Developing a cell-free RNA liquid biopsy

A proposal submitted in pursuit of a PhD awarded by UC Santa Cruz and supervised by Assistant Professor Daniel H. Kim.

Roman Elliott Reggiardo
0000-0001-6309-249X · rreggiar
Department of Bioengineering, University of California, Santa Cruz
· Funded by prev. NIH T32 HG008345; NIDDK/KUH F99 DK131504; UCSC CITL

Abstract

Of the 200,000+ Americans who will be diagnosed with lung or bronchial cancers in 2021, ≥156,000+ (78%) will already be at stages of disease that limit survival probability to between 6% and 33%. Current lung cancer screening technology is covered by Medicare and has been demonstrated to reduce mortality by 20%, in addition to greatly reducing the costs for chemotherapy and radiation treatment. Unfortunately, low utilization of screening via low- dose computed tomography (LDCT) prevents the widespread reduction of lung cancer mortality. This is due in part to fears over the misdiagnosis of lung nodules and the exposure to radiation needed for LDCT – in other words: the potential for harm by the screening without clear benefit. To ultimately achieve more accurate diagnoses earlier in disease progression, I aim to develop and preliminarily validate a non-invasive, plasma biopsy of lung adenocarcinoma (LUAD) tumor-derived RNA.

Recent advances in next-generation sequencing (NGS) have enabled the creation of multiple clinically-actionable 'liquid biopsies' that utilize biofluid, typically blood plasma or serum, as a medium to detect circulating tumor-derived DNA (ctDNA) biomarkers released from apoptotic (dying) tumor cells. In contrast to LDCT, liquid biopsies pose no risk to the patient population and provide molecular detail of disease status. **I propose to explore the highly dynamic transcriptomic signal secreted from all cells, alive or dying, in healthy or diseased tissues, by developing novel analytical approaches to studying RNA available in human blood plasma.**

Specific Aims

DNA-based liquid biopsies under-perform in earlier stages of disease, rely on indirect measures of RNA transcription, and detect signal from dying cells. To overcome these limitations and advance the field of non-invasive diagnostics I propose to develop the foundation for an **RNA liquid biopsy**. With limited exceptions, existing extracellular RNA (exRNA) approaches have sequenced only short and/or protein-coding RNAs and have produced few clinical diagnostic candidates. The Kim lab has integrated isolation of extracellular vesicles (EVs) from biofluids to enable robust isolation of long, un-degraded polyA+ RNA molecules. I am devoloping a novel analysis pipeline that enables detection of an abundant class of RNA, Transposable Elements (TEs), that we have found to be robust disease-specific exRNA molecules enriched in EVs. These innovative methods, along with my multidisciplinary training in RNA biology, bioinformatics, and diagnostic modeling, have strongly positioned this thesis proposal to begin to characterize disease-relevant exRNA and ultimately develop a prototype RNA liquid biopsy via three specific aims:

Aim 1: Identify mutant KRAS-induced transcriptional changes in early tumorigenesis and after clinical inhibition
Progress: completed, Output: 1st author manuscript accepted & co-1st in-review, published book chapter.
We find oncogenic KRAS signaling upregulates noncoding transcripts throughout the genome, many of which arise from transposable elements (TEs). These TE RNAs exhibit differential expression, preferentially in extracellular vesicles, and are regulated by KRAB zinc-finger (KZNF) genes epigenetically silenced in mutant KRAS cells and *in vivo* lung adenocarcinomas. Moreover, mutant KRAS induces the epigenetic activation of Interferon Stimulated genes (ISGs), suggesting a link between ZNF and TE dysregulation and intrinsic immune signaling. Finally, we demonstrate the potential for monitoring KRAS-induced transcriptional dysregulation in EVs isolated from cells treated with small molecule KRAS inhibitor AMG 510. Our results reveal the broad scope of intracellular and extracellular RNAs regulated by oncogenic KRAS signaling in early and treatment stages of tumorigenic transformation.

Aim 2: Develop TE- and Intron-aware RNA-seq analysis to comprehensively assess plasma exRNA expression
Progress: near completion, Output: 1st author manuscript submitted, collaboration in progress.
Analysis of TE expression in response to oncogenic KRAS signaling led to discovering enrichment, and robust differential expression, of these RNAs within extracellular vesicles. Along with analytical advances in the liquid biopsy field, this inspired the development of a multi-pronged transcriptional quantification approach using *Salmon*. I demonstrate the utility of this approach for detecting *in vivo* exRNA signals derived from pancreatic cancer and COVID-19 patients, respectively.

Aim 3: Validate a Transposable Element-aware RNA liquid biopsy in LUAD cohort
Progress: preliminary results, Output: 1st author manuscript in progress, another anticipated.
This aim will employ two complementary approaches to assess the value of TE-aware RNA analysis for LUAD diagnosis: 1) TE-aware recompute of relevant TCGA and GTEx tissues and 2) TE-aware analysis of a new, 120-member (50% LUAD) cohort patient blood plasma sequenced according to our established approach and orthogonally diagnosed with PET/CT imaging. RNA sequencing data from each cohort will be used to train parsimonius and interpretable machine learning classifiers to detect patients with LUAD based on exRNA expression alone. This aim is supported by F99-DK131504 (RER) and CDMRP-LC190293 (DKIM).

Significance

Critical gap for cancer diagnostic tools: Despite decreasing cancer death rates overall, over 156,000 Americans will be at high risk of dying from Lung cancers because they are being diagnosed primarily after stage 1. The expected near-term increase in late-stage diagnoses due to the COVID-19 pandemic serves only to emphasize the pressing need for accurate, accessible, and non-invasive diagnostic tools [1]. Furthermore, because of the recent advent of cancer therapies for even the most recalcitrant of oncogenes, such as KRAS, there is clear synergistic potential for genomic assays that can detect cancer with specificity for therapeutically targetable oncogenes [2,3,4,5]. This thesis investigates the potential for RNA liquid biopsy to address the unmet need for accurate, accessible, and safe non-invasive Lung Adenocarcinoma (LUAD) diagnosis.

The most mature liquid biopsy candidates, including one product available through Clinical Laboratory Improvement Amendments (CLIA) waiver, share common features: 1) they are based on isolation and sequencing of cell-free DNA (cfDNA), 2) they utilize detection of epigenetic marks, and 3) they are dependent on the Illumina sequencing-by-synthesis platform [6,7,8,9,10]. While these platforms have demonstrated success, there remain limitations to DNA-based technologies that curb potential performance in earlier stages of disease, where sensitivity is near 25% for Lung cancer [10]. A sizable fraction of cfDNA release requires apoptosis/necrosis of tumor cells that limits overall availability of tumor-associated cfDNA outside of those events and, due to the limited copy number of genomic DNA, these tests rely on robust expansion of the tumor cell population [11,12].

Our liquid biopsy directly addresses these shortcomings by utilizing abundant, dynamic, and physiologically relevant RNA molecules encapsulated in extracellular vesicles (EVs) secreted into the bloodstream: exRNA. Furthermore, current standard of care (SoC) for LUAD detection, low dose computed tomography (LDCT), has proven to reduce associated mortality but remains under-utilized and moderately dangerous in high-risk populations [13,14,15,16]. We anticipate this platform will address two critical drawbacks of current SoC for LUAD early detection: 1) a simple blood draw poses virtually no risk to the patient population, unlike radiation exposure from low-dose CT, and 2) the expected utilization of a blood-draw assay is much higher than approaches with significant radiation exposure risk and complex infrastructure [17].

Rigor of Prior Research: While the work on cfDNA has accelerated dramatically to include epigenetic and fragment-based liquid biopsies, cell-free and exRNA have only recently gained more traction [17,18,19]. Many pioneering studies have characterized small cell-free RNAs, yielding new insights into their promise as biomarkers of disease [20]. Recent work has also sought to capture non-invasive transcriptional signal by inference from cfDNA, highlighting both the need and the reticence to pursue non-small extracellular RNA due to fears of degradation [21]. Similar motivations have inspired a handful of groups to begin examining cell-free RNAs greater than 200 nt in length, yielding relevant insights into the potential utility of extracellular messenger RNAs (mRNAs) for monitoring health and diagnosing disease. For example, recent studies using mRNA have predicting gestational age and preterm delivery [22], Alzheimer's disease [23,24], and characterized markers in Lung and Breast cancers [25], providing proof-of-concept that long extracellular RNAs such as mRNAs have diagnostic potential in exRNA liquid biopsies.

Our work has shown that long noncoding (lnc-) RNAs and Transposable Element (TE) RNAs are preferentially released from cells in EVs, and that they are specifically upregulated in the context of hyperactive RAS signaling [26]. We performed RNA-seq on EV RNAs by adapting a protocol that Dr.Kim previously helped develop for input single-cell RNA-seq [27]. I also developed a custom, TE-aware bioinformatics pipeline to enable robust and accurate quantification of TE and lnc-exRNAs [28]. Furthermore, preliminary data have helped us identify and begin to employ an additional, efficient component of the diagnostic assay: target enrichment of TE RNAs that will enable low-cost, accurate,

and highly interpretable disease detection [17,19,29]. This proposal will rigorously evaluate diagnostic performance by greatly increasing the sample size and statistical power available to validate the assay, currently limitations of our prior research.

Critical barrier to progress: While the Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx)[30] projects have provided transcriptional references for tumor- and tissue-specific expression, and projects like the Human Cell Atlas have done a similar service to the single-cell RNAseq community[31], most exRNA datasets are comprised of small- and/or micro-RNA sequencing[30]. While there are documented, large scale cfRNA experiments conducted in the private sector these data sets are not available to academic scientists. Furthermore, those that are available have been hyper-optimized to include only signal of interest to the authoring entity [25]. This prevents the expansion of the field and limits the application of machine learning approaches to solving the complex diagnostic problems we are faced with. In contrast, the cohort assembled for Aim 3 will be orthogonally diagnosed using cutting edge imaging approaches and provide a powerful multi-modal dataset to the RNA liquid biopsy community.

How methods and technologies in the field will be changed: This aim reflects a significant advancement in analytical scope of RNA liquid biopsies: I will quantify transcripts and RNA reflected both in GENCODE annotations and TE databases. Additionally, I will use a novel cohort to perform cutting edge diagnostic analyses to not only construct classifiers that identify likely LUAD diagnosis but also identify distinct RNA markers that drive the classification. This presents a unique opportunity to both add resources to the nascent RNA liquid biopsy field and contribute meaningful analytical perspectives to future studies.

Impact of proposal: Successful completion of this proposal will make significant contributions to the understanding of RNA secreted from cells in extracellular vesicles and the potential for their application to disease diagnostics. In particular, the nature of extracellular TE RNA and the presence of intronic signal observed in healthy and diseased patients will provide new avenues for non-invasive diagnostics.

Contributions to Science & Evidence For Succesful Completion of Thesis

Positions & honors

- Summer 2018: Secondary Mentor, UCSC Summer Internship Program (SIP)
- 09/2018-09/2019: Trainee, UCSC BME NHGRI T32
- 2019-2021: Teaching Assistant, BME 110, 263, 22/23L, 163/263, 178/278
- Summer 2021: AI/ML & Bioinformatics Intern, Bluestar Genomics
- Fall 2021-Present: AI/ML & Bioinformatics Consultant, Bluestar Genomics
- Fall 2021-Present: Fellow, NIH NIDDK (KUH) F99/K00
- Fall 2021-Present: Fellow, UCSC Center for Innovative Teaching & Learning (CITL)
- Spring 2022-Present: Member, UCSC Genomics Institute Diversity, Equity, and Inclusion Committee
- Summer 2022: Baskin Engineering Anti-racist research fellow

Research performed in thesis lab

My thesis work focuses on the transcriptional dynamics of non-coding RNAs, particularly long non-coding RNA (lncRNA) and Transposable Element (TE) RNA in RNA liquid biopsies. I am investigating the context- and tissue-specific expression of lncRNAs and TE RNAs in human clinical samples, and in vitro models of lung and pancreatic cancers. My advisor, Assistant Professor Daniel Kim, is a recognized expert on non-coding RNA biology who is applying his expertise to help us establish a new generation of cell-free RNA liquid biopsies that detect RNA molecules exported in extracellular vesicles: exRNA. Together, we have established a platform that enables the detection and assessment of EV RNA and has shown significant promise as both a diagnostic and monitoring tool.

Aim 1: **LncRNA Biomarkers of Inflammation and Cancer**
Status: Published Book Chapter
Long noncoding RNAs (lncRNAs) are promising candidates as biomarkers of inflammation and cancer. LncRNAs have several properties that make them well-suited as molecular markers of disease:(1) many lncRNAs are expressed in a tissue-specific manner,(2) distinct lncRNAs are upregulated based on different inflammatory or oncogenic stimuli,(3) lncRNAs released from cells are packaged and protected in extracellular vesicles, and (4) circulating lncRNAs in the blood are detectable using various RNA sequencing approaches. Here we focus on the potential for lncRNA biomarkers to detect inflammation and cancer, highlighting key biological, technological, and analytical considerations that will help advance the development of lncRNA-based liquid biopsies. I am lead author on this published book chapter[32].

Aim 1: **Epigenomic reprogramming of transposable element RNAs and IFN-stimulated genes by mutant KRAS**
Status: **In press**, *Cell Reports*
This project reflects the outcomes of **Aim 1** in this proposal. RAS genes are the most frequently mutated oncogenes in cancer. However, the effects of oncogenic RAS signaling on the noncoding transcriptome are unclear. We analyzed the transcriptomes of human airway epithelial cells transformed with mutant KRAS to define the landscape of KRAS-regulated noncoding RNAs. We found that oncogenic KRAS upregulates noncoding transcripts throughout the genome, many of which arise from transposable elements. These repetitive noncoding RNAs exhibit differential RNA editing in single cells, are released in extracellular vesicles, and are known targets of KRAB zinc-finger proteins, which are broadly down-regulated in mutant KRAS cells and lung adenocarcinomas. Moreover, mutant KRAS induces IFN-stimulated genes through both epigenetic and RNA-based mechanisms. Our results reveal that mutant KRAS remodels the noncoding transcriptome through epigenomic

reprogramming, expanding the scope of genomic elements regulated by this fundamental signaling pathway. I am lead author on this manuscript, currently in **press** at Cell Reports[26].

Aim 1: **Extracellular RNA signatures of mutant KRAS(G12C) lung adenocarcinoma cells**
Status: In-revision
This project reflects the outcomes of **Aim 1** in this proposal. Extracellular RNAs (exRNAs) are actively secreted from cells in membrane-bound extracellular vesicles (EVs). Diverse classes of RNAs are secreted as exRNAs, including messenger RNAs (mRNAs), long noncoding RNAs (lncRNAs), and transposable element RNAs (TE RNAs). However, the full composition and clinical utility of exRNAs secreted in response to oncogenic signaling are unknown. Here we use both affinity- and nanofiltration-based EV isolation approaches to show that mutant KRAS(G12C) signaling results in the secretion of specific lncRNAs, TE RNAs, and mRNAs, some of which are prognostic for lung adenocarcinoma (LUAD) patient survival. We found that inhibition of KRAS(G12C) signaling broadly reprograms the noncoding transcriptome, as evidenced by a substantial increase in TE RNA secretion. KRAS(G12C) inhibition also increased the abundance of secreted lncRNAs and retained intron-containing transcripts, while decreasing the mRNA content of EVs. Oncogenic KRAS(G12C) signaling was required for the secretion of mRNAs from a set of 20 genes that are significantly associated with unfavorable clinical outcomes in LUAD. Our study suggests that both coding and noncoding RNAs that are secreted in EVs may serve as KRAS(G12C)-specific signatures for diagnosing lung cancer. I am co-lead author on this manuscript, in revision after review at PNAS[28].

Aim 2: **Comprehensive transposable element-aware characterization of full-length cell-free RNA**
Status: In-preparation
This project reflects the outcomes of **Aim 2** in this proposal. By utilizing expression calculated from TE- and Intron aware annotations and TCGA data, we demonstrate the potential for non-invasive RNA liquid biopsy to be used as both a Pancreatic Cancer diagnostic and a monitoring tool for extended COVID-19 illness. Ultra-low input RNA sequencing libraries of Human blood plasma capture disease-specific RNA expression, particularly from TE RNA and unspliced, full length RNA. I am lead author on a manuscript near completion.

KRAS regulates HERVH noncoding RNAs during human pluripotent stem cell differentiation
Status: In-preparation
I am analyzing the transcriptomes of randomly differentiating human induced pluripotent stem cells (iPSCs) with a 95% knock down of KRAS. I am using a combination of bulk, single cell, and Nanopore direct RNA sequencing to interrogate the transcriptional consequences of the lack of KRAS signaling. Bulk and Nanopore sequencing analysis have implicated a strong upregulation in LTR7-HERVH transposable elements and long non-coding RNA that have LTR7 derived promoter regions. Single cell RNA seq captures a dampened ability to differentiate towards neuronal lineages. I have presented this work at the Bay Area Stem Cell Conference, the NHGRI Research Training and Career Development meeting, and the UCSC stem cell journal club and am co-lead author on an upcoming manuscript.

Research internship with Bluestar Genomics

I completed an internship and consultancy with a biotechnology company that is pushing the field of cancer liquid biopsies forward: Bluestar Genomics. While there, I made significant contributions to their machine-learning platform that have resulted in a performance improvement prior to submission of the diagnostic tool for CLIA approval. I have gained critical experience working with large, clinical datasets prepared to validate a diagnostic platform at scale.

Validation of a Pancreatic Cancer Detection Test in New-Onset Diabetes Using Cell-Free DNA 5-Hydroxymethylation Signatures
Status: Pre-printed
Pancreatic cancer (PaC) has poor (10%) 5-year overall survival, largely due to predominant late-stage

diagnosis. Patients with new-onset diabetes (NOD) are at a six-to eightfold increased risk for PaC. We developed a pancreatic cancer detection test for the use in a clinical setting that employs a logistic regression model based on 5-hydroxymethylcytosine (5hmC) profiling of cell-free DNA (cfDNA). I made significant contributions to modeling performance under the supervision of UCSC BME PhD alum David Haan sufficient to earn co-authorship on a pre-printed manuscript[18].

Research peformed in collaboration with other groups

Demirci Lab – Stanford University

The Demirci Lab invented the ExoTIC technology used to isolate extracellular vesicles in our KRAS inhibitor work above and the manuscript described below. Our continued collaboration is supported by a DoD grant to develop non-invasive lung cancer biopsies that represents a key part of my thesis proposal.

Aim 3: **Extracellular Vesicle RNA biomarkers for early-stage Lung Adenocarcinoma**
Status: In-preparation (see preliminary data) Preliminary results of **Aim 3** where early-stage lung adenocarcinoma patients with matched clinical characterization are shown to be identifiable using non-invasive exRNA sequencing. We further validate the platform described in previous sections, showing the disease-specific expression of coding/non-coding genes is detectable and informative using an exRNA assay. Furthermore, we characterize these signals in relation to the clinical measurements provided for the nodules/tumors detected in the patient population. I am co-lead author on a manuscript in preparation.

Forsberg Lab – University of California, Santa Cruz

In addition to being a collaborator, Dr. Forsberg is the Co-Mentor on my F99/K00 award. I anticipate continued collaborations with the Forsberg lab resulting in multiple authorships.

Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells
Status: Published
Hematopoietic stem cells (HSCs) have the capacity to differentiate into vastly different types of mature blood cells. The epigenetic mechanisms regulating the multilineage ability, or multipotency, of HSCs are not well understood. To test the hypothesis that cis-regulatory elements that control fate decisions for all lineages are primed in HSCs, we used ATAC-seq to compare chromatin accessibility of HSCs with five unipotent cell types. We observed the highest similarity in accessibility profiles between megakaryocyte progenitors and HSCs, whereas B cells had the greatest number of regions with de novo gain in accessibility during differentiation. Despite these differences, we identified cis-regulatory elements from all lineages that displayed epigenetic priming in HSCs. These findings provide new insights into the regulation of stem cell multipotency, as well as a resource to identify functional drivers of lineage fate. I am co-author on this published manuscript[33].

Dynamics of Chromatin Accessibility during Hematopoietic Stem Cell Differentiation into Progressively Lineage-Committed Progeny
Status: In-review
This followup to the work described above explored the epigenetic regulation of hematopoietic progenitors and their lineage. I am co-author on a manuscript in review.

Chan Lab – University of Pittsburgh

The Chan lab has provided us with clinical blood plasma used in manuscripts described above. They have agreed to provide us with 100 additional samples from lung injury, COPD, and other pulmonary disease patient donors. These data will provide an alternative route for **Aim 3**.

Extracellular RNA biomarkers of Pulmonary Arterial Hypertension
Status: In-preparation (see preliminary data)
We find exRNA expression can identify Pulmonary Arterial Hypertension (PAH). Furthermore, we utilize robust clinical measurement and characterization to identify exRNA signals that correlate with clinical features of pulmonary health. This demonstrates the potential for non invasive RNA liquid biopsies in diagnosing and characterizing PAH and potentially other non-malignant pulmonary pathologies. I am co-lead author on a manuscript in submission.

Teaching, mentorship, and community

I've TA'd six courses across seven quarters during my graduate career at UCSC, run a week-long introduction to the *R* programming language for High School students, and leveraged these experiences to become the BME department's 2021-22 CITL Graduate Pedagogy fellow. **I now have support and expectations to facilitate TA-training for incoming BME PhD and Master's students** at the beginning of their graduate careers in Fall 2022.

It has been an additional privilege and joy to help mentor 6 (3 current) undergraduate BME/MCD students during my PhD thus far. Previous students David Lenci and Devin Virassammy graduated and are enrolled in master's programs while their peer Trevor Fujimoto graduated and now holds a scientist position at Bristol Meyers Squibb. Current students Daniel Arriaza and Madeline Chertkow are supported by the Koret Scholarship, their peer Queenie Li is a former participant in the Treehouse Undergraduate Bioinformatics Immersion (TUBI) program and doing an internship this summer with Invitae – a genetic testing company.

Most recently, I was selected as a **Baskin Engineering Anti-Racism** research fellow that will enable my investigation of inequities in the development of liquid biopsies like the one proposed in this thesis. I will produce a report for the school of engineering by the end of Summer 2022.

These experiences, and the support I receive from my F99 fellowship, enabled my successful application to the UCSC Genomics Institute Diversity, Equity, and Inclusion committee as a graduate student representative. In addition to working a broad range of DEI issues affecting the GI community, **I am the lead facilitator of the UCSC GI Bioinformatics Short Course for summer 2022** that will be a free, introductory course on bioionformatics and programming for 24 California community college students.

Research support

I've spent the majority of my PhD seeking independent research support. These efforts culminated in the receipt of a significant NIH award, but began with numerous rejections:

- NSF GRFP: Not Funded
- TRDRP Pre-Doc: Not Discussed
  - Resubmission: Scored, not funded
- NIH F31: Not Discussed
  - Resubmission: Not Discussed
- NIH F99/KOO: **Funded**

Ultimately, this award offers me 2 years of support to finish my PhD (covering the last aim in this proposal explicitly) and 4 years of post-doctoral funding to try and establish myself as an independent investigator.

Ongoing:
F99 DK131504 R. Reggiardo (PI) 9/2021 – 9/2023
Cell free RNA liquid biopsies: The goal of this proposal is to verify an RNA liquid biopsy platform and transition it to have specific utility for the study, and non-invasive assessment, of the hematopoietic bone marrow (BM) niche.
Role: Principal Investigator

Completed:
T32 HG008345 D. Haussler & R. E. Green (Co-PIs) 9/2018 – 9/2019
The UCSC Genomic Sciences Graduate Training Program is an innovative graduate training program that combines cutting-edge computational biology training in a diverse biomedical science and engineering environment.
Role: Trainee

Timeline

The F99 award provides **2 years** of pre-doctoral funding. As of this proposal, I have been supported for nearly **1 year**. I will need to complete thesis work and defend by **Summer 2023** in order to meet the expected timeline and begin **K00**-supported post-doctoral work.

Research Approach

Aim 1: Identify mutant KRAS-induced transcriptional changes in early tumorigenesis and after clinical inhibition

Introduction: Most of the human genome is noncoding and transcribed into RNA[34]. Moreover, about half of the human genome is comprised of transposable elements (TE) [35], and TEs contribute substantially to the noncoding transcriptome [36]. TE RNAs [37] and other classes on noncoding RNAs are often altered during cancer [38] and epigenetic reprogramming [27], where activation of RAS signaling leads to repression of microRNAs [39] and upregulation of long noncoding RNAs (lncRNAs) [27], respectively. In lung cancers, RAS mutations are present in a third of lung adenocarcinomas [40] and serve as driver mutations that initiate tumorigenesis [41]. Although RAS genes are among the most frequently mutated oncogenes in cancer [42], how oncogenic RAS signaling regulates the noncoding transcriptome remains unknown. What has become clear quite recently is the druggable nature of mutant KRAS, a discovery many years in the making [2,3,4]. As these new inhibitors begin to see the clinic, it is imperative we both understand the features and signals of succesful inhibition (or resistance) and utilize them as potent tools for investigating RAS signaling in a highly tunable manner [43,44,45].

Research Strategy: To investigate the role of mutant KRAS in reprogramming the transcriptome during early stages of cellular transformation, we characterized the composition of both intracellular and extracellular RNA, including protein-coding RNA, lncRNA, and TE RNA, using human airway epithelial cells ("AALE") [46] and human bronchial epithelial cells ("HBEC") [47] with constitutively active mutant KRAS. Furthermore, we used a traditional KRAS-driven lung cancer model ("H358") with endogenous mutant KRAS G12C expression exposed to novel KRAS G12C inhibitor AMG510 to interrogate the impact, and specific signatures, of clinical KRAS inhibition on extracellular RNA expression [28].

Overarching Goals: This aim sought to use RNA and Assay for Transposase-Accessible Chromatin (ATAC) seq to characterize the transformative power of mutant KRAS in the earliest stages of oncogenesis. By utilizing this distinct cell line, not a cancer cell line but able to be transformed by mutant KRAS alone, we hoped to identify a transcriptional program indicative of early tumorigenesis and thus early stage cancers. In this context, the extracellular RNA detected holds promise as potentially informative biomarkers for early-stage diagnosis – a critical need in the clinic. Establishing a greater understanding of transcriptional dynamics caused by inhibition of mutant KRAS is thus a natural followup; a key overall goal of this research is to enable the translation of the findings into actionable technology.

Research Design: Intra- and/or extracellular RNA sequencing data was generated from the cell lines described above with additional ATAC sequencing data prepared from the AALE model. Extracellular vesicles were isolated with either a commercial, column-based kit or the Demirci Lab's (Research Experience) microfluidic filtration-based approach. AALE, HBEC, and H358 cells and respective, secreted extracellular vesicles were processed in triplicate.

Aim 1.1 Determine mutant-KRAS dependent transcriptional landscape

Approach: RNA and ATAC sequencing data were generated from AALE cells with an 'empty' plasmid vector and those transfected with a mutant KRAS G12D containing plasmid. These libraries were processed with standard approaches, using *Salmon* [48] for transcript quantification and the NF-core pipeline for ATAC peak calling [49]. *Salmon* was run reference indices built from GENCODE transcripts

[50] and Repeat-masked insertions hosted by the UCSC genome browser. Differential expression was performed with DESeqII [51].

RNA sequencing data were also generated from extracellular vesicles isolated from cell culture media of H358 cells treated with either AMG 510 or DMSO loading control. Vesicles were isolated with two methods for comparison of efficacy.

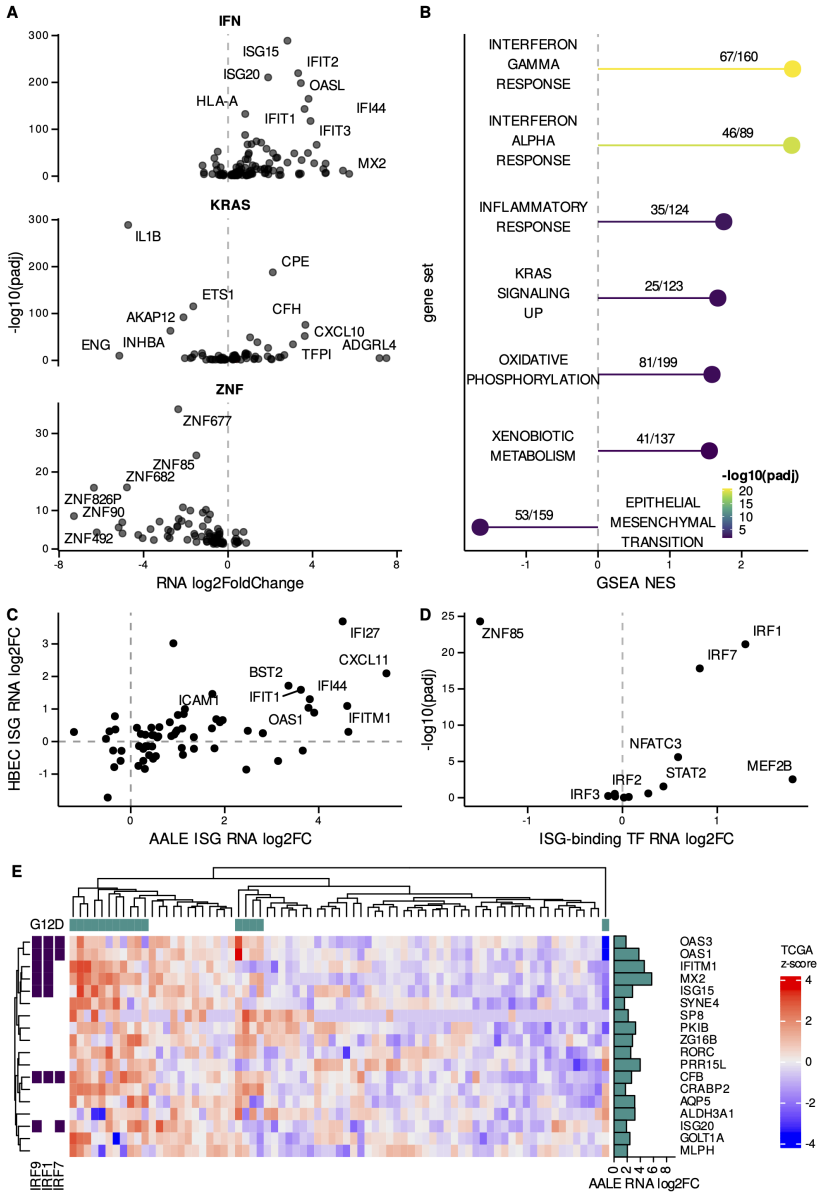

Figure 1: **A.** Volcano plots depicting significant differential expression observed in key gene sets (interferon response alpha/gamma: IFN, KRAS signaling up: KRAS, zinc-finger genes: ZNF). **B.** Significant Gene Set Enrichment Analysis (GSEA) results observed in mutant KRAS AALE differentially expressed genes ranked by adjusted p-value (padj), normalized enrichment score (NES), and annotated with the number of genes observed out of the total genes in each gene set. **C.** Differential expression of IFN-stimulated genes in mutant KRAS AALEs compared to mutant KRAS HBECs. **D.** Differentially expressed transcription factors (TF) with binding motifs enriched in differentially expressed ISG promoter regions. **E.** Hierarchical clustering of expression Z-score in TCGA LUAD RNA-seq data for ISGs upregulated in mutant AALE and exhibiting strong segregation in TCGA LUAD samples based on KRAS G12D mutation status; presence of IRF9/1/7 binding motifs in promoter regions of labeled ISGs.

Results: We compared the transcriptomes of AALE cells transduced with control lentiviral vector to AALEs that were transduced by mutant KRAS-containing lentiviral vector and performed differential expression analysis [Fig 1]. We identified thousands of significantly differentially expressed protein-coding RNAs (n=1028 upregulated, n=1194 downregulated), including ISGs, KRAS signaling genes, and zinc-finger genes [Fig 1A], as well as hundreds of significantly differentially expressed lncRNAs (n=116

upregulated, n=163 downregulated), revealing the broad extent to which mutant KRAS reprograms the transcriptome. GSEA revealed that the three most significantly enriched pathways were the interferon (IFN) alpha and gamma responses, as well as the hallmark inflammatory response (Figure 1B), along with increased KRAS signaling from mutant KRAS(G12D), increased metabolic gene expression, and decreased expression of epithelial-to-mesenchymal transition (EMT) genes [Fig 1B]. We observed strong concordance between mutant KRAS-induced ISGs in AALE and HBEC cells [Fig 1C], confirming our previous results. We then examined the promoter regions (+/- 500bp) of upregulated ISGs and identified motifs enriched in comparison to non-differentially expressed (DE) ISGs [Fig 1D]. To determine the in vivo relevance of our findings in both mutant KRAS AALE and HBEC cells, we examined ISG expression in mutant KRAS(G12D) lung adenocarcinomas (LUAD) from The Cancer Genome Atlas (TCGA), which revealed a subset of ISGs that were upregulated in KRAS(G12D) tumors when compared to normal lung samples with wild-type KRAS [Fig 1E]. These results reveal that mutant KRAS signaling activates an intrinsic ISG response in lung cells both in vitro (AALE, HBEC) and in vivo (TCGA LUAD).

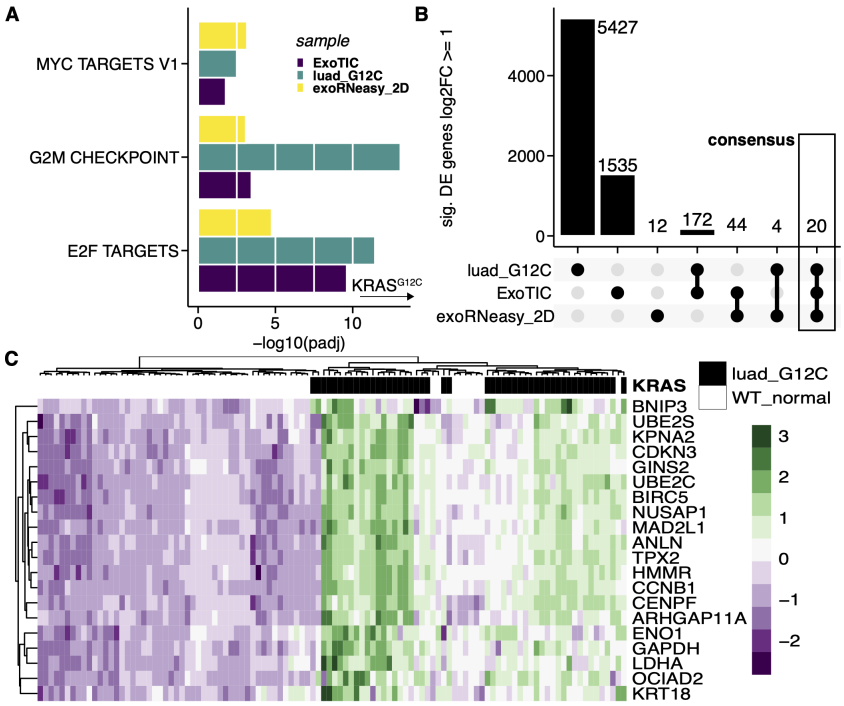

Figure 2: **A.** Bar plot of -log10 transformed adjusted p-value produced for each Hallmark gene set in Gene Set Enrichment Analysis across exoRNeasy, ExoTIC, and TCGA LUAD data sets. **B.** Upset plot quantifying overlap of upregulated genes (log2 fold-change >= 1) in exoRNeasy, ExoTIC, and TCGA LUAD differential expression. The labelled consensus set is used in the following panels. **C.** Heatmap with hierarchical clustering of scaled and centered count values for the 20 genes contained in the consensus overlapping set observed in B.

Furthermore, to determine the relevance of the exRNA signatures detected in both the ExoTIC and ExoRNeasy platforms, we utilized RNA-seq counts from the TCGA LUAD cohort. Initial DE and GSEA approaches identified shared enrichment of 3 hallmark gene sets (MYC TARGETS V1, E2F TARGETS, and G2M CHECKPOINT) and a consensus upregulation of 20 genes across the in vitro and in vivo datasets [Fig 2A/B]. Hierarchical clustering of the LUAD cohort using this 20-gene signature produced robust separation between the G12C LUAD tumor samples and the healthy (WT) lung tissue samples [Fig 2C].

Aim 1.2 Characterize extracellular RNA composition of extracellular vesicles released from cells with mutant KRAS

Approach: Extracellular Vesicles were extracted using the ExoRNeasy affinity-based isolation protocol from AALE cell culture medium in both WT and KRAS G12C contexts. Nanoparticle tracking analysis

(NTA) was used to measure size of vesicles isolated from the AALE cell culture medium and exRNA was sequenced. Analysis was performed as described in 1.1.

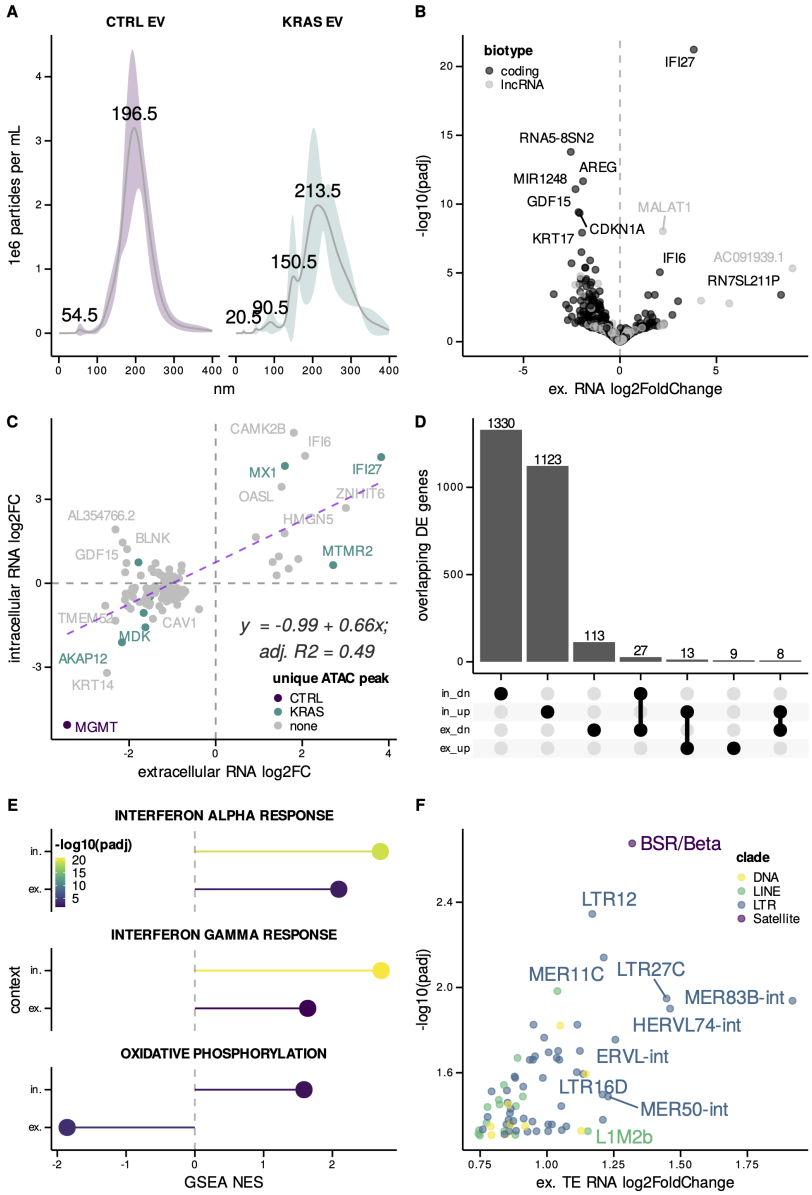

Figure 3: **A.** Size distribution of extracellular vesicles (EV) isolated from control (CTRL) and mutant KRAS AALEs. **B.** Volcano plot of differentially secreted GENCODE protein-coding RNAs and lncRNAs between mutant KRAS and CTRL AALE EVs. **C.** Scatter plot comparing differentially expressed genes between intracellular and extracellular mutant KRAS AALE RNA-seq libraries; linear regression fit with formula and goodness of fit displayed. **D.** Upset plot summarizing overlap of differentially expressed upregulated (up) and downregulated (dn) genes across intracellular (in) and extracellular (ex) contexts. **E.** Significantly enriched gene sets detected in both intracellular (in) and extracellular (ex) contexts. **F.** Differential secretion of TE RNAs in EVs from mutant KRAS AALEs when compared to control AALE EVs.

Results: Extracellular vesicles isolated from mutant KRAS AALEs were comprised of different sized vesicles that were ~90nm, ~150nm, and ~213nm in diameter, while vesicles from control AALE media were predominantly ~196nm in size [Fig 3A]. RNA isolated and sequenced from these vesicles exhibited mutant KRAS-dependent differential expression of both protein-coding genes (n=17 upregulated, n=140 downregulated) and lncRNA (n=5 upregulated, n=8 downregulated) [Fig 3B]. We also observed significant correlation between differentially expressed ISGs in our intracellular (in) and extracellular (ex) RNA-seq datasets that largely agreed with intracellular epigenetic changes (IFI6, MX1, IFI27, and OASL) [Fig 3C,D]. Furthermore, Gene Set Enrichment Analysis (GSEA) revealed that IFN alpha and IFN gamma signatures were enriched in both intracellular and extracellular RNA [Fig 3E], indicating that extracellular RNAs reflect intracellular ISG changes due to mutant KRAS signaling. We

found significant upregulation of predominantly LTR RNAs such as LTR12, MER11C, and LTR27C, along with LINE, DNA, and Satellite repeat RNAs in mutant KRAS AALE extracellular vesicles [Fig 3F].

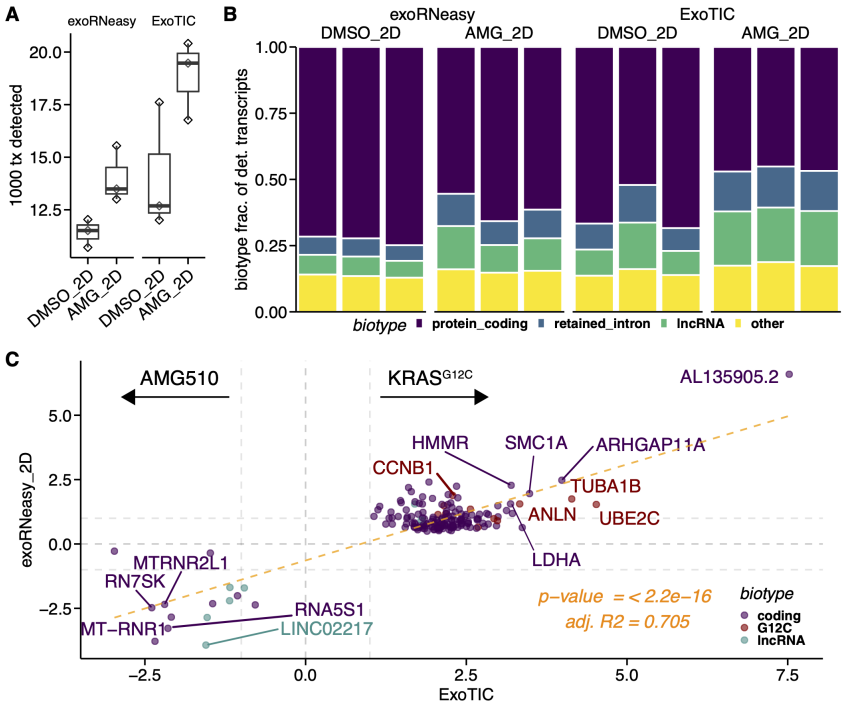

Figure 4: **A.** Distribution of number of transcripts detected above threshold of 5 normalized counts in both ExoTIC and exoRNeasy platforms. **B.** Stacked bar plot displaying the fraction of detected transcripts annotated as protein coding, retained intron, lncRNA, or other GENCODE biotypes. **C.** Scatter plot comparing log2-scale fold-changes between AMG and DMSO treatment using the ExoTIC (x-axis) and exoRNeasy (y-axis) platforms. Colors represent GENCODE biotypes lncRNA, protein-coding, or membership in the G12C-induced gene set from Xue et al [45].

When examining H358 EVs with and without KRAS inhibition, ExoTIC demonstrated greater transcriptional complexity and variability than the equivalent using ExoRNeasy. Both platforms exhibited no significant alterations in exRNA complexity upon AMG 510 treatment[Fig 4A]. RNA-seq captured primarily protein-coding transcripts, although this majority was significantly decreased in ExoTIC-derived exRNAs, which demonstrated an increased abundance of lncRNAs, retained introns, and other noncoding RNA biotypes [Fig 4B]. Despite these differences, the two EV isolation platforms had modest agreement in differentially expressed (DE) genes between DMSO (positive log change) and AMG (negative log change) conditions [Fig 4C], with 64 shared significantly upregulated genes (padj <= 0.05, log2FoldChange >= 1).

Aim 1.3 Identify potential molecular events that are RAS-dependent in early stage tumorigenesis

Approach: With vast DE observed in RNA seq, we decided to focus on two of the strongest events: upregulation of IFN response genes (ISGs) and downregulation of Zinc Finger (ZNF) genes [Fig 1A]. To explore this in more detail, we utilized ATAC-seq to help orthogonally identify regulatory changes distinctly represented by epigenetic events.

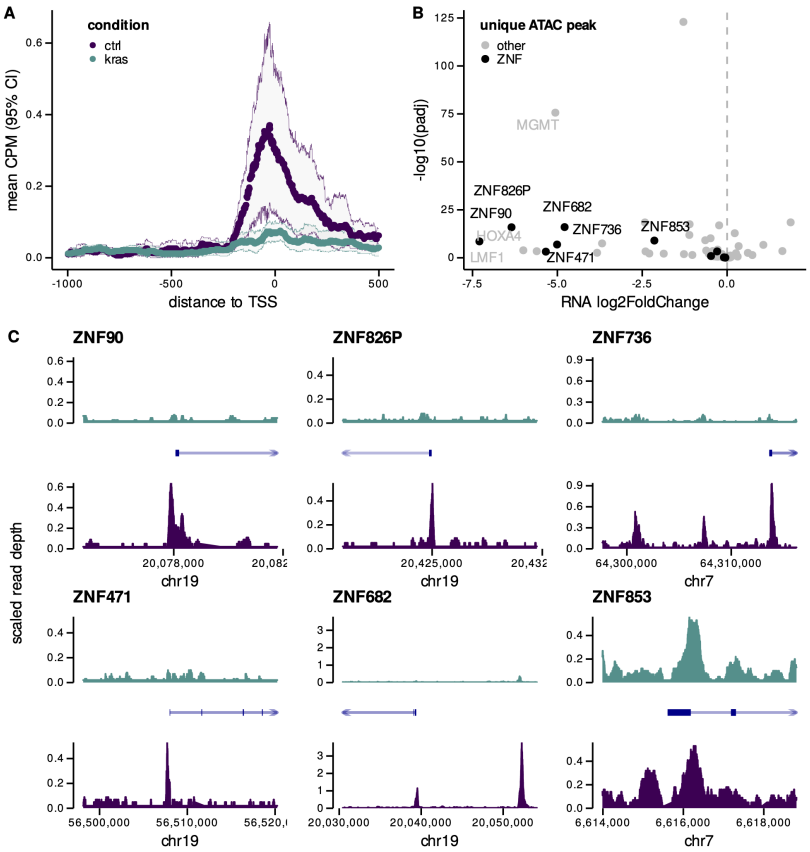

Figure 4: **A.** Mean ATAC-seq counts per million (CPM) (95% CI) in promoter regions of downregulated KZNFs (< -4.5 log2FoldChange) in both mutant KRAS and control (CTRL) AALEs. **B.** Differential expression of KZNFs with 'unique' peaks near TSS (only present in mutant KRAS or control AALEs). **C.** ATAC-seq coverage in both KRAS and CTRL AALEs for subset of KZNFs with unique peaks detected near TSS.

Results: To determine the extent to which mutant KRAS signaling epigenetically silences KZNF expression, we examined ATAC-seq data for all significantly downregulated KZNF loci, which revealed that mutant KRAS signaling substantially reduces chromatin accessibility at TSS regions [Fig 4A]. When we examined genes with 'unique' ATAC peaks that were only present in control AALEs but disappeared in mutant KRAS AALEs, we found that many of these genes were KZNFs that were significantly downregulated [Fig 4B]. Six of these downregulated KZNFs, ZNF90, ZNF826P, ZNF736, ZNF471, ZNF682, and ZNF853, had peaks unique to control AALEs [Fig 4C]. Furthermore, we identify these ZNF genes as downregulated in a subset of KRAS G12C containing tumors in LUAD TCGA [Fig 5A], indicating a broadly relevant KRAS-driven event with dramatic regulatory potential.

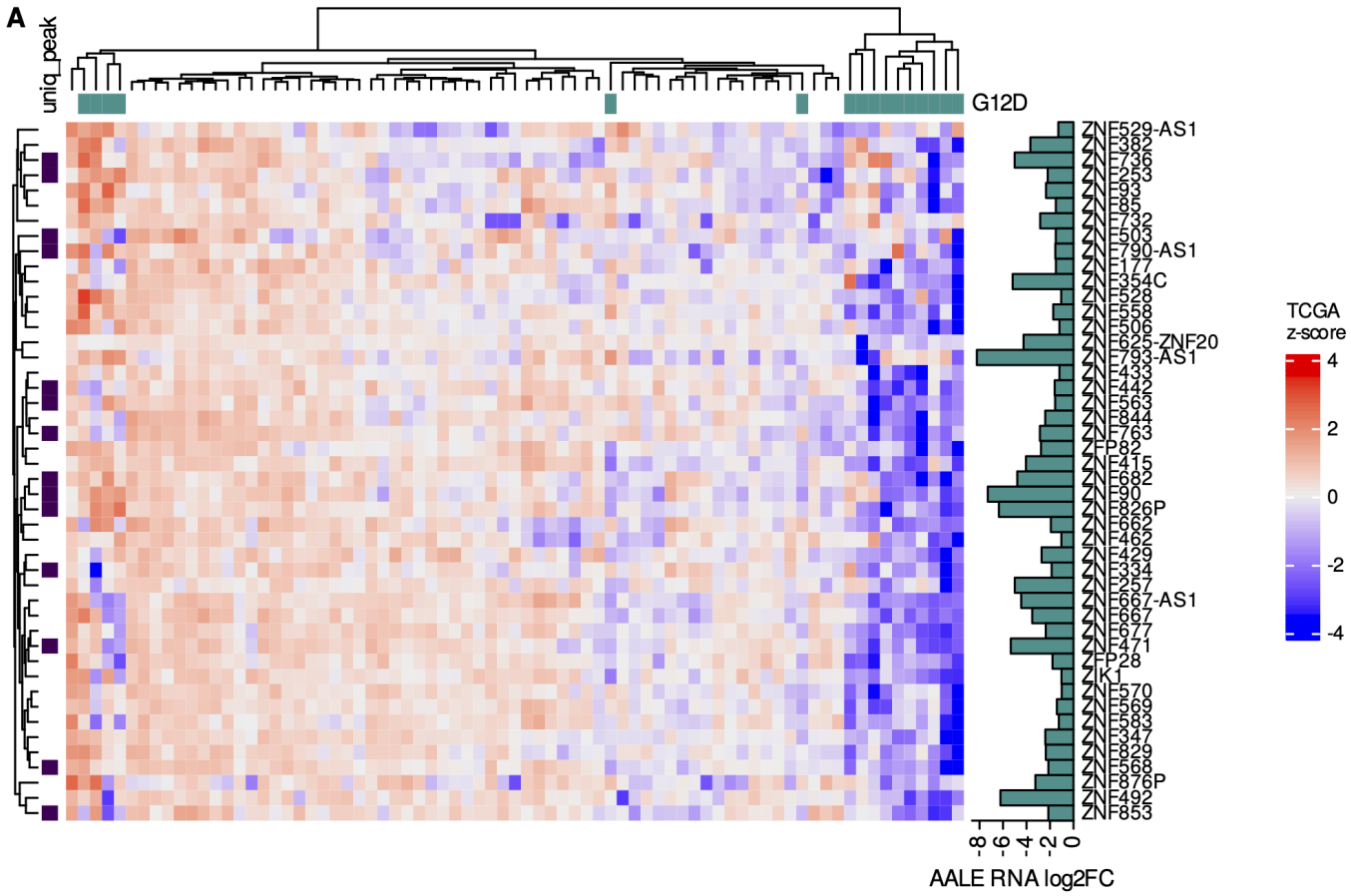

Figure 5: **A.** Hierarchical clustering of expression z-scores in TCGA LUAD RNA-seq data for KZNF genes downregulated in mutant KRAS AALEs; KZNFs with unique peaks in their promoter regions in control AALEs are labeled.

Conclusions and Future Directions

Collectively, our findings reveal the transcriptomic and epigenomic impact of oncogenic KRAS signaling on TE RNAs and ISGs. Our study suggests that KZNF repression by mutant KRAS signaling leads to de-repression of TE RNAs, triggering an intrinsic ISG response [Fig 5].
This model is supported by broad and significant downregulation of these same KNZFs in mutant KRAS-driven lung adenocarcinomas in vivo. Our conclusions are based on deeply sequencing and analyzing the intracellular and extracellular transcriptomes and epigenomes of mutant KRAS-transformed lung cells.

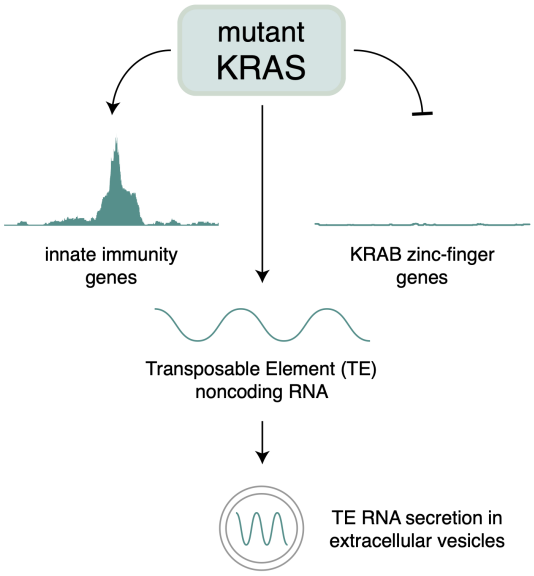

Figure 5: **A.** Simple model of KRAS-driven de-repression of TE RNAs and activation of intrinsic ISC response.

We also present further evidence for the utility of extracellular RNAs in detecting intracellular RNA changes in cancer cells. Notably, we show the secretion of specific TE RNA and ISG signatures that are aberrantly upregulated in mutant KRAS lung cells. The enrichment of TE-derived noncoding RNAs in extracellular vesicles released from mutant KRAS cells highlights their potential utility as RNA biomarkers for diagnosing RAS-driven cancers.

We expand upon this finding by demonstrating the potential for exRNA to be used in assessing response to clinical cancer treatment, where we reliably identify differentially expressed RNAs closely associated with oncogenic KRAS signaling. Upon silencing, we see ablation of these events in favor of evidence for either retention of RAS-associate mRNA to enable resistance to treatment or significant loss of expression.

Together, these findings expand the known role of oncogenic KRAS and establish strong evidence for the relevant of exRNA in cancer.

Aim 2: Develop TE- and Intron-aware RNA-seq analysis to comprehensively assess plasma exRNA expression

Introduction: Current liquid biopsy approaches focus primarily on cell-free DNA and well characterized coding genes [17,18,19,21]. This proposal seeks to expand the liquid biopsy diagnostic repertoire: validating extracellular RNA of coding, long non-coding, and transposable element biotypes as potentially useful clinical markers. Transposable Elements (TEs) comprise nearly 50% of the human genome, are systematically expressed in disease contexts, and are consistently excluded from most RNA-seq analysis. A single exRNA TE was found to be indicative of Alzheimer's disease in pre-symptomatic patients [24], and I have demonstrated that many more are systematically enriched in pancreatic cancer and Sars-CoV-2 infection. Analysis of both TEs in general and the particular nature of extracellular RNAs has presented numerous challenges and questions. Among them – the specificity of expression, the difficulty of multi-mapping reads, the heterogeneous cargo of vesicles, and the nature of the nucleic acids represented. This effort led us to identify challenges we could and could not address; locus-level TE expression, for example, is an active area of research but ultimately not essential to a diagnostic assay and difficult to get funded. Instead, I focused on establishing an approach that could generate novel feature sets that utilized the entire sequencing library available for each sample/patient as, with canonical transcriptome quantification, we consistently only utilized 20-50% of reads in disease contexts.

Research Strategy: Utilizing *Salmon* and its suite of analytical approaches, I implement a multi-pronged approach to characterizing RNA expression [48,52,53]. This includes both aligning to GENCODE annotated transcripts, repeats, and intronic regions [50,53]. This pipeline has been constructed within robust existing frameworks for bioinformatics analysis that exist in the *R* ecosystem [52]. As both a realization of purpose and an attempt at demonstrating the usefulness of the pipeline, I use it to analyze 30 clinical plasma samples from healthy donors, pancreatic cancer patients, and COVID-19 infected individuals.

Overarching Goals: We seek to establish an analytical approach that produces robust and performant feature sets for diagnostic modeling. To achieve this, I have tried to generate as complete an accounting of transcriptional activity detected in plasma exRNA as possible. This, we hope, will serve as the foundation for a robust liquid biopsy enabled by machine learning classifiers trained on some combination of the feature sets calculated by the approach.

Research Design: ExRNA isolated from EVs extracted from blood plasma of healthy donors (10), pancreatic cancer patients (10), and COVID-19 infected patients (10) was sequenced to a depth of approximately 5 million reads. These libraries are processed with the sample where a set of discovery analyses are undertaken: differential expression, unsupervised clustering, investigation of library properties and behavior. We demonstrate the potential for this approach to inform diagnoses.

Aim 2.1 Construct pipeline to detect and characterize total RNA content of extracellular RNA sequencing libraries

Approach: RNA sequencing reads (FASTQ) were trimmed with *Trimmomatic* and had quality assessed using *FASTQC* and visualized using *MultiQC*. Samples with outlier repetitive reads and/or library depth were removed. Initial alignment was performed with *STAR* using the GRCh38.p13.genome.fa hosted by the GENCODE consortium. Quantification was performed using *Salmon* with 3 separate transcriptome annotations:

1. The GENCODE consortium Hg38 reference annotation (v 39)
2. The above reference and repeat elements
3. A ‘process/intron aware’ reference generated by adapting the published guide [53] to bulk RNA-seq data

This approach, built in part with *tximeta* [52] enables the generation of robust, portable, and extensible reference annotations, sequencing meta data, and quantification information. There remain key additions to this pipeline and the resulting analysis that are discussed in **2.3**.

Aim 2.2 Perform pipeline-enabled analysis on cohort of clinical extracellular RNA samples

Sample collection, preparation, and sequencing:

1.5 ml of blood plasma from pancreatic and COVID-19 patients were used as a starting material from which EVs and RNA were isolated 10 blood plasma served as a control for both pancreatic and COVID-19 plasma samples, which were also processed using the same kit. cDNA was synthesized from exRNA and libraries were generated for Illumina sequencing using a NextSeq 500.

Approach: Sequenced samples were processed with the pipeline described in **2.1** assessed for transcriptional events with potential to distinguish diseased patients from healthy donors.

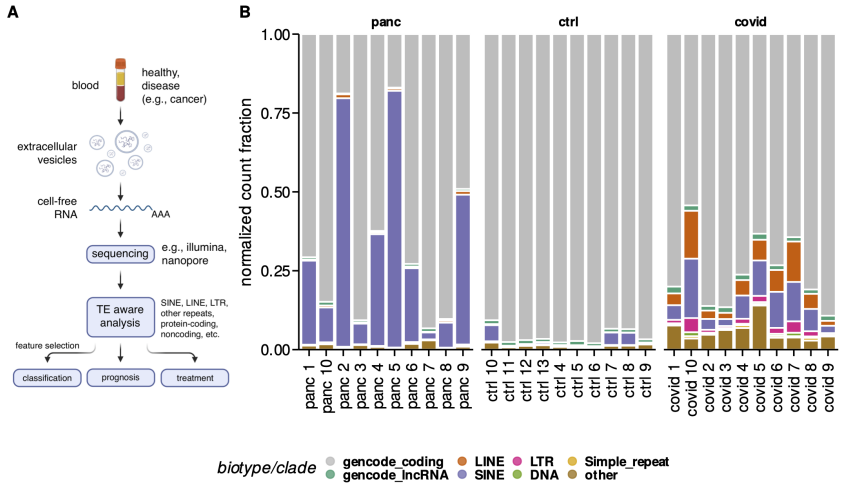

Figure 6: **A.** Simplified diagram of approach used in **Aim 2**. **B.** Distribution of biotype representation (by normalized count) in RNA seq quantifications for samples from each cohort, colored by GENCODE-annotated biotype or Repeat Masker-annotated TE clade.

Results: RNA-sequencing of RNA encapsulated in extracellular vesicles released into human blood plasma is the foundation and intended application of the pipeline developed in **aim 2** [Fig 6A]. An overall assessment of biotype representation in libraries prepared from each patient context demonstrated a specific and pronounced distribution of TE families: panc samples possessed variable but highly robust SINE expression, covid samples less dramatic but more consistent LINE expression [Fig 6B].

In order to assess the relevance of exRNA transcriptional signatures to Pancreatic Ductal Adenocarcinoma (PAAD), we recomputed TCGA RNA-sequencing data in a TE-Aware manner. This enabled Kaplan-Meier analysis using both GENCODE-annotated genes and Repeat Masked elements. We observed significant association with poorer survival outcomes using subsets of genes corresponding to hallmark gene sets significantly enriched in panc plasma GSEA [Fig 7A/C], upregulated GENCODE genes [Fig 7E], and downregulated TE RNA [Fig 7G].

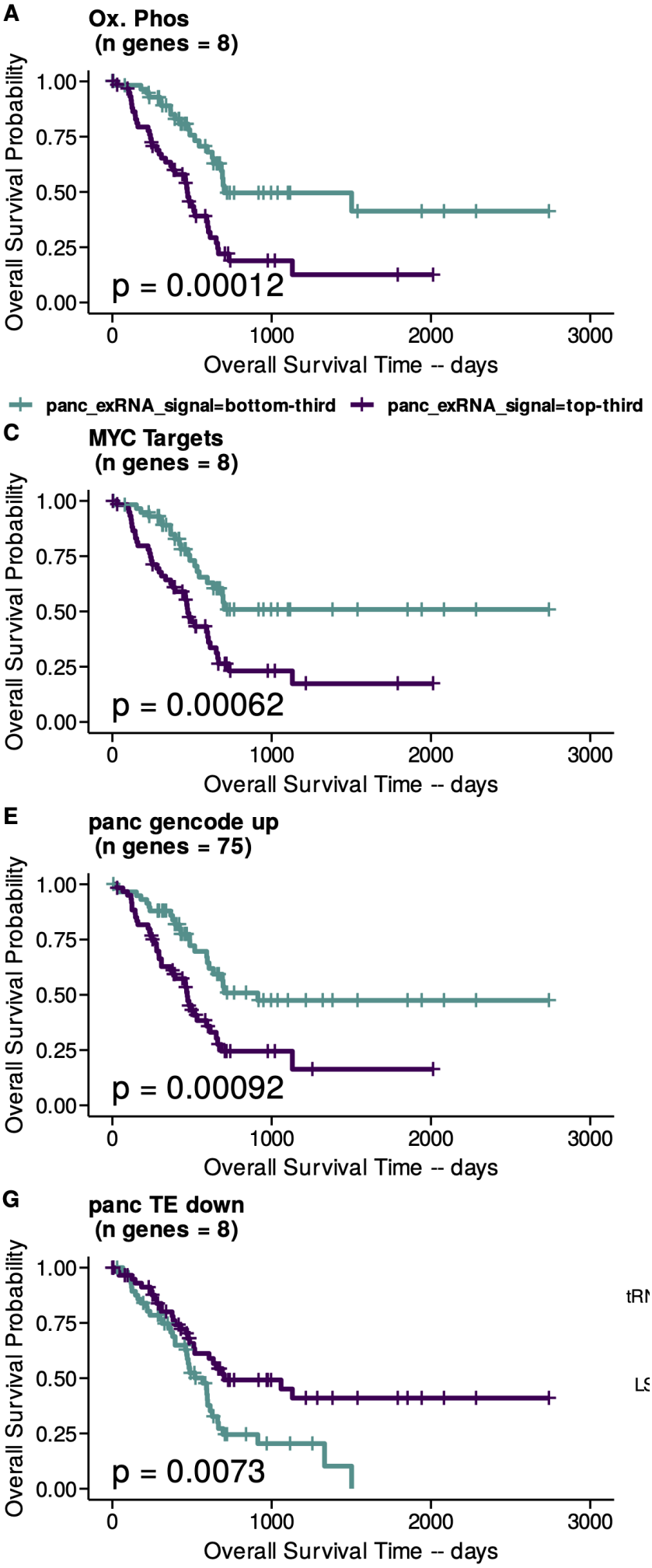
A
Ox. Phos
(n genes = 8)
Overall Survival Probability
1.00
0.75
0.50
0.25
0.00
p = 0.00012
0
1000
2000
3000
Overall Survival Time -- days
panc_exRNA_signal=bottom-third
panc_exRNA_signal=top-third
C
MYC Targets
(n genes = 8)
Overall Survival Probability
1.00
0.75
0.50
0.25
0.00
p = 0.00062
0
1000
2000
3000
Overall Survival Time -- days
E
panc gencode up
(n genes = 75)
Overall Survival Probability
1.00
0.75
0.50
0.25
0.00
p = 0.00092
0
1000
2000
3000
Overall Survival Time -- days
G
panc TE down
(n genes = 8)
Overall Survival Probability
1.00
0.75
0.50
0.25
0.00
p = 0.0073
0
1000
2000
3000
Overall Survival Time -- days

Figure 7: **A. - G** Kaplan-Meier survial curves calculated for gene expression of top and bottom third segments of TCGA PAAD in terms of expression for a specific set of genes enriched or depleted in pancreatic cancer patient blood plasma.

Aim 2.3 Complete, package, and benchmark the pipeline

Approach:
While I don't anticipate dramatic changes to the underlying principles of the approach, I do want to expand upon the interaction between different references and the depth to which TE alignments are explored. I propose the inclusion of the following modules to the pipeline:

1. **TE Age Estimation**: A compelling feature that may aid classification of liquid biopsy samples is estimated age of observed enriched/depleted TE families. Utilizing Lift Over [54,55], I will identify potential insertions of origin for enriched/depleted TEs and determine their orthogonality across the most recent primate lineages [56]. I can estimate (95% CI) the age of TEs observed and ablated in individual samples, creating low dimensional feature to compliment modeling approaches.
2. **Intron & TE overlap**: Preliminary analysis suggests there is not correlation between TE abundance and intronic read abundance. However, it may be quite informative to explore the overlap in the Intron and TE features in order to better understand the biological underpinning of intronic sequence content in exRNA data. In our vesicle selected approach we observe context-specific 'un-spliced' reads [Fig 9A].

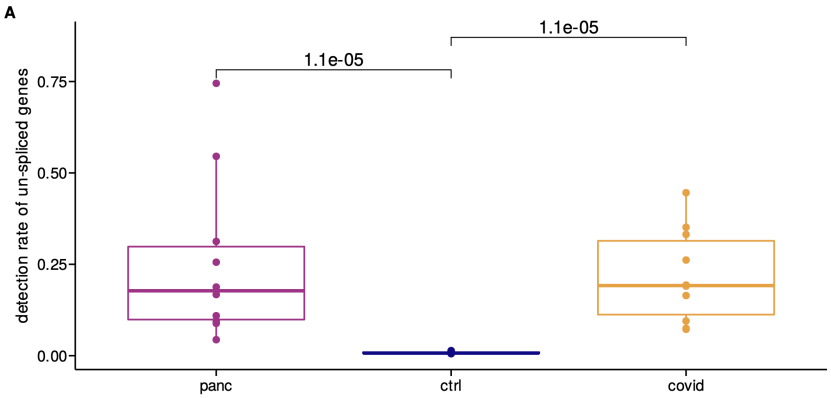

Figure 9: **A.** Rate of unsplice/spliced transcripts in intron-aware quantification

Overall Results:
As demonstrated in [Fig. 6] and [Fig. 7], this approach enables robust identification of coding and non-coding signatures present in exRNA liquid biopsy libraries. Furthermore, the analysis demonstrates promising unsupervised clustering results [Fig. 10A,B]. A key advantage of the approach is its library utilization, which is significantly enhanced compared to TE-naive quantification [Fig. 10D]. This advantage is clearly explained by examining library mapping performance which, as seen in [Fig. 10C], is biased towards multi-mapping reads in disease-contexts.

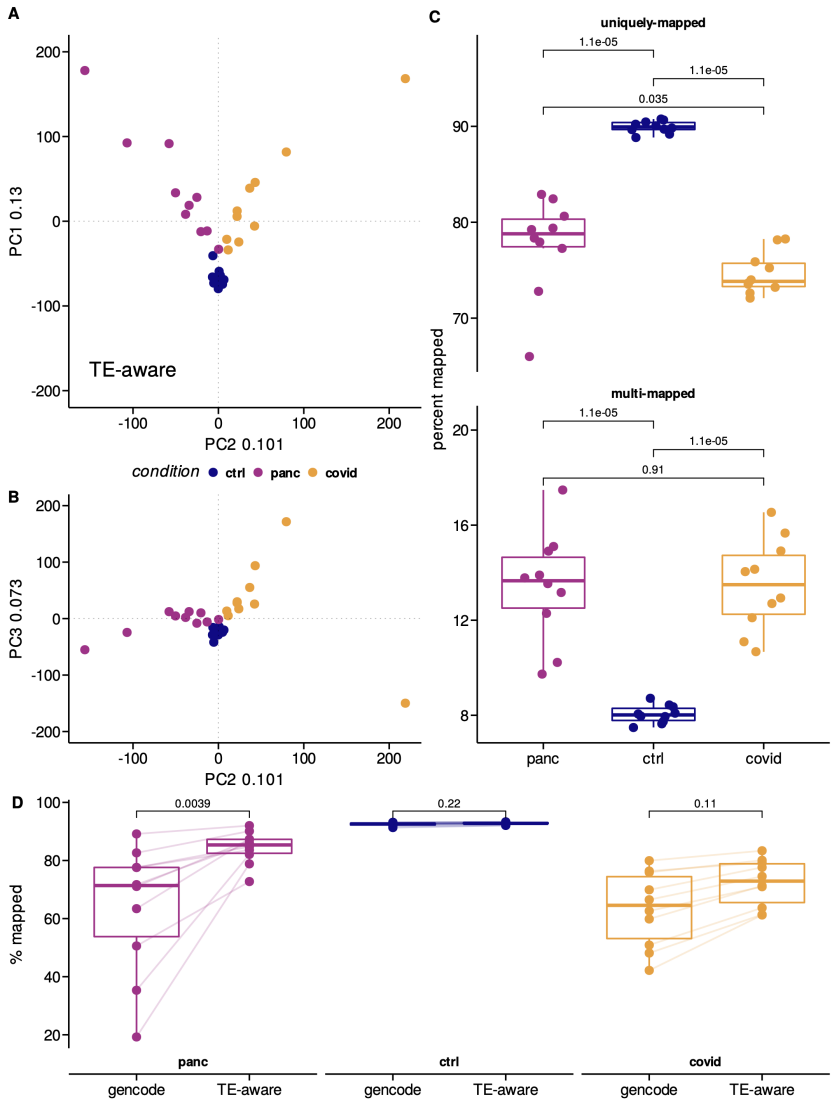

Figure 10: **A.** PCA dimensions 1 & 2 using normalized counts across all three cohorts: panc, ctrl, covid. **B.** PCA dimensions 2 & 3 using normalized counts across all three cohorts: panc, ctrl, covid. **C.** Distribution of uniquely- and multi-mapped reads, as determined by STAR alignments, in libraries prepared from each cohort (Wilcoxon). **D.** Comparison of Salmon read mapping rates between use of gencode-only reference annotation and TE-aware reference annotation (Wilcoxon).

Potential Problems & Alternative Strategies:
This approach sacrifices some specificity in mapping to enable a curtailed, yet still feature-rich, dataset to be generated from exRNA sequencing that is able to take distinct advantage of the biases inherent to the approach. While I do not anticipate that loss of specificity, particularly in TE mapping, to damage performance, it may be compelling and useful to explore the potential origins of TE reads and determine how multi-mapping biases may be affecting quantification.

Aim 3: Validate a Transposable Element-aware RNA liquid biopsy in LUAD cohort

Introduction: Preliminary data generated in our lab suggest context specific enrichment of RNA expression is observable in human blood plasma via sequencing of RNA isolated from extracellular vesicles. We hypothesize that these transcriptional events can be used to classify, and thus diagnose, disease state in patients with high sensitivity and specificity. Furthermore, we suspect that TE RNA will have significant diagnostic value due to their overwhelming genomic abundance and systematic variation observed in our preliminary studies. In this aim, I propose the execution of an exRNA liquid biopsy in patients diagnosed with lung adenocarcinoma.

Research Strategy: Our recent findings revealed that Pancreatic Adenocarcinoma (PAAD) and COVID-19 patients exhibit highly disease-specific TE RNA signatures in the plasma. We hypothesize that TE RNAs and TE-derived lncRNAs serve as tissue- and disease-specific biomarkers in human plasma. To produce a highly robust assessment of RNA liquid biopsy performance for Lung Cancer detection, we will recruit a 120 member, balanced cohort of lung cancer patients and healthy donors. Each donor will be assayed with an improved, standard of care diagnostic: Fluorodeoxyglucose-positron emission tomography / Computed Tomography (PET/CT) as well as the exRNA liquid biopsy described in **Aim 2** [Fig. 12]. These data sets will be used to train diagnostic classifiers using features derived from both PET/CT and the exRNA biopsy, demonstrating the efficacy of each approach individually and an integrated diagnostic approach. Furthermore, we will be able to compare the TE-aware exRNA liquid biopsy data to a recompute of the TCGA LUAD dataset currently underway in the lab, providing a robust contextuliazation of the signal observed in our clinical cohort.

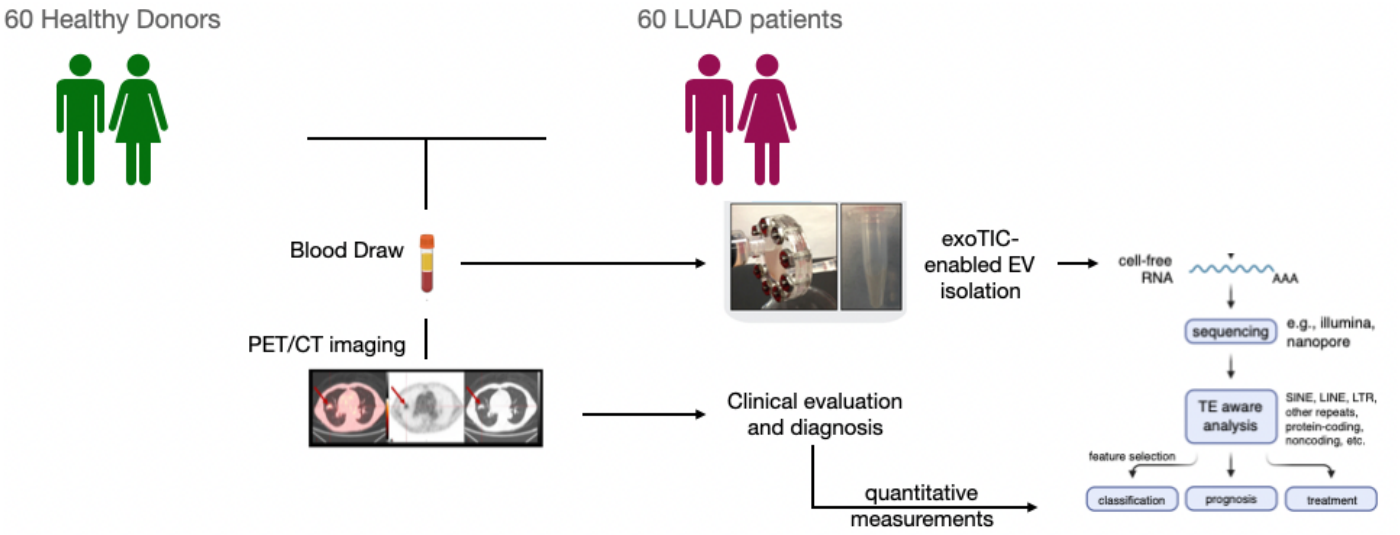

Figure 11: **A.** Diagram of study design and diagnostic assay execution.

Overarching Goals: In this aim, we propose the sequencing of a statistically powerful number of heathy and LUAD patient plasma, overcoming a significant weakness of previous studies in this proposal. By performing the assay in parallel with a standard of care for lung cancer diagnosis, I will be able to confidently assess performance of the exRNA liquid biopsy for real-world application. Finally, the multi-modal diagnostic dataset will offer the opportunity to devise a new, integrated diagnostic approach combining signal from exRNA and PET/CT.

Research Design: We will sequence exRNA isolated from exosomes extracted from blood plasma in cohorts of healthy (60) and lung cancer (60) patients using protocols described previously and enhanced within this proposal. PET/CT data will be processed by collaborators and quantitative imaging variables will be collected for each patient. Binary classification models using Logistic Regression (lasso) and/or Gradient-boosted decision trees (GBDT) will be trained on TE-and-Process-aware features generated from the exRNA liquid biopsy. PET/CT features will be incorporated in a separate model as either a stacked ensemble with the exRNA features or used as interaction terms with the exRNA features. Models will be evaluated for sensitivity on cross-validated training samples and held out validation samples during performance optimization. Final estimated performance of optimized models will be reported on a hold out set [Fig. 11]. *Sample size and power analysis*: Significant efforts will be made to ensure a sampling of individuals diverse in age, sex, and ethnicity are selected for analysis in this study. We anticipate the sampling procedure of 120 patients at 50% confirmed disease status to provide 85% power with 5% one-sided error to detect effect sizes within 0.5 standard deviations.

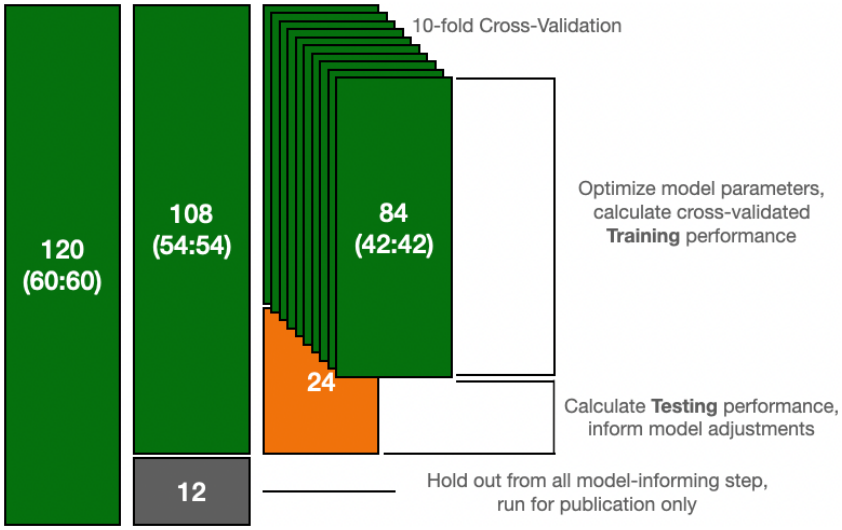

Figure 12: **A.** Diagram of study sampling, training, and testing design.

Aim 3.1 Profiling exRNA in lung adenocarcinoma cancer patient plasma and classifying with exRNA liquid biopsy

Sample collection, preparation, and sequencing:
See **Aim 2**, PET/CT will be performed and data processed by collaborators.

Analysis: Quantification will be performed as described in **Aim 2**. Pseudo-aligned reads will be aggregated and normalized using DESeqII which will use the resulting counts in calculations to determine differential expression across conditions with consideration of available metadata (age, sex, batch, etc). Differential expression output will be used to generate pre-ranked gene lists for Gene Set Enrichment Analysis (GSEA) and Genomic Regions Enrichment of Annotations Tool (GREAT). Normalized data will also be used to perform clustering to identify subgroups driven by RNA expression and features therein that reliably discriminate between healthy and disease patient populations. Principal Component Analysis (PCA) will be employed to reduce the dimensionality of the data set and identify features that contribute to sample heterogeneity and clinical features that correlate with the derived Principal Components. Components generated using PCA will be further reduced with UMAP to identify distinct sub-groups within the healthy and disease populations based on linear and non-linear covariance in the feature space.

Modeling Approach: I will implement machine learning models that enable binary classification via optimized sets of gene expression features to determine the disease label (LUAD or healthy) of each patient sequenced in the study (60 LUAD, 60 healthy). In particular, I will seek to develop simple, explainable models using three algorithms: Random Forests (RF), Gradient Boosted Decision Trees (GBDT), and Least Absolute Shrinkage and Selection Operator (Lasso) regularized Logistic Regression. Unlike deep-learning classification approaches, these all allow the construction of models that enable direct identification of features (genes) that contribute to classification and thus estimation of plausible biological context [57]. RF and GBDT are both implementations of ensemble learning – building and combining many weak-learning models – through successive bootstrapping of the training data (bagging, RF) or successive optimization of iterative models (boosting, GBDT).

These approaches, and other tree-based algorithms, enable highly explainable classification due to the logical sequence of decisions used to classify. Lasso is a form of penalized regression that optimizes against complicated models that rely on many features (genes) in favor of simpler models that achieve similar performance [58]. Lasso is a standard in diagnostic modeling, will produce explainable classifications on par with RF and GBDT, and identify strong predictors from the many potential RNA biomarkers [10.1155/JBB.2005.147]. All approaches have been demonstrated to work on sample sizes equivalent to the those proposed in this aim [23,59].

A split-sample approach will be taken with the dataset where stratified subsets of 70%, 20%, and 10% of the input data will be created for training, testing, and validation respectively. Data is split as described to allow for statistical assessment of model performance and validation of generalizable classification performance when classifying on samples the model has not seen before. Model paramaters and 'training performance' will be calculated through ~10 fold cross-validation of the chosen model on the 70% split of data. This optimized model will be used to calculate 'testing performance' by classifying the 20% split of data, the performance of which will be taken into consideration for further model optimization. Final 'validation performance' will be estimated by using the final model(s) to classify the 10% split of data. The metric we will primarily evaluate to determine performance will be sensitivity (true positive rate) at a specific level of specificity (1 - false positive rate). For example, we are interested in models that perform at >= 95% specificity – and thus, we will calculate 95% confidence intervals of sensitivity when only allowing <= %5 false positivity.

Aim 3.2 Determine diagnostic performance of exRNA liquid biopsy, PET/CT assay, and integrated diagnostic approach

Sample collection, preparation, and PET/CT:
We will be provided with quantitative variables derived from the PET/CT assay performed and assessed by collaborators.

Analysis: Expected variables are as follows:
(1) Nodule borders (spiculated, lobulated, smooth, irregular),
(2) Location (upper lobe vs lower lobes, 75 percent of cancers are in the upper lobes),
(3) Calcification pattern (eccentric, central, popcorn, stipples, laminated),
(4) Size,
(5) Growth rate,
(6) Density
In addition to the disease classification decision derived from the PET/CT assay. These will be constructed into feature sets, explored with preliminary data analysis to determine distributions across the populations and/or the presence of batch effects. Data will be transformed accordingly to enable quantitative analysis with potential drop-out from healthy donors having minimal non-zero quantitative variables.

Modeling Approach:

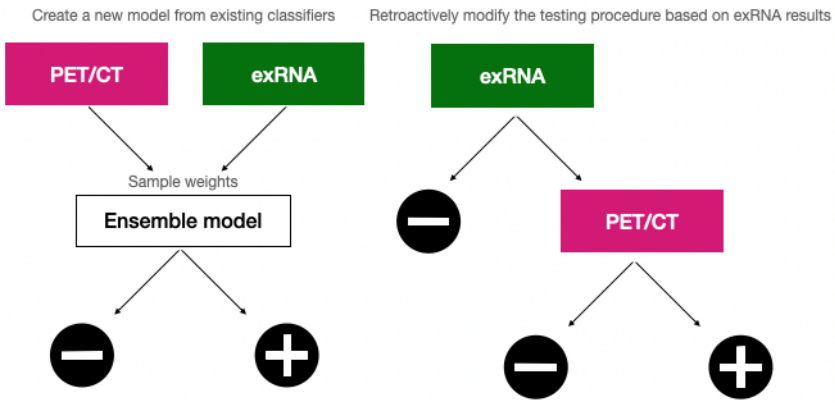

Figure 13: **A.** Diagram of stacking/ensemble model approach.

Stacked/ensemble model: A possible implementation of the exRNA liquid biopsy is to use it in sequence with existing standard of care, PET/CT [Fig 13]. To model the impact this approach would have on diagnostic performance, we will 'stack' the collaborator-generated PET/CT model and the exRNA model(s) generated in **3.1** and see if the combined classification information improves sensitivity. This will involve training a new model on the output probabilities/model scores ('y hat' for

Lasso) from both approaches. Additionally, we will perform a retroactive analysis where we use the exRNA liquid biopsy classification to inform whether or not the PET/CT is 'performed' or in this case, used to make a diagnostic decision. Here, we will calculate the potential sensitivity and specificity for the cohort given the decision to diagnose with PET/CT is pre-empted by a decision to use that information or not depending on the exRNA biopsy.

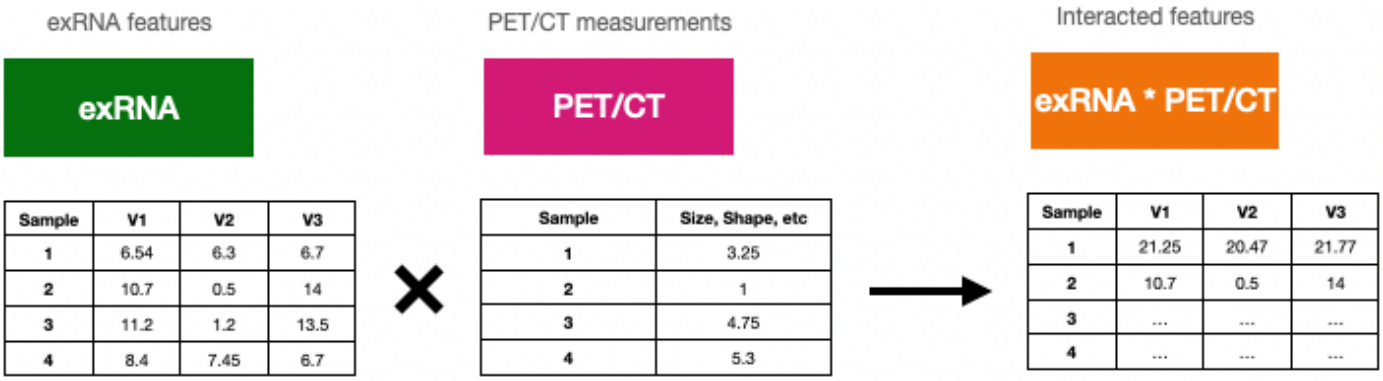

Figure 14: **A.** Diagram of interaction term model approach.

Interaction model(s): Another potential integration is to interact the exRNA features with the PET/CT features [Fig. 14]. I've demonstrated the efficacy of feature interaction previously with Bluestar genomics [18]. Essentially, we will calculate linear combinations between PET/CT features and exRNA features, transforming the exRNA features with an independent PET/CT variable to leverage underlying relationhsips that can inform classification. This will benefit model performance if, for example, two if the relationship between exRNA features and disease status is dependent on a PET/CT variable (ex. Size). This transforms the standard logistic regression formula for patient *p* with *i* exRNA features *F* and PET/CT interactor *cp* from:

$$\hat{y}_p = B_0 + \sum_{i=1}^{n} B_i * F_i$$

to

$$\hat{y}_p = B_0 + c_p(\sum_{i=1}^{n} B_i * F_i)$$

with there also being a possibility of calculating the interaction as a quotient, exponential, or transforming the taking the logarithm of interactor *cp*. We will compare interacted models to standard models by comparing sensitivity estimates and sample probability distributions.

Aim 3.3 TCGA-enabled molecular contextualization

Motivation: While the dataset generated in **Aim 3** will be more robust than previously assembled by our group, it still remains a very limited sample size for such a variable disease. In order to enhance the contextual understanding of what our exRNA liquid biopsy is detecting, we have recomputed LUAD and LUSC TCGA datasets in the TE-aware manner described in **Aim 2**. This will enable to training of additional models, the exploration of consensus TE expression that may be reflected in exRNA, and the mining of robust meta data to determine possible clinical covariates that drive the observed signals.

Approach: TCGA data is already being recomputed on the AnViL TERRA service. Count data will be processed as described elsewhere in **Aim 3** and used to characterize TE expression in tumor samples.

Furthermore, TE-aware classification models can be trained as described in **3.1** matched normal or GTEx samples (also being recomputed) and used to clasify exRNA samples as a test of signal cross-over/relevance between the two different contexts. If needed, this can also serve as a feature engineering opportunity where we identify RNA enriched in disease samples as potential features for our exRNA model.

Expected Outcomes: Employing powerful and reliable statistical learning approaches to identify features that can contribute to parsimonious models will enable diagnostic benchmarking of this exRNA liquid biopsy platform and better understanding of the underlying biology that enables diagnosis. These features, and their resulting models, will form the foundation for more targeted diagnostic approaches to be built on. I will further reveal the unique dynamics of TE RNA in distinct clinical contexts, including an effort to revitalize TCGA data with enhanced TE-aware quantifications and the largest TE-aware exRNA analysis to date. This analysis will enable more detailed molecular characterization of the exRNA datasets generated in the aim. Both newly generated and re-analyzed data produced in Aim 1 will serve as valuable resources to the genomic and computational biology communities as they seek to enhance the clinical utility of RNA sequencing in diagnosing disease and identify optimal treatment options. I will make this data available in a well annotated format in the UCSC Xena Browser.

Potential Problems & Alternative Strategies: Based on preliminary data generated in our lab, I do not anticipate problems with generating sufficiently deep and complex RNA-sequencing libraries. It is possible that some samples behave as either technical or biological outliers that may damage model training – these will be detected by quality-control clustering at each stage of analysis using both hierarchical cluster and principal component analysis (PCA). It is also possible that technical batches are found in sequencing, these will be identified either through a priori identification or in clustering and quantified using surrogate variable analysis which will generate quantitative descriptions of the batch effect(s) that can be accounted for in DESeqII normalization and differential expression computation.

References

1. **Cancer Statistics, 2021**
Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, Ahmedin Jemal
CA: A Cancer Journal for Clinicians (2021-01) https://doi.org/ghs9tj
DOI: 10.3322/caac.21654 · PMID: 33433946

2. **The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity**
Jude Canon, Karen Rex, Anne Y Saiki, Christopher Mohr, Keegan Cooke, Dhanashri Bagal, Kevin Gaida, Tyler Holt, Charles G Knutson, Neelima Koppada, … JRussell Lipford
Nature (2019-10-30) https://doi.org/ddhm
DOI: 10.1038/s41586-019-1694-1 · PMID: 31666701

3. **Phase I study of AMG 510, a novel molecule targeting KRAS G12C mutant solid tumours**
R Govindan, MG Fakih, TJ Price, GS Falchook, J Desai, JC Kuo, JH Strickler, JC Krauss, BT Li, CS Denlinger, … DS Hong
Annals of Oncology (2019-10) https://doi.org/gpznhd
DOI: 10.1093/annonc/mdz244.008

4. **Phase 1 study evaluating the safety, tolerability, pharmacokinetics (PK), and efficacy of AMG 510, a novel small molecule KRAS^{G12C} inhibitor, in advanced solid tumors.**
Marwan Fakih, Bert O'Neil, Timothy Jay Price, Gerald Steven Falchook, Jayesh Desai, James Kuo, Ramaswamy Govindan, Erik Rasmussen, Phuong Khanh H Morrow, Jude Ngang, … David S Hong
Journal of Clinical Oncology (2019-05-20) https://doi.org/gpznhz
DOI: 10.1200/jco.2019.37.15_suppl.3003

5. **Targeting KRAS Mutant Cancers with a Covalent G12C-Specific Inhibitor**
Matthew R Janes, Jingchuan Zhang, Lian-Sheng Li, Rasmus Hansen, Ulf Peters, Xin Guo, Yuching Chen, Anjali Babbar, Sarah J Firdaus, Levan Darjania, … Yi Liu
Cell (2018-01) https://doi.org/gcxd2b
DOI: 10.1016/j.cell.2018.01.006 · PMID: 29373830

6. **5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages**
Chun-Xiao Song, Senlin Yin, Li Ma, Amanda Wheeler, Yu Chen, Yan Zhang, Bin Liu, Junjie Xiong, Weihan Zhang, Jiankun Hu, … Stephen R Quake
Cell Research (2017-08-18) https://doi.org/gbsgxg
DOI: 10.1038/cr.2017.106 · PMID: 28820176 · PMCID: PMC5630676

7. **Ultra-deep next-generation sequencing of plasma cell-free DNA in patients with advanced lung cancers: results from the Actionable Genome Consortium**
BT Li, F Janku, B Jung, C Hou, K Madwani, R Alden, P Razavi, JS Reis-Filho, R Shen, JM Isbell, … GR Oxnard
Annals of Oncology (2019-04) https://doi.org/ggdcv7
DOI: 10.1093/annonc/mdz046 · PMID: 30891595 · PMCID: PMC6503621

8. **High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants**
Pedram Razavi, Bob T Li, David N Brown, Byoungsok Jung, Earl Hubbell, Ronglai Shen, Wassim Abida, Krishna Juluru, Ino De Bruijn, Chenlu Hou, … Jorge S Reis-Filho
Nature Medicine (2019-11-25) https://doi.org/gg5rv4
DOI: 10.1038/s41591-019-0652-7 · PMID: 31768066 · PMCID: PMC7061455

9. **Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA**
Gulfem D Guler, Yuhong Ning, Chin-Jen Ku, Tierney Phillips, Erin McCarthy, Christopher K Ellison, Anna Bergamaschi, Francois Collin, Paul Lloyd, Aaron Scott, … Samuel Levy
Nature Communications (2020-10-19) https://doi.org/gjnwkd
DOI: 10.1038/s41467-020-18965-w · PMID: 33077732 · PMCID: PMC7572413

10. **Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA**
MC Liu, GR Oxnard, EA Klein, C Swanton, MV Seiden, Minetta C Liu, Geoffrey R Oxnard, Eric A Klein, David Smith, Donald Richards, … Donald A Berry
Annals of Oncology (2020-06) https://doi.org/dqx7
DOI: 10.1016/j.annonc.2020.02.011 · PMID: 33506766 · PMCID: PMC8274402

11. **Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics**
Ariana Rostami, Meghan Lambie, Caberry W Yu, Vuk Stambolic, John N Waldron, Scott V Bratman
Cell Reports (2020-06) https://doi.org/gpzs6h
DOI: 10.1016/j.celrep.2020.107830 · PMID: 32610131

12. **Serial profiling of cell-free DNA and nucleosome histone modifications in cell cultures**
Vida Ungerer, Abel J Bronkhorst, Priscilla Van den Ackerveken, Marielle Herzog, Stefan Holdenrieder
Scientific Reports (2021-05-04) https://doi.org/gkczxr
DOI: 10.1038/s41598-021-88866-5 · PMID: 33947882 · PMCID: PMC8096822

13. **Screening for Lung Cancer — 10 States, 2017**
Thomas B Richards, Ashwini Soman, Cheryll C Thomas, Brenna VanFrank, SJane Henley, MShayne Gallaway, Lisa C Richardson
MMWR. Morbidity and Mortality Weekly Report (2020-02-28) https://doi.org/gnj774
DOI: 10.15585/mmwr.mm6908a1 · PMID: 32106215 · PMCID: PMC7367073

14. **Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial**
Harry J de Koning, Carlijn M van der Aalst, Pim A de Jong, Ernst T Scholten, Kristiaan Nackaerts, Marjolein A Heuvelmans, Jan-Willem J Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, … Matthijs Oudkerk
New England Journal of Medicine (2020-02-06) https://doi.org/ggjqfm
DOI: 10.1056/nejmoa1911793 · PMID: 31995683

15. **Principles of Cancer Screening**
Paul F Pinsky
Surgical Clinics of North America (2015-10) https://doi.org/gnj773
DOI: 10.1016/j.suc.2015.05.009 · PMID: 26315516 · PMCID: PMC4555845

16. https:https://progressreport.cancer.gov

17. **Detection and characterization of lung cancer using cell-free DNA fragmentomes**
Dimitrios Mathios, Jakob Sidenius Johansen, Stephen Cristiano, Jamie E Medina, Jillian Phallen, Klaus R Larsen, Daniel C Bruhm, Noushin Niknafs, Leonardo Ferreira, Vilmos Adleff, … Victor E Velculescu
Nature Communications (2021-08-20) https://doi.org/gmkf7w
DOI: 10.1038/s41467-021-24994-w · PMID: 34417454 · PMCID: PMC8379179

18. **Validation of a Pancreatic Cancer Detection Test in New-Onset Diabetes Using Cell-Free DNA 5-Hydroxymethylation Signatures**

David Haan, Anna Bergamaschi, Gulfem D Guler, Verena Friedl, Yuhong Ning, Roman Reggiardo, Michael Kesling, Micah Collins, Bill Gibb, Adriana Pitea, … Samuel Levy
Cold Spring Harbor Laboratory (2021-12-29) https://doi.org/gpzz7z
DOI: 10.1101/2021.12.27.21268450

19. **Genome-wide cell-free DNA fragmentation in patients with cancer**
Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, … Victor E Velculescu
Nature (2019-05-29) https://doi.org/ggrw24
DOI: 10.1038/s41586-019-1272-6 · PMID: 31142840 · PMCID: PMC6774252

20. **Reassessment of Exosome Composition**
Dennis K Jeppesen, Aidan M Fenix, Jeffrey L Franklin, James N Higginbotham, Qin Zhang, Lisa J Zimmerman, Daniel C Liebler, Jie Ping, Qi Liu, Rachel Evans, … Robert J Coffey
Cell (2019-04) https://doi.org/gfxzwx
DOI: 10.1016/j.cell.2019.02.029 · PMID: 30951670 · PMCID: PMC6664447

21. **Inferring gene expression from cell-free DNA fragmentation profiles**
Mohammad Shahrokh Esfahani, Emily G Hamilton, Mahya Mehrmohamadi, Barzin Y Nabet, Stefan K Alig, Daniel A King, Chloé B Steen, Charles W Macaulay, Andre Schultz, Monica C Nesselbush, … Ash A Alizadeh
Nature Biotechnology (2022-03-31) https://doi.org/gpzz7x
DOI: 10.1038/s41587-022-01222-4 · PMID: 35361996

22. **Noninvasive blood tests for fetal development predict gestational age and preterm delivery**
Thuy TM Ngo, Mira N Moufarrej, Marie-Louise H Rasmussen, Joan Camunas-Soler, Wenying Pan, Jennifer Okamoto, Norma F Neff, Keli Liu, Ronald J Wong, Katheryne Downes, … Stephen R Quake
Science (2018-06-08) https://doi.org/gdp276
DOI: 10.1126/science.aar3819 · PMID: 29880692 · PMCID: PMC7734383

23. **Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing**
Shusuke Toden, Jiali Zhuang, Alexander D Acosta, Amy P Karns, Neeraj S Salathia, James B Brewer, Donna M Wilcock, Jonathan Aballi, Mike Nerenberg, Stephen R Quake, Arkaitz Ibarra
Science Advances (2020-12-11) https://doi.org/ghqs9s
DOI: 10.1126/sciadv.abb1654 · PMID: 33298436 · PMCID: PMC7821903

24. **Presymptomatic Increase of an Extracellular RNA in Blood Plasma Associates with the Development of Alzheimer's Disease**
Zhangming Yan, Zixu Zhou, Qiuyang Wu, Zhen Bouman Chen, Edward H Koo, Sheng Zhong
Current Biology (2020-05) https://doi.org/gpzs6j
DOI: 10.1016/j.cub.2020.02.084 · PMID: 32220323

25. **A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection**
Matthew H Larson, Wenying Pan, Hyunsung John Kim, Ruth E Mauntz, Sarah M Stuart, Monica Pimentel, Yiqi Zhou, Per Knudsgaard, Vasiliki Demas, Alexander M Aravanis, Arash Jamshidi
Nature Communications (2021-04-21) https://doi.org/gj3w25
DOI: 10.1038/s41467-021-22444-1 · PMID: 33883548 · PMCID: PMC8060291

26. **Epigenomic reprogramming of repetitive noncoding RNAs and IFN-stimulated genes by mutant KRAS**

Roman E Reggiardo, Sreelakshmi Velandi Maroli, Haley Halasz, Mehmet Ozen, David Carrillo, Erin LaMontagne, Lila Whitehead, Eejung Kim, Shivani Malik, Jason Fernandes, … Daniel H Kim
Cold Spring Harbor Laboratory (2020-11-04) https://doi.org/gpzs6m
DOI: 10.1101/2020.11.04.367771

27. **Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming**
Daniel H Kim, Georgi K Marinov, Shirley Pepke, Zakary S Singer, Peng He, Brian Williams, Gary P Schroth, Michael B Elowitz, Barbara J Wold
Cell Stem Cell (2015-01) https://doi.org/f8mb98
DOI: 10.1016/j.stem.2014.11.005 · PMID: 25575081 · PMCID: PMC4291542

28. **Extracellular RNA signatures of mutant KRAS(G12C) lung adenocarcinoma cells**
Reem Khojah, Roman E Reggiardo, Mehmet Ozen, Sreelakshmi Velandi Maroli, David Carrillo, Utkan Demirci, Daniel H Kim
Cold Spring Harbor Laboratory (2022-02-24) https://doi.org/gpzz72
DOI: 10.1101/2022.02.23.481574

29. **A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes**
Charlotte Soneson, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D Robinson, Shobbir Hussain
Nature Communications (2019-07-31) https://doi.org/gpzs6k
DOI: 10.1038/s41467-019-11272-z · PMID: 31366910 · PMCID: PMC6668388

30. **exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids**
Oscar D Murillo, William Thistlethwaite, Joel Rozowsky, Sai Lakshmi Subramanian, Rocco Lucero, Neethu Shah, Andrew R Jackson, Srimeenakshi Srinivasan, Allen Chung, Clara D Laurent, … Aleksandar Milosavljevic
Cell (2019-04) https://doi.org/gfxzww
DOI: 10.1016/j.cell.2019.02.018 · PMID: 30951672 · PMCID: PMC6616370

31. **The Human Cell Atlas**
Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, …
eLife (2017-12-05) https://doi.org/gcnzcv
DOI: 10.7554/elife.27041 · PMID: 29206104 · PMCID: PMC5762154

32. **LncRNA Biomarkers of Inflammation and Cancer**
Roman E Reggiardo, Sreelakshmi Velandi Maroli, Daniel H Kim
Long Noncoding RNA (2022) https://doi.org/gpz9hw
DOI: 10.1007/978-3-030-92034-0_7 · PMID: 35220568

33. **Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells**
Eric W Martin, Jana Krietsch, Roman E Reggiardo, Rebekah Sousae, Daniel H Kim, ECamilla Forsberg
Epigenetics & Chromatin (2021-01-06) https://doi.org/gh2gxz
DOI: 10.1186/s13072-020-00377-1 · PMID: 33407811 · PMCID: PMC7789351

34. **Landscape of transcription in human cells**
Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, … Thomas R Gingeras
Nature (2012-09) https://doi.org/f36xtm

DOI: 10.1038/nature11233 · PMID: 22955620 · PMCID: PMC3684276

35. **Initial sequencing and analysis of the human genome**
, Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, ...
Nature (2001-02-15) https://doi.org/bfpgjh
DOI: 10.1038/35057062 · PMID: 11237011

36. **Transposable elements reveal a stem cell-specific class of long noncoding RNAs**
David Kelley, John Rinn
Genome Biology (2012) https://doi.org/ggh7d6
DOI: 10.1186/gb-2012-13-11-r107 · PMID: 23181609 · PMCID: PMC3580499

37. **Transposable elements in cancer**
Kathleen H Burns
Nature Reviews Cancer (2017-06-09) https://doi.org/gbj4fc
DOI: 10.1038/nrc.2017.35 · PMID: 28642606

38. **The Role of Non-coding RNAs in Oncology**
Frank J Slack, Arul M Chinnaiyan
Cell (2019-11) https://doi.org/ggm2wq
DOI: 10.1016/j.cell.2019.10.017 · PMID: 31730848 · PMCID: PMC7347159

39. **Repression of the miR-143/145 cluster by oncogenic Ras initiates a tumor-promoting feed-forward pathway**
Oliver A Kent, Raghu R Chivukula, Michael Mullendore, Erik A Wentzel, Georg Feldmann, Kwang H Lee, Shu Liu, Steven D Leach, Anirban Maitra, Joshua T Mendell
Genes & Development (2010-12-15) https://doi.org/bksrsx
DOI: 10.1101/gad.1950610 · PMID: 21159816 · PMCID: PMC3003192

40. **Comprehensive molecular profiling of lung adenocarcinoma** *Nature* (2014-07-09) https://doi.org/f6h32b
DOI: 10.1038/nature13385 · PMID: 25079552 · PMCID: PMC4231481

41. **Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras**
Erica L Jackson, Nicholas Willis, Kim Mercer, Roderick T Bronson, Denise Crowley, Raymond Montoya, Tyler Jacks, David A Tuveson
Genes & Development (2001-12-15) https://doi.org/df52xx
DOI: 10.1101/gad.943001 · PMID: 11751630 · PMCID: PMC312845

42. **RAS Proteins and Their Regulators in Human Disease**
Dhirendra K Simanshu, Dwight V Nissley, Frank McCormick
Cell (2017-06) https://doi.org/gbmdsr
DOI: 10.1016/j.cell.2017.06.009 · PMID: 28666118 · PMCID: PMC5555610

43. **Epithelial-to-Mesenchymal Transition is a Cause of Both Intrinsic and Acquired Resistance to KRAS G12C Inhibitor in KRAS G12C–Mutant Non–Small Cell Lung Cancer**
Yuta Adachi, Kentaro Ito, Yuko Hayashi, Ryo Kimura, Tuan Zea Tan, Rui Yamaguchi, Hiromichi Ebi
Clinical Cancer Research (2020-09-08) https://doi.org/gp54ng
DOI: 10.1158/1078-0432.ccr-20-2077 · PMID: 32900796

44. **Mechanisms of Resistance to KRASG12C-Targeted Therapy**
Neal S Akhave, Amadeo B Biter, David S Hong
Cancer Discovery (2021-04-05) https://doi.org/gp54nh

DOI: 10.1158/2159-8290.cd-20-1616 · PMID: 33820777 · PMCID: PMC8178176

45. **Rapid non-uniform adaptation to conformation-specific KRAS(G12C) inhibition**
Jenny Y Xue, Yulei Zhao, Jordan Aronowitz, Trang T Mai, Alberto Vides, Besnik Qeriqi, Dongsung Kim, Chuanchuan Li, Elisa de Stanchina, Linas Mazutis, … Piro Lito
Nature (2020-01-08) https://doi.org/gg6xb7
DOI: 10.1038/s41586-019-1884-x · PMID: 31915379 · PMCID: PMC7308074

46. **Immortalization and transformation of primary human airway epithelial cells by gene transfer**
Ante S Lundberg, Scott H Randell, Sheila A Stewart, Brian Elenbaas, Kimberly A Hartwell, Mary W Brooks, Mark D Fleming, John C Olsen, Scott W Miller, Robert A Weinberg, William C Hahn
Oncogene (2002-06-27) https://doi.org/bsdv7w
DOI: 10.1038/sj.onc.1205550 · PMID: 12085236

47. **Immortalization of Human Bronchial Epithelial Cells in the Absence of Viral Oncoproteins**
Ruben D Ramirez, Shelley Sheridan, Luc Girard, Mitsuo Sato, Young Kim, Jon Pollack, Michael Peyton, Ying Zou, Jonathan M Kurie, JMichael DiMaio, … John D Minna
Cancer Research (2004-12-15) https://doi.org/c8gfbf
DOI: 10.1158/0008-5472.can-04-3703 · PMID: 15604268

48. **Salmon provides fast and bias-aware quantification of transcript expression**
Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, Carl Kingsford
Nature Methods (2017-03-06) https://doi.org/gcw9f5
DOI: 10.1038/nmeth.4197 · PMID: 28263959 · PMCID: PMC5600148

49. **nf-core/atacseq: nf-core/atacseq v1.2.1 - Iron Centipede**
Harshil Patel, Phil Ewels, Alexander Peltzer, Drew Behrens, Gisela Gabernet, Mingda Jin, Mashehu, Maxime Garcia
Zenodo (2020-07-29) https://doi.org/gp55dh
DOI: 10.5281/zenodo.3965985

50. **GENCODE 2021**
Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, … Paul Flicek
Nucleic Acids Research (2020-12-03) https://doi.org/gp54nf
DOI: 10.1093/nar/gkaa1087 · PMID: 33270111 · PMCID: PMC7778937

51. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**
Michael I Love, Wolfgang Huber, Simon Anders
Genome Biology (2014-12) https://doi.org/gd3zvn
DOI: 10.1186/s13059-014-0550-8 · PMID: 25516281 · PMCID: PMC4302049

52. **Tximeta: Reference sequence checksums for provenance identification in RNA-seq**
Michael I Love, Charlotte Soneson, Peter F Hickey, Lisa K Johnson, NTessa Pierce, Lori Shepherd, Martin Morgan, Rob Patro
PLOS Computational Biology (2020-02-25) https://doi.org/ggszwj
DOI: 10.1371/journal.pcbi.1007664 · PMID: 32097405 · PMCID: PMC7059966

53. https:https://combine-lab.github.io/alevin-tutorial/2020/alevin-velocity

54. **The Human Genome Browser at UCSC**
WJames Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler
Genome Research (2002-05-16) https://doi.org/fpf5rm
DOI: 10.1101/gr.229102 · PMID: 12045153 · PMCID: PMC186604

55. **Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser**
BJ Raney, TR Dreszer, GP Barber, H Clawson, PA Fujita, T Wang, N Nguyen, B Paten, AS Zweig, D Karolchik, WJ Kent
Bioinformatics (2013-11-13) https://doi.org/gb5g3r
DOI: 10.1093/bioinformatics/btt637 · PMID: 24227676 · PMCID: PMC3967101

56. **Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo**
Thomas A Carter, Manvendra Singh, Gabrijela Dumbović, Jason D Chobirko, John L Rinn, Cédric Feschotte
eLife (2022-02-18) https://doi.org/gqb4tf
DOI: 10.7554/elife.76257 · PMID: 35179489 · PMCID: PMC8912925

57. **The Elements of Statistical Learning**
Trevor Hastie, Robert Tibshirani, Jerome Friedman
Springer Series in Statistics (2009) https://doi.org/cd7nhz
DOI: 10.1007/978-0-387-84858-7

58. **Statistical Learning from a Regression Perspective**
Richard A Berk
Springer Texts in Statistics (2016) https://doi.org/ghngt7
DOI: 10.1007/978-3-319-44048-4

59. **Non-invasive characterization of human bone marrow stimulation and reconstitution by cell-free messenger RNA sequencing**
Arkaitz Ibarra, Jiali Zhuang, Yue Zhao, Neeraj S Salathia, Vera Huang, Alexander D Acosta, Jonathan Aballi, Shusuke Toden, Amy P Karns, Intan Purnajo, … Michael Nerenberg
Nature Communications (2020-01-21) https://doi.org/gpg897
DOI: 10.1038/s41467-019-14253-4 · PMID: 31964864 · PMCID: PMC6972916