence intervals touching the 0.0 line indicate nonsignificant differences in the cell disease scores between the compared groups. Single cell disease scores were scaled from 0 to 1 using min-max scaling to allow easier comparison of differences. Exact Tukey and ANOVA results can be found at our github under (SCRNA-GWAS-Benchmarking/analysis/0A_scGWAS_scDRS/Sensitivity).

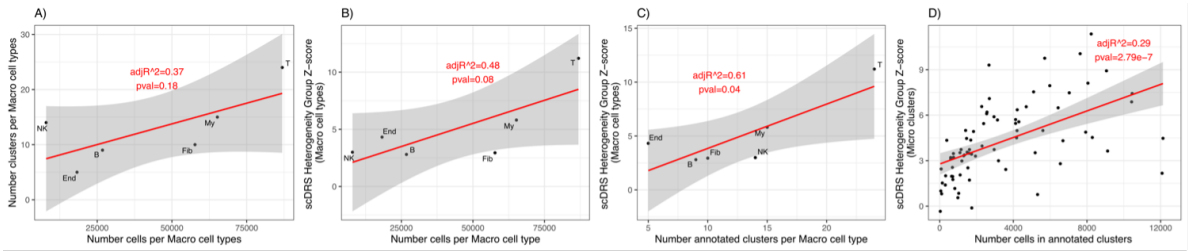

Figure B.14: Linear regression between heterogeneity score and number of clusters/cells in RA scRNA-seq data. The adjusted R2 and model p-value (F-test) are included. Left: the number of cells in Macro-cell types (T-cell, B-cell, Myeloid (My), NK, Fibroblast (Fib), Endothelial (End)) and their number of annotated clusters **(A)** and group scDRS disease score heterogeneity z-scores **(B)** (N=6). **C.** Number of annotated clusters in large-cell types and group scDRS disease score heterogeneity z-scores (N=6). **B.** the number of clusters in large-cell types and their group scDRS disease score heterogeneity z-scores. **D.** Number of cells in annotated clusters, and their group heterogeneity z-scores (N=77). Details of the linear regression results can be found in the jupyter notebook heterogeneity.ipynb on github. MAGMA window of 50-35kb was used.

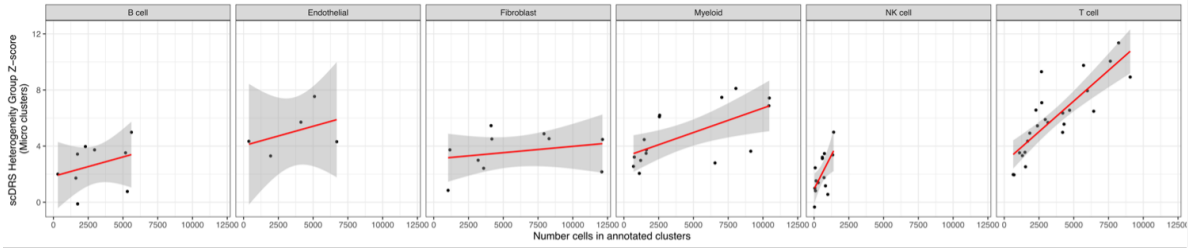

Figure B.15: Linear regression between heterogeneity score and number of cells in each cluster annotated from RA scRNA-seq data, separated by cell type. The adjusted R2 and model p-value (F-test) are included. Details of the linear regression results can be found in the jupyter notebook heterogeneity.ipynb on github. MAGMA window of 50-35kb was used.

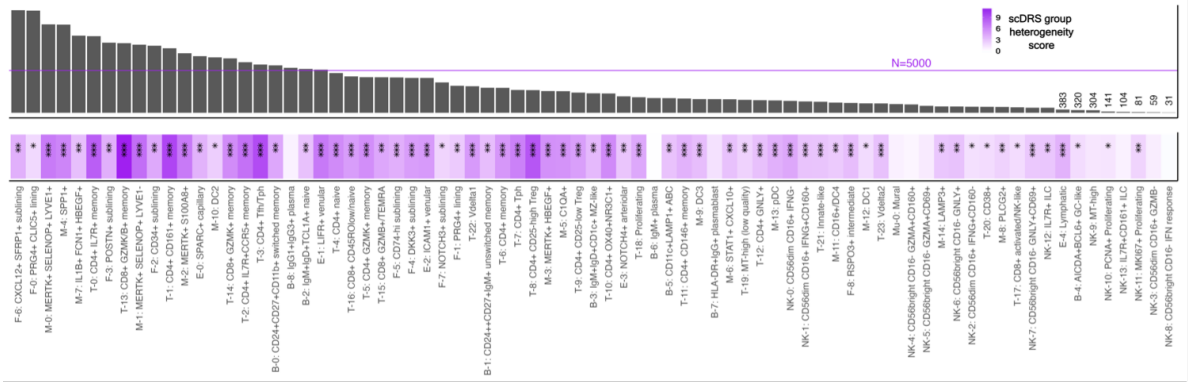

Figure B.16: The scDRS group heterogeneity scores [of disease scores] of cell clusters from RA and the size of said clusters. Any cluster below 500 cells is noted by the number of cells while the rest have a N=5000 bar for reference with the largest number being 12k. Significance legend: *P$<$ 0.05, **P$<$ 0.01, ***P$<$ 0.001. MAGMA window of 50-35kb was used.

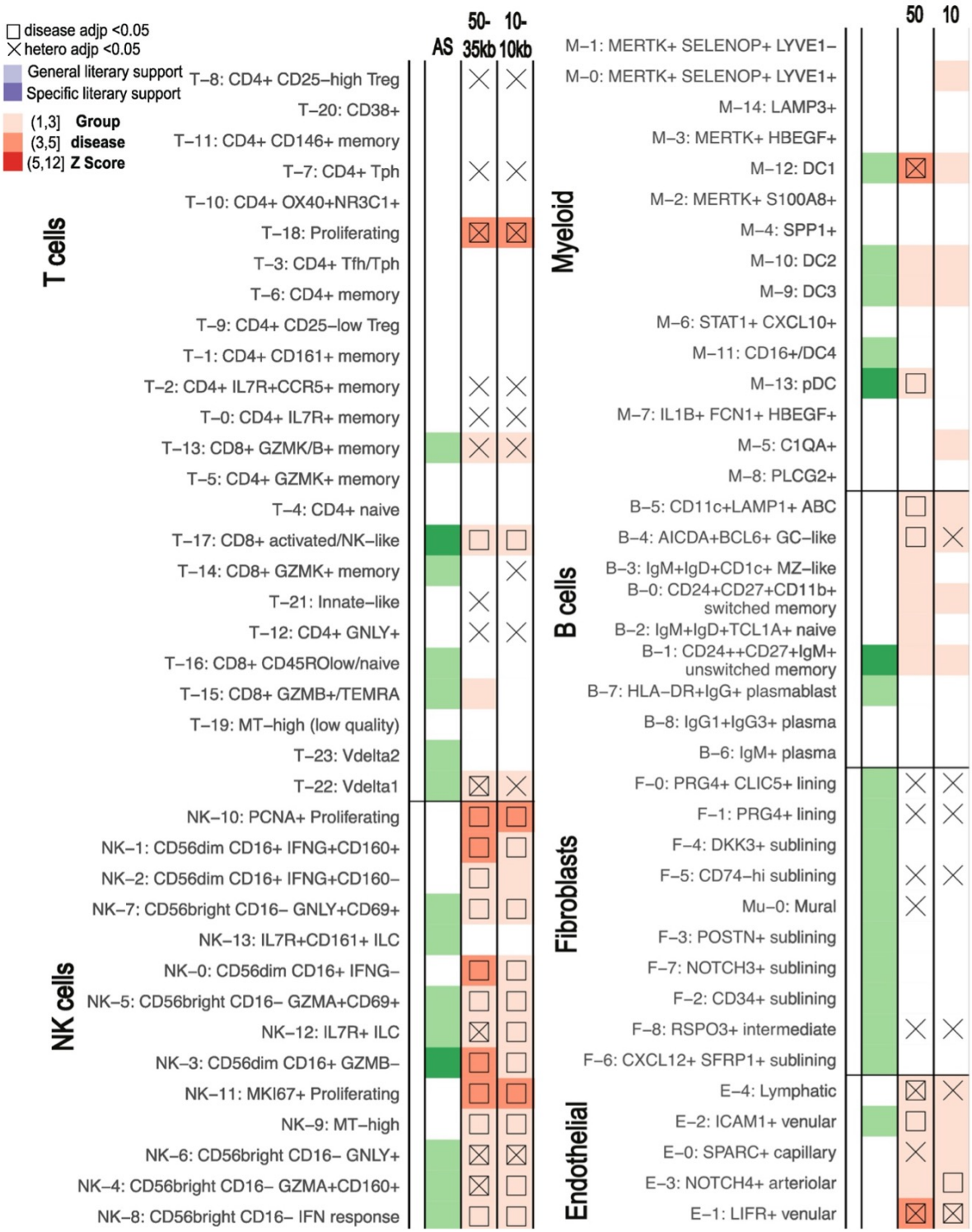

Figure B.17: **MAGMA window comparisons for ankylosing spondylitis**. scDRS results of significant clusters for AS using 50-35kb and 10-10kb windows. scDRS defines significant clusters with a group disease Z-score as shown in the gradient legend. AS = ankylosing spondylitis.

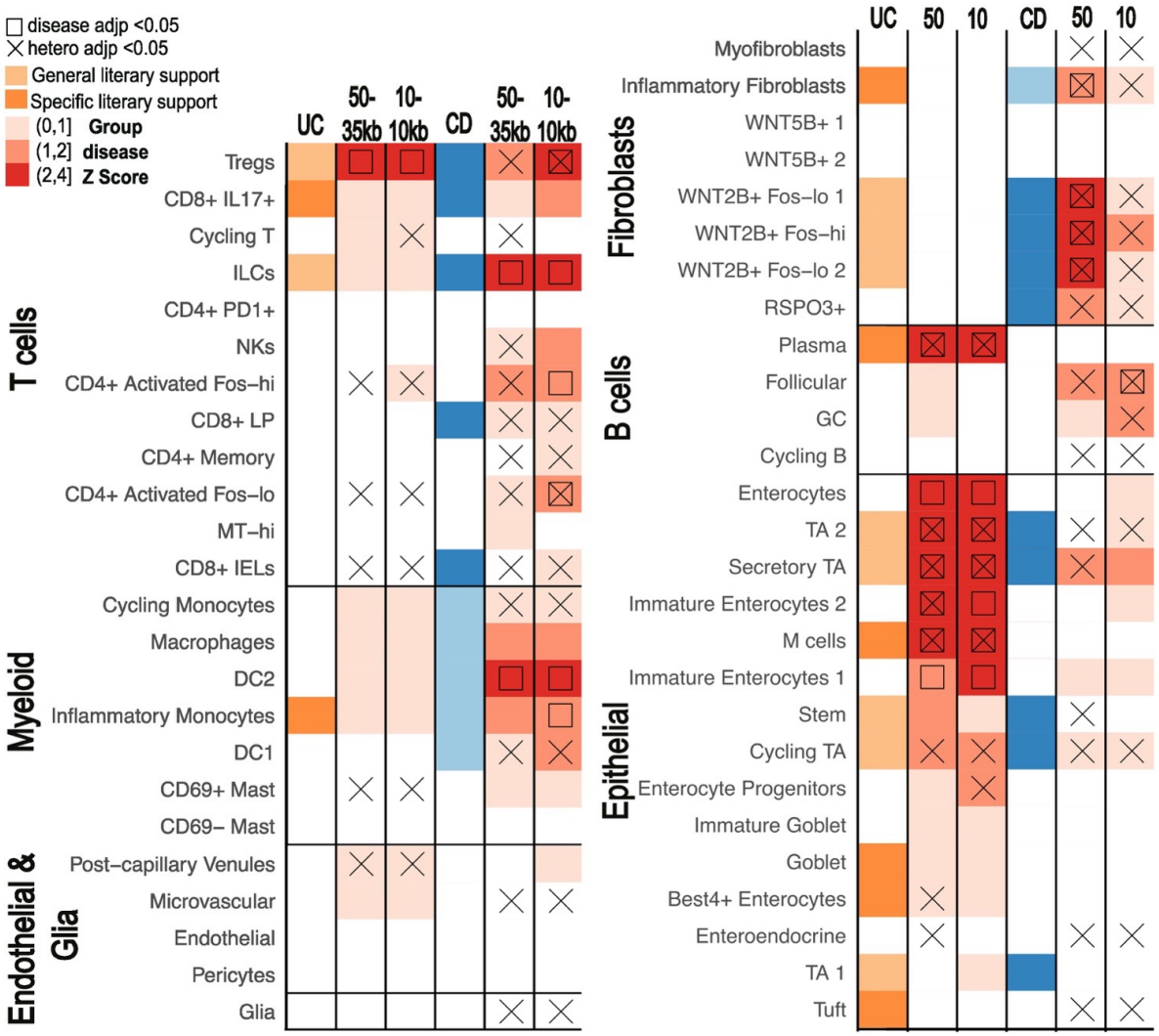

Figure B.18: **MAGMA window comparisons for ulcerative colitis and crohn's disease.** scDRS results of significant clusters for UC and CD using 50-35kb and 10-10kb windows. scDRS defines significant clusters with a group disease Z-score as shown in the gradient legend. Significance legend: *P< 0.05, **P< 0.01, ***P< 0.001. UC=ulcerative colitis, CD=crohn's disease.

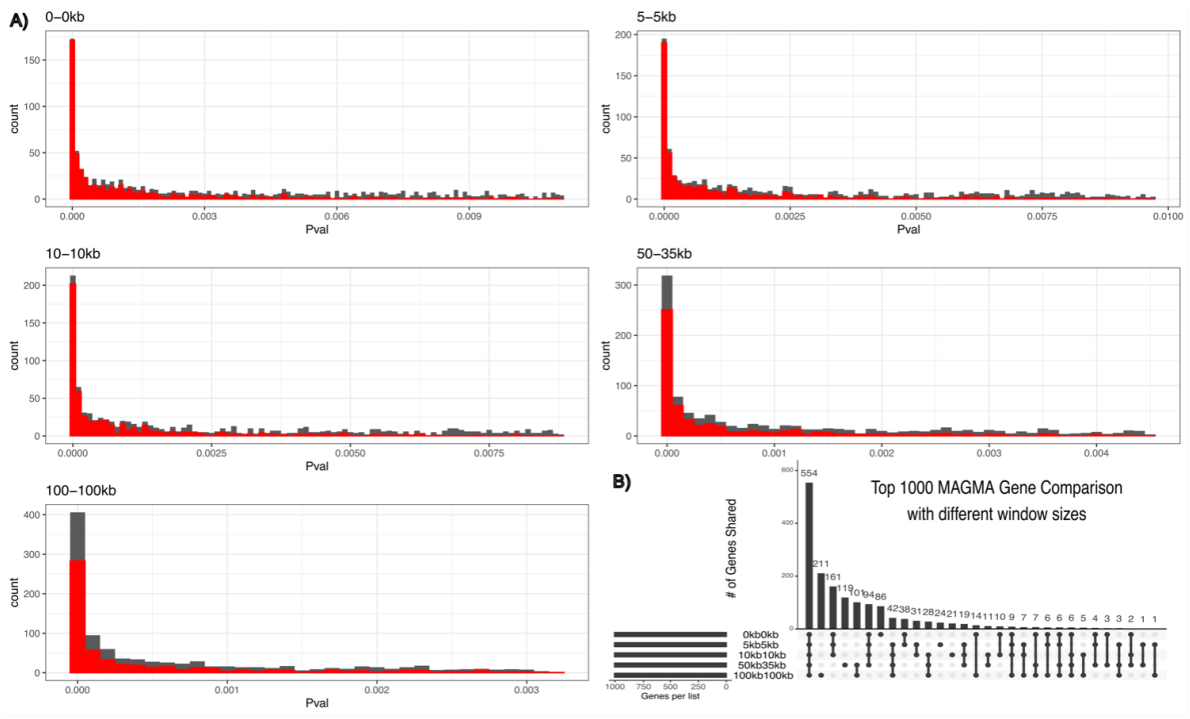

Figure B.19: **A.** The number of genes shared between the top 1000 genes deemed significant by MAGMA that also were found in scRNA-seq, with 5 different windows (0kb-0kb, 5kb-5kb, 10kb-10kb, 50kb-35kb, 100kb-100kb) **B.** The distribution of p-values of the the same top 1000 genes in B) with the distribution of the shared 554 genes highlighted in red.

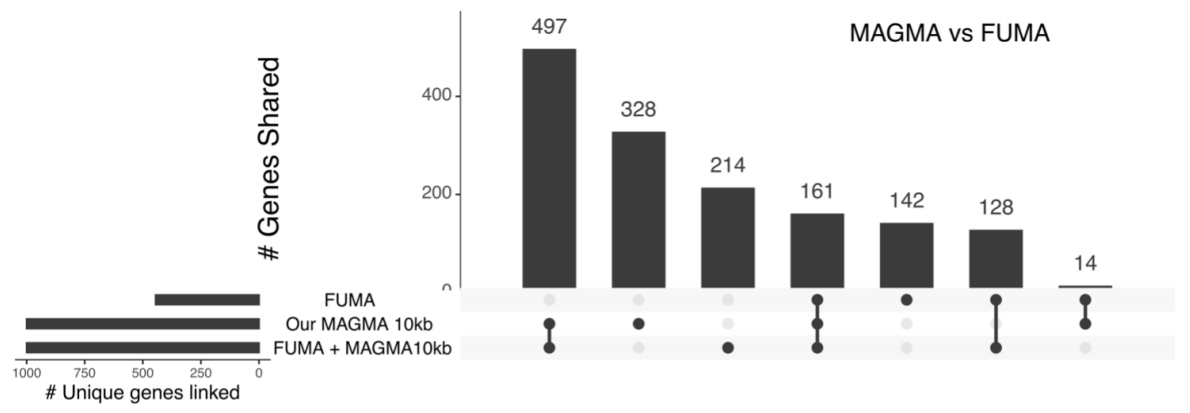

Figure B.20: UpSet plot of the mapped genes according to FUMA (FUMA), MAGMA run on FUMA's final summary statistics (FUMA+MAGMA10kb), and MAGMA run on our final summary statistics (Our MAGMA 10kb).

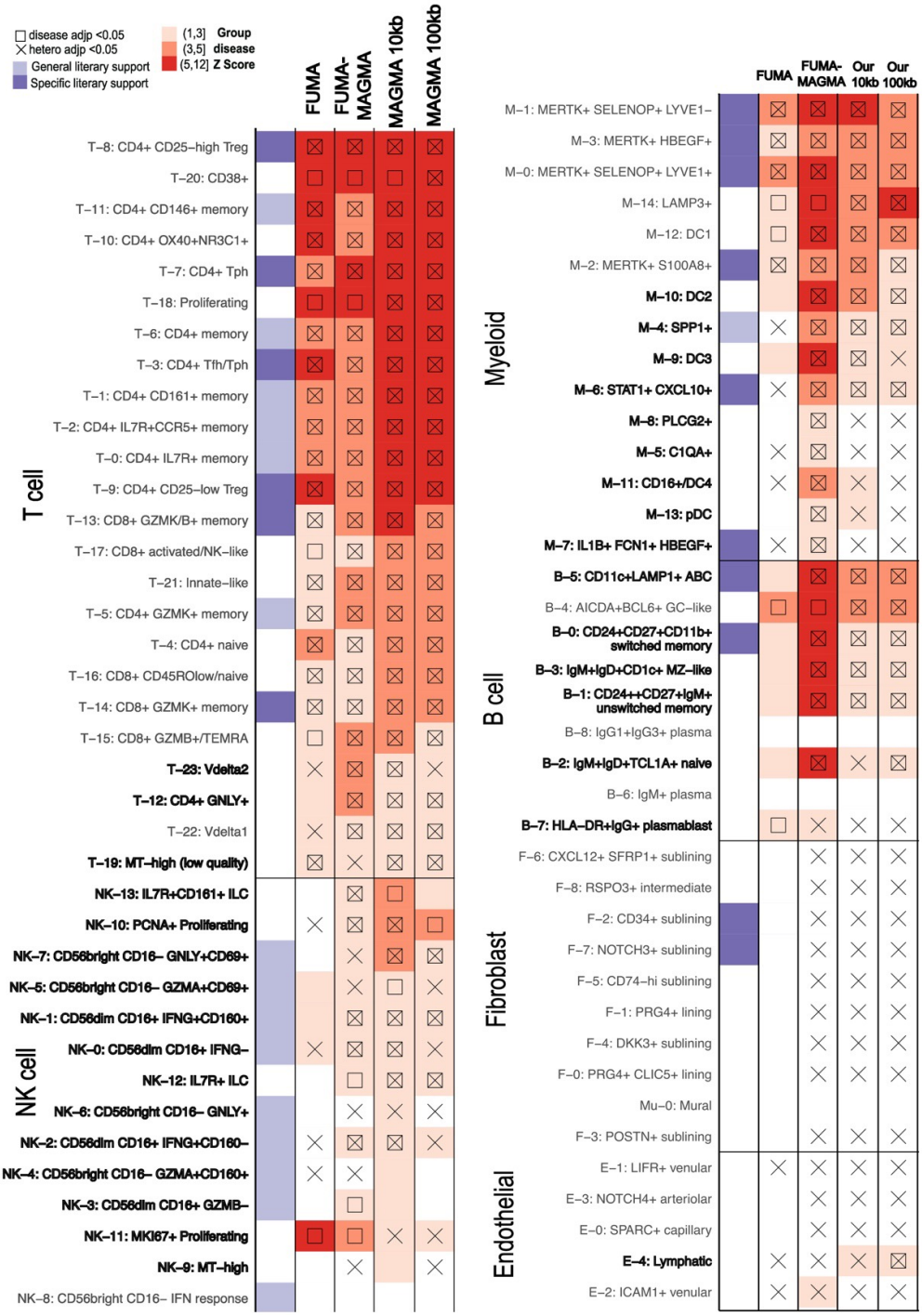

Figure B.21: scDRS results of significant clusters from rheumatoid arthritis calls using inputs from FUMA generated SNP analysis using MAGMA 10kb window mapping (FUMA-MAGMA), using FUMA based mapping including 10kb window, eQTL, and 3D chromatin interaction mapping (FUMA), and the 10kb/100kb MAGMA windows from Figure 5. Cell states with literary support are highlighted in purple. Cell states with differences in disease significance calls are bolded.

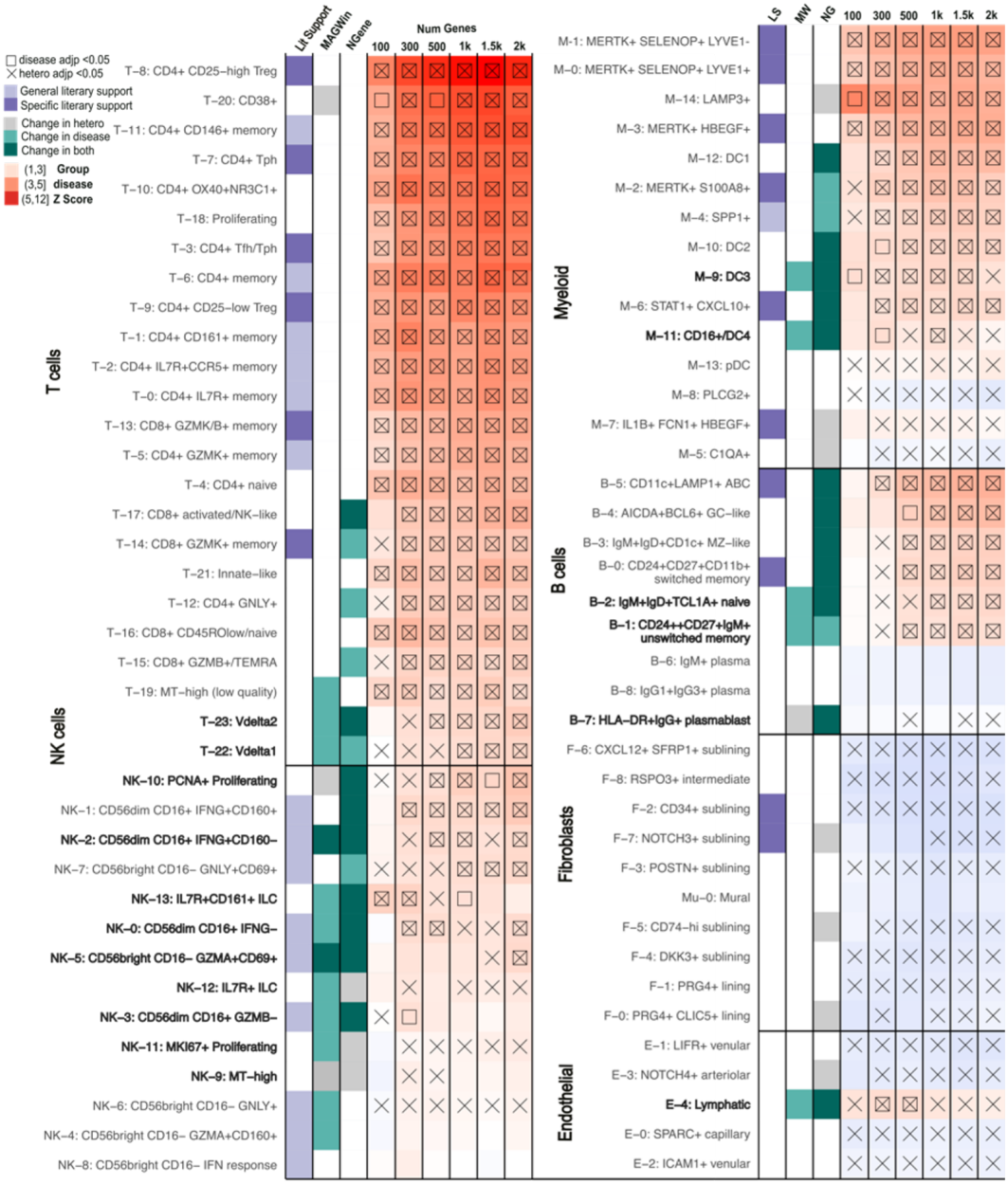

Figure B.22: scDRS results for RA of cell-states that show different levels of significance with a different number of top ranked MAGMA genes (100, 300, 500, 1000, 1500, and 2000). Cell states with significant disease scores and heterogeneity scores are marked by a box or an X, respectively. General literary support means that a cell type with multiple cell states is supported by the literature while specific means a specific single cell state was supported. Cell states with changes in just scDRS disease score, heterogeneity score, or both significance calls across MAGMA windows (MAGWin) or number of top-ranked genes (NGene) are marked with grey or turquoise squares. Cell states with changes both across MAGMA windows and number of top-ranked genes are bolded.

Appendix C

Appendix C. Supplemental Material to Improving calls of differentially transcribed enhancers and their upstream regulators

C.1 Contribution Statement

Drs. Mary Allen and Robin Dowell conceived of the original project. Dr. Dowell and I designed the algorithms for Mu Counts and the edited format of LIET. Dr. Dowell conceptualized the Leading Edge and I designed and implemented the final algorithms of it used in this work. All three of us conceptualized how to evaluate the algorithms. I implemented the algorithms, tested them, and analyzed data. Dr. Jacob Stanley confirmed accuracy of code, math, and changes to LIET and provided general feedback. I wrote the following Supplemental Methods and Notes which were reviewed by the other authors.

C.2 Supplemental Notes

C.2.1 Impact of classic statistical parameters

Although most parameter options for classic statistical tools had limited impact on recall or precision, the mean-dispersion estimation methods of DESeq2 had large implications on the tREs called significant according to F1 scores in all cell types (Supplemental Figure C.23). Most consistently, the DESeq2 mean-based estimation method led to the worst recall and largely skewed p-value distributions; both markers of poor statistical capture of the data (Figure 1A, Supplemental Figure C.24). While DESeq2 local-based dispersions increased recall and alleviated the skew

of p-value calls, EdgeR and Limma had the most consistent p-value distributions without skew (Supplemental Figure C.24).

C.2.2 Flexible motif scanning cutoffs

We considered that TFEA uses a fixed motif scanning cutoff for all motifs, which results in some motifs having far more called instances due to their sequence being more likely by random chance alone. To address this bias, we enabled different p-value cutoffs for motif scanning to be used per transcription factor motif. Default cutoffs for each motif were optimized as a function of the number of motifs found across all tREs of the genome, resulting in different adjusted-pvalue cutoffs per motif (details in Supplemental Methods). This eliminated the gross over-calling inherent with some motifs.

C.3 Methods

All code for analyses are available at (`https://github.com/Dowell-Lab/Improving_tRE_Analysis_Paper/`) which will subsequently be referred to as **github:**.

C.3.1 PRO-seq Analysis

C.3.1.1 Trimming, Mapping, and Quality Control

Samples were mapped to the hg38 genome and NCBI RefSeq annotations were used (hg38 release GCF 000001405.40-RS 2023 03). All samples were trimmed and mapped using an in-house NextFlow pipeline (https://github.com/Dowell-Lab/Nascent-Flow), run with NextFlow (version 20.07.1). Briefly, fastq files were trimmed for adapter sequences and low quality bases using BBMap (version 38.05) and aligned to reference genomes with HISAT2 (version 2.1.0). Downstream mapped read files (CRAM files and IGV-compatible TDF files) were generated with Samtools (version 1.8), Bedtools (version 2.28.0), and IGVtools (version 2.14.1). Samples were then assessed for quality using metrics from the following software packages: FastQC (version 0.11.8), HISAT2 (version 2.1.0), Preseq (version 2.0.3), RSeQC (version 3.0.0), and BBMap (version 38.05).

C.3.1.2 Identifying bidirectional transcripts

This approach was used for all datasets (separately for each dataset defined in Supplemental Table 1) unless otherwise noted. Regions of bidirectional nascent run-on transcription were identified using Tfit and dREG. For both, we removed multimapped reads and reads with low mapping quality score with the following code (all caps indicate a bash variable is being used):

```
samtools view -@ 16 -h -q 1 |${SRR}.bam | grep -P
```

where ${SRR} refers to the prefix of the bam file (usually the SRR key). For Tfit, we used the in-house NextFlow pipeline (https://github.com/Dowell-Lab/Bidirectional-Flow). Final analyses used 3′ bedgraphs that were generated with the -3 flag of bedtools coverage. Briefly, Tfit was run in a two step process, first with the template matching module to identify sites of bidirectional transcription, then these regions were used for input to fit the precise RNA polymerase behavior. For dREG, we followed the recommended pipeline (per https://github.com/Danko-Lab/dREG) and generated BigWig input files by converting the filtered BAM files using bedtools bamToBed to BED files, which where then converted to bedGraph format with bedtools genomecov, and finally BigWig files from bedGraphToBigWig.

Final consensus regions of bidirectional transcription were identified using *muMerge* version 1.1.0 (https://pypi.org/project/mumerge/). *muMerge* probabilisticly determines the most likely midpoint of transcriptional initiation for the bidirectional RNAs (mu or μ). Briefly, Tfit and dREG bidirectional calls were first mumerged separately across all replicates. For p53, these regions were then merged again with each celltype noted as "conditions." For the GR and TNF cells, these regions were merged again with each paper being noted as "conditions." As done previously[223], the dREG and Tfit *muMerge* files were then combined such that calls above 2.5kb were removed and Tfit calls were used for any regions overlapping by at least 40% (relevant code found in (https://github.com/Dowell-Lab/Bidir_Counting_Analysis).

C.3.1.3 Counting reads over regions

The counting pipelines used in this work can be found and run with a nextflow pipeline at (https://github.com/Dowell-Lab/Bidir_Counting_Analysis). Briefly, a bed file of consensus regions (bidirectionals) must be provided where the midpoints of regions are considered the centers of bidirectionals (e.g. from *muMerge*). Users define a fixed window from the midpoints of regions over which they want to consider tREs or gene TSS bidirectionals (e.g. 500 means total 1kb region). Gene bodies are counted over to exclude the gene TSS bidirectionals from gene counts as described previously [223]. The pipeline addresses overlapping transcription from nearby bidirectionals via the algorithm Mu_Counts. Briefly, bidirectional transcripts are counted strand-specifically where the transcript regions are from the center of the bidirectional (mu (μ)) to which ever is shorter: the fixed window length or the μ of the next closest bidirectional. A full visualization of the pipeline can be found at Supplemental Figure C.3. Counting over the fixed window length alone is also done so users have access to both. In both cases, to address overlapping transcription from genes, first, bidirectionals overlapping genes with at least a user-defined percentage of the gene-body region transcribed (according to bedtools coverage, default 70%) on both strands are removed. Counts for bidirectionals with transcribed genes overlapping one strand are replaced with the stranded counts from the strand without overlap multipled by 2. We then ensure that only the regions within genes predicted to not have overlapping transcription from bidirectionals are counted over. The longest isoforms of genes are used for final counts where the regions of bidirectionals with counts more than a user defined limit are removed from the region considered for that gene for counting.

C.3.2 Differential Expression Benchmarking

The following packages were used with R version 4.4.0 (2024-04-24) on platform x86_64-apple-darwin20: DESeq2_1.44.0, edgeR_4.2.1, limma_3.60.4.

C.3.2.1 Simulation based

Relevant code and figures for this section can be found at (**github:**/Simul_Bench_DE)

Refraction Analysis Two samples with $> 100M$ non-duplicated reads were used to assess the impact of tRE vs gene TSS counts with decreasing depths: SRZ1554311 and SRR1145801 ([43, 254]). SRZ1554311 is a combination of technical replicates with the curation detailed in [223]. tREs were identified from Tfit and *muMerge* was run on the two samples to ensure no overlapping bidirectionals. Full counting fixed windows of 600bp and 1kb were used. Bams with uniquely mapped and non-duplicate reads were subsampled using the -bs flag of samtools view with unique seeds for each subsample to ensure random variability.

Dispersion Visualization Estimated mean-dispersion trends and coefficients were calculated with DESeq2 default settings using the functions DESeq2DataSetFromMatrix, estimateSizeFactors, estimateDispersions, and nbinomWaldTest. This was performed on all three celltypes with Nutlin-3a/DMSO data (MCF7, HCT116, SJSA) and PRO-seq samples from HeLa cells perturbed with either dox-inducible shRNA Ints11 or dox-inducible shRNA control (latter had read depths above 70M and quality control scores of 1 according to DBNascent) ([14, 7, 5, 223]. Log fold changes between conditions (shRNA control vs Ints11 and DMSO vs Nutlin) or biological replicates were visualized for gene TSS bidirectionals, gene bodies, and tREs. Counts were collected from genes as described above, while gene TSS bidirectionals and tREs used a 600bp, unstranded, fixed window. Counts for tREs overlapping a transcribed gene (>30 summed counts from samples) on both strands were removed, and if on only one of the strands were counted by multiplying the non-convoluted strand counts by 2. Any features with less than 21 counts between all samples within the same cell type were removed before analysis.

Power Analysis R package powsimR version 1.2.4 was installed from their github repository (https://github.com/bvieth/powsimR) according to their instructions. The HeLa cell samples and counts from Dispersion Visualization section were used for power analysis ([14]). Features were filtered at 16 rather than 21 counts. Parameters of the data were estimated by powsimR using a TMM normalization (estimateParam function). Differential transcription data was simulated with 25 simulations, 5% of features being differentially transcribed, and six different number of replicates (2, 3, 5, 10, 15, 20). The following normalization and differential expression software were

used: Median Ratio Normalization and DESeq2, TMM and EdgeR-LRT (likelihood ratio test), TMM and EdgeR-QL (quasi-likelihood test), TMM and Limma-Trend, TMM and Limma-Voom. All simulations produced very consistent results so only DESeq2 with Median-Ratio normalization is shown.

C.3.2.2 p53 Differential Expression Benchmarking

Relevant code and figures for this section can be found at (**github:**/Bench_DE) in subdirectories Truth_Sets and Before_LE.

Tested Methods The following combinations (total) were used to identify differentially transcribed regions: First, as a baseline, transcribed bidirectionals (> 20 counts within cell type) with positive log fold changes were randomly assigned to "significant" or "not significant" with equal probability. For classic tools, the most common three tools were considered according to their relevant documentation. Details for these methodologies can be found in the documentation and papers corresponding to the appropriate packages [149, 198, 130]. DESeq2 was performed with local, mean, or parametric dispersion methods, Wald or Likelihood Ratio significant tests, and Ratio, positive counts, and iterative normalization methods. EdgeR with Locfit, Movingave, LOESS, and Locfit.mixed dispersion methods (robust or not), Likelihood Ratio, quasi-likelihood, and quasi-likelihood-robust significance tests. Limma with voom or trend dispersion methods and eBayes or eBayes-robust significance tests. Both Limma and EdgeR used trimmed mean of M-values (TMM), TMM with singleton pairing (TMMwsp), upper-quartile, or relative log expression (RLE) for normalization methods. Finally, to further consider the impact of normalization on results, we considered virtual spike-in normalization factors [153]. Adjusted p-value cutoffs of $1e-70, 1e-50, 1e-30, 1e-20, 1e-18, 1e-16, 1e-14, 1e-12, 1e-10, 1e-8, 1e-6, 1e-4, 1e-2, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95$, and 0.99 were used.

Defining the Truth Sets The expected "True Positives" were transcribed bidirectionals split into 4 sections for each cell type: BOTH (overlapping the appropriate cell type-specific ChIP peaks as called in their original, respective publications and having their bidirectional cen-

ters (μs) be within 500bp of the appropriate TF motif), EITHER (motif or CHIP qualifications), CHIP, and MOTIF. ChIP Peaks originally published in hg19 coordinates were converted to hg38 coordinates using UCSC Liftover (https://genome.ucsc.edu/cgi-bin/hgLiftOver) as available Spring 2024. To assess potential bias from motif prediction for p53, HOCOMOCOv12 motif P53.H12CORE.0.P.B was used with FIMO p-value significance cutoff of either p-value $1e-6$ or $1e-5$. The expected "True Negatives" were identified for each cell type as transcribed bidirectionals (total counts > 20 within cell type) not within 10kb of a ChIP peak, with below 11 standardized ($counts_{experimental} - counts_{control}$) ChIP peak reads, and with centers (μs) at least 8kb away from both p53 HCOCOMOCOv12 motifs (P53.H12CORE.0.P.B and P53.H12CORE.1.S.C) (FIMO p-value $1e-5$). The jupyter notebooks for getting these Truthsets can be found at P53_Classic_Bench_DE/Assess_TPs_H12.ipynb and Assess_TNs_H12.ipynb.

Since ChIP peaks and motifs do not necessarily confer transcription[7], we also considered the three cell types as replicates to allow six replicates per condition rather than two. We considered the same tested differential expression methods explained above and had two groups to use as additional "True positive" sets: "Combined Union" and "Combined Intersect." The former are tREs called significant by any tested methods (N=1640). The latter are tREs called significant all tested methods except necessarily TMM_eBayes_Trend and TMM_eBayes-robust_Trend (for final Combined Intersect N=399). These latter two methods were not required since they led to only 38 tREs being called across all methods.

Precision was calculated as TP/(TP+FP) where FP was all regions called significant that were considered a True Negative defined as above. Recall was calculated as the number of True Positives called by the tool divided by the number of True Positives defined above. Area under precision-recall curves usually provides a more robust evaluation of classification methods by considering all significance cutoffs. The differing ranges of precision and recall across platforms, however, prevented fair evaluation with this metric. Specifically, DESeq2 showed far lower maximum recall values compared to EdgeR and Limma for MCF7 and SJSA, even when a p-adjusted cutoff of 0.99 was included (Supplemental Figure C.25). Adjusted pvalue cutoffs of 10^{-70}, 10^{-50}, 10^{-30}, 10^{-20},

10^{-18}, 10^{-16}, 10^{-14}, 10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, and 0.99 were considered.

C.3.3 Length based Benchmarking

Relevant code and figures for this section can be found at (**github:**/Length_Bench).

C.3.3.1 Defining the Truth Length Sets

Long-read nascent run-on sequencing currently provides the best, high-throughput length estimation of enhancer-associated transcripts, but has few published experiments [57, 56, 194]. Therefore, we evaluated the ability of tools to predict the length of enhancer-associated transcripts (as defined by long-read nascent run-on sequencing) from short-read nascent run-on sequencing data. Five long-read nascent run-on sequencing fastq files (same control states) for K562 were downloaded from SRA with relevant information found in Supplemental Table 2 ([57]). Fastqs already had adapters and poly A/I tails removed according to Guppy as used by the original authors ([57]). Due to the low-depth of nascent long-read samples, and since individual reads rather than counts would be used, all fastqs were combined into a single fastq file before mapping. Following the code used by the original authors, reads were mapped to hg38 using minimap2 (primarily designed for error-prone long reads) with the following parameters (words in all caps refer to bash variables):

```
minimap2 -acx map-ont -t 16 -k14 \
--sam-hit-only ${FASTA} ${FASTQ} \
| samtools sort -o ${BAM}
```

where ${FASTA} points to the fasta file for the hg38 genome, ${FASTQ} refers to the long-read fastq file downloaded, and ${BAM} refers to the named bam file to serve as output.

As a truth set for comparison, 411 tRE associated transcripts with support from ENCODE Phase 3 (ENCFF464BRU) were manually annotated as transcriptionally isolated in both long-read and short-read data, having at least two long-reads (mapping quality $\geq$ 30) supporting a clear

transcript end position, and significant depth from previously published nascent run-on short read samples in K562 controls (SRA SRR4454567/8/9/70) [250, 164]. The most downstream end of long reads (minimum mapping quality of 30) within the annotated transcript region was used as the "true end" of the transcripts.

C.3.3.2 Linking RNA calls across tools

Since a single tool might call multiple bidirectionals or transcripts within a region of interest, we identified the transcripts for each tool most appropriately aligning to the long-read supported transcripts with the following methods.

We used *muMerge* to get consensus bidirectional center calls for the relevant tools: dREG and Tfit. Briefly, coverage filtered (at least 9 counts per tRE predicted) output by dREG and Tfit were given to *muMerge* along with a metadata file grouping DMSO and heatshock samples together. For this work, a new flag was added to *muMerge* to allow the original positions of the regions and samples in which they're found to also be saved (–orig_names). The code run was

```
python ${SRC}/mumerge.py -i {METADATA} -o {OUT_PREFIX} \
--orig_names
```

where ${SRC} refers to the directory containing the cloned repository for *muMerge*, ${METADATA} is the file with the metadata information for the samples (e.g. replicates and conditions), and ${OUT_PREFIX} is the prefix for all output to be saved with. Users may get similarly coverage filtered regions and details on parameter layouts using the verbose branch of *muMerge*: `https://github.com/Dowell-Lab/mumerge/tree/verbose`.

Tfit has been shown to have the highest accuracy in calling the center of bidirectional transcription and using the 3′ bedgraphs particularly improves the calls (this work) ([223, 278]). Therefore, the final centers of bidirectional transcription (μs) were based on Tfit 3′ calls. Bedtools closest was run to find the distances between Tfit 3′ calls and dREG and Tfit calls. A bidirectional was then linked to the original Tfit 3′ bidirectional of interest if the μs (center of bidirectional) were

within 500bp of one another.

Homer calls were matched in two ways: 1) requiring a transcript's $5'$ end to be within 400bp of the Tfit $3'$ μ, or 2) requiring a transcript to overlap region between Tfit $3'$ μ and downstream 100bp and in the case of multiples, choose the one where the $5'$ end was closest to μ. The latter method includes transcripts that have poor calls for the transcript start but still capture the transcript. dREG and Tfit (full read) calls were matched to Tfit $3'$ calls according to the μs being within 500bp.

C.3.3.3 Noise generation

To determine the regions over which to add noise, we used 1.5kb downstream of the true length annotation (according to long read) with a minimum of 3.5kb downstream of the bidirectional center (μ) [padded regions]. If a strand of a bidirectional does not have any long reads, then 2kb is used. All regions were manually examined to ensure significant reads, short or long, were not omitted and nearby transcripts were not accidentally included. To assess the impact of noise, we added reads in random locations (sampling with replacement) across the padded regions. The number of reads added was different percentages of the number of original reads mapped to the same region (filtering for supplementary or secondary alignment): 20, 40, 60, 80, 100, 120, and 140. To mimic full reads, the random position was elongated to the $5'$ end by 74bp to reach the original 75bp read length. To mimic the $3'$ ends of reads in a bedgraph, the position alone was used. These reads (formatted as bedgraphs) were then merged with the original multi-mapped filtered bedgraphs using bedtools version 2.28.0.

C.3.3.4 Homer

Homer v5.1 was run with multi-mapped filtered bams or bedgraphs (full reads) ([84]). First, tag directories were made using 'makeTagDirectory' and the '-keepAll' flag to ensure the same input was used across all tools. Then, nascent run-on sequencing de novo transcript identification was run with (bash variables in all caps)

```
findPeaks ${TAGDIR} -style groseq -o ${OUT_FILE} \
-minBodySize 150 -tssSize 50 -bodyFold 3 \
-endFold 5 -uniqumap hg38-50nt-uniqmap
```

where ${TAGDIR} refers to the TagDirectory from the filtered bams/bedgraphs, ${OUT_FILE} is the output path to save the peaks. Different parameter combinations like endFold 7, bodyFold 4, and tssFold 4 were also tested with the above showing the best results and therefore used for benchmarking. More detailed explanation of the Homer groseq algorithm can be found at http://homer.ucsd.edu/homer/, hereby called **homer:** at `http://homer.ucsd.edu/homer/ngs/groseq/groseq.html` where you can also download the hg38-50nt-uniqmap folder at `http://homer.ucsd.edu/homer/data/uniqmap/uniqmap.hg38.50nt.zip`. The latest version available at the time of this work was October 26, 2018.

C.3.3.5 dREG

dREG uses single-position based bigwigs. To convert the 3′ bedgraphs to bigwigs, we used bedGraphToBigWig v4. We then ran dREG v1.0 through https://dreg.dnasequence.org/ using default parameters (R version 4.3.2 (2023-10-31)). The dREG lengths are directly based on dREG output.

C.3.3.6 Tfit

For initial analysis, Tfit was run with both 3′ bedgraphs or full read bedgraphs using Tfit_focus branch of the pipeline in `https://github.com/Dowell-Lab/Bidirectional-Flow/tree/Tfit_focus`. This nextflow pipeline has been optimized to run Tfit genome-wide. To allow unnecessary computational burden, for noise based calls, we had Tfit only search over the regions of study by feeding Tfit the padded regions used for noise production as preliminary regions. Tfit was run using the shell script available at `https://github.com/Dowell-Lab/Bidirectional-Flow/blob/main/bin/tfit_model.sh` (commit 89f2fd1, words in all caps are bash variables) :

```
${TFIT_DIR}/tfit_model.sh -t ${TFIT_PATH} \
-c ${TFIT_CONFIG} -b ${BG} -k ${PADDED_REGIONS} \
-p ${PREFIX} -n 32
```

where ${TFIT_DIR} refers to the path where the tfit_model.sh script is, ${TFIT_PATH} is the path to the Tfit software, ${TFIT_CONFIG} is the path to the config file for Tfit, ${BG} is the bedgraph used by Tfit, ${PADDED_REGIONS} is a bed file with the padded regions used for noise calls (considered preliminary regions over which to look for bidirectionals), and ${PREFIX} is the prefix used to name output. Details regarding the parameters can be found at https://github.com/Dowell-Lab/Bidirectional-Flow/. Tfit release version 1.2 (Repository version) was used and can be found at https://github.com/Dowell-Lab/Tfit/releases/tag/v1.2. Importantly, in this version, the source code incorrectly labels τ as λ. To avoid confusion, all cases where Tfit's λ parameter is used, but is actually τ, we label as τ here.

Tfit lengths were considered according to three options:

(1) The original output length from Tfit (when going from μ to the edge of the region): $\tau + \sigma$,

(2) $\tau + \sigma + \frac{\text{footprint}}{2}$, since the footprint is not originally considered in the length output of the region from Tfit, and

(3) $\frac{\text{footprint}}{2} + |x - \mu|$ where x is the 95th percentile of the EMG, numerically solved for given μ, σ, τ

$$CDF(x \mid \mu, \sigma, \tau) = 0.95$$

where $CDF(\cdot)$ is the cumulative distribution function of the EMG and τ, μ, σ are parameters of the EMG.

C.3.3.7 LIET

LIET v1.0.0 takes both a pad file to determine the complete regions over which to look and annotation file to use as priors in its Bayesian modeling. We used the same pads used for

noise production (1.5kb downstream of the true length annotation (according to long read) with a minimum of 3.5kb downstream of the bidirectional center (μ)). Input annotations for LIET were entered with the $5'$ location being the bidirectional center from *muMerge* (μ) + Tfit footprint/2 as the start and 400bp downstream of this position as the $3'$ location, labeling all regions as positively stranded for simple tracking purposes. Exact LIET parameters and priors used are found in Supplementary Table 1. Originally, LIET was designed to run the full LIET model on the sense strand (positive according to the annotation) and an EMG model on the antisense strand. The software was edited to run either the EMG or LIET model on either strand. Similarly, percentiles of the probability distributions (with background removed) were calculated. Briefly, weights making up the total weight (w) were recalculated (to w') after removing background (w_b) as described by Equation C.1. The probability distribution functions of each strand using these recalculated weights were then calculated based on a domain of $(-10^6, 10^6)$. As defined in Equation C.2, we compute the genomic positions corresponding to each target percentile (q) from these pdfs by identifying where the strand-specific cumulative distributions reach that value.

$$\text{Given } \mathbf{w} = [w_1, w_2, \ldots, w_{n-1}, w_b], \quad \text{define } \mathbf{w}' = \left[\frac{w_1}{\sum_{i=1}^{n-1} w_i}, \frac{w_2}{\sum_{i=1}^{n-1} w_i}, \ldots, \frac{w_{n-1}}{\sum_{i=1}^{n-1} w_i}\right] \tag{C.1}$$

where:

- w = original vector of weights for the LIET model,
- w′ = recalculated vector of weights for the LIET model.

$$\begin{aligned} \text{CDF}_p(x) &= \sum_{i=1}^{x} \text{PDF}_p(i) \\ \text{CDF}_n(x) &= \sum_{i=1}^{x} \text{PDF}_n(i) \\ \text{position}_p(q) &= \arg\min_x |\text{CDF}_p(x) - q| - 10^6 \\ \text{position}_n(q) &= \arg\min_x |\text{CDF}_n(x) - q| - 10^6 \end{aligned} \tag{C.2}$$

where:

- $\text{CDF}_p(x)$ = cumulative distribution for the positive (p) strand,
- $\text{CDF}_n(x)$ = cumulative distribution for the negative (n) strand,
- $\text{position}_p(q)$ = closest position to percentile q on the positive strand (0-based),
- $\text{position}_n(q)$ = closest position to percentile q on the negative strand (0-based).

C.3.3.8 Calculating consensus lengths

To calculate consensus lengths based on replicates similar to *muMerge*, we considered multiple methods. First, we took the average. Next, we took a weighted average according to weights of coverage or in the case of LIET, the number of reads assigned outside of background:

$$\text{Weighted Average } 3' \text{ position} = \frac{\sum_{i=1}^{n} x_i \cdot c_i}{\sum_{i=1}^{n} c_i}$$

where x_i is the $3'$ position of the transcript in sample i and c_i is either full coverage of the transcript according to feature counts or the weight of transcription $(1 - w_b)$ according to LIET in sample i.

C.3.4 Length based p53 Differential Expression Benchmarking

Relevant code and figures for this section can be found at (**github:**/Bench_DE/Length_DE).

C.3.4.1 Mu_Counts

This description follows what is shown in Supplemental Figure C.3. The full pipeline (as a Nextflow pipeline) and additional descriptions are accessible at `https://github.com/Dowell-Lab/Bidir_Counting_Analysis/tree/main`. The first step involves filtering inputs. First, the tool takes in consensus regions for bidirectional transcripts where the midpoint is assumed to be the initiation point of bidirectional transcripts and the length is a confidence interval around it (like from *muMerge*). Therefore, we remove bidirectionals that are likely called due to technical noise (those with confidence intervals above 3.5kb) and filter bams/crams to remove multimapped reads. This module produces consensus files with the widths needed for future analyses and unique

names (based on parameters). The next major step is identifying Gene TSS Bidirectionals; these show distinct transcriptional patterns from tREs and correspond to gene PROMPTs rather than enhancers. Bidirectionals coordinating to the transcription start site (TSS) of genes are identified by overlapping bidirectionals with a small window (parameter TSS_WIN - default 25bp) with 1kb regions around gene TSSs (parameter tss_1kb_file, also available for GR38.p14). Gene TSS bidirectionals are assigned so that a single gene isoform only has one TSS bidirectional, with the one whose midpoint is closest to the TSS used. Multiple gene isoforms are allowed to share the same TSS bidirectional if their TSSs are within 50bp of each other. The third step involves addressing overlapping transcription from both genes and other tREs. Nascent transcription of genes continues downstream of annotations and many tREs are found within introns. Therefore, gene bodies considered transcribed (parameter COUNT_LIMIT_GENES=70% isoform is covered with reads) along with the region 10kb downstream of said isoform are overlapped with the nonTSS bidirectionals (tREs). Any nonTSS bidirectional overlapping active gene transcription on both strands is removed since deconvolution cannot confidently occur. NonTSS bidirectionals overlapping gene transcription on one strand have counts replaced with those from the strand with no overlapping transcription, doubled. We then address overlapping transcription from bidirectionals with other bidirectionals using Mu_Counts. If the parameter COUNT_WIN for a tRE means the tRE is now overlapping with another tRE region, the neighborhing tREs will be counted so that the maximum distance of each RNA is the μ of the nearest neighboring tRE. RNAs for tREs are then counted separately according to the strand (e.g. counts on positive/negative strand for a bidirectional's positive/negative RNA are counted separately before being combined). Finally, we address gene counts that improperly contain counts from overlapping bidirectionals. Azofeifa et al[11] showed that gene counts can be largely disrupted by including responding bidirectionals. Therefore, the pipeline optionally removes the regions of bidirectionals with counts above parameter COUNT_LIMIT_BIDS.

Grouping Test Sets We hypothesized that length would have testable implications on two key groups of calls: Isolated (bidirectionals without possible convolution from nearby transcription)

and "Overlapping" pairs ("True positive" and "True negative" bidirectionals that can influence each other's calls with convoluting transcription). Isolated bidirectionals were identified as those with no other tREs with 5.5kb of them, no overlap with genes, and at least 500bp upstream of a gene TSS and 10kb downstream of a gene annotated termination site (only considering genes that had total counts > 200). Overlapping pairs were identified as "True positives" and "True negatives", as defined below, with μs within 5kb of each other.

Defining the Truth Sets In order to not bias the truth sets to a specific tool, we took a slightly different approach to defining the truth sets for this length-based analysis compared to the analysis focusing on differential transcription alone. First, we wanted to limit how much the position of mu potentially bias truth sets as Homer cannot consider mu. Therefore, we considered "True Positives" if a 50bp region around Tfit $3'$ μ overlapped a ChIP peak with a p53 motif within it. To ensure we had high enough statistical power when considering overlapping "True Positives" and "True Negatives," we used less stringent requirements for these truth sets when considering overlapping tREs. Expected "True Positives" were considered according to p53 ChIP peaks containing TP53 HOCOMOCOv12 motifs with a p-value cut off of 1e-5. Expected "True Negatives" were transcribed bidirectionals (total counts > 20) without motifs or ChIP peaks within 2kb rather than 10kb to allow consideration of "Overlapping" pairs within 5kb of each other.

C.3.5 TFEA and Leading Edge Analysis

Transcription Factor Enrichment Analysis from https://github.com/Dowell-Lab/TFEA (version v1.1.1) was run with the ranked files (according to log fold change, then adjusted p-values) from the different parameter-tool combinations. The FIMO (from Meme v5.0.3) scanning background was set to uniform. Otherwise, default parameters were used.

C.3.5.1 Leading Edge Methodologies

The final algorithm is integrated into the most updated version of TFEA (v2.0.1): `https://github.com/Dowell-Lab/TFEA`. Relevant code and figures from using the leading edge for

identifying significantly changing tREs in p53 and GR datasets can be found at for this section can be found at (**github:**/Bench_DE) in subdirectories Get_LE and Compare_LE.

Simply speaking, the leading edge is interpreted the position at which any elements with higher p-values (hence less statistically significant changes) are no longer considered as contributing to the transcription factor being enriched. Two methods to find the leading edge were used:

Let $E(t)$ represent the cumulative enrichment as a function of the ranked tREs t.

(1) **Matched Background:** The first position (so leftmost if positive enrichment score and rightmost if negative) where the slope of the cumulative enrichment score curve is equal to or below that of the background enrichment line.

Let $B(t)$ represent the background enrichment line (a straight line from $E(0)$ to $E(T)$).

$$\text{Slope of enrichment:} \quad E'(t)$$
$$\text{Slope of background:} \quad B'(t) = \frac{E(T) - E(0)}{T}$$

Define the matched background position t^* as:

$$t^* = \begin{cases} \min\{t : E'(t) \leq B'(t)\} & \text{if } auc(E(T)) > 0 \text{ (positive enrichment)} \\ \max\{t : E'(t) \geq B'(t)\} & \text{if } auc(E(T)) < 0 \text{ (negative enrichment)} \end{cases}$$

(2) **Plateaued Enrichment:** Captures the position at which the cumulative enrichment changes have stopped steadily changing due to enrichment. The second derivative of cumulative enrichment ($E''(t)$) shows two cases: a monotonic stabilization towards 0 or a non-monotonic function with up to six extrema (high oscillation) before stabilization. With non-monotonic cases, we use the first extrema as a conservative leading edge. With monotonic stabilizing cases (no extrema within the first 40% of tREs), we calculate the elbow of the curve as shown below.

$$t^* = \begin{cases} \arg\min_{t<0.4T} E'''(t) = 0 & \text{(Early Oscillation)} \\ \text{elbow}(E''(t)) & \text{(Monotonic stabilization)} \end{cases}$$

where T is the total number of tREs and the elbow is computed as:

$$\text{elbow}(E''(t)) = \arg\max_t \left(\text{distance from line } \ell(t) \text{ from } E''(0) \text{ to } E''(T)\right)$$

Smoothing Method:

To reduce sensitivity to noise, multiple smoothed iterations of the cumulative enrichment curve $E(t)$ are generated using B-spline interpolation with degree $k = 5$. The final leading edge is determined by taking the median position across all spline fits. The initial smoothness parameter and subsequent smoothness sequences for the spline were optimized empirically by assessing about 100 case-scenarios: across several perturbations (P53, TNF/DEX, WSP, UPM, shRNAs) and cell types (HCT116, MCF7, SJSA, BEAS2B, different primary samples of small airway epithelial cells, ESC), bidirectional numbers ranging from 15,000 to 120,000, and transcription factors with motifs covering from 0-50% of tREs. Importantly, subsequent smoothness assessment (described in next section) ensures that the algorithm is robust to this initial smoothness parameter.

Let:

- N_{motif}: number of tREs with motif calls
- s_0: initial smoothness value, defined by a tiered rule based on N_{motif}
- s_i: spline smoothness parameter for iteration i

Initial smoothness s_0:

$$s_0 = \begin{cases} 3 \times 10^{-12} & \text{if } N_{\text{motif}} > 30000 \\ 4 \times 10^{-12} & \text{if } N_{\text{motif}} > 20000 \\ 5 \times 10^{-12} & \text{if } N_{\text{motif}} > 10000 \\ 2 \times 10^{-11} & \text{if } N_{\text{motif}} > 8000 \\ 3 \times 10^{-11} & \text{if } N_{\text{motif}} > 7000 \\ 4 \times 10^{-11} & \text{if } N_{\text{motif}} > 5000 \\ 5 \times 10^{-11} & \text{if } N_{\text{motif}} > 4000 \\ 7 \times 10^{-11} & \text{if } N_{\text{motif}} > 3000 \\ 8 \times 10^{-11} & \text{if } N_{\text{motif}} > 2500 \\ 9 \times 10^{-11} & \text{if } N_{\text{motif}} > 2000 \\ 4 \times 10^{-11} & \text{if } N_{\text{motif}} > 1500 \\ 1 \times 10^{-10} & \text{if } N_{\text{motif}} > 1000 \\ 2 \times 10^{-10} & \text{otherwise} \end{cases}$$

Smoothness sequence:

Subsequent smoothness values are decreased to allow less smoothing:

$s_{i+1} = s_i + 5 \times 10^{-(n_i - 1)}$, where n_i = power of the current smoothness parameter

This continues until the number of extrema for $E''(t)$ ($E'''(t) = 0$) exceeds a threshold (with a maximum of 20 iterations):

$$\text{Max } \left(E'''(t) = 0\right) = \begin{cases} 6 & \text{if } N_{\text{tRE}} > 60,000 \\ 4 & \text{otherwise} \end{cases}$$

Importantly, the ultimate leading edge is robust to these thresholds (4, 5, or 6), but we found that these separations ensured the greatest breadth of smoothness parameters considered while avoiding noise.

In case of under-smoothed start:

If the initial s_0 already exceeds the allowed number of extrema, then it is increased by $3 \times 10^{-(n_0)}$ until the constraint is satisfied:

$$s_0 = s_0 + 3 \times 10^{-(n_0)}, \quad \text{where } n_0 = \text{ power of the initial smoothness parameter}$$

C.3.5.2 Other Updates to TFEA:

Relevant code and figures for this section can be found at (**github:**/Improving_FP_calls). **TF-specific FIMO significance cutoffs:** To determine good default adjusted p-value cutoffs for FIMO (using Meme v5.0.3) for each motif, all 847,522 bidirectionals identified in [223] were scanned for motifs across their 3kb total regions ($\pm$ 1.5kb from μ) and using p-value cutoffs of $1e-4, 1e-5, 1e-6$, and $1e-7$. If a motif was called in between .5% (4,237) and 7% (59,326) bidirectionals based on one specified p-value cutoff, that p-value was considered the default cutoff for the TF to use. If less than .5% of bidirectionals had a motif at $1e-7, 1e-6, 1e-5$, or $1e-4$, the default p-value cutoff was changed to $1e-6, 1e-5, 1e-4$, and $1e-3$, respectively. If greater than 5% of bidirectionals had a motif at $1e-4, 1e-5, 1e-6$, and $1e-7$, the default p-value cutoff was changed to $1e-5, 1e-6, 1e-7$, and $1e-8$, respectively. If greater than 40% of bidirectionals had a motif at $1e-7$, the default p-value was changed to $1e-9$. Transcription factors with more than 20,000 tRE motifs for tREs found within the Nutlin-3a and DMSO dataset with $1e-6$ (ZN121, ZN135, ZN441, ZN560, ZN613, ZN770) were assigned best pvalues of $1e-10$. The code for this analysis can be found at Other_TFEA_Updates/Assess_pval_motifs.ipynb. TFEA was edited to take in a file with the desired p-value motifs with the option –fimo_thresh and the default values calculated above are provided as a file in the github repository of TFEA.

Leading Edge metrics for significance calls: False positives and true negatives were defined as the TFs with the highest enrichment scores that had no well-characterized linkage to the perturbation and were called significant or not according to the GC-corrected adjusted-pvalue (< 0.01), respectively. TFs with enrichment scores below 0.05 were not considered as they could be easily filtered out. Three metrics were calculated regarding the tREs with changes in enrich-

ment that were higher than that from background. Background slope was calculated as the maximum(cumulative enrichment score) - minimum(cumulative enrichment score) / Number of tREs. Frac_Background, as focused on in the remainder of the text, is the fraction of tREs with their slope higher than background. The fraction of tREs in the interquartile range of ranks with slopes higher than background (Frac_Back_Q13) and the difference between slopes (True - Background) (Diff_Back) were also calculated and showed comparable trends to Frac_Background.

To identify the quantile of ranked tREs where the cumulative enrichment score increased the fastest in magnitude, tREs were binned into 15 quantiles (around 3000 tREs per bin). The quantile with the maximum absolute value slope when using the median spline tested was returned as the Max_Quant. TFs with Max_Quant values in the middle of the ranked list (6,7,8,9,10) were considered False positives by the LE.

C.3.5.3 Wood Smoke Particle LE and TFEA-LE Analysis:

Relevant code and figures for this section can be found at (**github:**/WSP). **Matching ATAC-seq peaks and PRO-seq tREs:** To allow complete comparison between ATAC-seq and PRO-seq, bidirectionals were first called within each sequencing approach with Tfit before being mapped to each other. 80bp windowed PRO-seq bidirectionals (muMerged according to condition (BID)) were mapped to 1kb windowed ATAC-seq peaks (muMerged according to condition (ATAC)) with bedtools closest (words in all caps refer to variables):

```
bedtools closest -k 4 -d -D "ref" -a ${BID} -b ${ATAC} > ${OUT}
```

where ${BID} refers to the bed file of muMerged bidirectionals from PRO-seq and ${ATAC} refers to the muMerged ATAC-peak bed file. PRO-seq bidirectionals with μs within 2kb of an ATAC peak μ were kept with the closest corresponding ATAC peaks removed. 74% of PRO-seq bidirectionals (50,850) mapped to 60% (50,512) of ATAC-peaks. Otherwise, ATAC peaks and PRO bidirectionals were kept and noted as only occurring in one method. All regions were used for downstream analysis.

Running TFEA: To ensure the fairest comparison between ATAC-seq and PRO-seq,

we only considered Non-GeneTSS bidirectionals/peaks as defined by Counting_Bid_Analysis. For TFEA and leading edge, rankings from EdgeR-TMM-QL were used but DESeq2 and other EdgeR results showed the same trends noted in this work. TFEA was run with and without assuming a uniform background for FIMO and led to no clear difference in results.

Considering Leading Edge Metrics for TF Calls: To ensure results weren't solely based on the number of motifs, TF calls were first filtered to have between 600 and 10000 tREs with the motif. To be considered a call supported by the leading edge metrics, the Match-Background leading edge had to be before the midpoint and after the Plateaued Enrichment leading edge, the Fraction of tREs above background had to be below 0.46 for PRO-seq and 0.51 for ATAC-seq, and the Max_Quant could not be 6,7,8,9,10 (interquartile range inclusive). TFs were then split into the following categories:

- **All**: GC-corrected Padj < 0.001, uncorrected Padj < 0.001, and meets LE requirements
- **GC_only**: GC-corrected Padj < 0.001, uncorrected Padj ≥ 0.001, does not meet LE requirements
- **UNC_only**: GC-corrected Padj ≥ 0.001, uncorrected Padj < 0.001, does not meet LE requirements
- **GC_LE**: GC-corrected Padj < 0.001, uncorrected Padj ≥ 0.001, and meets LE requirements
- **UNC_LE**: GC-corrected Padj ≥ 0.001, uncorrected Padj < 0.001, and meets LE requirements

The jupyter notebook with this analysis can be found at WSP/Plot_MB_curves.ipynb

The TF calls were assessed for directionality by comparing GC-corrected Enrichment scores. If the magnitude of scores were both > 0.05, they were assessed as being in the same or opposite directions ($\pm$) between ATAC-seq and PRO-seq of the same time points, or when using gene TSS bidirectionals (according to gene differential rankings) compared to tREs in the same condition.

C.4 Supplemental Figures

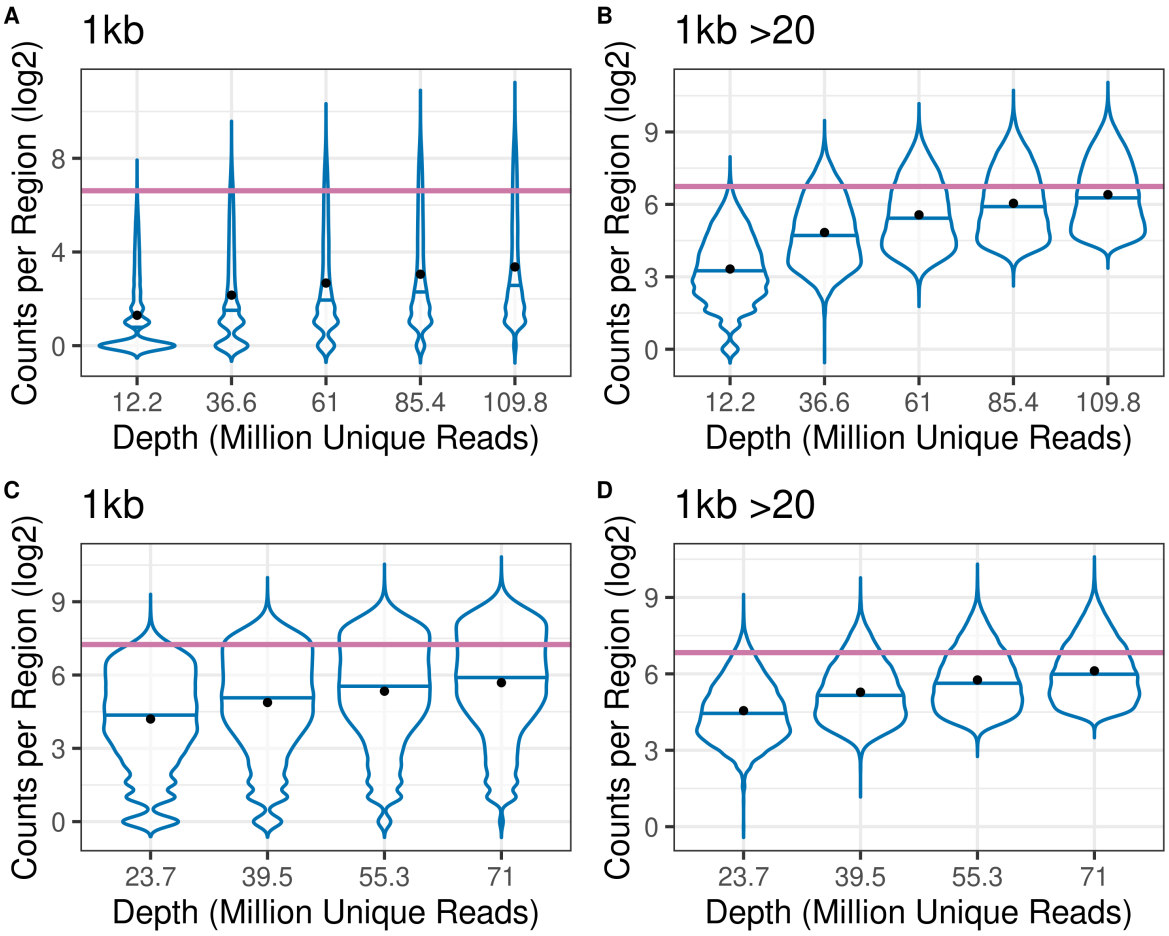

Figure C.1: **About 100 million unique reads are needed for tREs to reach the same median counts obtained by gene TSS bidirectionals when only using about 40 million unique reads across two independent samples.**. Distribution of counts, including all tREs with counts above 0 (A+C) or 20 (B+D) at full depth visualized as violin plots. The median counts for gene TSS bidirectionals at 36.6 million (A+B) or 39.5 million (C+D) unique reads is shown as a pink line. Data in A+B is SRZ1554311 and Data in C+D is SRR1145801.

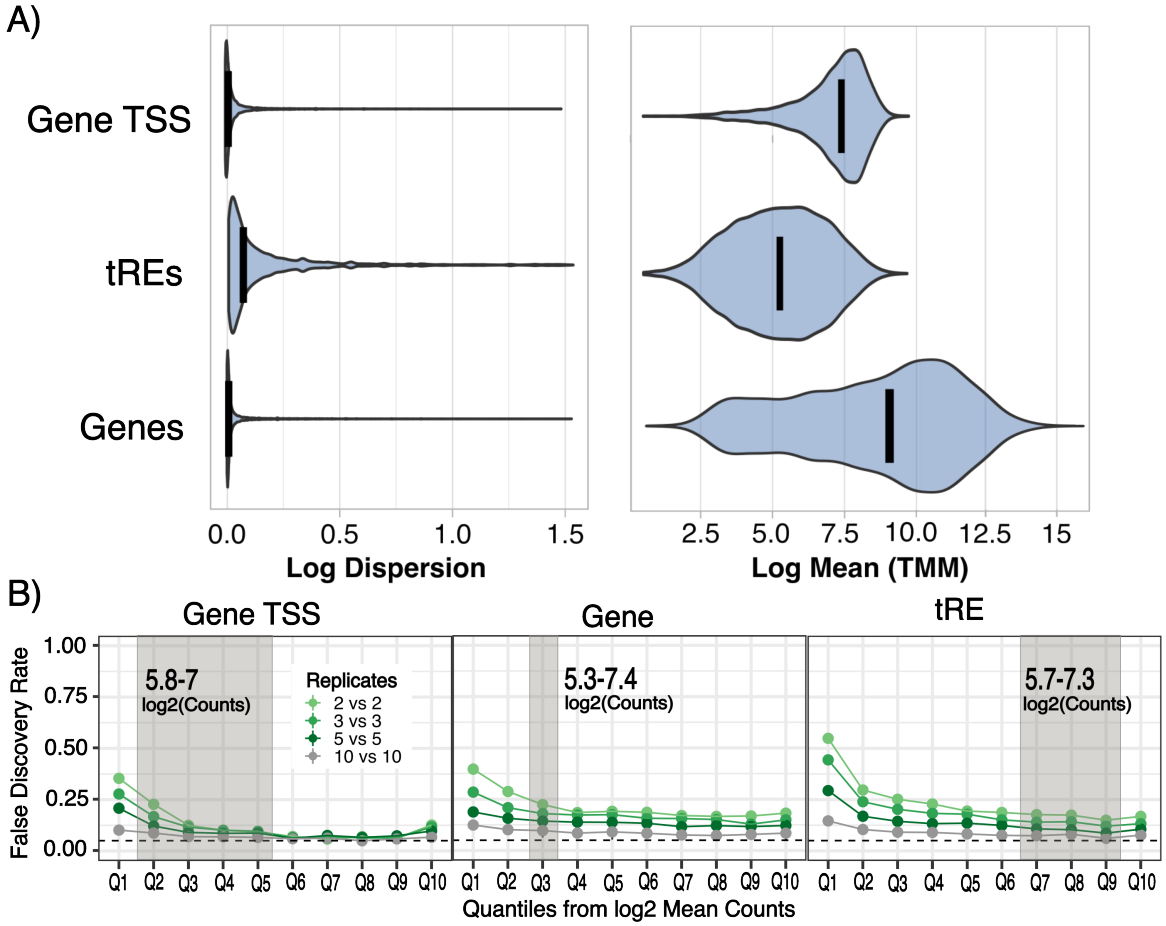

Figure C.2: **Simulated data with mean-dispersion trends comparable to p53 data show that tREs require much higher number of replicates to achieve comparable false discovery rates as genes and gene TSS bidirectionals. A.** Distributions were estimated by powsimR based on real PRO-seq count data of the following features from two biological replicates (details in Supplementary Methods). **B.** False Discovery Rates of gene TSS bidirectionals, genes, and tREs across increasing replicate numbers (colors) and quantiles (Q1-Q10) according to average counts (log2). Due to the different mean distributions across the feature types, the quantiles do not represent the same count levels. The grey box represents a section of features with comparable average counts (around 5.5-7). Shade of green reflects replicate numbers per condition.

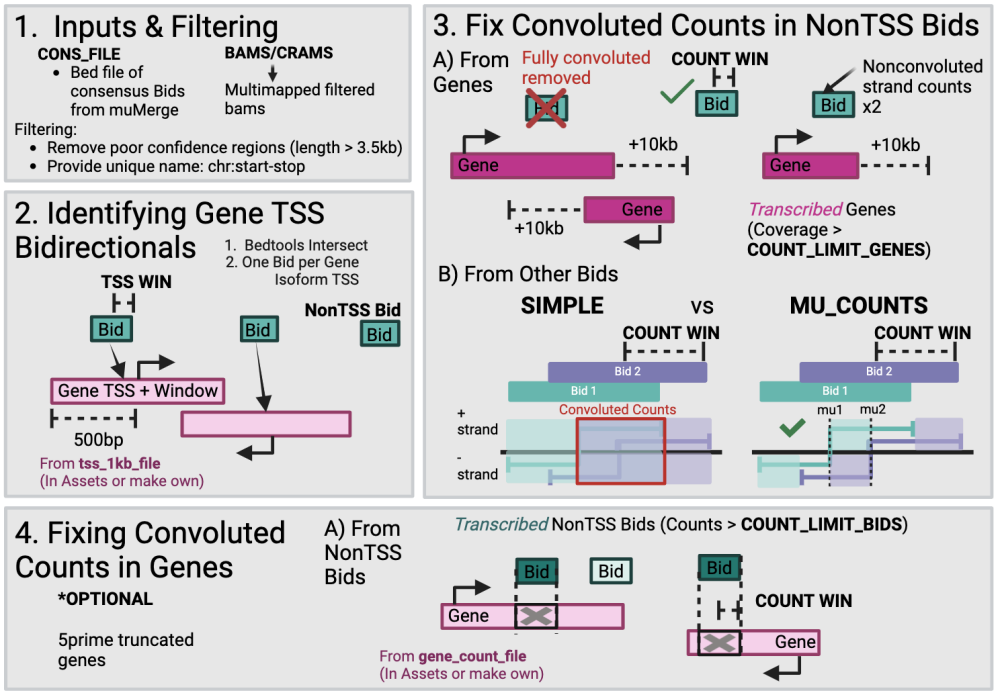

Figure C.3: **Visual representation of the Mu_Counts pipeline.** Full description in Supplemental Methods **Step 1**-Filtering Inputs. We remove bidirectionals that are likely called due to technical noise and filter bams/crams to remove multi-mapped reads. This produces consensus files with the widths needed for future analyses and unique names. **Step 2**-Identifying Gene TSS Bidirectionals. Bidirectionals coordinating to the transcription start site (TSS) of genes are identified by overlapping bidirectionals with a small window (parameter TSS_WIN - default 25bp) with 1kb regions around gene TSSs (parameter tss_1kb_file, also available for GR38.p14). **Step 3A**-Addressing overlapping transcription from genes. Gene bodies considered transcribed (COUNT_LIMIT_GENES=70% isoform is covered with reads) along with the region 10kb downstream of said isoform are overlapped with the nonTSS bidirectionals (tREs). Any nonTSS bidirectional overlapping active gene transcription on both strands is removed since deconvolution cannot confidently occur. NonTSS bidirectionals overlapping gene transcription on one strand have counts replaced with those from the on-convoluted strand, doubled. **Step 3B**-Addressing overlapping transcription from bidirectionals with other bidirectionals (Mu_Counts). If the parameter COUNT_WIN for a tRE means the tRE is now overlapping with another, the neighborhing tREs will be counted so that the maximum distance of each RNA is the μ of the nearest neighboring tRE. RNAs for tREs are then counted separately according to the strand (e.g. counts on positive/negative strand for Bid 2's positive/negative RNA are combined). **Step 4**-Fixing gene counts due to overlapping bidirectionals. The pipeline optionally removes the regions of bidirectionals with counts above parameter COUNT_LIMIT_BIDS.

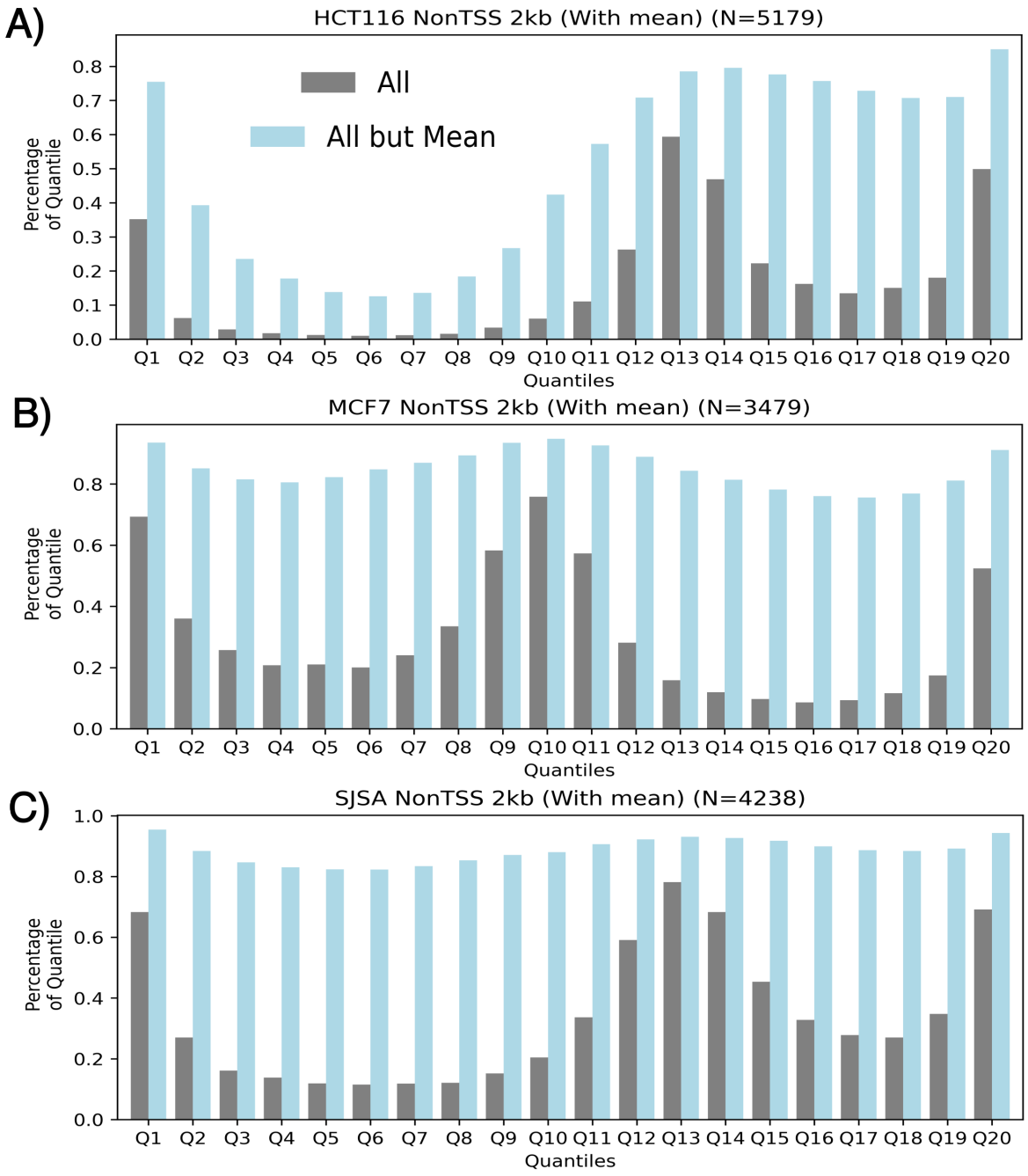

Figure C.4: **The features with highest statistical confidence for differential transcription are consistent across classic differential expression tools for all celltypes.** The percentage of tREs within the same ranked quantile according to all tested tool-parameter combinations (grey) or all excluding combinations using the Mean-based dispersion estimation (light blue) for (A) HCT116, (B) MCF7, and (C) SJSA. tREs are ranked according to direction of change and adjusted p-value where the poles have the greatest statistical significance and the middle tREs have little to no change.

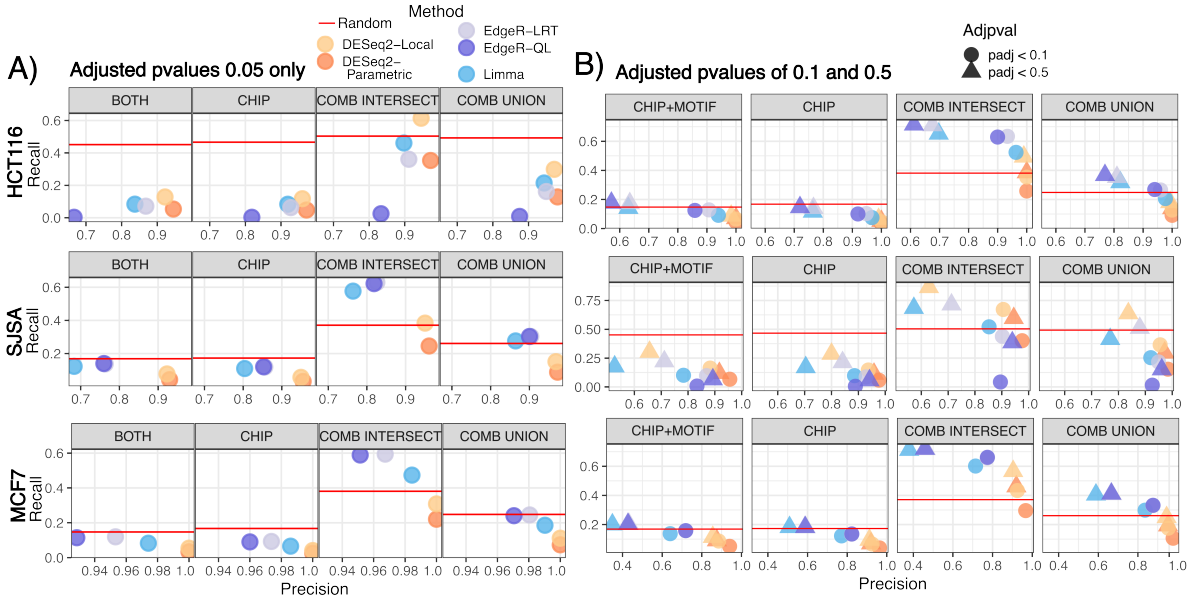

Figure C.5: **High adjusted p-value cutoffs of 0.1 or 0.5 are required to reach recall of p53 truth sets enabled from random calling.** Recall and Precision for Nutlin-3A (p53) responding tREs when using five different classic statistical method combinations. True positives are based on p53 ChIP peaks (CHIP) or peaks with p53 motifs (BOTH) or calls achieved by all tools/parameter combinations (COMB INTERSECT) or any tool/parameter combination (COMB UNION) when considering cell types as replicates. False positives are based on calls without both motif and P53 ChIP peak. Red lines indicate recall from randomly assigning tREs with positive fold changes as a true call. A) Adjusted p-values < 0.05 are used. B) Adjusted p-value cutoffs of 0.1 or 0.5 are used.

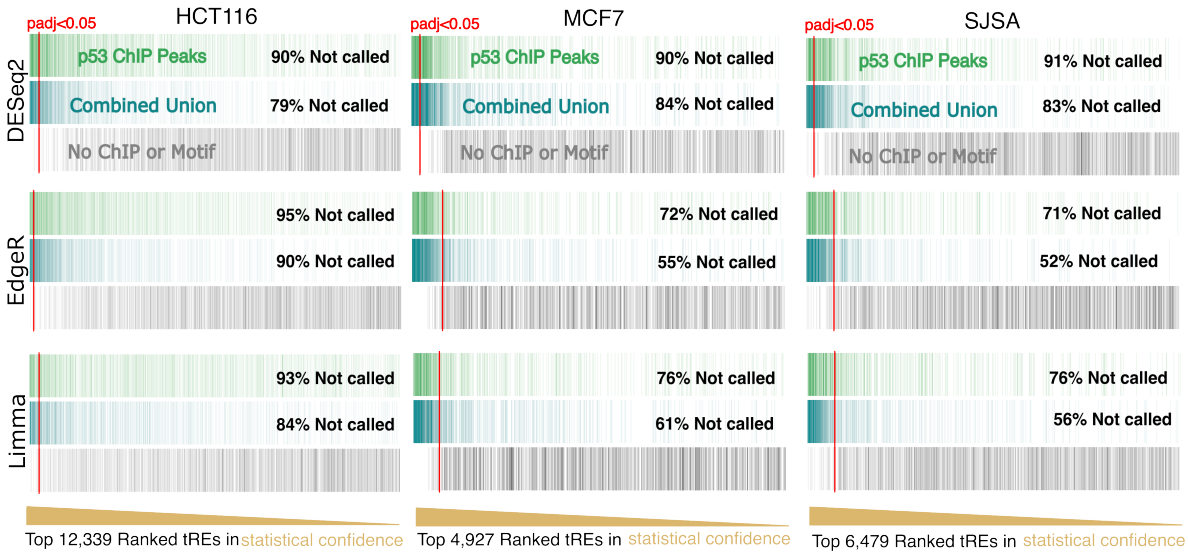

Figure C.6: **True positives are often mixed with statistical confidence of expected true negatives, regardless of cell type and tool-based rankings.** The top tREs with positive log fold change for p53 ranked according to adjusted p-values values (proxy for statistical confidence) for each cell type. The tREs overlapping p53 ChIP peaks for the appropriate cell type are colored green (top), called when using all cell types as replicates ("Combined Union") are colored turquoise (middle), and those with no clear linkage to Nutlin-3a or P53 are colored grey (bottom). A red line corresponds to the position at which all tREs to the left are called significant at p-adjusted value cutoff of 0.05 (padj$<$ 0.05). HCT116 is expected to have very conservative classic results for EdgeR and Limma due to one of the samples having significantly lower overall transcription levels than the rest of the samples [5].

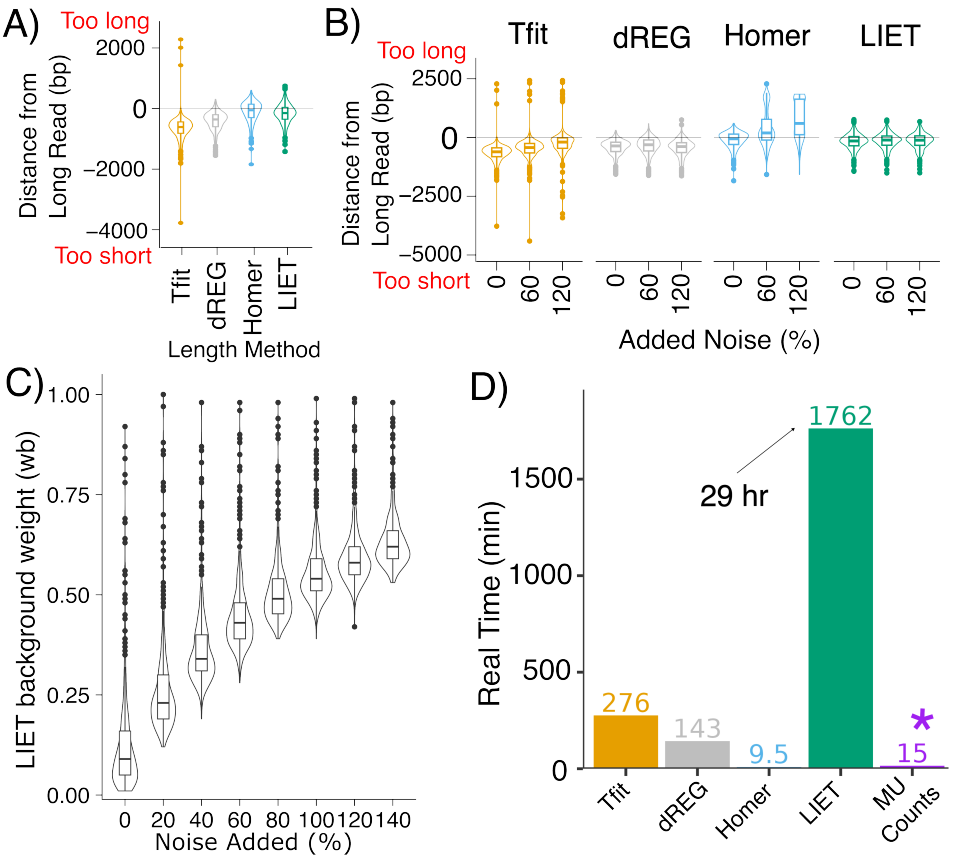

Figure C.7: **Despite taking the longest, tRE-adapted LIET consistently produces lower error in tRE length after considering overlapping transcription.** Results of the four tRE identification methods (Tfit (yellow), dREG (grey), Homer (blue), LIET (green) on length prediction across 411 tREs, both without (**A**) and with (**B**) noise added. Distance from Long Read serves as a proxy for length prediction error (details in Supplementary Methods). A negative value refers to the prediction being too short, and a positive to the prediction being too long. Results for A and B that consider all twelve methods (as described in Supplementary Methods) and all noise levels can be found at (**github:**/Length_Bench/Comparison/Compare_Lengths.ipynb). All results correspond to the short-read data from SRA SRR4454567. Similar results were found when calculating consensus lengths from tools. **C.** LIET background weight (Bayesian prior (posterior estimated graphed here)) effectively captures noise. Results from other samples and LIET-model adaptations can be found in the same notebook above. **D.** Real time taken to run each of the length-predicting methods on one sample, or in the case of mu-Counts in all samples at once. Homer, the fastest approach for a single sample, takes 25 minutes to run on all samples at once.

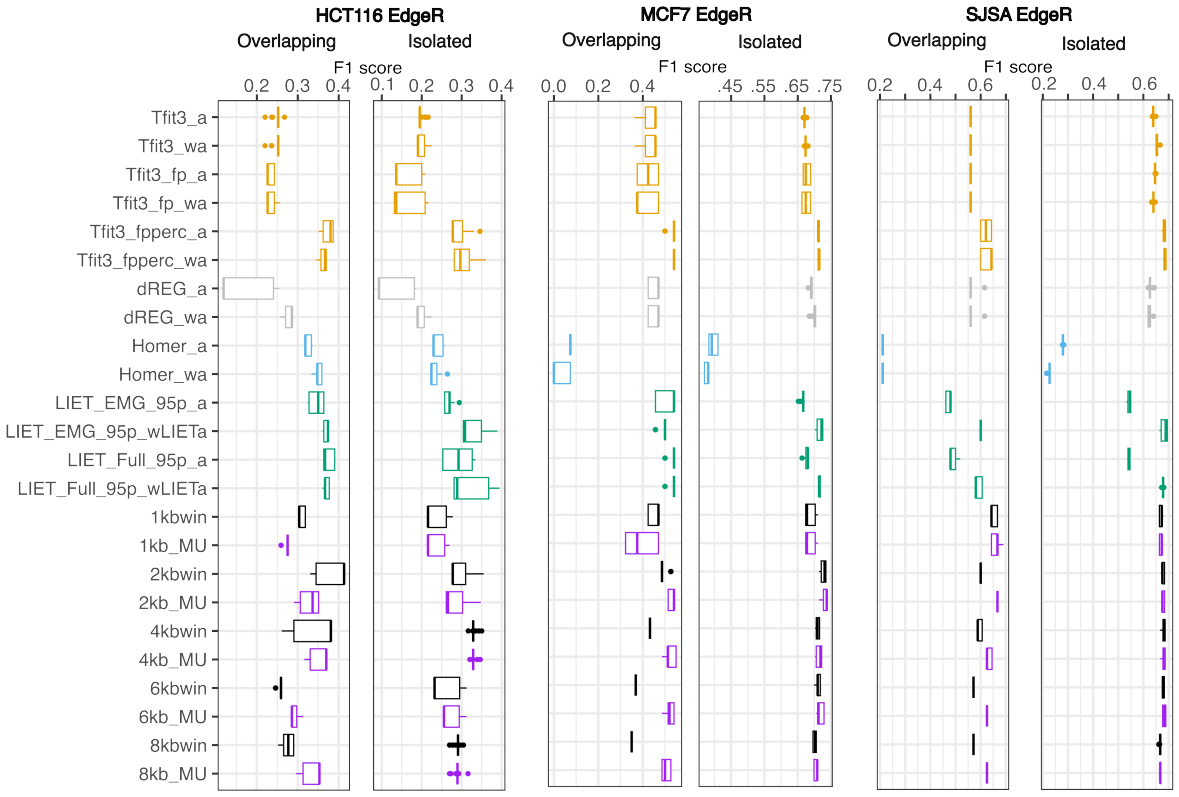

Figure C.8: **tRE-focused LIET and Mu_Counts both show the highest F1 scores for overlapping and isolated tREs across all cell types.** F1 scores using p53 ChIP-peaks as the truth set for HCT116, MCF7, and SJSA when using EdgeR and counts from windows defined by multiple different methods. Methods colored as Figure C.7. Fixed_win refers to a fixed window and Fixed_win_mu (and suffix win) refers to the Mu_Counts (and suffix MU) pipeline being used with the provided initial fixed window. Average is noted by _a, weighted average (based on counts) is noted by _wa. Tfit includes fp and fpperc where the footprint is added to Tfit length, with or without the 95th percentile of the EMG, respectively. EMG_95p means that the 95th percentile of LIET using just the EMG was used. wLIET means that the weighted average was used with weights based on $wLIET = 1 - w_b$ (w_b =background weight in LIET). Details on other methods can be found in Supplementary Methods section. All comparisons (e.g. cell types and platforms) can be found at (**github:**/Bench_DE/Length_DE/p53_Len_Compare_Vis.ipynb).

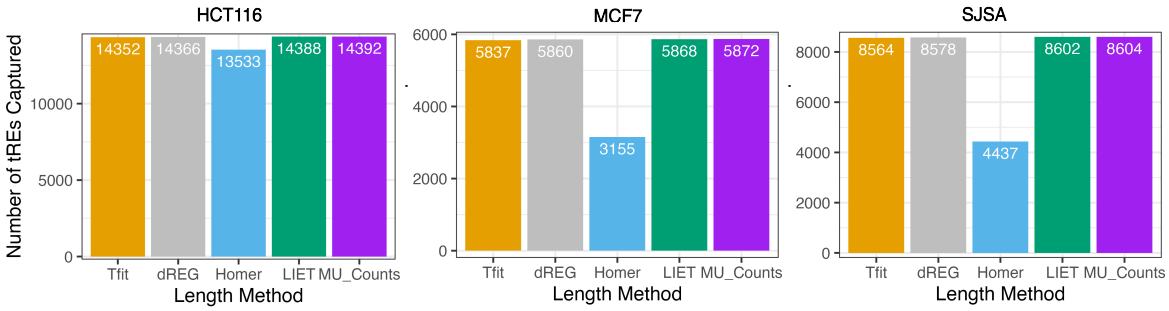

Figure C.9: **Homer provides length for significantly less tREs compared to its counterparts.** The number of tREs used for length-based differential transcription assessment whose lengths could be estimated by the relevant methods: Tfit (Tfit with 3′ bedgraphs), dREG, Homer, LIET (with 1- w_b used for consensus calculation in LIET (details in Supplemental Methods)), and Mu_Counts (so all). The total tREs are based on the consensus tREs determined as described in Supplementary Methods, so that all methods are considering the same "universe" of tREs. Importantly, LIET uses pre-annotated search positions on which to model tREs while Homer cannot take predefined regions on which to consider its model. Therefore, LIET and Mu_Counts have a huge advantage to consider any pre-specified tREs (as defined by tRE identification tools like Tfit and dREG).

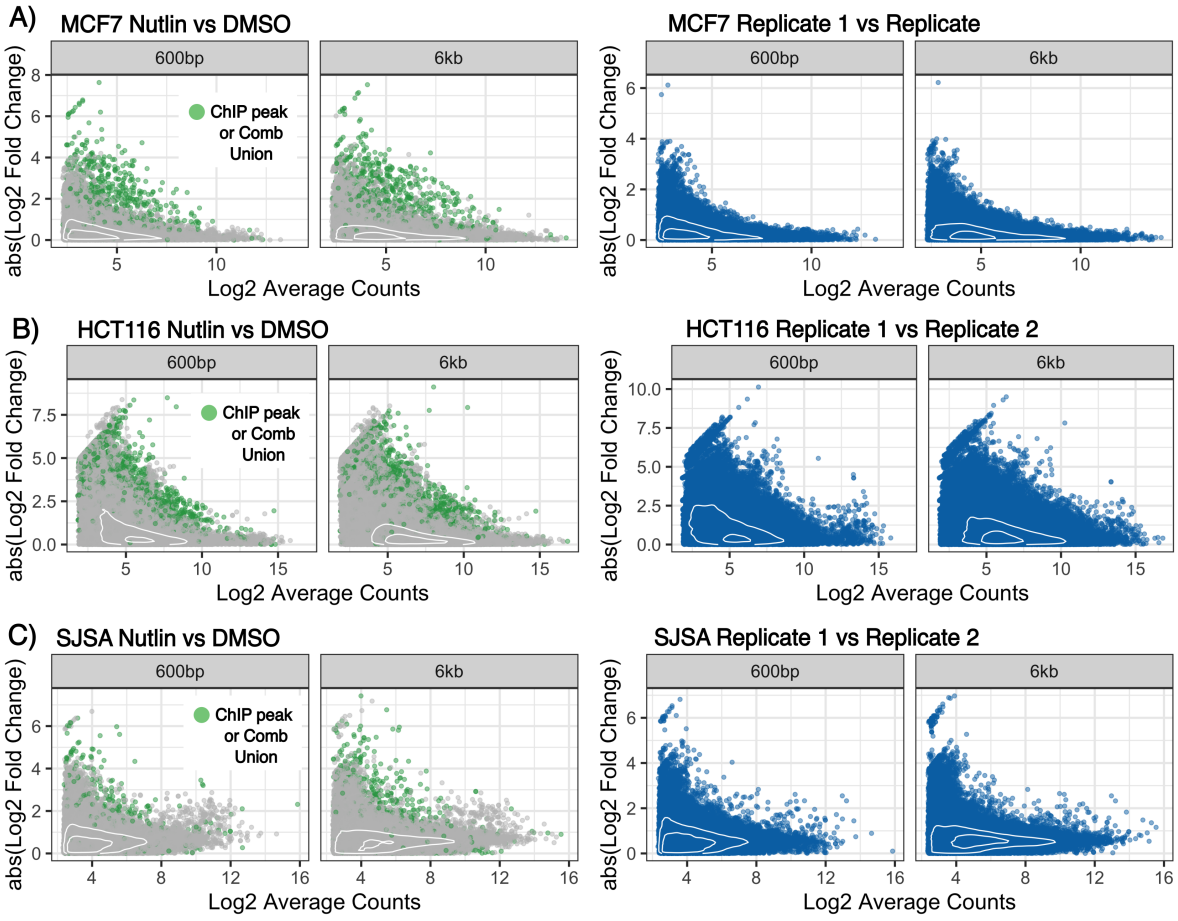

Figure C.10: **Mean-dispersion trends remain similar, leading to low statistical confidence for p53 data, despite length correction with Mu_Counts.** Absolute log fold change of tREs between Nutlin-3a and DMSO (left, green/grey) and biological replicates (right, blue) with tREs that are supported by ChIP peaks or with combined cell types ("Comb Union") are highlighted in green. Results are considered for A) MCF7, B) HCT116, and C) SJSA when using counts from 600bp fixed windows (left) and 6kb Mu_Counts (right).

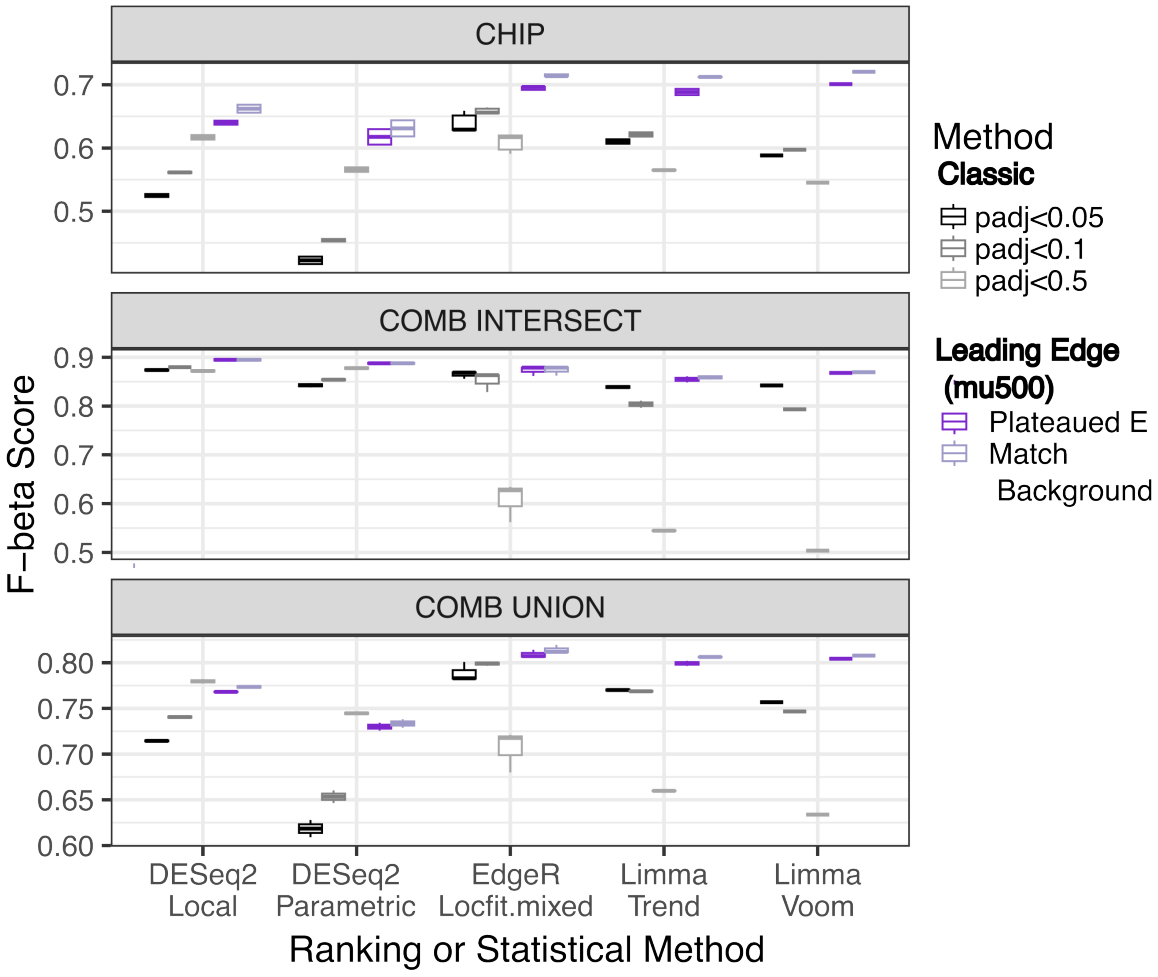

Figure C.11: **The leading edge methods confer improved balance of precision and recall based on F1-beta scores.** F1-beta scores of p53 responsive tREs based on their proximity to corresponding p53 ChIP peaks. These results are from SJSA using Mu_Counts with a max window size of 2kb since they were well-representative of all results. Other cell type and window size results are comparable but can be found at (**github:**/Compare_LE/p53_Compare_LE_stats.ipynb). Graphs with recall and precision mapped as scatter plots are also available at the same notebook. Due to the extreme conservativeness of DESeq2, adjusted p-values of 0.5 were occasionally able to allow a small increase in recall compared to leading edges while maintaining precision above 0.9.

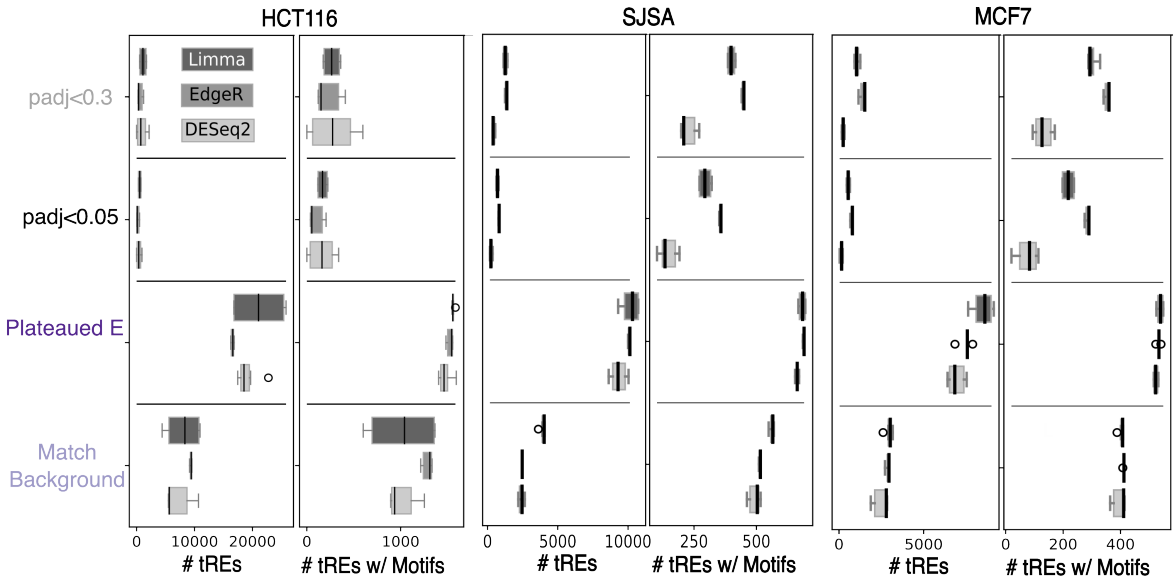

Figure C.12: **Leading Edge positions are consistent across ranking platforms.** Boxplots of leading edge positions for both leading edge methods and classic statistical tools (padj < 0.3 and padj < 0.05) when considering all tREs or just those with the corresponding TF motifs within 1.5kb of the tRE μs (midpoints). The variability of Limma-Voom vs Limma-Trend leading edge results in HCT116 is not observed with any other cell type or condition.

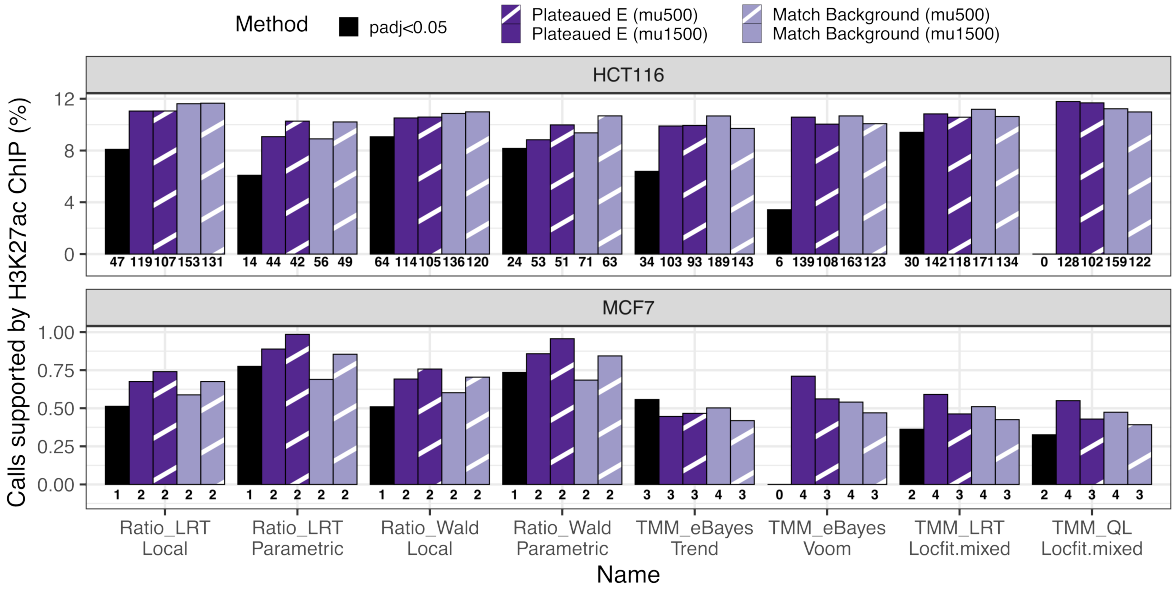

Figure C.13: **Leading edge calls have higher enrichment of H3K27ac support.** Percentage of calls for HCT116 (top) and MCF7 (bottom)supported by ChIP peaks. (Note H3K27ac unavailable for Nutlin-3a in SJSA cells). A H3K27ac peak must be only called after Nutlin-3a has been added to media (1hr for HCT116 and 2.5hr for MCF7). p05 refers to the classic statistical approach cutoff for various methods (column labels), Plateaued E for adding tREs within the "Plateaued Enrichment" leading edge that have a p53 motif within 500bp (mu500) or 1.5kb (mu1500) of their midpoint. Same for "Match Background" leading edge.

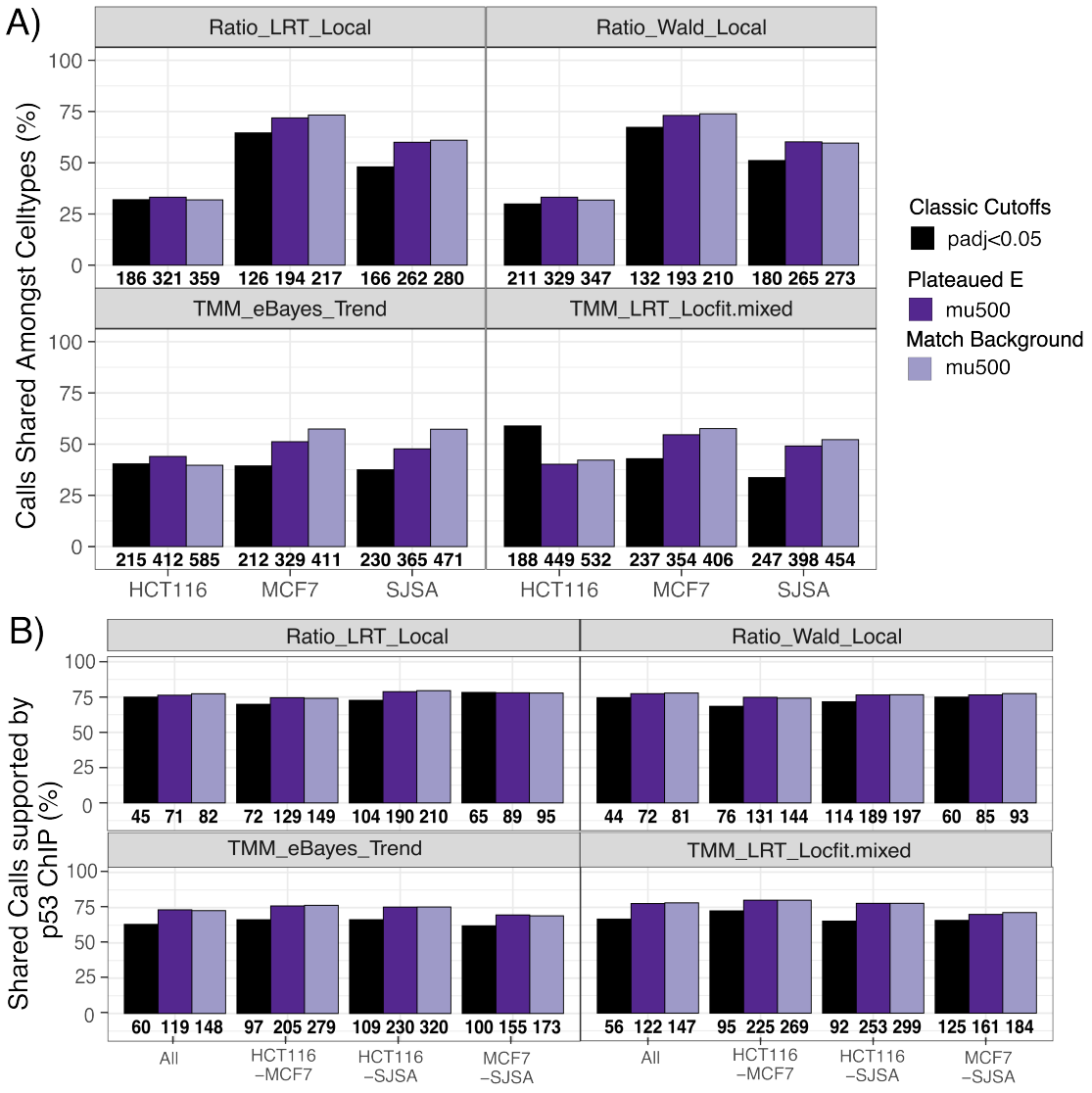

Figure C.14: **Despite leading edge calling more p53-responsive tREs shared across cell types, these calls are generally supported by p53 ChIP peaks A.** Percentage of tREs for each cell type that are called in at least one other celltype (i.e. shared) according to classic tools (padj<0.05) (noted in titles - e.g. Ratio_LRT_Local), or leading edge methods including calls with p53 motifs within 500bp of tRE midpoints (Plateaued E (mu500) and Match Background (mu500)) **B.** Percentage of calls shared between all celltypes or pairwise combinations that overlap p53 ChIP peaks shared across cell types. In all cases, the bottom numbers refer to the N of the bars while the y values graphed are the percentages of the total calls within each method. Results from all classic tool-parameter combinations as well as leading edge with tREs containing motifs within 1.5kb can be found at (**github:**/Bench_DE/Compare_LE/p53_Compare_Celltypes.ipynb)

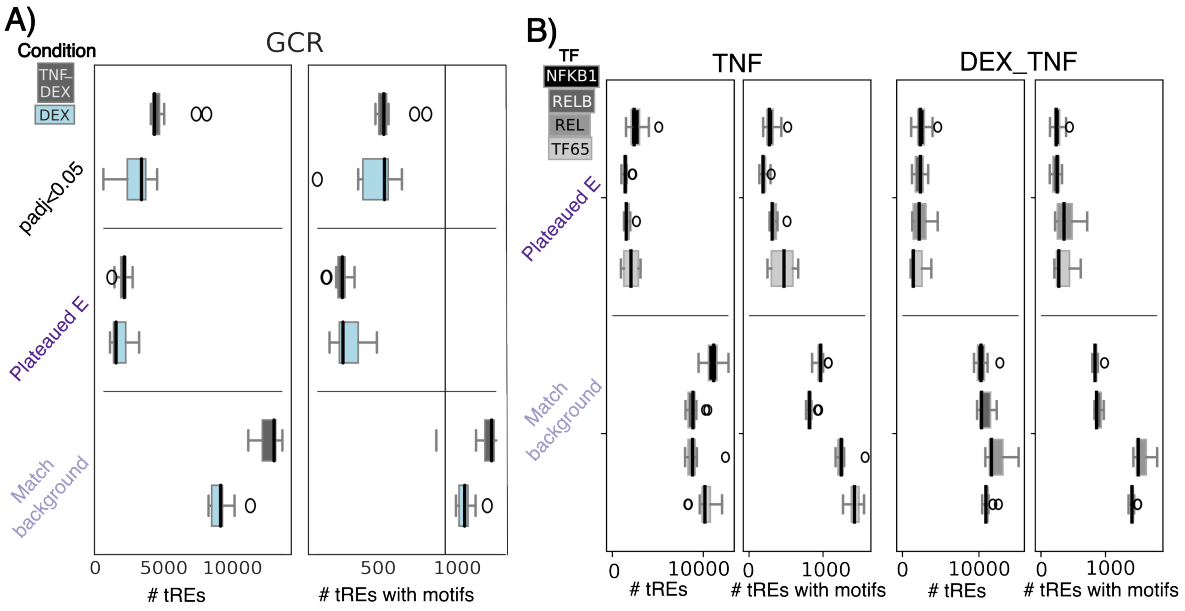

Figure C.15: **Leading Edge positions for GR and NFκB TFs are consistent across ranking methods.** Boxplots of leading edge positions in A) dexamethasone (DEX), B) TNF, or C) dexamethasone and TNF (DEX_TNF) when considering all tREs (left) or just those with the corresponding TF motifs (GR, NFKB1, RELB, REL, TF65) within 1.5kb of the tRE μs (right). Boxplots represent leading edge or classic statistical methods with padj<0.05.

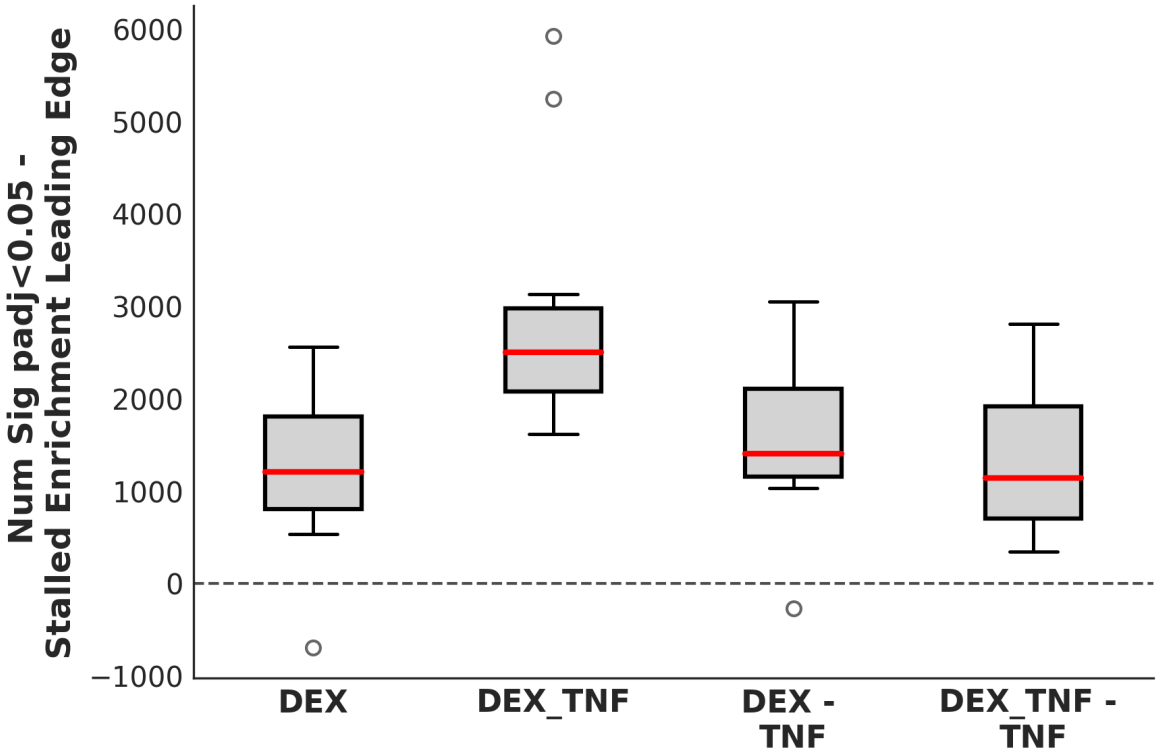

Figure C.16: **Unlike with Nutlin-3a and p53, classic tools call more tREs significant than the GR leading edge for DEX and TNF perturbed cells.** Boxplots of the difference between the number of significant calls according to adjusted p-values < 0.05 and all tREs (regardless of motif) within the Plateaued Enrichment leading edge. Dexamethasone (DEX), dexamethasone with TNF (DEX_TNF), and these conditions with the TNF only significant (adjp < 0.05) calls removed (-TNF). The only case where the classic statistical approach does not call more significant tREs is for Limma (eBayes significance test and Trend dispersion estimation).

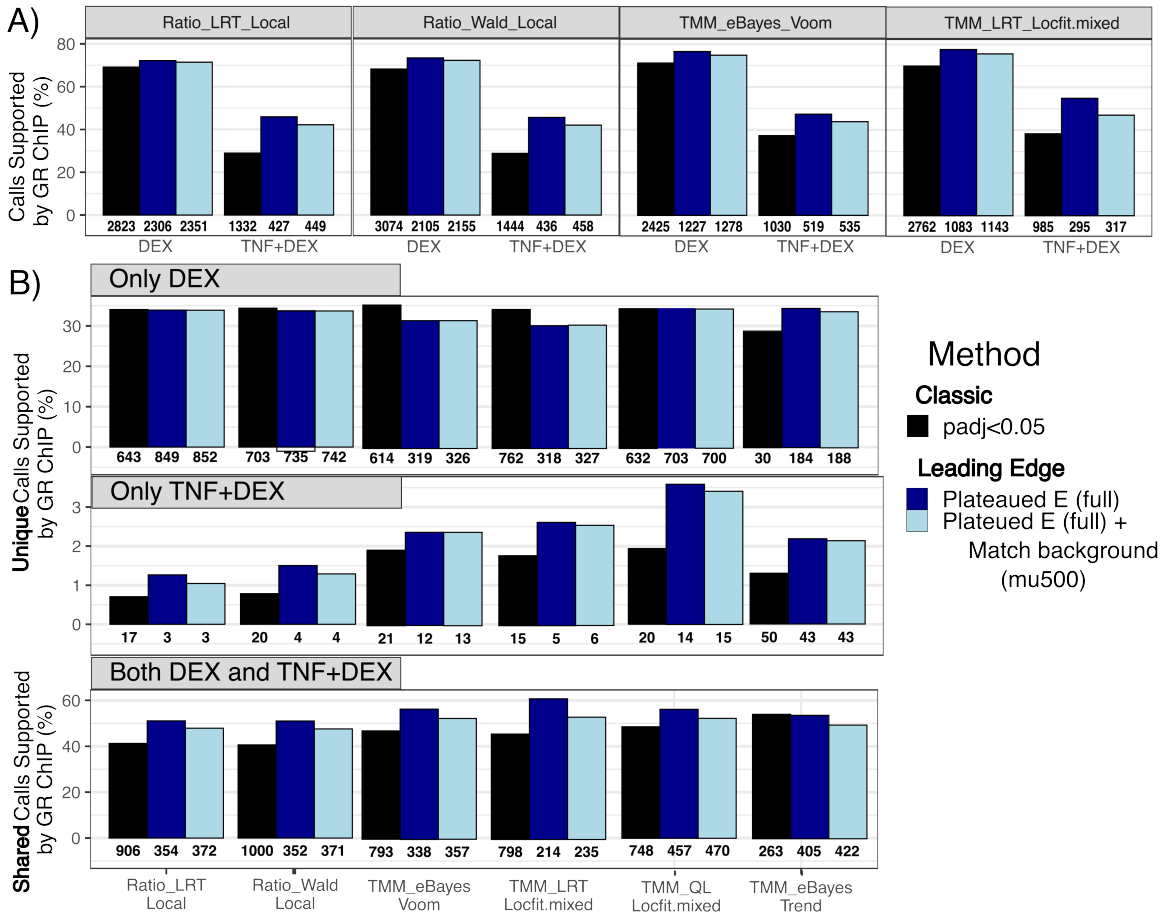

Figure C.17: **Leading Edge based calls for GR for are equally or more enriched in GR ChIP peaks than calls with classic tools. A.** The percentage of calls supported by ChIP-seq peaks of GR for cells treated with dexamethasone (DEX) or both dexamethasone and TNF (TNF+DEX). padj<0.05 refers to the classic approach. Plateaued E (full) means all calls within the Plateaued E leading-edge are considered. Plateaued E (full) + Match Background (mu500) adds tREs within the Match Background leading-edge that have a GR motif within 500bp of their midpoints. Tool-parameter combinations are representative of all results. **B.** The percentage of calls considered either unique to DEX or TNF+DEX perturbed cells and percentage of calls considered shared by both perturbations supported by equivalent comparisons with GR ChIP. Tool-parameter combinations are representative of all results. Calls from leading edge with the classic approach and motif calls (e.g. mu500) are not shown since they give almost equivalent results to the classic statistical approach. Full results including these and all tool-parameter combinations can be found at (**github:**/Bench_DE/Compare_LE/GR_Compare_Conditions.ipynb).

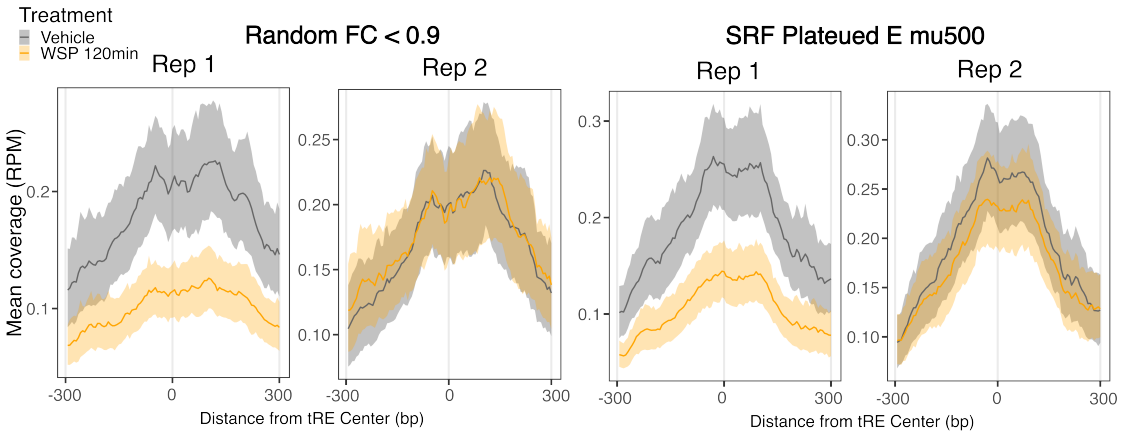

Figure C.18: **Leading edge tREs show decreased accessibility in both replicates while the same number of random tREs with fold changes below 0.9 do not.** Mean reads per-million of the two ATAC-seq vehicle and 120 minute WSP replicates of tREs randomly selected from tREs with DESeq2 fold changes below 0.9 (Random FC < 0.9) vs those with an SRF motif within the Plateaued E Leading edge (SRF Plateaued E mu500) (both N=98). Only 1 tRE is found in both sections. Results for SRF Match Background mu500 (with its own random set) look almost identical. The tREs that have FC below 0.9 but are outside the Plateaued E LE show lower decreased accessibility with overlaps of confidence intervals for combined replicates, unlike the tREs of the same number just barely within the leading edge. These graphs can be found at (**github:**/WSP/Graph_WSP_metaplots.ipynb).

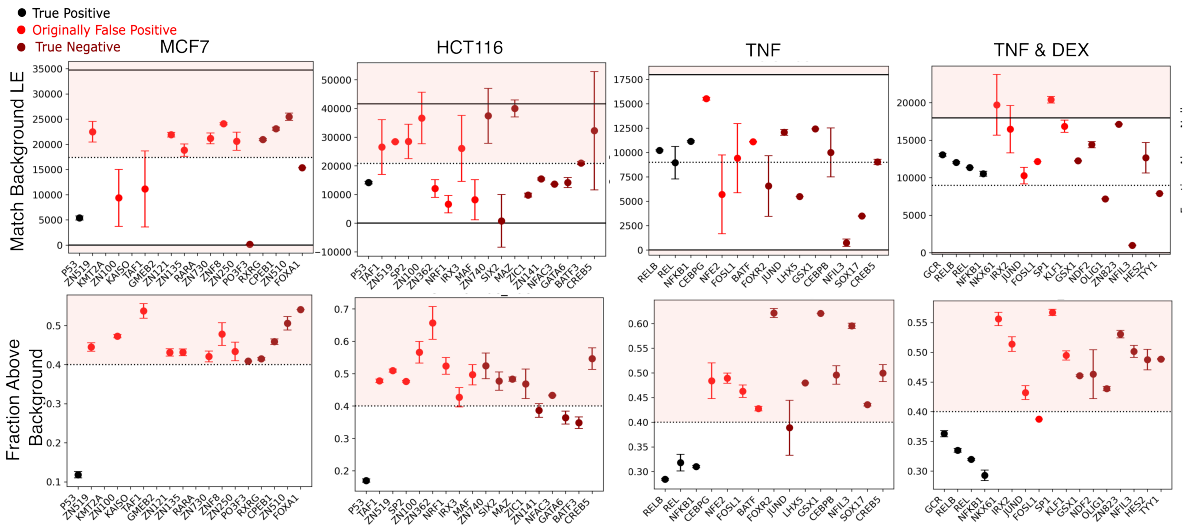

Figure C.19: **Leading-edge related values serve as robust secondary metrics of false positive TFEA calls.** TFs are color-coded as known expected calls (black), improperly called significant (light red), or properly called non-significant enriched (dark red). Dots indicate the number of tREs within a Match Background leading edge (top) or fraction of tREs with cumulative enrichment scores above that expected from background (bottom). The quarter point of tREs and below 0 (top) or fraction above 0.4 (bottom) are highlighted. HCT116 and MCF7 refer to Nutlin-3a (p53) datasets for these celltypes. TNF and TNF & DEX refer to lung cells perturbed with TNF or both TNF and dexamethasone. Only these are shown for brevity; other celltypes and perturbations can be found at (**github:**/Improving_FP_calls/LE_FP.ipynb).

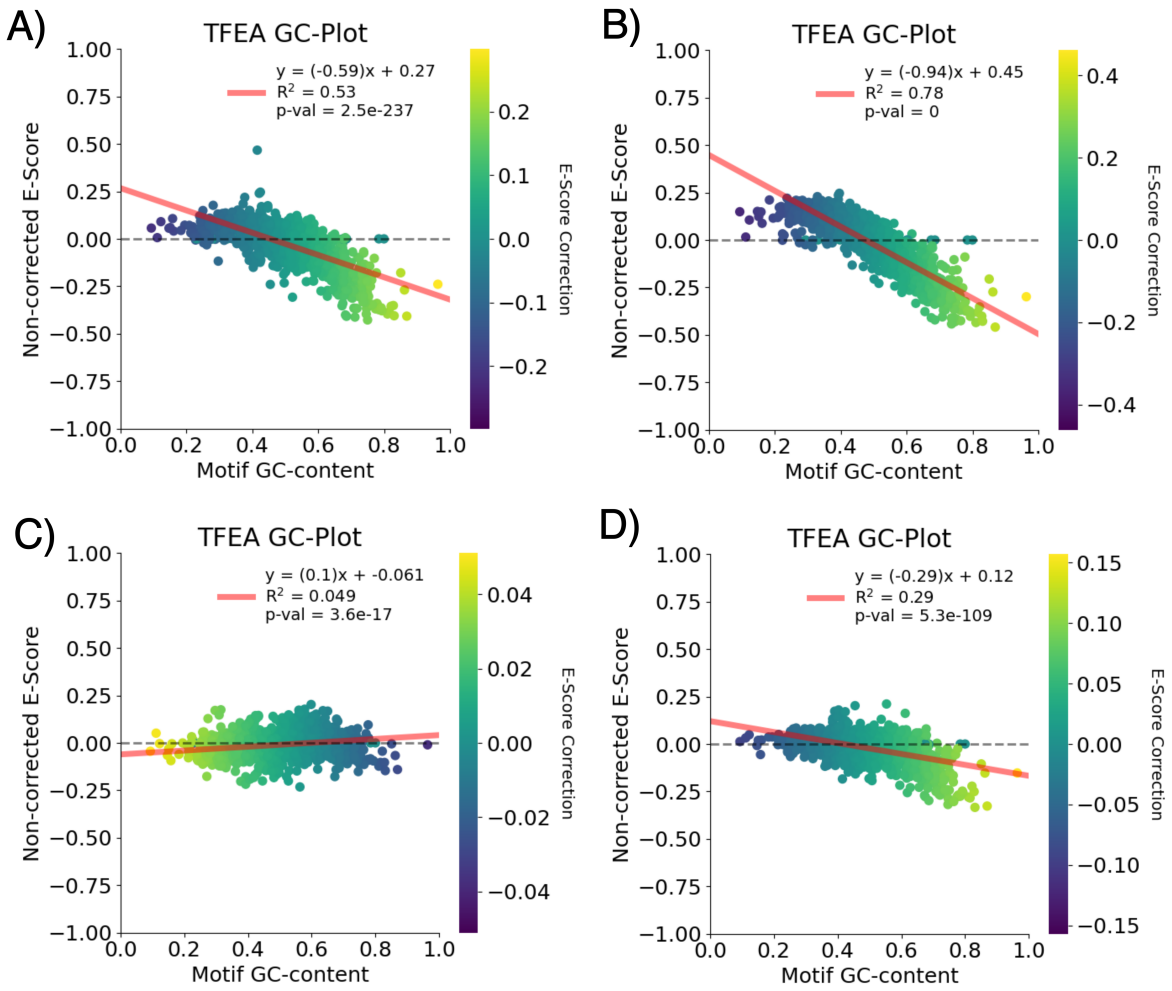

Figure C.20: **PRO-seq and to a lesser extent ATAC-seq data for woodsmoke particles have strong GC-bias.** GC-correction graphs from TFEA for A) PRO-seq 30min, B) PRO-seq 120min, C) ATAC-seq 30min, D) ATAC-seq 120min. Each dot represents a TF with its original score (non-corrected E-score, y-axis) against its GC-content (y-axis) and colored by the new E-score after correction.

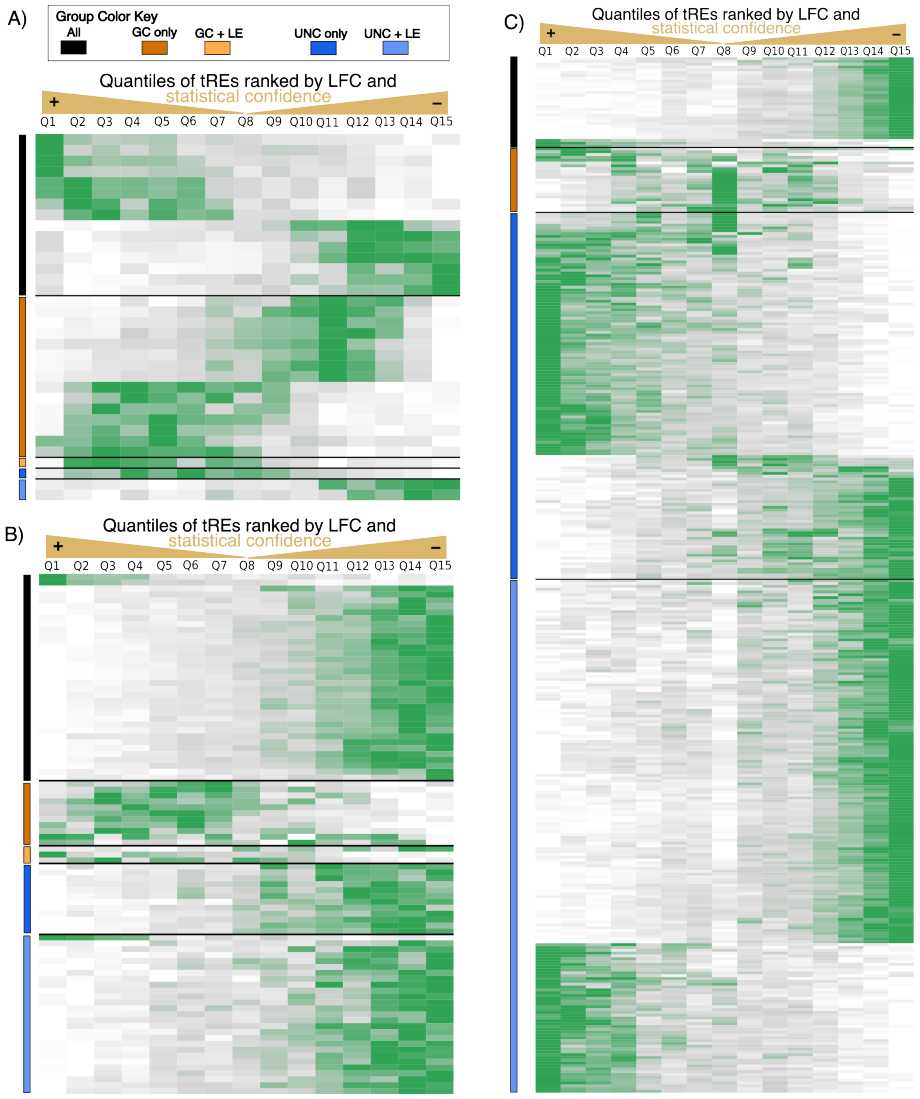

Figure C.21: **Leading edge metrics successfully remove false positive TF calls in both ATAC and PRO-seq, regardless of GC correction** Quantile enrichment plots for A) ATAC-seq WSP at 30min vs Veh, B) ATAC-seq WSP 120min vs Veh, and C) PRO-seq WSP 120min vs Veh. Transcription factor calls are split into All (black, LE metrics support, GC-corrected and uncorrected adjusted p-values < 0.001), GC only (red, GC-corrected but not uncorrected adjusted p-value < 0.001 and not supported by LE-metrics), GC-LE (orange, supported by GC-correction and LE-metrics but not uncorrected adjusted p-value), UNC only (dark blue, supported by only uncorrected adjusted p-value), UNC+LE (light blue, supported by only uncorrected adjusted p-value and LE-metrics). Slope at quantiles of tREs (each quantile has 2,944 tREs). Ns: ATAC-seq 30min All=15, GC only=15, GC+LE=1, UNC only=1, UNC+LE=2; ATAC-seq 120min All=39, GC only=11, GC+LE=3, UNC only=12, UNC+LE=29; PRO-seq 120min All=33, GC only=24, GC+LE=0, UNC only=137, UNC+LE=207.

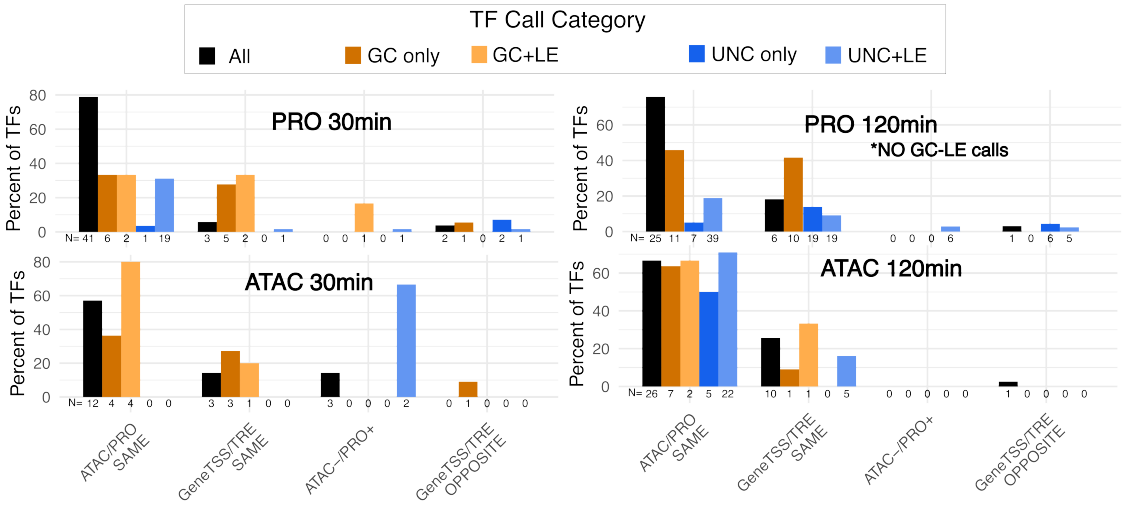

Figure C.22: **Direction of TF significant calls between ATAC and PRO or gene TSS bidirectionals and tREs was mostly enriched when including leading-edge focused metrics.** Percentage of TFs from each of the TF call categories that have enrichment scores going the same direction in ATAC-seq and PRO-seq (ATAC/PRO SAME) or with gene TSS bidirectionals and tREs (GeneTSS/TRE SAME), or suggest closing of chromatin but increased transcription (ATAC-/PRO+) or opposite directions between gene TSS bidirectionals and tREs (GeneTSS/TRE OPPOSITE). All refers to TFs called significant by GC-correction, no correction, and LE metrics. GC only and GC+LE refer to TFs called with GC-corrected but not uncorrected significance, and called without or with LE metric support, respectively. UNC only and UNC+LE refer to TFs called with uncorrected but not GC-corrected significance, and called without or with LE metric support, respectively. PRO-seq 120 minutes did not have any TFs supported by both GC-correction and Leading edge metrics (NO GC-LE calls).

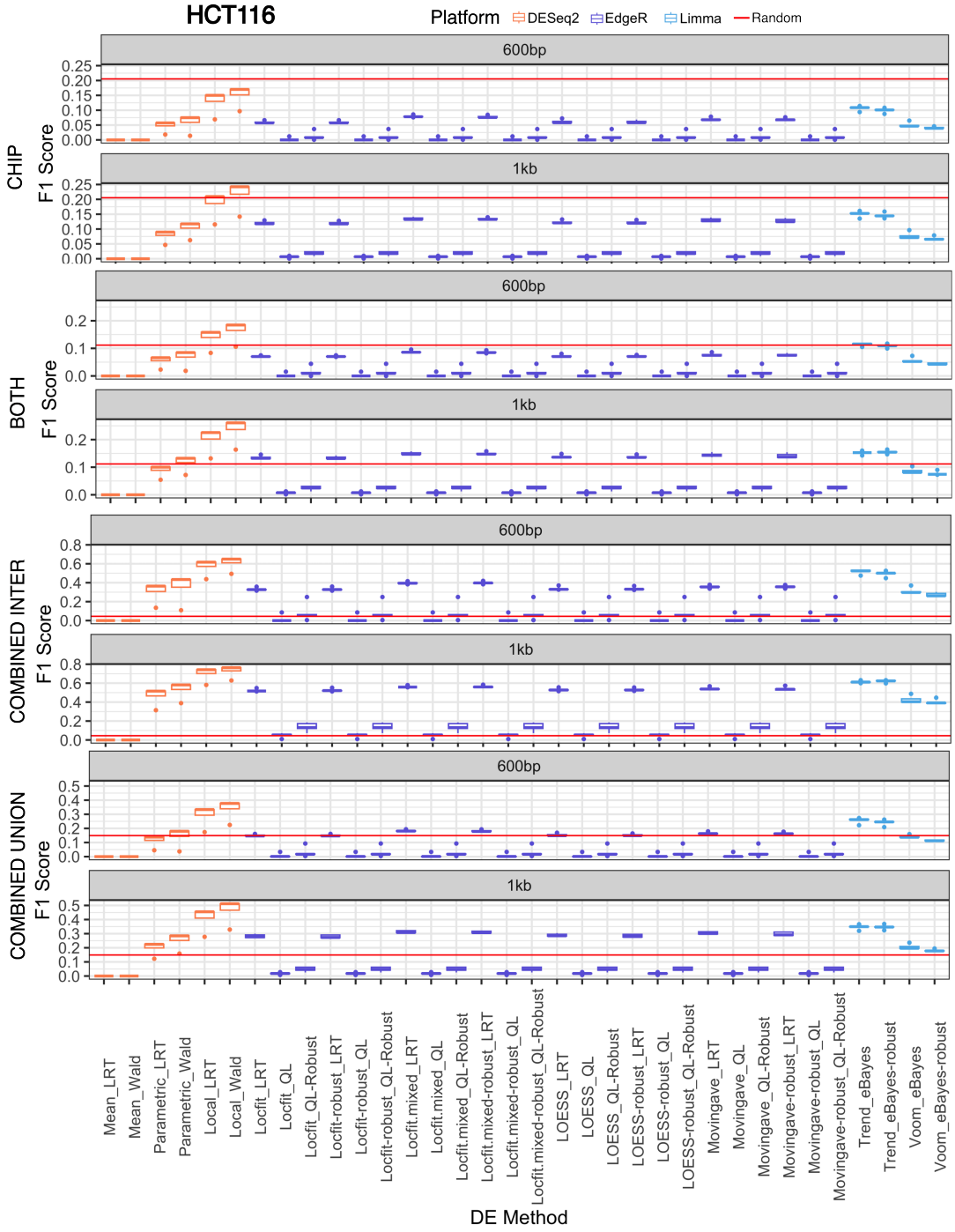

Figure C.23: **Low F1 scores, specifically for the DESeq2-mean parameter combination, are consistent across varying truth sets across cell types.** A. HCT116 cells (Continued on the following page.)

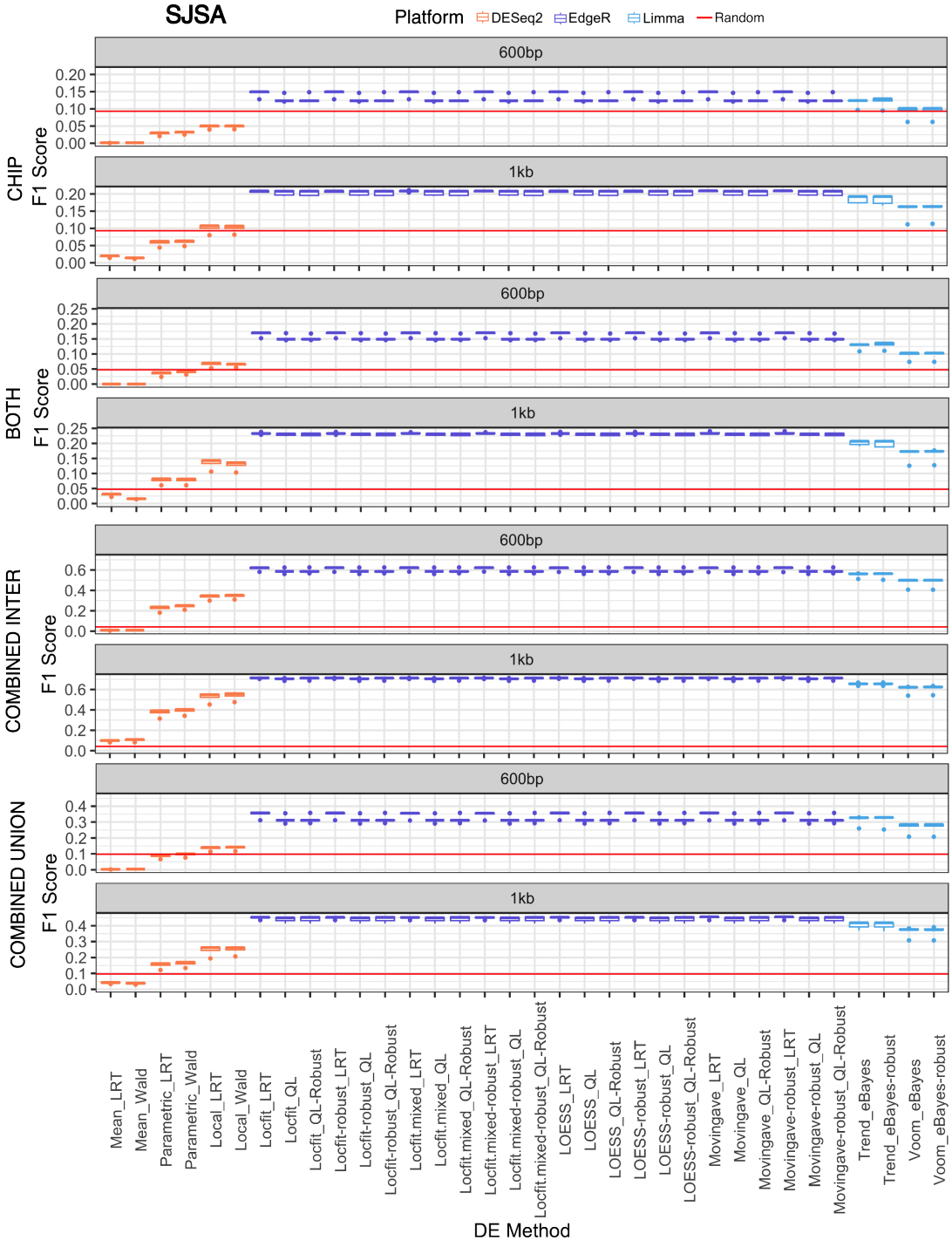

Figure C.23: B. SJSA cells (Continued on the following page.)

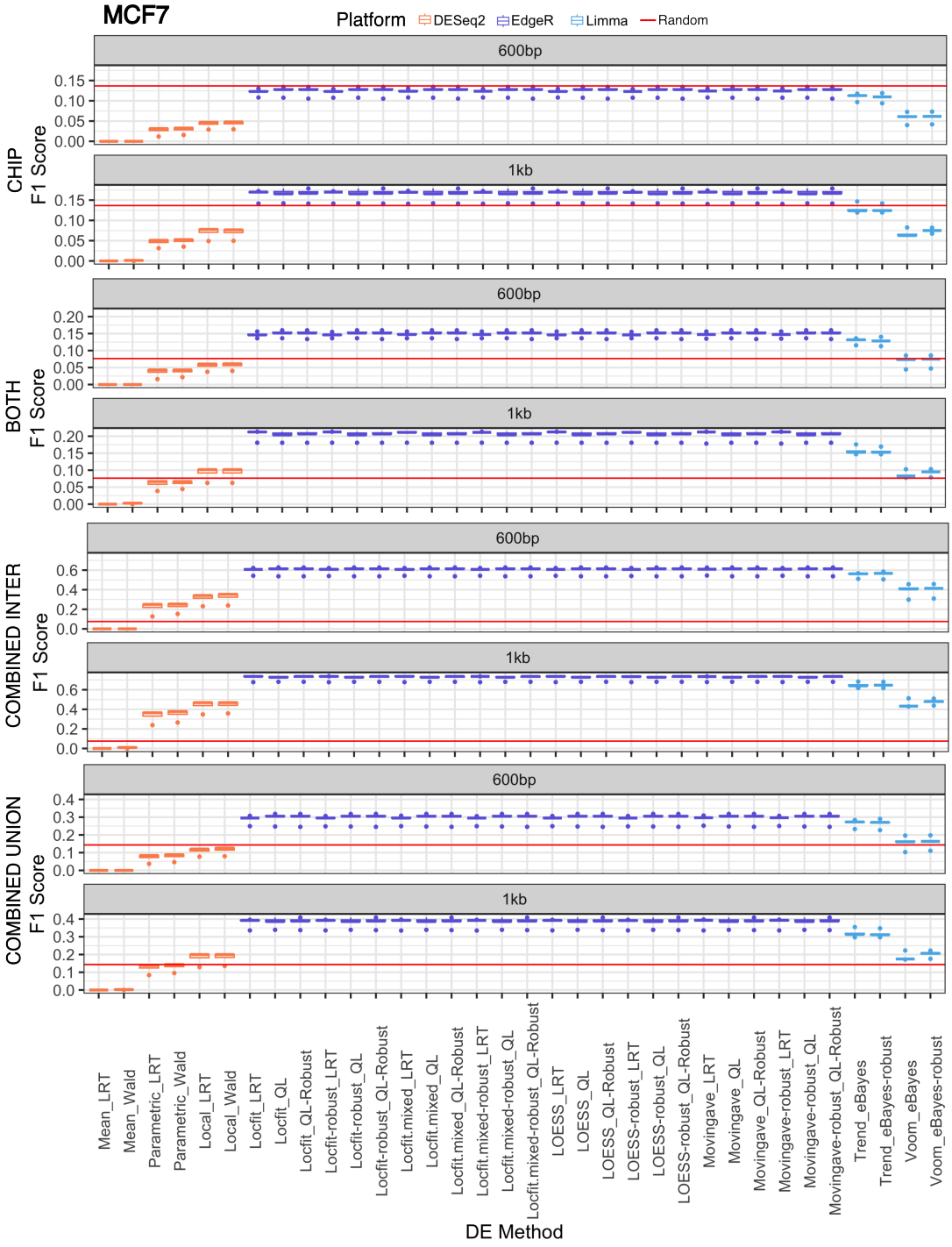

Figure C.23: C. MCF7 cells. (A-C) F1 scores for all tested dispersion-significance test combinations are shown as boxplots colored according to the platform (DESeq2, EdgeR, or Limma). The red horizontal lines refer to the median F1 score calculated when randomly assigning features as significant five times. 600bp and 1kb refer to the window sizes used over which to count tREs. BOTH, CHIP, Combined Intersect, and Combined Union refer to the true positive sets used (details in Supplementary Methods).

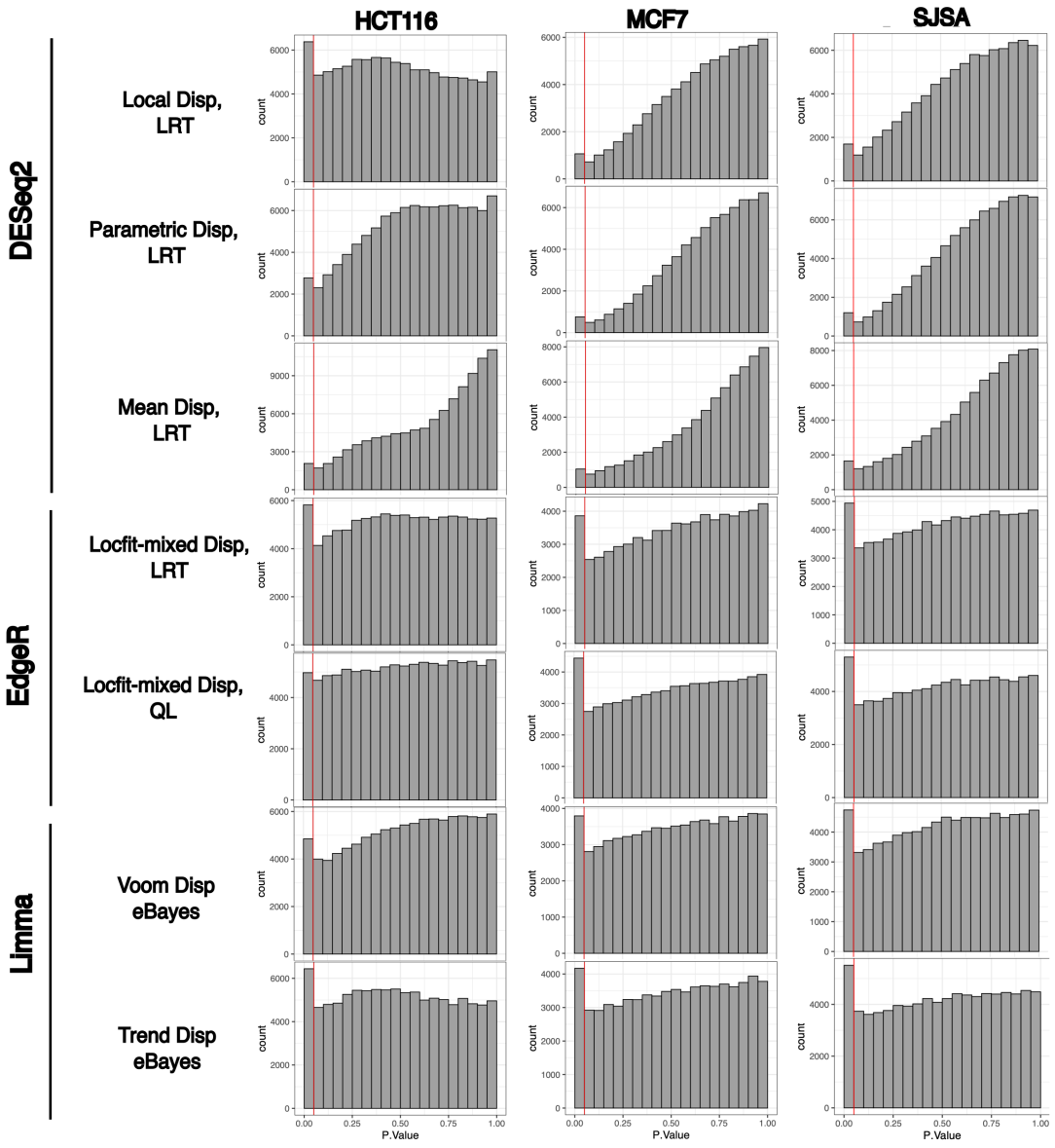

Figure C.24: **Distribution of p-value scores (non-adjusted) reflects unclean analysis particularly for DESeq2, and mean-dispersion estimation.** The red line indicates the 0.05 position (common p-value cutoff). Built-in normalization factors (Ratio for DESeq2, TMM for EdgeR and Limma) were used here but use of virtual-spike in normalization factors show almost identical results (data not shown) [153]. DESeq2 Mean Dispersion (Disp) shows bumps in p-values common with poor dispersion calling. EdgeR and Limma showed the strongest consistency in a peak at the 0-0.05 p-value position with a generally uniform distribution after. All outputs were from using Mu_Counts with a 2kb fixed window to ensure biased counting approaches did not explain comparisons. To simplify visualization, only one of tested combinations with almost identical results are shown (e.g. DESeq2 Wald vs LRT, QL-Robust vs QL).

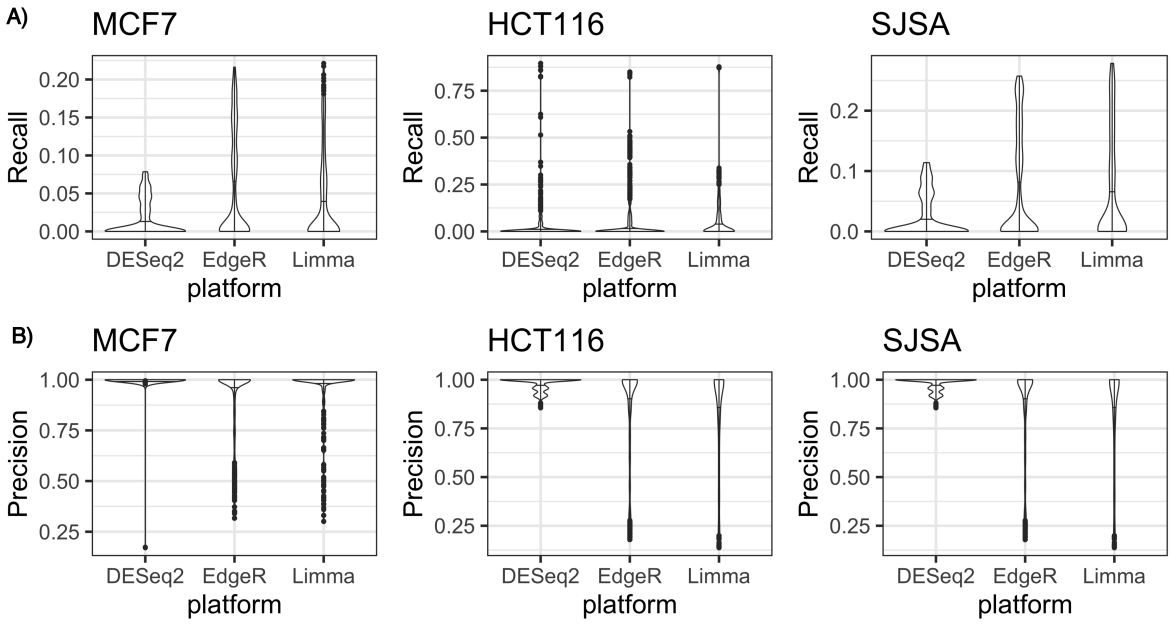

Figure C.25: **AUC-PR metrics would be inappropriate due to variable ranges of precision and recall across tools and parameter combinations for p53 based analysis.** Precision and recall for the three tools (DESeq2, EdgeR, and Limma) were tested using a broad range of parameters and p-value cutoffs (see Supplementary Methods), and results are shown as boxplots. MCF7, HCT116, and SJSA refer to the cell types considered. The true positive set used was based on ChIP identified p53 calls (details in Supplementary Methods).

Appendix D

Appendix D. Supplemental Material to Air pollutant multiomics improves functional annotation of SNPs associated with lung disease

D.1 Methods

D.1.1 Cell culture

[awaiting updates from Sarah]. Cell culture of primary small airway epithelial cells (smAECs) was performed as described previously[79]. Briefly, deidentified primary human small airway epithelial cells (< 2 mm bronchiole diameter) were obtained from the National Jewish Health Biobank from donors with no history of deployment or chronic lung disease. Cells were first cocultured with irradiated NIH/3T3 (ATCC) fibroblasts in F-medium containing 1 μM Y-27632 (APEX Bio) and then passaged onto tissue culture dishes double-coated with Type I bovine collagen solution (Advanced Biomatrix) and grown to confluence in BronchiaLife Epithelial Medium with the complete LifeFactors Supplement Kit (Lifeline Cell Technology). All exposures were performed in this medium. All cells were maintained in 5% CO2 at 37°C.

D.1.2 PRO-seq

[awaiting updates from Sarah]. Cells required treatment and harvest on two separate days to generate enough nuclei to perform Precision Run-on Sequencing (PRO-seq) on two replicates per condition [untreated control (Ctrl) vs. UPM]. For each set of replicates, Sm36 cells (small airway epithelial cells from donor 36) were seeded on 5- × 15-cm double collagen-coated plates per treatment and grown to confluence (3–4 days), then left untreated (Ctrl) or treated with

UPM (100 μg/cm2) for 30 or 120 minutes. Cells were harvested and nuclei prepared as reported previously[210]. PRO-seq libraries for all replicates were then prepared simultaneously by subjecting one aliquot of $1x10^7$ nuclei/sample to 3-min nuclear run-on reactions in the presence of Biotin-11-CTP (PerkinElmer) following our previously detailed protocol [68]. Uniquely indexed libraries were pooled and sequenced on an Illumina NextSeq instrument using 75 bp single-end reads by the BioFrontiers Sequencing Facility at the University of Colorado Boulder.

D.1.3 ATAC-seq

[This might not be needed but for now updated description from previous publication]. Nasal epithelial cells were grown to confluence in 6-well tissue culture dishes and treated with vehicle (PBS), UPM, or WSP for 30 or 120 min. Cells were rinsed and scraped in ice-cold PBS, then around 50,000 cells from each treatment were pelleted and processed in duplicate for Omni-ATAC-Seq using the protocol developed by [42]. Uniquely indexed libraries were pooled and sequenced on an Illumina NextSeq using 37 bp paired-end reads by the BioFrontiers Sequencing Facility at the University of Colorado Boulder.

D.1.4 Genome

Unless noted otherwise, hg38 genome was used and Refseq annotations from GrCh38.p14 were used. Putative transcripts were included for enhancer annotation (whether intronic, exonic, intergenic) but not for differential expression analysis. Putative genes were not included for counting/differential expression analysis, but were included in all other cases (e.g. linking tREs to exons/introns).

D.1.5 Fastq processing

Unless otherwise noted, samtools v1.8, bedtools v2.28, python v3.6.3, bbmap v38.05, fastqc v0.11.8, hisat2 v2.1.0, preseq v2.0.3, and igvtools v2.3.75 were used. In all cases, fastq files were qc'd with FastQC, trimmed for adapters, mapped to hg38 using hisat2, and resulting SAM/BAM

files qc'd before being converted to TDF format using igvtools for visualization. Scripts for these steps can be found in Preprocessing. RNA-seq fastqs were qc'd, trimmed, and mapped to hg38 using the Nextflow pipeline https://github.com/Dowell-Lab/RNAseq-Flow (commit 6e73ba2). Fastq files from PRO-seq were qc'd, trimmed, and mapped to hg38 using the Nextflow pipeline https://github.com/Dowell-Lab/Nascent-Flow (commit 42add22). ATAC-Seq fastq files were trimmed for adapters, minimum length, and minimum quality using the bbduk tool from the BBMap Suite (v38.73) with arguments "ref=adapters.fa ktrim=r qtrim=10 k = 23 mink=11 hdist=1 maq=10 minlen=20." Quality control was monitored both pretrim and post-trim for all samples using FastQC. Trimmed reads were mapped to the human genome (hg38; downloaded from the University of California Santa Cruz genome browser on September 16, 2019, with corresponding hisat2 index files) using hisat2. Resulting SAM files were converted to sorted BAM files using samtools (v1.9). Read coverage was then normalized to reads per million mapped using a custom python script, and files were converted to TDF format using igvtools (v2.5.3) for visualization in IGV. Reads were de-duplicated in BAM files for counting.

D.1.6 Getting enhancer coordinates from PRO-seq

Tfit v1.0 [12] was run for each experiment using Nextflow pipeline Bidirectional-Flow with the updates outlined in (Townsend et al., 2025) (https://github.com/Dowell-Lab/Bidirectional-Flow (branch Tfit_focus, commit cb74496)). 3-prime bedgraphs were used as input into Tfit using the –tfit_3prime parameter. Consensus bidirectionals across the three perturbation experiments (WSP, UPM, ADP) were identified using muMerge v1.1.0 (https://github.com/Dowell-Lab/mumerge). LIET-EMG (https://github.com/Dowell-Lab/LIET/tree/LIET_EMGtoo, branch LIET_EMGtoo, commit f3b1b31) was then using muMerge μs. Annotations provided to LIET included the start being the μs from muMerge, and ends being 200bp downstream. Paddings were 3kb up and downstream unless there was another enhancer mu within that region in which case the distance between enhancers was used, with a minimum of 500bp. Any cases where the 95th percentile of the LIET model extended beyond the pad had paddings decreased by 200bp in the relevant direction

(upstream/downstream), with a minimum pad of 300bp. As described in previous work [246], we used a weighted average (weighted by coverage not attributed to background) of the 95th percentile of the LIET-EMG model to get consensus lengths from the particulate experiments. Code for all these steps can be found in the Preprocessing section of the Github.

D.1.7 Differential expression analysis

For RNA-seq, featurecounts (subread v1.6.2) was used with parameters -0, -s 1, -t "exon" according to the Refseq GTF. For PRO-seq, genes and bidirectionals were counted over using the Nextflow pipeline at https://github.com/Dowell-Lab/Bidir_Counting_Analysis (commit 4076721) using count type "MU_COUNTS" and the count_limit_bids of 60. This pipeline was run separately for each experiment (ADP, UPM, WSP) but using the same mumerged regions (cons_file). Briefly, this pipeline identified gene TSS bidirectionals from tREs, and addressed overlapping regions in counting (Townsend, 2025). Genes were counted over using 5' truncated regions as detailed in (Sigauke et al., 2025; Townsend et al., 2025). For ATAC-seq, 1kb regions were used (500bp around mus). Featurecounts was again used, counting over bams with duplicated reads de-duplicated. Code and results for counting can be found in the Github under Counting.

All differential expression analysis was done using DESeq2 (v1.44.0) with PRO-seq and RNA-seq size factors calculated based on a list of housekeeping genes for the sake of consistency (size factors based on all gene counts were very comparable). For ATAC-seq analysis of enhancers, the size factors were based on DESeq2 size factors when using merged enhancer regions to ensure no double counting (non-strandedness of ATAC-seq means counts can't be deconvoluted between overlapping regions), but the analysis was done using counts from the non-merged regions. Overenrichment analysis of GO terms was done using org.Hs.eg.db (v3.19.1) and clusterProfiler::enrichGO (v4.12.6), using all possible ontology terms, and a multiple correction adjustment using the Benjamini-Hochberg approach. Due to the large number of significantly enriched terms, we then summarized significant terms into categories. The full GO results and their annotated categories can be found in Supplemental Table GO. WSP and UPM PRO-seq both showed extremely robust responses, likely

increasing the false positive rate. Therefore, we used more conservative adjusted p-value cutoffs (1e-10 for genes and 0.001 for enhancers) to reduce false positive calls and ensure downstream functional analyses were focused on regions with high confidence of change. General results do not change when increasing stringency of cutoffs to $1x10^{-20}$. In all other cases, an adjusted p-value cutoff of 0.05 was used. Code, notebooks, and results from all differential expression analyses can be found in the Github under Diff_Exp.

D.1.8 Differential transcription factor motif enrichment (TFEA-LE)

TFEA (from Github https://github.com/Dowell-Lab/TFEA/tree/Lead_edge, branch Lead_edge, commit 76ef85a) was used with HOCOMOCO motifs from v12. TFEA was run separately for TSS bidirectionals and tRE bidirectionals, and for each timepoint and condition (e.g. WSP 30min vs Veh). TF motif p-value cutoffs were edited as suggested by original authors to get the estimated number of TF motif instances to be between 700 and 5000 to ensure p-values were not shrunk due to simply high number of motif instances. TF Motif ZN362.H12CORE.0.P.C was not included in analysis because it had motif instance numbers above 15,000. The code for this and final p-values can be found in the Github. Final TF significant calls were done based on Townsend et al. 2025: motif instances between 600 and 10000, Fraction above background ≤ 0.46, GC-corrected adjusted p-value ≤ 0.01, and the Match background leading edge is above the plateaued enrichment leading edge but below half the number of tested enhancers.

SRF and AhR/AhRR bidirectionals were assigned by considering those enhancers within the plateaued enrichment leading-edge from TFEA-LE that also had a motif instance within 500bp of the enhancer μ. This was done for each timepoint and condition (e.g. WSP 30min vs Veh).

Full code for these analyses can be found at TFEA/.

D.1.9 Comparing enhancer and TF calls

Enhancers and genes were assigned to different timing categories based on their response across the 30min and 120min timepoints, when available. Early rise (or fall) features were those that

were significantly increasing (or decreasing) at the 30min mark compared to vehicle and decreasing (or increasing) at the 120min mark compared to 30min and were not significantly different between vehicle and 120min. Plateau rise (or fall) features were those that were significantly higher (or lower) in both 30min and 120min compared to vehicle with no significant change between 30min and 120min. Late rise (or fall) features were those that significantly increased (or decreased) expression compared to Vehicle and 30 min, but were not significantly changed at 30min compared to Vehicle. For RNA-seq, the 30min was replaced by the 2hr (earliest time point) and 120min with 4hr (latest time point).

D.1.10 qRT-PCR

[Confirming this with Arnav] smAEC cells were grown to confluence in 6-well tissue culture dishes as previously described[78] and treated with vehicle (PBS), UPM, or WSP for 3 or 6 h. Cells were harvested in TRIzol (Life Technologies) and RNA purified by PureLink RNA Mini Kit (Life Technologies) prior to qRT–PCR, performed with normalization to RPL19 as previously detailed([210]. Sequences of primers used for qRT–PCR analysis are provided in Supplemental Table Primers.

D.1.11 ELISA

Hyaluronic Acid Enzyme-linked immunosorbent assay (ELISA). The ELISA kit was purchased from R and D Systems (DY3614). This kit detects low, medium and high molecular weight hyaluronic acid. Assay plates are coated with primary capture antibody. After sample addition, biotinylated detection antibody is added. Streptavidin-conjugated peroxidase is then added, and quantification of hyaluronic acid is performed using a colorimetric assay. Absolute quantification is calculated by comparison with a standard curve and adjusted for sample dilutions.

D.1.12 Genomic association analyses

We analyzed data from the All of Us Research Program, a diverse cohort including participants of European, African, Asian-American, and other ancestries. Asthma and chronic obstructive pulmonary disease (COPD) were identified from harmonized electronic health records mapped to the OMOP Common Data Model, which standardizes ICD-9/10 and SNOMED codes into unique condition concept IDs. Asthma cases were defined by $\geq$ 1 occurrence of OMOP concept IDs (317009, 40483397, 4308356, 4191479, 4125022, 46270030, 45766728, 256448, 4250128, 46270028, 4123253, 4279553, 42535716, 4309833, or 46273454). COPD cases were defined using concept IDs 25573, 4110056, and 257004. Participants with both conditions were excluded. The final analytic set included 5,843 COPD, 20,454 asthma, and 158,404 controls. Genetic variants with $\geq$ 90% completion and MAF $\geq$ 0.05 were retained; individuals with $>$ 10% missing genotypes were excluded. Logistic regression using SNP dosage tested COPD vs. controls and asthma vs. controls, adjusting for sex, age, pack-years, and 16 principal components. A meta-analysis combining COPD and asthma results was conducted using sample size–weighted p-values to assess shared genetic effects.

D.1.13 Predicting transcriptional networks

To predict the transcription factors associated with a given tRE, we considered the presence of binding motifs, whether the TF was considered active based on global motif response (TFEA-LE), and ChIP-seq data from ENCODE. To predict the target genes of a given tRE, we considered quantitative trait loci (QTL), and nearest gene coordinates. We also recently showed that target genes show nascent transcription correlated with that of their tREs; so we also considered the genes with the highest correlation of nascent transcription across lung cell nascent sequencing samples[79, 210, 223]. Github Repo https://github.com/Dowell-Lab/bidir_gene_pairs/tree/windowed_correlations (branch windowed_correlations, commit 32eea2ddd206c83e35c5876744ba08bc7683b0a8) was used. TPM-normalized counts were used.

D.1.14 Fine mapping and candidate SNP selection

We ranked SNPs for manual annotation based on whether the meta-p-value was significant and if the odds ratio was in the same direction for both asthma and COPD. SNPs were further annotated from both external (accessed via APIs) and internal data in a code pipeline built to be reusable for later analyses. The code collecting all this data from APIs or our own datasets can be found in SNP_Analysis/. The full annotations can be found in Supplemental Table SNPs with descriptions of columns found in Supplemental Table SNP Columns. Importantly, we noticed that the web-server data was sometimes more up-to-date than that provided by the API databases. Therefore, we still encourage users to use the direct web-servers to explore a SNP of interest. For this work, APIs were accessed September 12, 2025.

Bidirectional Type: The bidirectional(s) in which the SNP was found for our dataset were labeled as either intergenic, exonic, intronic, Gene TSS, or LIET Gene TSS Intron/Exon/Intergenic. Gene TSS bidirectionals were first annotated according to the above Nextflow pipeline for counting. This pipeline aims to assign one active TSS bidirectional to each annotated gene isoform by choosing the bidirectional with a μ closed to the annotated TSS. We next assigned the non-TSS bidirectionals from this pipeline to intergenic, introns, exons, or gene TSS regions based on 100bp regions around the μs. Such regions were overlapped with exons, introns, and Gene TSS regions of the hg38.p14 genome using bedtools intersect with options (e.g.

```
bedtools intersect -wo -f .51 -a ${BID} -b ${EXON }
```

) so that the bidirectional is assigned to either introns or exons, but not both. Gene TSS regions are 1kb (500bp up and downstream of the Gene TSS). Sometimes the μ of a bidirectional is intergenic but LIET-based lengths of the bidirectional indicate that it overlaps the TSS region. We considered that these bidirectionals might be unannotated alternative start sites and therefore labeled them as LIET Gene TSS bidirectionals with extra labels of Intron/Exon/Intergenic to indicate which other region the bidirectional falls into. The code and more detailed results from this analysis can be found in Preprocessing/LIET/Get_LIET_results_bed_5.15.25.ipynb.

Disease Associations: SNPs associated with diseases in external analyses were noted based on GWASCatalog (Cerezo et al., 2025), ClinVar [127], and OpenTargets (API v25.4.4) [22]. Full annotations included any variants with predicted linkage disequilibrium above 0.8 for OpenTargets and if the posterior probability was above 0.05 and absolute beta value above 0.1. The 44 SNPs considered externally associated with a phenotype were all the lead SNPs. When available, we saved the trait, beta, posterior probability (of the SNP), types of association, predicted targets of SNP, scores for the predicted target linkage, locus size (number of SNPs possibly contributing to the Beta value), and whether the lead SNP for the association is the same as the SNP of interest.

Molecular Effects: We first calculated the distance from the SNP to the μ of overlapping bidirectionals since the μ is predicted to estimate the transcription initiation site. SNPs overlapping TF Binding sites based on ENCODE were annotated based on the OpenTargets (API v25.4.4) [22, 164], with the transcription factor, and cell types and tissues in which the binding site saved. OpenTargets also overlaps variants with features annotated by ENCODE's chromHMM (e.g. Enhancer, Genic-Enhancer, TssAFlnk, etc.)[164]. We saved the regions the SNPs overlapped with, and the celltypes and tissues where the overlapping ChromHMM is found based on the OpenTargets API. Molecular effects (e.g. intron_variant, missense_variant) were saved along with the variant, predicted effect, method for prediction (e.g. Ensembl VEP), amino acid changes or distance to a footprint (e.g. TF motif instance), and predicted targets/effect scores if available again from OpenTargets API. Full annotations included any variants with predicted linkage disequilibrium above 0.8 for molecular effects and chromHMM predictions.

Target Genes: First, genes whose 5-prime and 3-prime ends were the most proximal were recorded. Second, genes with the highest correlation of transcription with the enhancer(s) were recorded based on all non-cancerous lung cells, or just Beas-2B cells or smAEC cells. [GIVE THE Ns for each]. The correlations of the most proximal genes were also recorded. Finally, we recorded quantitative trait loci and 3D Chromatin-based linkages (e.g. within loop, anchor-to-anchor) to genes based on the Open Targets[22] and 3DSNPv2 APIs (Quan et al., 2022). For Open Target-based findings, full annotations included any variants with predicted linkage disequilibrium above

0.8 for these, with cases of the variant of interest being the lead SNP being noted. The tissues for the interactions, and if available, betas, posterior probabilities (of SNPs), and p-values were recorded. The transcriptional changes in particulate matter responses for all potential target genes were recorded.

D.2 Supplemental Figures

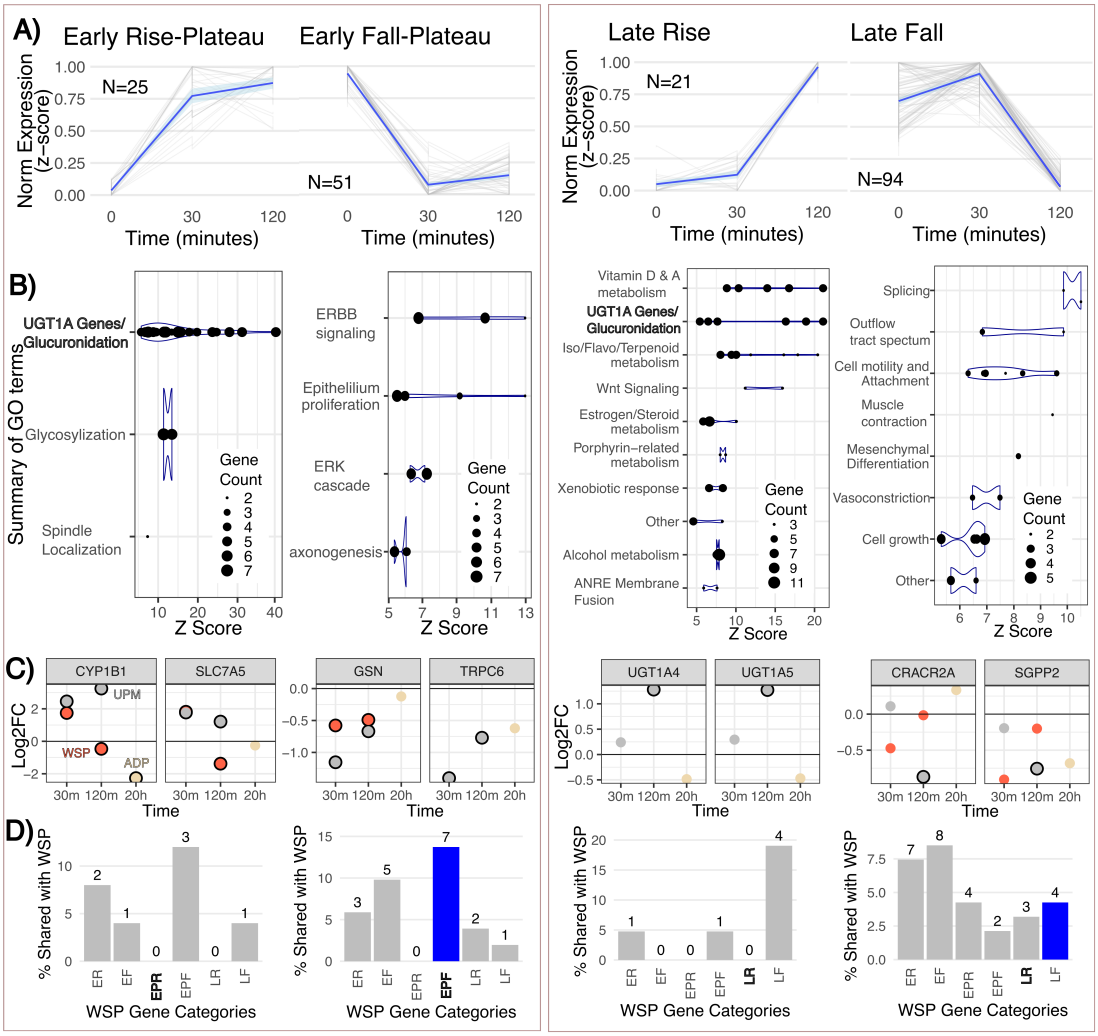

Figure D.1: **Plateau and Late Response Genes A.** Line plots of the mean (blue, 95% confidence interval shaded) and individual (grey) normalized transcriptional changes of genes statistically significant (adjusted-p-value $< 1x10^{-10}$) matching different timeline categories (details in Methods). Norm Expression (z-score) refers to normalized counts min-max scaled. 0 minutes refers to Vehicle response. **B.** Summary of Gene Ontology (GO) terms of the matching genes (exact terms and grouping found in Supplemental Table GO) where each term is a GO term called significant with the size of the dot corresponding to the number of significant genes matching the GO term. **C.** Log2 fold change of normalized counts of the genes with the greatest statistical significance in change for UPM among three different perturbations (UPM, wood smoke particles (WSP), or Afghan dust particles (ADP)) compared to vehicle at their available timepoints. Cases where the change had an adjusted p-value $< 1x10^{-6}$ were outlined in black. *UGT1A4/5* were not significantly transcribed in WSP conditions. **D.** Percentage of significant UPM genes that were found across the timeline categories of WSP responsive genes (adjusted-p-value $< 1x10^{-10}$). ER=Early Rise, EF=Early Fall, EPR=Early Rise-Plateau, EPF=Early Fall-Plateau, LR=Late Rise, LF=Late Fall. The bar corresponding to the same time point as UPM for WSP is highlighted in blue. Numbers above bars correspond to the number of UPM genes shared. Number of WSP genes in each category are as follows: ER=416, EF=1527, EPR=164, EPF=1148, LR=879, LF=549.

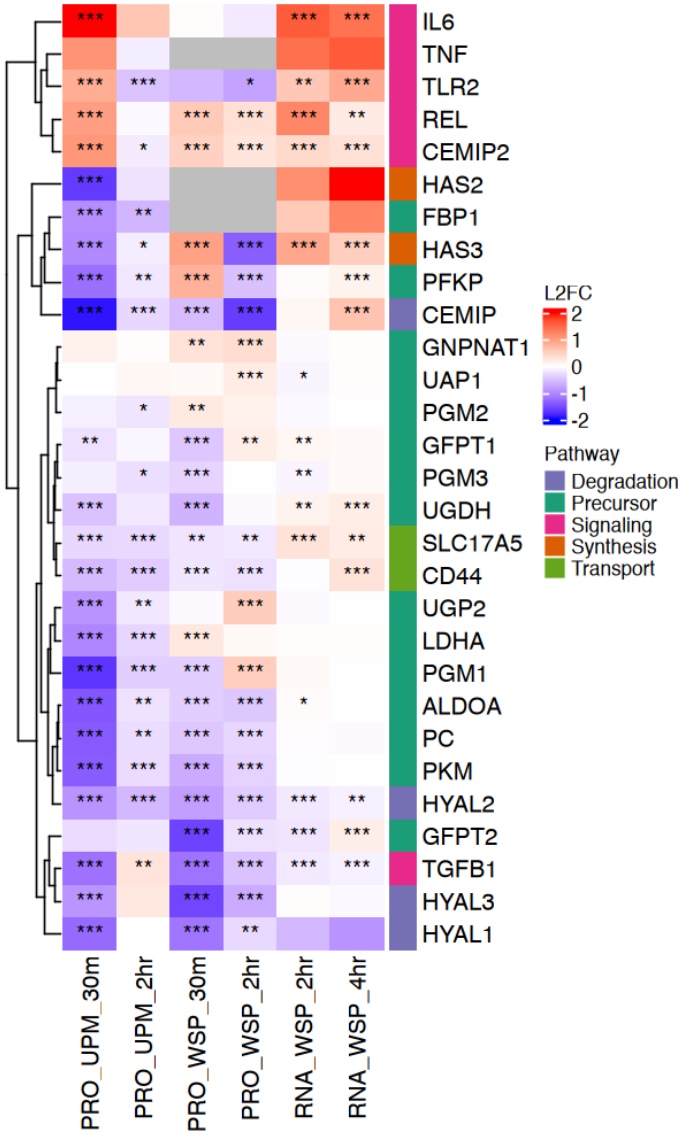

Figure D.2: **Hyaluronan-related genes show similar regulation across WSP and UPM and steady-state** Log2 fold change (Log2FC) compared to vehicle of smAEC cells perturbed with UPM (urban particulate matter) or Beas-2B cells perturbed with WSP (wood smoke particles) of genes involved in hyarluronan-related processes (Pathway). Adjusted p-values from DESeq2 are shown as * < 0.05, ** < 0.01, *** < 0.001. PRO refers to PRO-seq, RNA refers to RNA-seq. Grey means the gene was removed from analysis by DESeq2 due to low-expression outliers.

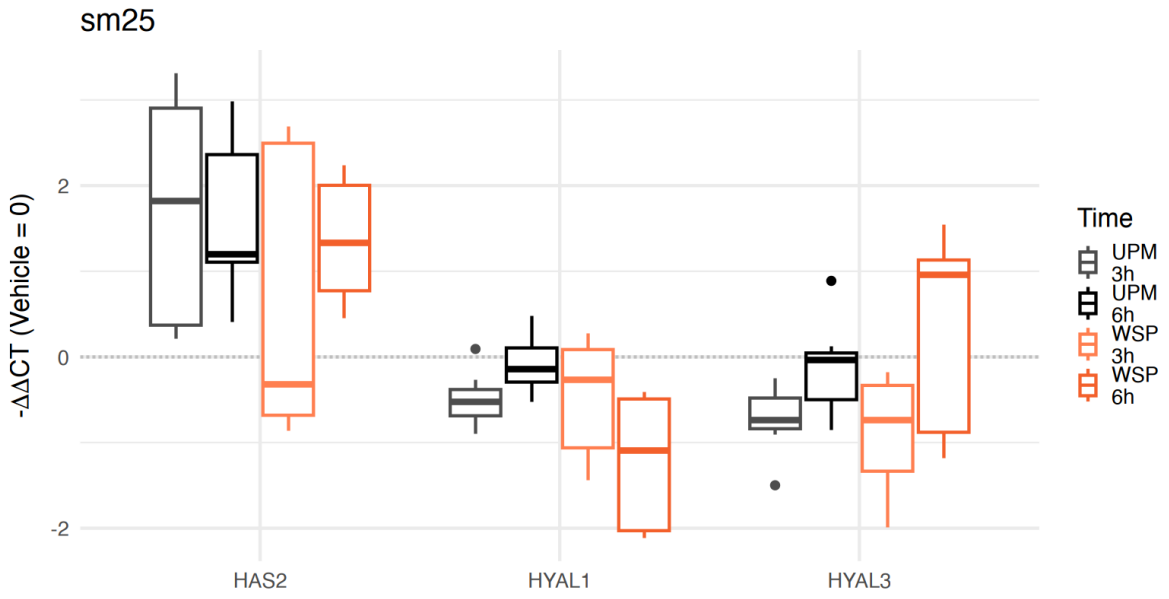

Figure D.3: **HYA synthesis genes generally go up upon pollutants while HYA degradation genes go down** qRT-PCR results for smAECs treated with UPM and WSP, with results from RNA extracted at 3h or 6h time-points after perturbation compared to Vehicle. There are 8 replicates per condition/time-point, split between two different experiments/days

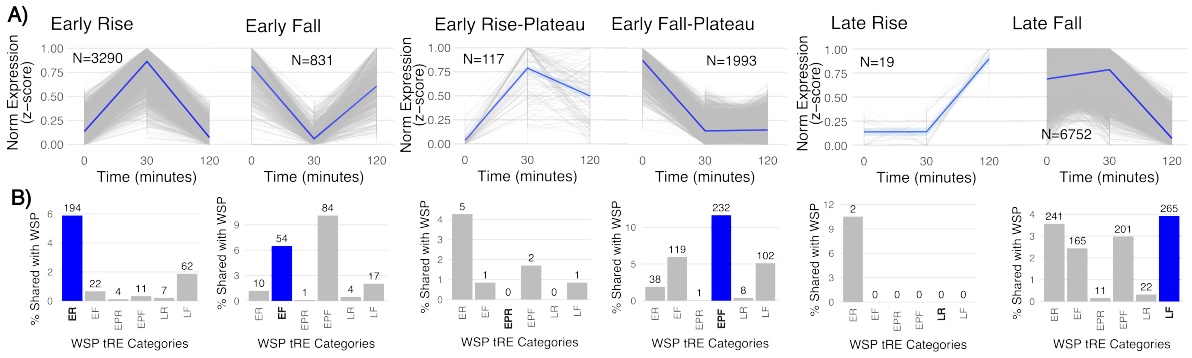

Figure D.4: **tREs show a massive early response comparable to genes and a unique large fall in expression at later points A.** Line plots of the mean (blue, with light blue 95% confidence interval) and individual (grey) normalized transcriptional changes of tREs statistically significant (adjusted-p-value < 0.001) matching different timeline categories (details in Methods). Norm Expression (z-score) refers to normalized counts min-max scaled. 0 minutes refers to Vehicle response. **B.** Percentage of significant UPM tREs that were found across the timeline categories of WSP responsive tREs (adjusted-p-value < 0.001). ER=Early Rise, EF=Early Fall, EPR=Early Rise-Plateau, EPF=Early Fall-Plateau, LR=Late Rise, LF=Late Fall. The bar corresponding to the same time point as UPM for WSP is highlighted in blue. Numbers above bars correspond to the number of UPM tREs shared.

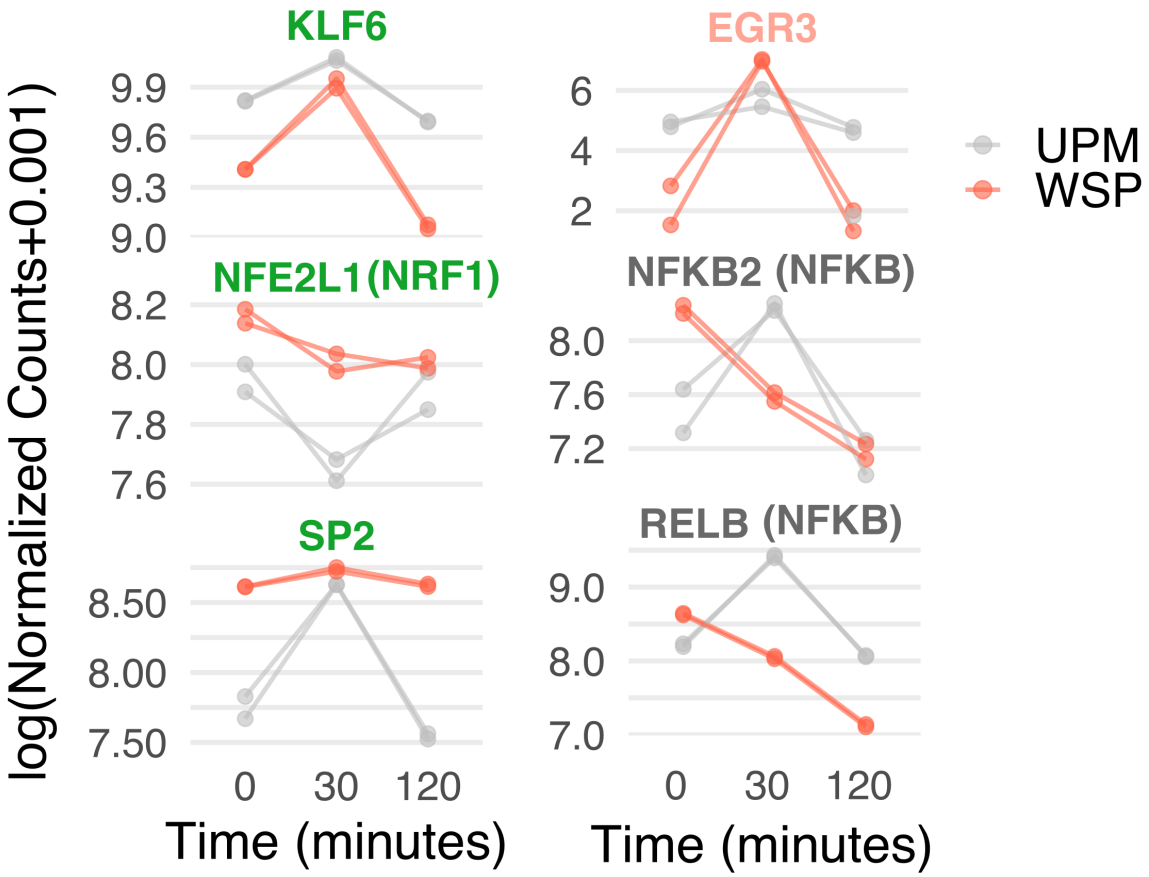

Figure D.5: **Transcription factor transcriptional changes follow changes captured by motif analysis with TFEA** KLF6, NRF1, and SP2 all have motifs enriched in tREs with decreased transcription levels at the 30-minute mark according to TFEA in both WSP and UPM (green) and also show significant change in transcription for both perturbations at 30 minutes. KLF6 and SP2 have been observed both as activators and repressors. EGR3 only had significant negative enrichment score in WSP and shows significant upregulation in WSP. Genes encoding for the NFKB TFs show upregulation at 30 minutes only in UPM and have positive motif enrichment in UPM.

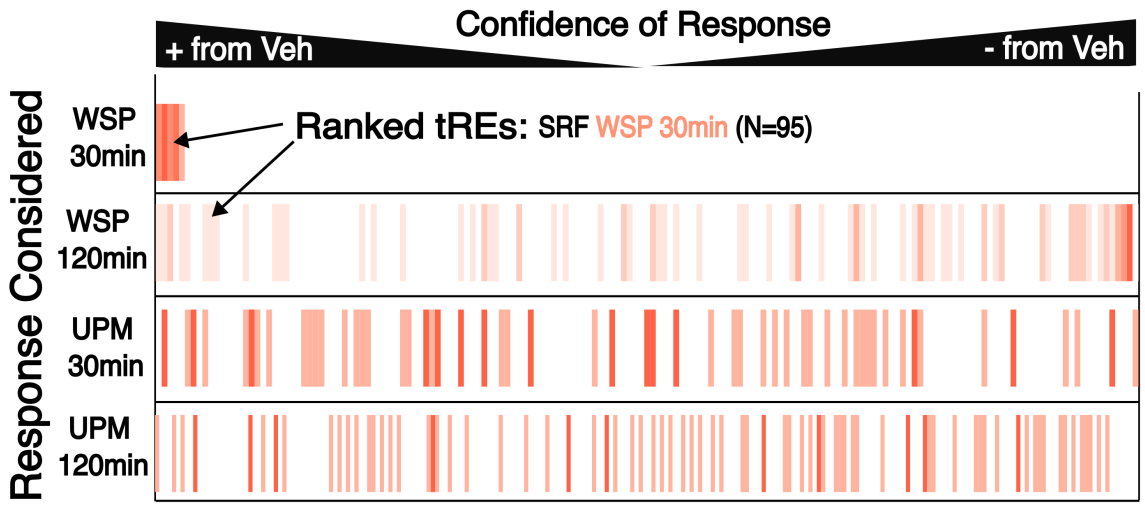

Figure D.6: **tREs attributed to SRF response in WSP response show no pattern of response in UPM** tREs called responding via SRF to WSP at 30 minutes are colored, and the x axes represent where all tREs rank in confidence of up(+) or down(-) regulation compared to Vehicle across 4 different responses: WSP 30min, WSP 120min, UPM 30min, UPM 120min. Left (or Right) most x-coordinates are tREs with the highest positive (or negative) log2fold change and lowest p-values for the listed response.

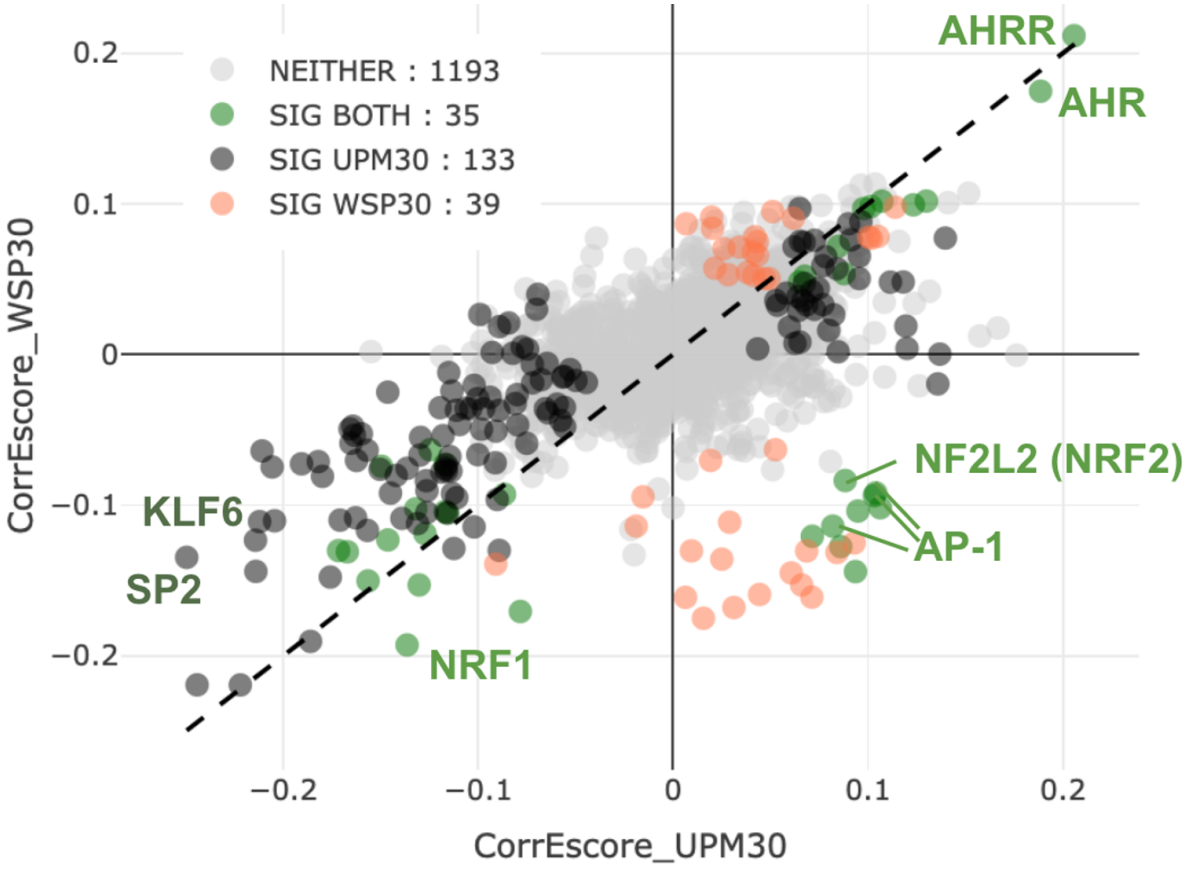

Figure D.7: **Mostly strong agreement of ATAC-seq TF responses between UPM and WSP** Scatterplot of TFEA GC-corrected enrichment scores for WSP and UPM 30 minutes compared to vehicle for ATAC-seq in nasal airway epithelial cells. Transcription factors are highlighted as being a call in neither UPM/WSP, only one or the other, or both (where significance requires adjusted p-value < 0.01 and Fraction above Background < 0.46 – see Methods). Both KLF6 and SP2 have fraction above backgrounds of 0.51 in WSP, indicating weaker but still significant enrichment.

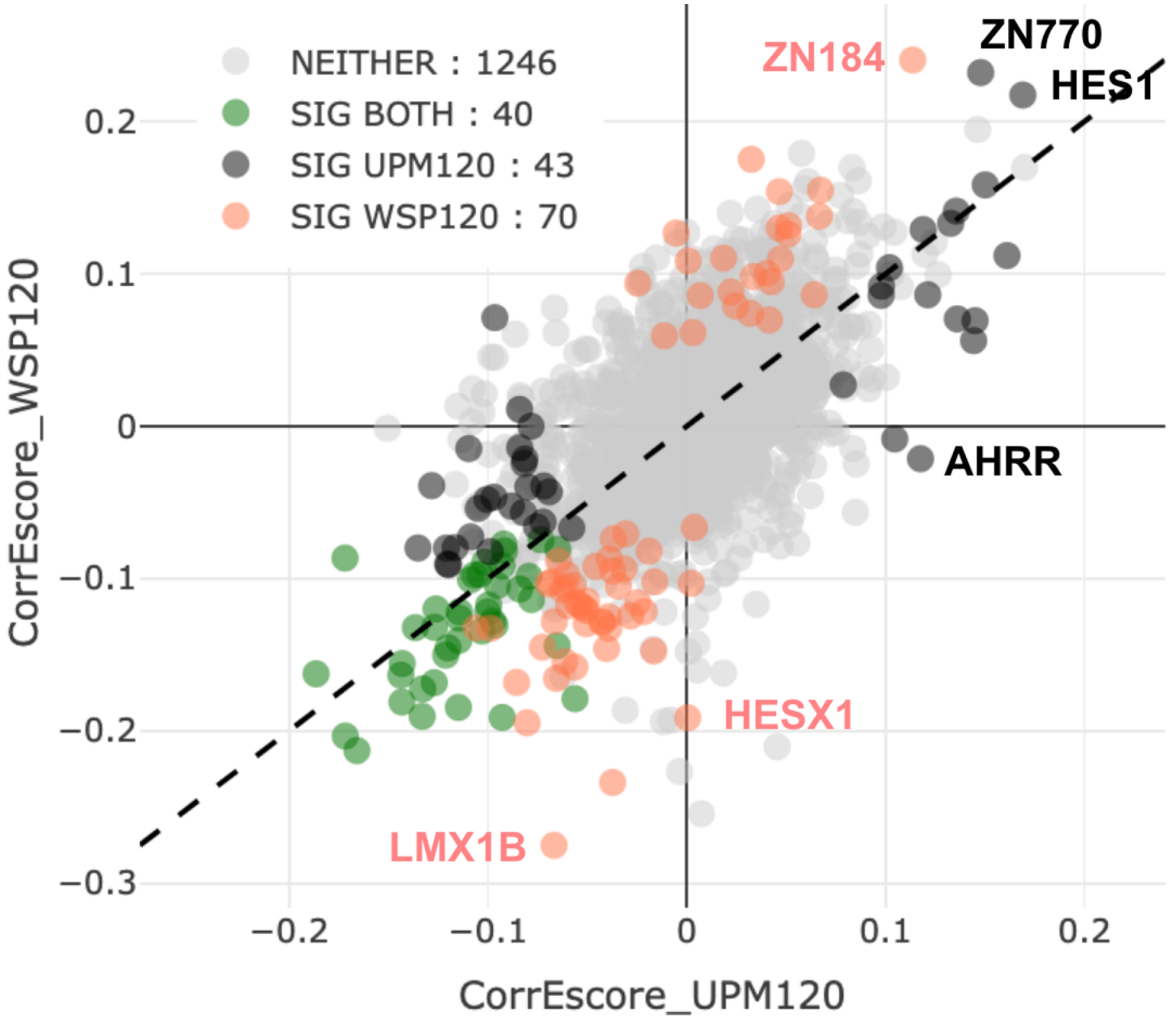

Figure D.8: **Mostly strong agreement of 120min TF responses between UPM and WSP** Scatterplot of TFEA GC-corrected enrichment scores for WSP and UPM 120 minutes compared to vehicle. Transcription factors are highlighted as being a call in neither UPM/WSP, only one or the other, or both (where significance requires adjusted p-value < 0.01 and Fraction above Background < 0.46 – see Methods).

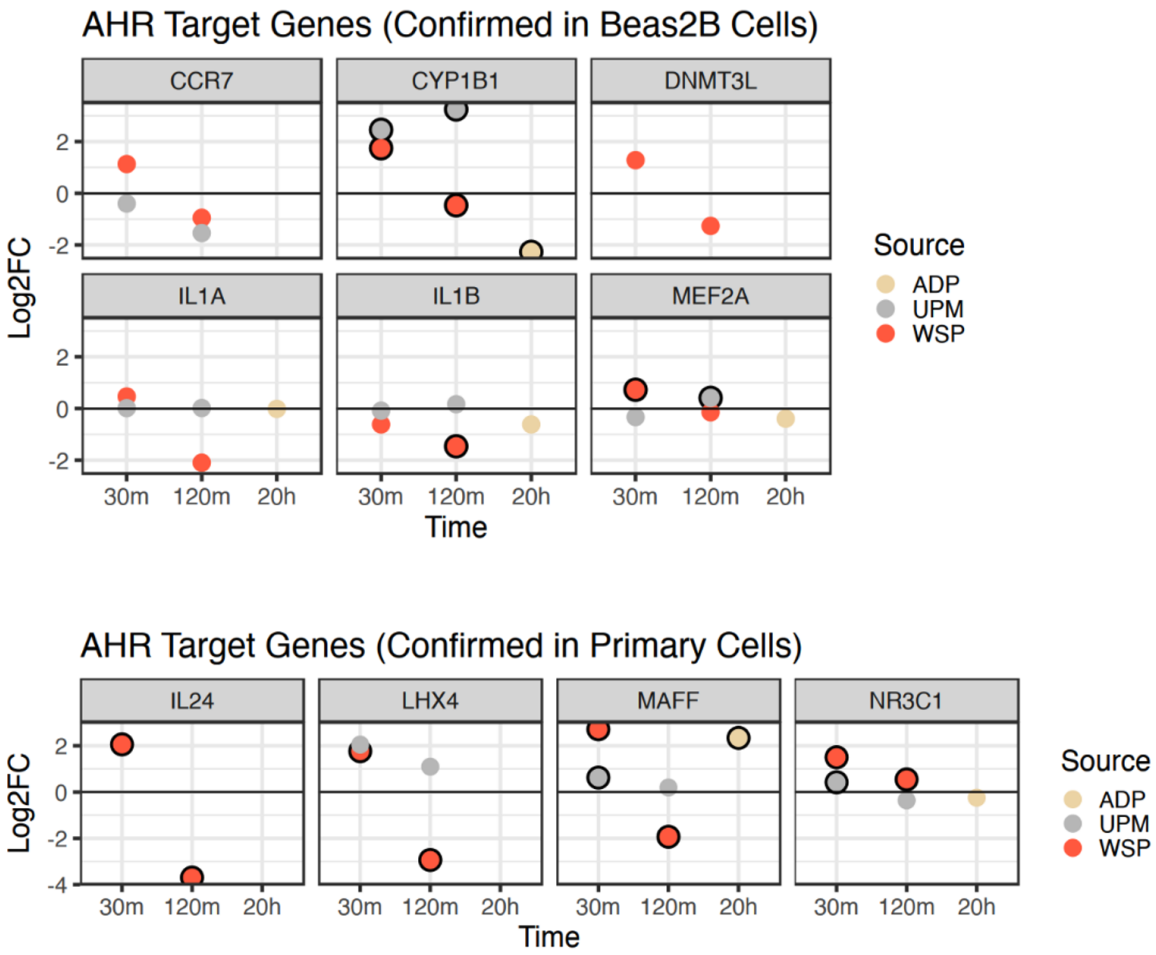

Figure D.9: **Previously confirmed AhR target genes show mostly similar responses at 30 minutes for UPM and WSP** Log2 fold change (Log2FC) compared to vehicle of cells perturbed with ADP (afghan dust particles), UPM (urban particulate matter), or WSP (wood smoke particles) of genes that were confirmed as AhR target genes based on ChIP-seq and sRNA knockdown of AhR in either Beas2B cells or small airway epithelial cells in Gupta et al. 2021. Dots with black outline indicate the gene has an adjusted p-value for fold change $< 1x10^{-10}$ for WSP and UPM and 0.01 for ADP.

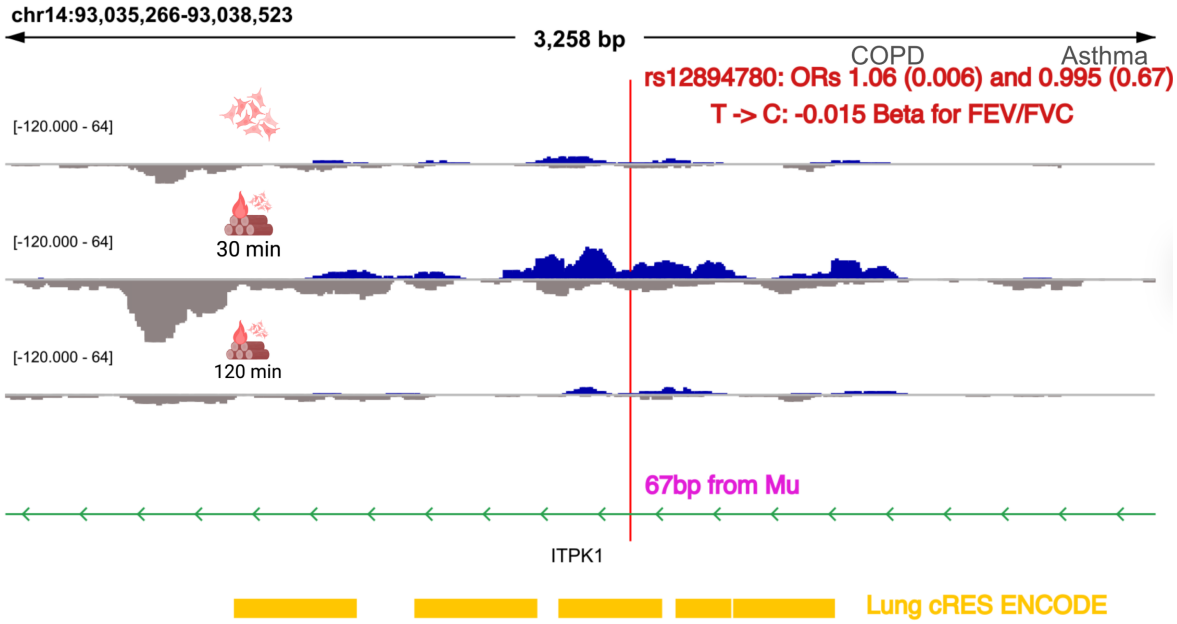

Figure D.10: **Replicated SNP rs1284780 (association with COPD in our and other studies) is an intronic tRE SNP within ITPK1** The SNP is 67bp away from the μ (midpoint) of a tRE showing robust early response to WSP. The odd ratios for COPD (left) and asthma (right) are listed with the association p-values.

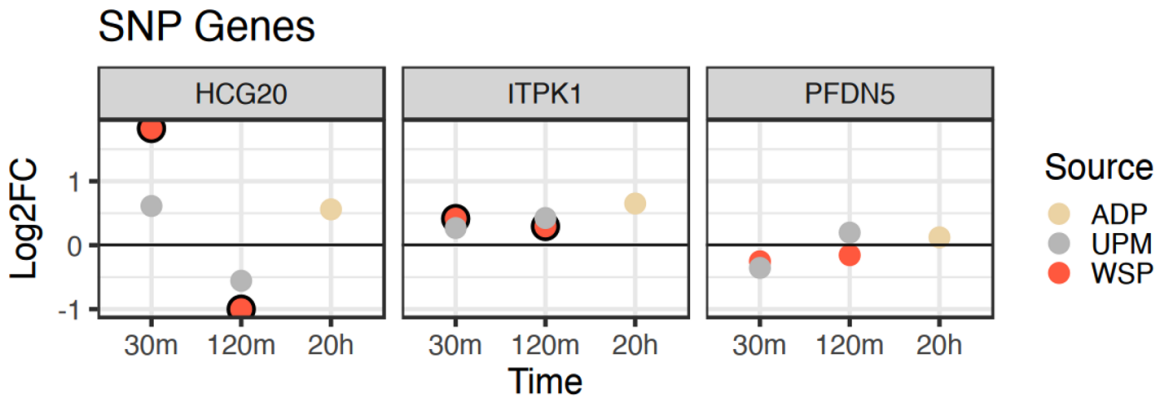

Figure D.11: **Three genes in which significantly associated SNPs are found tend to have changing transcriptional patterns shared across particulate perturbations.** DESeq-calculated log2 fold change (Log2FC) of genes within samples perturbed with one of three particulate matters (ADP=afghan dust particles, UPM=urban particulate matter, WSP=wood smoke particles) at three time points (30 or 120 minutes, 20 hours) compared to vehicle sample. Black outline surrounds genes that have an adjusted p-value lower than 0.01. PFDN5 has adjusted p-values all above 0.1. ITPK has adjusted p-values below 0.001 for UPM.

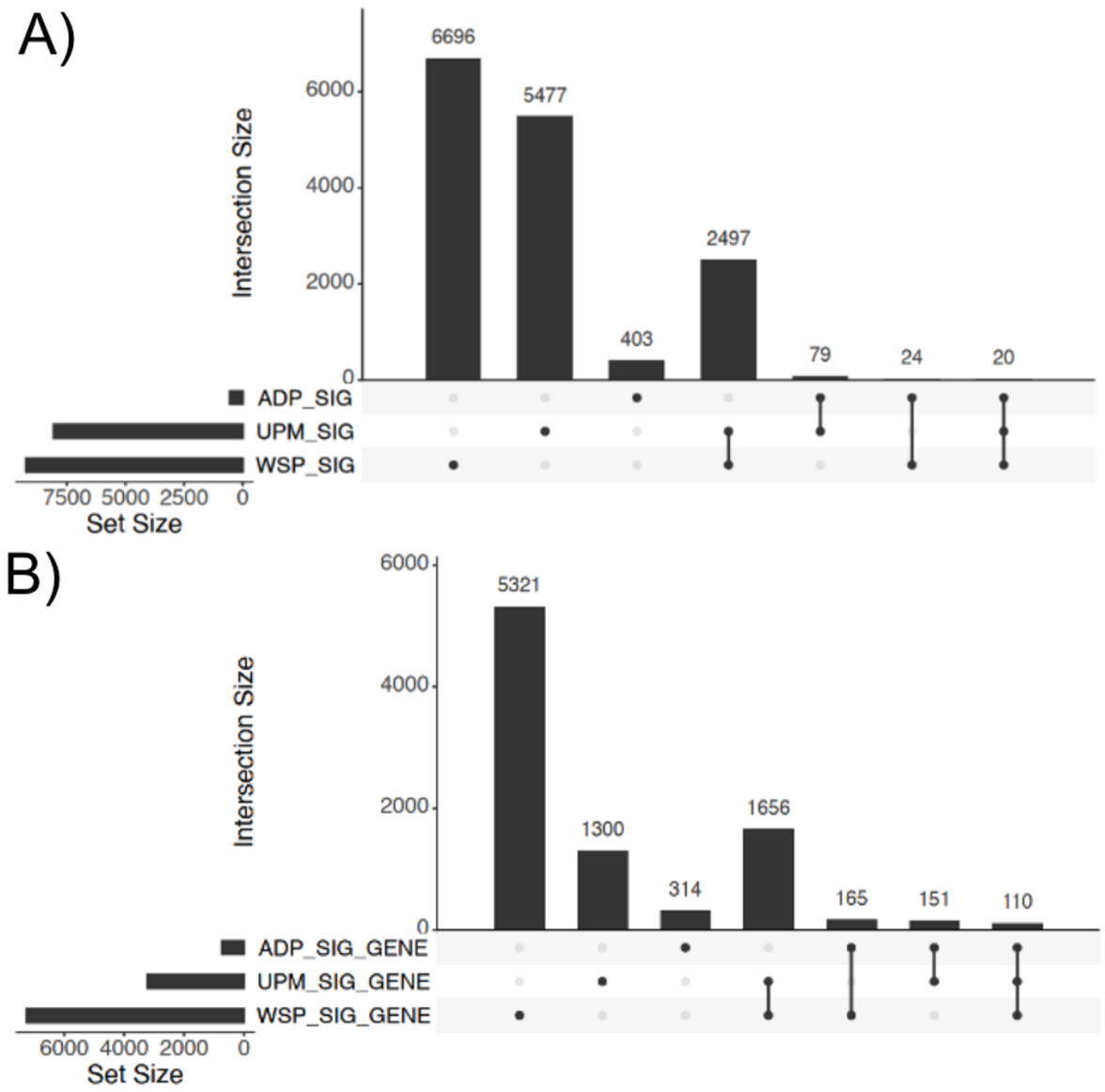

Figure D.12: **Most tREs and genes responding according to strict significance cutoffs are unique to perturbations** UpSet plots of A) tREs and B) genes called significantly responding at any time point for each perturbation compared to vehicle: ADP (Afghan dust particles 20h), UPM (urban particulate matter 30min or 120min), WSP (wood smoke particles 30min or 120min). Adjusted p-value cutoffs of 0.001 and 0.01 were used for tREs for UPM (N=8073) and WSP (N=9237), and ADP (N=526), respectively. Adjusted p-value cutoffs of $1x10^{-10}$ and 0.01 were used for genes for UPM and WSP, and ADP, respectively. Details are found in Methods.

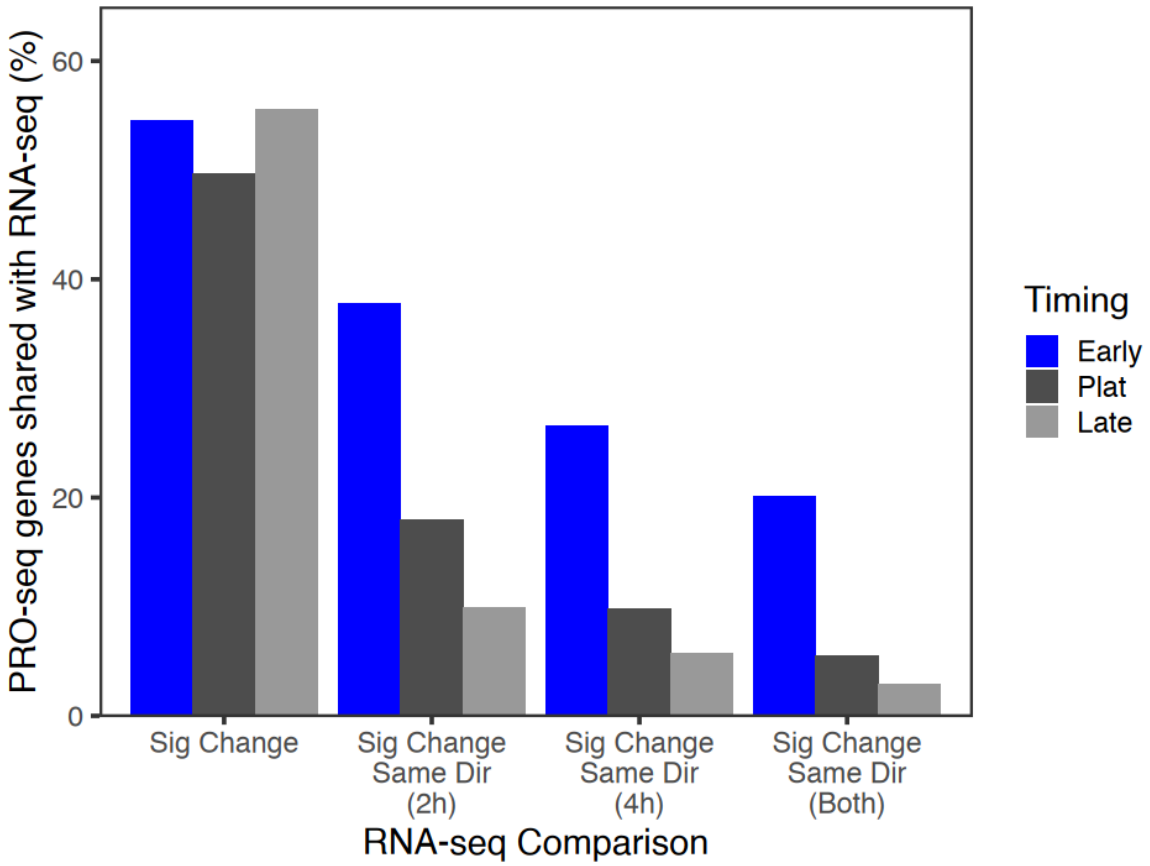

Figure D.13: **Harmonization between RNA-seq and PRO-seq responsive genes is not biased against early response genes** Percentage of genes responding in the three time categories in PRO-seq (early fall/rise (Early), early plateau fall/rise (Plat), late fall/rise (Late) that have a significant change in RNA-seq (Sig Change) in the same direction at the 2h, 4h, or both timepoints. Beas-2B cells perturbed with WSP were used for both.

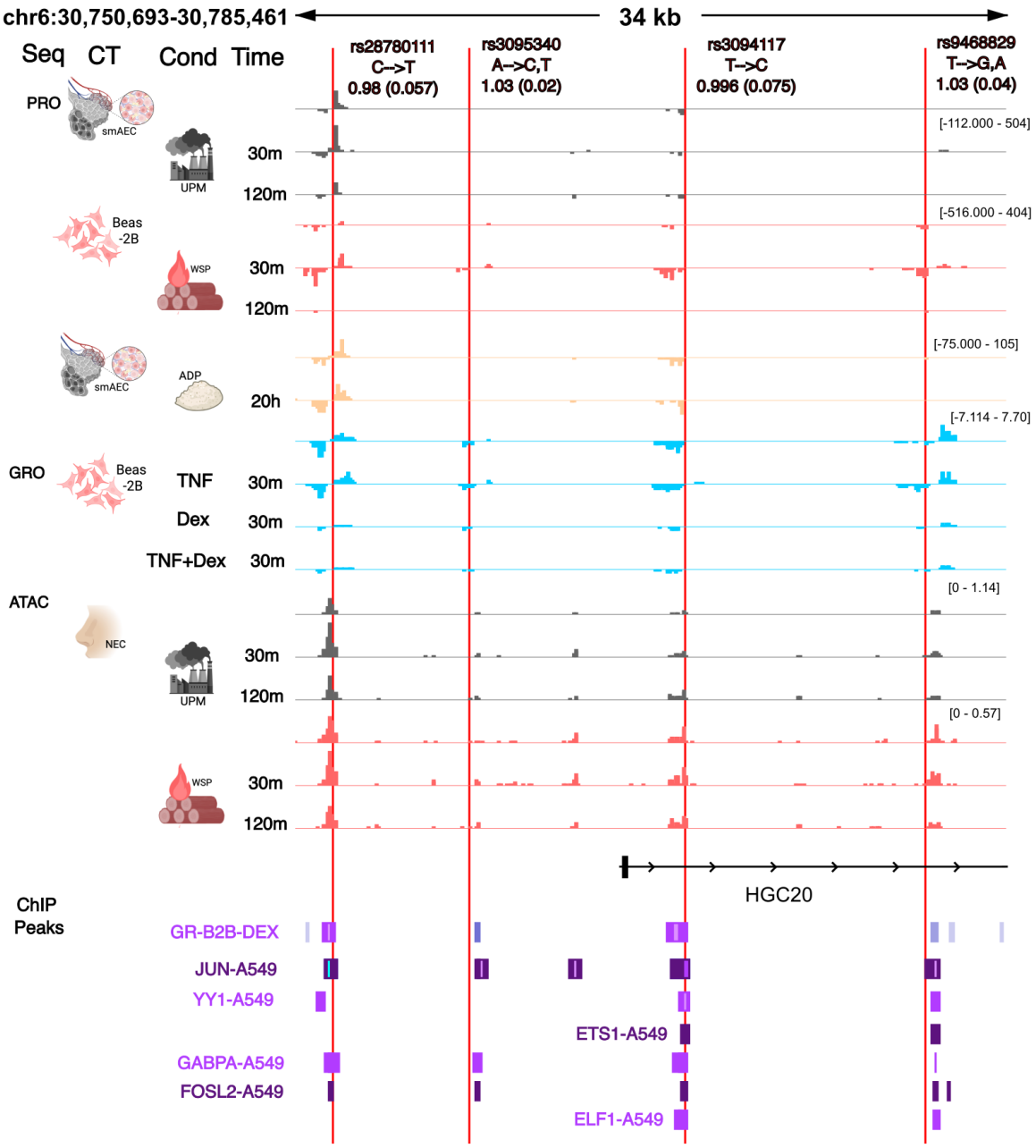

Figure D.14: **Previous SNP correlated with asthma is within group of tREs and other SNPs similarly associated**IGV tracks of PRO/GRO-seq or ATAC-seq (Seq column) of lung cells (small airway epithelial cells (smAEC), Beas-2B cells, nasal epithelial cells (NEC)) (CT column) perturbed with urban particulate matter (UPM), wood smoke particles (WSP), afghan dust particles (ADP), TNF, and/or dexamethasone (Dex) (Cond column) at varying time points (Time column). The numerical read distributions are not directly comparable outside of experiments due to different normalization approaches (scale noted on the right top of each experiment). ChIP peaks from A549 (ENCODE) of different transcription factors or from Beas2B cells perturbed with Dex (GR-B2B-DEX) are highlighted with varying colors for easier interpretation. SNPs are labeled by their rsid values and odds ratios for asthma (p-values) in our study.

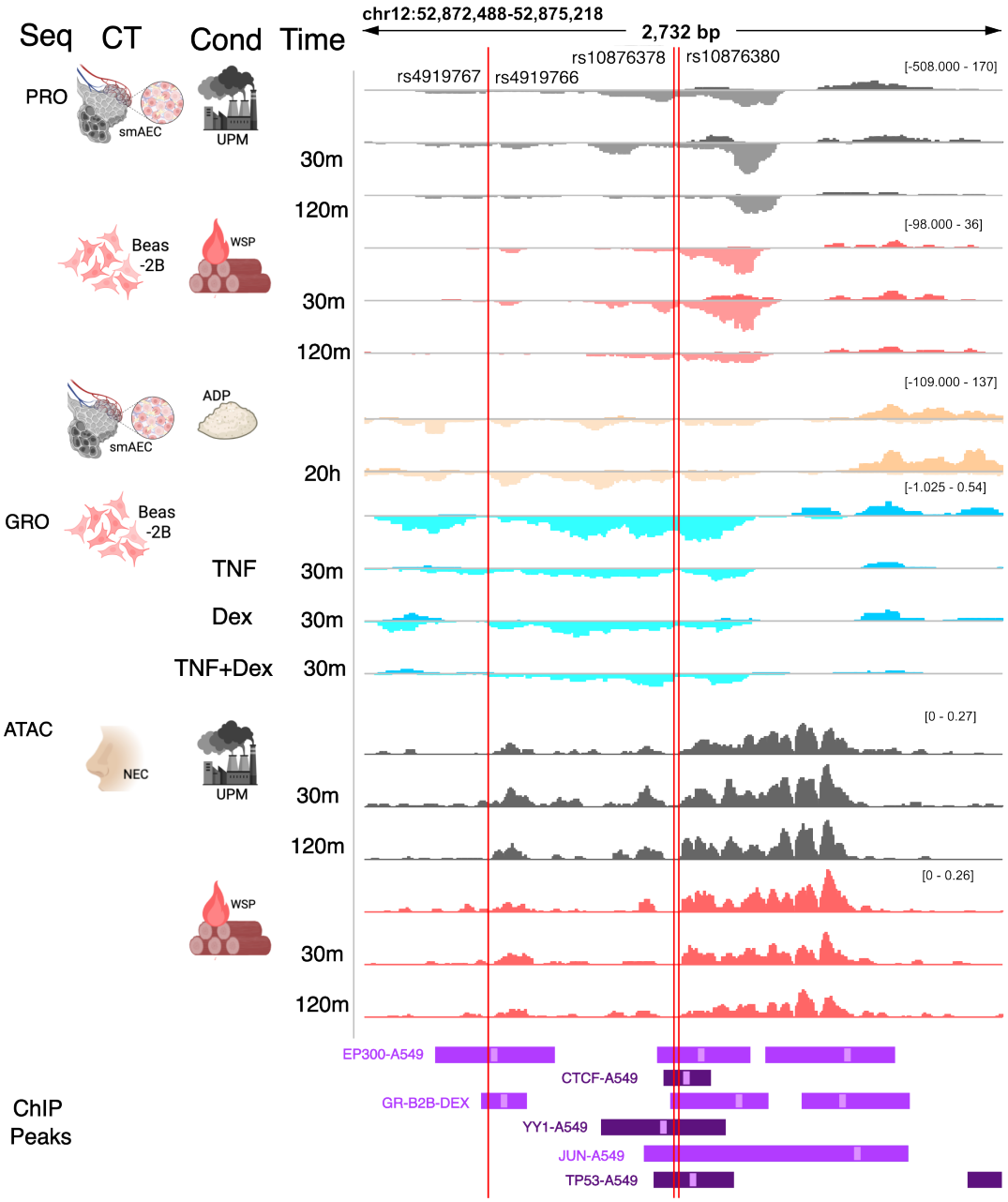

Figure D.15: **Chromosome 12 SNPs fall within enhancer group with consistent signal across conditions and omics data** IGV tracks of PRO/GRO-seq or ATAC-seq (Seq column) of lung cells (small airway epithelial cells (smAEC), Beas-2B cells, nasal epithelial cells (NEC)) (CT column) perturbed with urban particulate matter (UPM), wood smoke particles (WSP), afghan dust particles (ADP), TNF, and/or dexamethasone (Dex) (Cond column) at varying time points (Time column). The numerical read distributions are not directly comparable outside of experiments due to different normalization approaches (scale noted on the right top of each experiment). ChIP peaks from A549 (ENCODE) of different transcription factors or from Beas2B cells perturbed with Dex (GR-B2B-DEX). SNPs are labeled by their rsids.

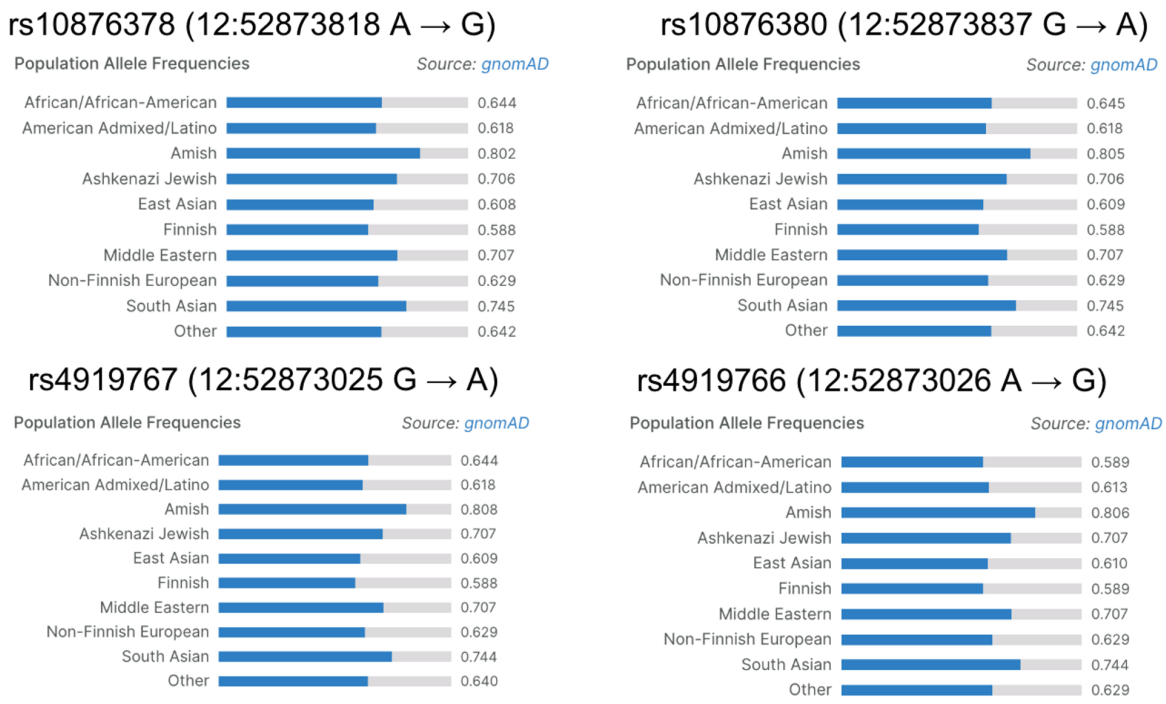

Figure D.16: **Four of the top SNPs indicating lower likelihood of COPD and asthma are more common across all ancestry populations and at similar levels.** Population allele frequencies from Open Targets webserver (https://platform.opentargets.org/) of each of the four SNPs falling within chromosome 12 and within the top 10 SNPs associated with both COPD and asthma.

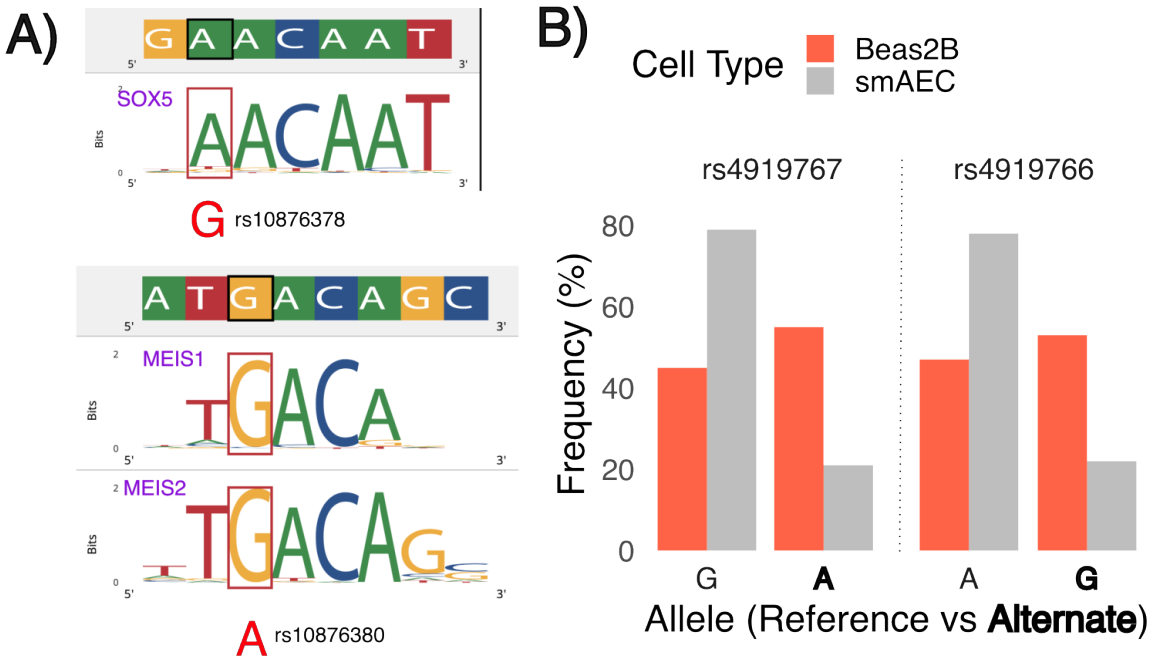

Figure D.17: **Four SNPs all disrupt canonical binding sites of transcription factors A.** SNP rs10876378 (chr12:52873818) indicates a change from reference A to G, which might disrupt the motif instance of SOX5. Finally, SNP rs10876380 might disrupt the motif instances of either MEIS1 or MEIS2. All images were collected before annotation from opentargets.platform.com. **B.** Frequency of alleles for each SNP in PRO-seq reads for Beas2B cells and smAECs.

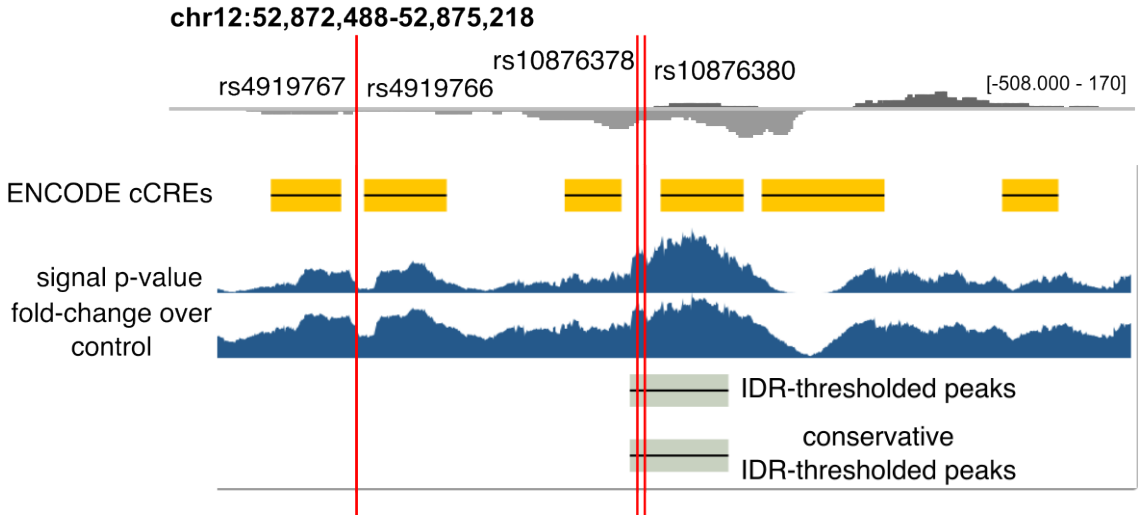

Figure D.18: **Small PRDM1 ChIP-seq peaks near PRDM1-motif SNPs.** Chr12 region SNPs along with ENCODE cCREs and ENCODE-produced data from PRDM1 ChIP-seq in A549 cells (ENCODE Experiment ENCSR977FEF) in their provided Genome browser, looking at rep1,2. Irreproducibility Discovery rate (IDR) ENCODE Genome browser rep1,2. Conservative IDR-thresholded peaks indicate the peak was found across a pair of true replicates (not just pseudoreplicates).