System-wide in vivo, multi-omics and computational approaches to identify mechanisms behind tumor-immune coevolution and RNA secretion

by
Bahar Zirak

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Alexander Marson
7F25CDCE383C4A8...

Alexander Marson
Chair

DocuSigned by:
Hani Goodarzi
[illegible]6487...

Hani Goodarzi

DocuSigned by:
Balyn Zaro
[illegible]D4D3...

Balyn Zaro

DocuSigned by:
Karin Pelka
7A0E406D780B453...

Karin Pelka

Committee Members

This dissertation is dedicated to my parents, who sacrificed everything, so I can have everything I ever needed and more.

And to my grandma, I think about you every day.

ACKNOWLEDMENTS

Firstly, I would like to acknowledge my thesis advisor, Dr. Hani Goodarzi. Your curiosity, dedication, and commitment to science and your lab are truly inspiring. When I first joined the lab, I was eager to explore new research areas and acquire new skills, but I was also nervous about entering an unfamiliar field. Your patience, encouragement, and support over the years created a safe environment for me to experiment, make mistakes, learn, and grow. The way you inspire and challenge your trainees to reach their full potential while providing unwavering support is incredibly motivating.

I'd also like to acknowledge my thesis committee members, Alexander Marson, Karin Pelka, and Balyn Zaro, for their valuable insights, support, and collaborations throughout this journey.

I would like to acknowledge Dr. Michael Rosenblum. I joined his lab as a research assistant after graduating from college, and it was the first place I felt a sense of belonging after immigrating to the US. I fell in love with research in the Rosenblum lab, and beyond that, I found lifelong friends and a warm community where I felt motivated and safe to pursue science. Thank you, Mike, for your mentorship and support throughout those transformative years.

I would like to thank my best friends, Lokesh (Lucky) Kalekar and Priscila Munoz Sandoval. Lucky, you are a brilliant and passionate scientist and mentor from whom I've had the opportunity to learn so much. Above all, you are the most supportive and encouraging friend, even from thousands of miles away.

Priscila, I am so fortunate that our paths have crossed multiple times: once at UC Berkeley, once in the Rosenblum lab, and now, six years later, we are finishing our PhD journey together at UCSF. You are an incredible scientist, mentor, and a wonderful friend. I'd also like to thank my lab mates and classmates Brian Woo and Nour Abdulhay. Going through grad school together and being in the same lab with you made this experience much more enjoyable and memorable. I'm so lucky to have met you both here in grad school and can't wait to see what the future holds for all of us.

Also, thank you to everyone in the Goodarzi lab for your mentorship and support over the years. Having a community like you made this journey truly enjoyable.

To Khashayar Zadeh, you have been a constant source of encouragement and support throughout the highs and lows of grad school, and I couldn't have asked for a better partner to have shared this part of my life and more with.

To Sharmin Zirak, thank you for being my sister, my cheerleader, and my best friend since the beginning. You have supported, encouraged and inspired me in ways I can't ever fully express.

Lastly, to my parents, Siavash Zirak and Maliheh Salari, I am deeply honored to be your daughter and grateful for the constant love and support you have given me throughout my life. Everything you have done has been for me and my sister. This dissertation is as much yours as it is mine.

CONTRIBUTIONS

The work presented in this dissertation was performed under the supervision of Dr. Hani Goodarzi.

Chapter 1: System-wide in vivo and multi omics approaches to identify molecular mechanisms in tumor immune coevolution, contains unpublished materials, following authors contributed to the study: Bahar Zirak, Camille Derderian, Abolfazl Arab, Vishvak Subramanyam, Regan Volk, Tanvi Joshi, Kristle Garcia, Jing-Yi Chung, Trey Charbonneau, Balyn Zaro, Hani Goodarzi

Chapter 2: Revealing the Grammar of Small RNA Secretion Using Interpretable Machine Learning, contains material from a published work as seen in: https://doi.org/10.1016/j.xgen.2024.100522

System-wide in vivo, multi-omics and computational approaches to identify mechanisms behind tumor-immune coevolution and RNA secretion

Bahar Zirak

ABSTRACT

Tumor progression is the major cause of death in cancer patients. Due to their higher chromosomal instability and other genomic alterations, tumors evolve rapidly in response to therapeutic interventions and other external pressures. The immune system is our first line of defense against cancer and its interaction with cancer cells can both constrain and promote tumor growth and metastasis. A heterogeneous tumor consists of subclones with different characteristics. During tumor progression the anti-tumor immune activity removes subclones that express highly immunogenic antigens and leaves behind cancer cells with high immune escape or immunosuppressive properties. This process is called tumor immunoediting [1].

Studying tumor progression requires reliable *in vivo* models that effectively capture the intricacies and complexities of this process. During the 1970s, Isaiah Fidler demonstrated that repeated passaging of cancer cells in mice can be used to emulate metastatic progression [2]. This in vivo selection model has been used by many different research groups (including us) to model tumor progression in a number of cancer models. Our group has utilized these in vivo selection models to study cell autonomous mechanisms of tumor progression. More recently, however, we have come to realize that by leveraging these *in vivo*-selection models we can focus on studying non-cell autonomous mechanisms. Building on this notion, here, we propose a generalization of *in vivo* selection that models the role of the immune system in shaping tumor evolution. Our “immune selection” model takes advantage of a panel of genetic mouse models with various degrees of immunocompetency to serve as hosts for established syngeneic tumor cell lines. We utilized these ‘immuno-selected’ derivatives, in conjunction with cutting-edge tools in genetic engineering and single-cell genomics, to study the tumor-immune co-evolution. We discovered that the interferon response pathway lies at the heart of tumor immune evasion. Additionally, we have uncovered novel molecular pathways responsible for conferring resistance to both antitumor immunity and immunotherapies. Targeting these pathways holds significant therapeutic potential, particularly when used in conjunction with immune checkpoint blockades (ICBs) and other forms of immunotherapy.

The second part of this thesis is focused on utilizing machine learning and computational tools to identify important molecular mechanisms in small RNA secretion. We developed ExoGRU, a deep-learning model for predicting secretion probabilities of small RNAs based on their primary sequence. We used ExoGRU to (i) identify mutations that abrogate the secretion of known cell-free small RNAs, and (ii) predict high confidence sets of synthetic sequences that are secreted or retained. We also used independent experimental approaches to validate our model’s prediction abilities. We discovered that the molecular signature needed for small RNA secretion lies in its primary sequence. Furthermore, we identified both previously known and novel RNA binding proteins (RBPs) crucial for facilitating this secretion.

In both projects discussed, we demonstrate the effectiveness of in vivo, high-throughput, multi-omics and computational tools in uncovering novel mechanisms, particularly in the evolution of tumor immunity and RNA secretion, areas traditionally challenging to explore with conventional methods.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: SYSTEM-WIDE IN VIVO AND MULTI OMICS APPROACHES TO IDENTIFY MOLECULAR MECHANISMS IN TUMOR IMMUNE COEVOLUTION

1.1. Introduction:

Research into cancer immune evasion and immune editing has been a focus of numerous groups since Robert Schreiber first introduced the concept. Studies have predominantly centered on the cancer immune microenvironment and tumor immune escape mechanisms, leading to breakthroughs like checkpoint inhibitors such as CTLA-4 and PD-1/PD-L1. While these treatments have shown remarkable success in certain cancers like melanoma, lymphoma, and myeloma, they often prove ineffective for all patients, particularly in the case of solid cancers.

With the advancement of high-throughput sequencing and large-scale genomics, we have made significant strides in uncovering the diverse pathways that cancers exploit to evade anti-tumor immunity, including checkpoint blockade and other immunotherapies. These studies have employed genome-wide or sub-genome in vivo CRISPR screens, utilizing various mouse models or different immune checkpoint blockade treatment conditions [3]. These investigations have shed light on a multitude of novel pathways that different cancer types utilize to resist or escape immune checkpoint blockade, including IFN response pathways and antigen processing, among others. While these screens have provided valuable information on cancer escape during tumor progression or treatment, they don’t capture the natural state of tumor cells and their evolutionary trajectory, as they induce perturbations. In this study we took advantage of an in vivo selection model in order to emulate the natural evolutionary trajectory of cancer cells.

Inspired by the in vivo selection model for cancer metastasis studies first introduced by Isaiah Fidler [2] and used in our own studies, we developed an immuno-selection model. This innovative approach adapts the in vivo selection method to specifically study non-cell-autonomous mechanisms, focusing on tumor-immune interactions and co-evolution. By repeatedly passaging mouse syngeneic cancer cell lines in different genetic mouse models with varying levels of immunocompetency, we aimed to uncover tumor evolutionary pathways under distinct immune conditions. This approach allowed us to better understand the mechanisms of immune-mediated anti-tumor resistance using various large-scale transcriptomics and genomics techniques.

We believe that our approach, beginning with learning from the evolutionary process without perturbing the system, coupled with the use of in vivo screens to validate our findings, has not only allowed us to identify familiar as well as novel molecular pathways that can only be captured by tracking evolutionary changes occurring in cancer cells under constant immune pressure.

1.2. Immunoselection: A Comprehensive Workflow For Investigating Cancer Immune Coevolution Through In Vivo And Computational Approaches

To gain deeper insights into the diverse evolutionary routes shaped by tumor-immune interactions, we devised an in vivo mouse syngeneic model, specifically focusing on the 4T1 TNBC and B16-F10 Melanoma cell lines. TNBC and melanoma exhibit markedly different responses to immunotherapy. Melanoma demonstrates a response rate of 30-50% to immune checkpoint blockade (ICB)[4] whereas TNBC's response rate is notably lower, falling below 20% [4].

For our in vivo mouse workflow, we separately injected 4T1 and B16 cell lines to mice with diverse immunocompetency backgrounds. The 4T1 cancer cells were introduced to a mouse panel comprising WT(Balb/cJ), NSG, RAG, and NU/J mice, while the B16-F10 cells were administered to a panel consisting of B6, Rag, and NSG mice (**Figure 1.1A**). To assess the impact of exposure to varying immune pressures over time, we conducted repeated in vivo passages of the resultant tumors in their respective backgrounds. This process was carried out for a total of four passages, allowing us to highlight the differences in tumor evolution within distinct microenvironments. To closely examine alterations in gene expression within tumor cells developed in distinct in vivo settings, we conducted bulk RNA sequencing on our comprehensive set of tumor cell lines, encompassing those derived from 4T1 and B16 parental cells, as well as cells from generation 1 (G1), G2, G3, and G4. **Figure 1.1B** summarizes the RNAseq data in a heatmap that compares gene expression between generation 4 of 4T1 and B16 tumors cultivated in diverse mouse strains. As observed the primary variance arises from the inherent characteristics of the cell lines themselves. Additionally, gene expression differences exist among G4 tumors derived from different mouse strains within each cell line.

We also looked more closely into G1-G4 tumors of each of these cell lines. Heatmaps in **Figure 1.2A** and **1.2D** highlight gene expression profiles across multiple generations and mouse strains in 4T1 and B16 tumors respectively. Both heatmaps underscore the divergent tumor profiles resulting from cultivation in different strains. The heatmaps illustrate genes with significant differences between tumors grown in various strains, as determined by a one-way ANOVA test ($p\text{-value} < 0.05$).

To investigate the evolved non-cell-autonomous pathways under different immune pressures we performed differential gene expression analysis using the Deseq2 package[5]. Volcano plots on **figure 1.2B** highlight the genes upregulated in generation 4 of Balb-selected vs Rag, NU/J and NSG-selected tumors respectively. Highlighted are various interferon signature genes (Isg15, Irf1, Ifi202b, Ift80, Ifitm3) antigen presenting pathway (Tap2, B2m, H2-D1), chemokine (Cxcl1, Cxcl5), complement system (C3), immunoproteasomes (Psmb8), genes involved in TGFb signaling (Smad3, Itgb5), and autophagy related genes (Bnip3,Ulk1). The gene expression analysis revealed a greater similarity in upregulated genes between Balb versus NU/J-selected and Balb versus Rag-selected tumors. This finding is consistent with the similarity in immune composition between Rag and NU/J-mice, both lacking T cell immunity. Notably, two genes, C3 and Bnip3, were consistently upregulated in Balb compared to all three conditions. Various genes under complement family have previously shown to be involved in tumor immunosuppression.[6–8] Furthermore, Bcl2-interacting protein 3 (Bnip3) promotes breast cancer tumor growth under hypoxic conditions through autophagy-dependent pathways [9]. Therefore, these genes could be contributing to progression of these tumors under innate and adaptive immune pressure. The majority of the upregulated pathways in Balb-selected mice are directly or indirectly implicated in immune response regulation, which underscores the success of our immune selection model. We further performed time course analysis using DESeq2 package to look at the expression of genes over time from G0 to G4, we coupled these analysis with iPAGE analysis[10] to highlight the pathways differentially expressed overtime in WT-selected tumor as opposed to NSG, Rag or NU/J-selected tumors. The results of these analyses are visually represented in the iPAGE heatmaps found in **Figure 1.2C**. Gene set enrichment analysis of this dataset revealed a significant upregulation of Interferon-mediated pathways in 4T1 tumors cultivated in wild-type (WT) mice in contrast to tumors that developed in NSG, Rag, and NU/J mice. The main producers of type I interferon are innate immune cells, particularly dendritic cells (DCs), and also T cells[11]. Additionally, IFN-gamma is predominantly produced by T cells [12]. This emphasizes the importance of both innate and adaptive immunity in shaping tumor immunity coevolution. It also underscores the effectiveness of our workflow in capturing differences arising from exposure to the different innate and adaptive immune responses.

We conducted a comparable analysis for the B16 dataset, as illustrated in the volcano plots presented in **Figure 1.2E**. Notably, numerous metabolic pathways, including ESD, Ugcg, Agtpbp1, and Tnrc6a, were found to be upregulated in WT-selected tumors, alongside hypoxia-related genes such as Tet1. Furthermore, immune and interferon-mediated genes such as Il7r, Ifnar, and Cd27 were observed to be upregulated. Metabolic pathways have been identified as crucial in cancer immune evasion. For instance, the upregulation of the Ugcg gene in WT-selected B16 tumors, observed consistently in both comparisons, has been shown to enhance glycolysis and oxidative phosphorylation across various cancer types [13,14]. This increase in metabolic pathways has been associated with metabolic competition, thereby limiting resources available for T cells and other immune cells [15]. Furthermore, the products of glycolysis, such as lactic acid, create a more acidic environment, which dampens the effect of anti-tumor immunity[16].

Figure 1.2F also shows the differentially expressed pathways over generations in WT-selected vs NSG and Rag-selected B16 tumors. Similar trends were observed, where the response to the type I interferon pathway was accentuated in B6 versus Rag tumors. However, in the comparison between B6 and NSG tumors, the IFN pathways were not emphasized; instead, metabolic pathways such as carbohydrate catabolic processes were more prominent.

Over generations, both cell lines have exhibited distinct immune and metabolic pathway expressions. Notably, the interferon response pathways emerge as a central mechanism upregulated in WT-selected tumors compared to immune compromised-selected tumors in both cell lines, which highlights the co-evolution of immune and tumor cells.

1.3. Immune Evasion Markers Predict Response To Immunotherapy

Our investigation into tumor evolution under varying immune pressures has yielded extensive amounts of data for exploration. Yet, we sought to bridge the gap between our findings and clinical relevance. To achieve this, we analyzed RNAseq data from the G4 of 4T1 model, focusing on genes that were differentially expressed in tumors grown in WT versus those in immunocompromised mice. We posited that tumors thriving in WT mice would activate or suppress specific pathways to adapt to an immune-enriched environment. Overlaying these genes onto pretreatment RNAseq data from the ISPY2 trial breast cancer patients allowed us to stratify patients into high and low immune evasion scores (**Figure 1.3A**). Subsequently, we investigated the survival outcomes and immunotherapy responses of these patient categories. Our findings revealed that patients with higher immune evasion scores exhibited lower responses to anti-PD1 therapy (**Figure 1.3B**) and an increased likelihood of metastatic relapse (**Figure 1.3C).** Conversely, patients with lower immune evasion scores demonstrated significantly higher rates of distant metastasis-free survival (**Figure 1.3D**) and overall survival (**Figure 1.3E**).

1.4. Single Cell Analysis Of Immune-Challenged Tumors, Discover Immune-Regulated Tumor Subcluster

We were intrigued to explore how tumors developed under varying immune conditions would interact with the immune system and employ different pathways to evade anti-tumor immunity. To investigate this, we inoculated WT mice with 4th generation cell lines derived from WT, Rag -/- and NSG selected tumors. After allowing two weeks for growth, we conducted single-cell RNA sequencing (scRNAseq) on both the tumors and their microenvironment. Employing scRNAseq not only provides insights into the intra-tumor heterogeneity of cancer cells but also enhances our comprehension of the tumor microenvironment. **Figure 1.4A** illustrates the workflow of our immune challenge pipeline, where fourth-generation cancer cells from 4T1 backgrounds are injected into WT mice. The resulting tumors are denoted with an 'IC' (immune-challenged) suffix added immediately after the strain of the mice in which they were previously developed from. The UMAP visualization in **Figure 1.4B** depicts overlapping Balb-IC, NSG-IC and Rag-IC 4T1 tumors consisting of 27498 cells. We aimed to distinguish tumor cell clusters by assessing Epcam expression to identify epithelial cells. This was complemented by high copy number variation, a characteristic trait of cancer cells. This approach helped differentiate cancer cells from normal epithelial cells. (**Figure 1.4C**). We then directed our analysis towards discerning the differences between tumor subclusters across various IC-tumors. Pathway analysis was conducted on Balb-IC, NSG-IC, and RAG-IC tumor clusters. Notably, immune-related pathways such as the Interferon Alpha and Gamma response pathways emerged as top upregulated pathways in Balb-IC tumors. As previously mentioned, the primary source of interferon alpha expression is innate immune cells, whereas T cells predominantly produce interferon gamma. It was intriguing yet logical to observe that the interferon alpha pathway was upregulated in tumors selected in wild-type (WT) mice compared to those in NSG mice lacking adaptive and most innate immunity. Conversely, interferon gamma was elevated in tumors selected in WT mice compared to Rag -/- mice, where the key difference lies in the absence of T and B cells (**Figure 1.4D**). This observation suggests that tumors experiencing sustained pressure from the innate and adaptive immune system across generations demonstrate heightened expression of these pathways in response.

To explore intratumor heterogeneity, we further sub clustered our cancer cells and examined the distribution of these subclusters across our samples. As depicted in **figure 1.4E** we observed differences in the proportion of each cluster between the various IC conditions. The heatmap in **figure 1.4E** illustrates the -Log$_{10}$P value representing the subclusters enrichments in Balb-IC vs NSG-IC or Rag-IC, calculated using Fisher's exact test. Next, we looked further into top variable genes in each cluster to better understand the gene expression differences between these subclusters. As depicted in the heatmap of **figure 1.4F**, various cell populations are grouped based on their top 5 genes. Notably, cluster 6 exhibited expression of interferon and immune stimulated genes, including Stat1, Ly6e, Gbp3, Oasl2, H2-T23, among others. Upon examining the distribution of this cluster in **figure 1.4E**, a significant reduction in its size was observed in NSG-IC tumors. This finding suggests that the cluster did not expand in the absence of immune pressure. On the contrary, this cluster exhibited expansion in Rag-IC and Balb-IC tumors, which were consistently subjected to heightened innate and adaptive immune pressure.

We conducted similar analysis on the B16 dataset by injecting fourth-generation tumors from WT, NSG, and Rag-selected B16 tumors into WT B6 mice, followed by scRNAseq analysis of resulting tumors. **Figure 1.5A** displays the UMAP, with an overlay of the B6-IC, Rag-IC, and NSG-IC groups, totaling 19792 cells. Additionally, melanocytes were identified using the Mitf gene, along with CNV analysis to distinguish melanoma cells (**Figure 1.5B**). As previously described, we performed pathway enrichment analysis on tumor clusters from different IC groups. Our pathway analysis revealed similar enrichment of the IFN-gamma response pathway in B6-IC tumors in comparison to NSG-IC and Rag-IC tumors (**Figure 1.5C**). To explore the heterogeneity of melanoma cells more comprehensively, we subdivided them into distinct subclusters as shown in **figure 1.5D**. The barplot illustrates the distribution of these clusters across the IC samples, revealing that all subclusters except cluster 9 are present across all samples even though with varying enrichments. We looked into the top highly 5 variable genes within each cluster and noted cluster 6 to be enriched in immune related and interferon signature genes such as Iigp1, Ifit3, Igtp, Psmb8 and Gbp2 (**Figure 1.5E**), similar to what we saw in 4T1 dataset. Notably, this cluster exhibits significant enrichment in the B6-IC compared to NSG-IC and Rag-IC samples (**Figure 1.5D**). Moreover, we observed elevated expression of these genes across all other subclusters in B6-IC compared to NSG-IC and Rag-IC (**Figure 1.5E**), with pronounced expression levels evident throughout the overall cancer cell cluster within this group (**Figure 1.5F**). This suggests that B6-IC tumors may be fostering the expansion of this cluster and markers, potentially selecting these cells under immune pressure as an immune escape mechanism.

1.5. Unveiling Tumor Resistance Mechanisms To Anti-Tumor Immunity And Immune Checkpoint Blockades Through In Vivo CRISPRi Screening

To delve into the functional mechanisms underlying tumors developed in immune-enriched WT mice, we conducted an in vivo CRISPRi screen targeting molecular pathways that had shown significant upregulation in these tumors. We assembled a library comprising approximately 200 genes encompassing those consistently upregulated across multiple tumor generations in Balb or B6 mice for the 4T1 and B16 libraries respectively. We also included genes that exhibited upregulation in the tumor cluster of Balb-IC or B6-IC tumors compared to the other IC conditions in our single cell data sets, with minimum LogFC >0.5 and P value<0.05. To further elucidate the role of these genes in tumor progression under varying immune pressures, we introduced the library into WT, NSG and RAG mice and subjected sgRNAs extracted from the resulting tumor to sequencing (**Figure 1.6A**). **Figures 1.6B-C** demonstrate the enrichment and depletion of 4T1 or B16 guides, respectively, in WT, RAG and NSG mice backgrounds. Upon initial examination, we noticed that a greater proportion of genes previously upregulated in the WT condition for B16 mice show decreased tumor fitness upon depletion in the WT mice compared to RAG or NSG. We didn't observe a similar pattern with the 4T1 screened genes. The volcano plot indicates a more balanced distribution of these depleted genes across WT, NSG, or RAG mice for 4T1 sgRNAs. This difference could be attributed to the potentially greater immunogenicity of B16 melanoma compared to 4T1 TNBC tumors, where most upregulated genes have arisen from the evolution of these tumors under a stronger immune response.

We then focused our attention to genes where their depletion resulted in a decrease in cellular fitness under anti-tumor immune pressure. We confirmed the presence of previously identified pathways involved in immune evasion, including Stat1 [17–19](and its target Plscr1 [20]), Sp140 [21], and Ptpn18 [22]. Additionally, we found genes like Pnpt1 and Rorc that are recognized for their involvement in various metabolic pathways, including Mitochondrial RNA processing and cholesterol biosynthesis [23,24], yet their role in immune regulation is not well understood. Moreover, our analysis has uncovered that Rab25, Tuba4a, and Cdh1, all associated with integrin trafficking and membrane organization, are part of immune sensitizing pathways upon depletion. Despite their previous associations with tumor progression in breast cancer or other cancer types, their contribution to immune evasion remains inadequately understood.[25–28]

In the B16-Ci screen, we noted depletion of sgRNAs linked to genes related to the antigen presentation pathway, such as Tap1, as well as components of the MHC I complex like H2-D1 and H2-T23, under immune pressure. Drawing from existing literature, we understand that tumor-induced upregulation of MHC class I molecules can dampen natural killer cell responses [29] , and IFN-induced MHC expression, especially H2-T23 can hinder CD8 T cell activity through the NKG2A/CD94 receptor pathway[3]. Therefore, its upregulation in B16 tumors is used as a mechanism of T cell suppression. We also observed depletion of several ATPase (Atp6v1f, Atp6v0b, Atp6v0c) or ATP synthase (Atp5g2) genes results in sensitization of the cancer cells to the immune system. These pathways are recognized components of proton transmembrane transport [30,31] and metabolic pathways such as oxidative phosphorylation [32]. Similar to the findings in 4T1 tumors, we also identified the depletion of Stat1 as a mechanism associated with immune sensitization in this context.

Overall, our findings suggest that both cell lines regulate immune responses through a combination of shared and distinct mechanisms. Specifically, while 4T1 cells exhibit regulated pathways involved in IFN regulation, metabolic pathways, and membrane trafficking, B16 cells have evolved pathways related to IFN regulation, antigen presentation, and oxidative phosphorylation. However, the depletion of Stat1 emerged as a central and shared mechanism between the two cell lines.

We further aimed to investigate how the fitness of the library transfected 4T1-Ci and B16-Ci cells changes under additional immune pressure from immune checkpoint blockade (ICB). We inoculated WT mice with the corresponding cell lines and administered anti-PD-L1, anti-CTLA4, a combination of both, or their corresponding IgG control antibodies (**Figure 1.6D**). Subsequently, we performed sequencing on the extracted guides from the resulting tumors. Significantly larger differences in terms of depleted or enriched genes emerged from tumors grown in the presence of ICB compared to tumors grown in Igg Ctrl treated WT mice, as opposed to the same comparison between tumors grown in WT versus NSG mice. We found a limited number of overlapping genes between the comparisons of WT versus NSG mice and ICB-treated versus untreated WT mice. However, the majority of depleted genes were distinct between the two comparisons, with a significant portion unique to the ICB treated group. (**Figure 1.6E-F**)

Analyzing data from B16 tumors treated with ICB revealed the recurrence of previously studied IFN signatures and immune-regulated genes such as Adar, Ifit1, Ifi27, H2-Aa, and

Fkbp6. Notably, many of these genes are associated with antigen presentation pathways (**Figure 1.6F**). For example, Adar1 depletion affected both 4T1 and B16 cell's fitness under ICB treatment. Adar1 edits endogenous dsRNA, preventing activation of the immune response via the MDA5 pathway. Since ADAR1 is induced by interferon, it helps tumors evade the interferon-mediated immune response triggered by ICB. Therefore, depleting ADAR1 enhances tumor sensitivity to ICB[33].

Furthermore, we noticed an interesting trend where the H2-K1's depletion, a component of MHC class I, resulted in an increase in fitness of B16 cells under all ICB treatments. We previously found that the depletion of MHC class I subsets (H2-D1 and H2-T23) in B16 cells grown in WT mice compared to those in NSG mice was immune sensitizing. This demonstrates that, while there are some similarities between antitumor immune escape and ICB resistance, significant differences remain. In the comparison between WT and NSG or Rag, the presence of CD8 cells leads to the upregulation of H2-T23 and H2-D1 in WT-selected tumors. These molecules can interact with CD94/NKG2A on CD8 cells and induce T cell exhaustion and facilitate immune escape. Conversely, in the comparison between ICB-treated and control groups in WT mice, there is reduced T cell exhaustion under ICB-treated conditions. As a result, cells that deplete H2-K1 gain a growth advantage since they are not targeted by the activated T cells in the tumor microenvironment. The increased expression of antigens and MHC molecules may thereby enhance the efficacy of the therapy. This underscores the complexity of both anti-tumor immune responses and immunotherapy evasion, highlighting both shared and distinct mechanisms at play.

1.6. In Vivo Validations Of Immunotherapy Resistance Pathways In 4T1 And B16 Cells

Next, we delved deeper into the novel and less-explored pathway, prioritizing it for further validation. Among these, Rab25 has been implicated in tumor progression and resistance to radiotherapy [25], while Rnaset2a is known to be induced by the Interferon and its lack of expression increases CD8 infiltration in leukoencephalopathy [34]. Yet, the precise roles of these genes in tumor immune evasion remain underexplored. We undertook separate depletion experiments for Rab25 and Rnaset2a in 4T1 cells (**Figure 1.7A**). The depletion of these genes led to the sensitization of 4T1 cells to ICB treatment, although no significant reduction in tumor size was observed in the untreated group. We also explored the role of ITM2C which was depleted in our B16-IC screen in PD-L1 as well as CTLA-4 treated mice. Integral membrane protein 2C (ITM2C) is a negative regulator of amyloid-beta peptide production, and may also play a role in TNF-induced cell death and neuronal differentiation [35,36]. It has also been found to be upregulated in multiple myeloma [37], however its role in tumorigenesis and cancer immune escape is not well understood. Through our in vivo validation we observed that ITM2C depletion results in sensitization of B16 tumors to CTLA4 therapy (**Figure 1.7B**).

To further understand the mechanism by which the Rab25 KD 4T1 tumors are sensitized to CTLA4, we conducted flow cytometry analysis to assess T cell infiltration in these tumors. While we didn't detect significant differences in CD3 and CD8 infiltration, we observed that the Rab25 KD cell line exhibited a higher percentage of CD4 effector T cells and a lower percentage of CTLA4 positive Tregs compared to the NTC tumors

(**Figure 1.7C**), indicating a less immunosuppressive tumor microenvironment in comparison to the NTC cell line.

1.7. Mechanisms Of Rab25 Resistance To Cancer Immunity And Immunotherapy

To delve deeper into the disparities between these two cell lines, we subjected 4T1 Rab25 KD, and NTC cell lines to bulk RNA sequencing. **Figure 1.8A** illustrates the differentially expressed genes in Rab25 KD 4T1 cell line compared to the NTC, particularly highlighting immune-related genes such as TLR7, and chemokines such as CXCL2 and CXCL3 upregulated in Rab25-KD. Additionally, we noted the downregulation of IFN-stimulated genes, such as Ifi208 and Ifi209, in Rab25 knockdown cells. This observation further underscores the significance of the interferon response pathway in mediating the resistance of the 4T1 tumors to immunotherapy.

We further performed pathway enrichment and network analysis on genes upregulated in Rab25 KD cells. (**Figure 1.8B**). In our analysis, we observed an increase in pathways associated with neutrophil migration and chemotaxis. Drawing from published data, it has been established that neutrophil accumulation within tumors can contribute to a favorable response to immunotherapy[38]. This suggests that heightened neutrophil activity might be a contributing factor to the enhanced sensitivity of Rab25 knockdown to immunotherapy. Furthermore, Rab25 is recognized for its role in regulating integrin membrane trafficking[26]. To deepen our understanding of the impact of Rab25-KD on membrane dynamics, we conducted mass spectrometry analysis to identify proteins enriched in the membranes of Rab25-KD and non-targeting control (NTC) cells. Interestingly, we observed a reduction in Itgav and Itgb6 levels in the membrane upon Rab25 depletion (**Figure 1.8C**). This prompted further investigation into the involvement of integrins in immune evasion. Our exploration revealed that integrins, particularly αvβ6 heterodimers, are major activators of TGF-β on the cell membrane. Activated TGF-β exhibits various immunosuppressive qualities, such as polarizing macrophages towards a more pro-tumorigenic phenotype and increasing Treg infiltration [39]. **Figure 1.8D** presents a protein interaction network analysis with a minimum interaction score of 0.7, indicating high confidence. This analysis reveals that both ITGB6 and ITGBV closely interact with TGFB1.

We were intrigued to pinpoint where the Rab25 expression initially appeared in our dataset. **Figure 1.8E** displays the expression of Rab25 across multiple generations in Balb and NU/J mice. Notably, Rab25 expression increases over time in WT mice, while its expression diminishes in NU/J mice, reiterating that Rab25 has been upregulated as a mechanism for tumor adaptive immune evasion. Moreover, to investigate the localization of Rab25 within the TME, we looked back into our scRNAseq data. **Figures 1.8F** and **1.8G** present UMAP and violin plots illustrating Rab25 expression across different clusters. Remarkably, Rab25 expression is highly specific to cancer cell clusters and is absent in other clusters, including normal breast tissue. This underscores Rab25 as a promising target for future cancer therapies. Finally, we investigated Rab25 expression in patients from the ISPY2 trial. We stratified patients based on their Rab25 expression levels derived from pre-ICB treated tumor RNAseq data into low and high groups. We observed that patients with higher Rab25 expression in their tumors have a lower probability of distant metastasis-free survival (P value=0.037) (**Figure 1.8H**). This highlights the significance of Rab25 expression not only in immune evasion of 4T1 mouse tumors but also potentially in resistance to immune checkpoint blockade (ICB) in patients.

1.8. Single Cell Analysis Of Tumor Immune Environment Reveals Mechanisms Of Immunosuppression In 4T1 And B16 Tumors

We aimed to delve deeper into the immune response within the tumor microenvironment (TME) of our immune-challenged tumors. **Figure 1.9A**, demonstrates the different major cell clusters identified using epithelial (Epcam), endothelial (Pecam1), fibroblast (Fbn1) and immune (Ptprc) cell markers. The barplot illustrates the distribution of non-immune cell clusters across different 4T1-IC conditions. Endothelial and cancer associated fibroblasts (CAFs) are larger in Balb-IC. **Figure 1.9B** also shows the percentage of live CD45 and CD3 infiltrating cells in the tumor, as determined by FACS staining and analysis. We used the FACS percentages because we had enriched these cells prior to running our samples on the single-cell RNA-seq platform. Therefore, we couldn’t rely on the scRNA-seq data for an accurate representation of these cells in the tumor microenvironment.

We then focused more on the immune cell cluster. **Figure 1.9C** illustrates the UMAP plot of the immune cell subclusters comprised of Tumor associated macrophages (TAMs), Neutrophils, Dendritic cells (DCs), B cells and Mast cells as identified using the cell specific markers as depicted in **Figure 1.9D**. The barplot in **figure 1.9C** shows the distribution of these clusters across the 4T1 Balb-IC, RAG-IC and NSG-IC conditions. Among these populations, tumor-associated macrophages appear to be more abundant in NSG-IC and Balb-IC samples compared to Rag-IC samples, while neutrophils are most prevalent in Rag-IC. We further looked into the NK/T cell cluster, undertaking sub-clustering to differentiate various NK and T cell populations based on specific markers (**Figure 1.9E-F**). In all conditions, the CD4-Treg cluster appeared to dominate the NKT subset. However, our analysis revealed a significant difference in the CD4 response, particularly within the CD4_Th1 and CD4_Th2 subsets, where the Balb-IC group exhibited the lowest percentage of these cells. Additionally, Balb-IC showed a higher proportion of exhausted CD4 and CD8 cells, suggesting a less activated and more exhausted T cell response in Balb-IC tumors.

Analysis of B16-IC tumors revealed both similarities and differences in the tumor microenvironment (TME) compared to 4T1-IC tumors. Cancer-associated fibroblasts (CAFs) were highly enriched in both B6-IC and NSG-IC tumors compared to Rag-IC tumors, while endothelial cells were significantly more abundant in B6-IC tumors. We also provided the percentages of infiltrated CD45 cells separately as these were enriched by FACS prior to scRNAseq (**Figure 1.10A**). We further focused on the immune cell cluster. Across all IC conditions, monocytes and macrophages predominated the immune microenvironment. Notably, the B6-IC tumors displayed a significant decrease in NK and T cell populations compared to NSG-IC, and a notable reduction in neutrophil clustering compared to other IC conditions (**Figure 1.10B**). The Dotplot in **figure 1.10C** illustrates the different immune cell cluster markers.

Further examination of the NK/T cell cluster in **Figure 1.10D** unveiled distinct NKT cell populations, each with a unique phenotype. The distribution of these clusters across different B16-IC conditions is depicted in the barplot, with specific cell markers identified in the dot plot in **Figure 1.10E**.

Analysis of the NKT subcluster revealed a distinctly immunosuppressive phenotype in the B6-IC sample. Initially, a lower NK cell population was observed in the B6-IC NKT population, aligning with our previous observation of upregulated expression of MHC molecules and antigen presentation in B6-IC cancer cells. This suggests a mechanism of immune evasion, as decreased NK cell infiltration in these tumors is likely due to increased MHC molecule expression. Additionally, lower numbers of CD8 stem cell-like cells were observed in the B6-IC NK/T cluster compared to Rag-IC and NSG-IC, indicating another potential mechanism of immune evasion by B6-IC cells, as these stem-like cells are crucial for anti-tumor immunity and immunotherapy response [40,41]. Conversely, B6-IC tumors exhibited higher numbers of exhausted CD8 cells marked by Pdcd1 expression.

In the CD4 clusters, the Treg population appeared larger in B6-IC and Rag-IC compared to NSG-IC. However, B6-IC had the lowest number of CD4-Th1 or CD4 effector cells. Collectively, these analyses suggest that the B6-IC NKT cluster exhibits a more immunosuppressive phenotype compared to the other IC conditions.

1.9. Discussion

To track tumors' natural evolution under different immune microenvironments, we utilized our immunoselection framework alongside large-scale transcriptomics and genomics tools. In both 4T1 and B16 cell lines, we compared pathways differentially expressed in tumors selected in immune-enriched versus immunocompromised environments. We observed that, among other immune pathways, interferon alpha and gamma pathways were highly upregulated in WT-selected tumors. Since innate and adaptive immune cells are the primary sources of these interferons, this underscores the strength of our in vivo model in capturing immune-tumor interactions and co-evolution. Previous studies have shown that interferons play a dual role in cancer progression. While acute high-dose responses of IFN are typically anti-tumor, low-dose long-term exposure can be pro-tumor [42]. Given that our model subjected these tumors to four generations of immune pressure, we hypothesize that the interferon response developed in these tumors is of the latter, pro-tumoral nature.

To better capture intratumor heterogeneity and TME differences, we inoculated WT mice with immune-enriched and immune-compromised selected tumors. Focusing on tumor subclusters, we identified a subcluster enriched in ISG immune-mediated gene signatures upregulated in WT-selected tumors. This cluster was present in both 4T1 and B16 datasets, but with notable differences: it expanded in both RAG-selected and WT-selected 4T1 tumors, whereas it was significantly larger in WT-selected B16 tumors and much smaller in RAG and NSG-selected B16 tumors. We hypothesize that B16 tumors are naturally more immunogenic than 4T1 tumors, with a higher presence of NK/T cells and CD8 cells in the TME. Due to their greater exposure to adaptive immunity, immune response mechanisms are more strongly selected for in B16 tumors.

Furthermore, our in vivo CRISPRi screen demonstrated that, in B16 cells, there was significant selection for antigen presentation pathways such as H2-T23, H2-D1, and Tap1, likely due to higher NK and CD8+ T cell infiltration in these tumors compared to 4T1 tumors. Previous studies have shown that MHC I, especially H2-T23, engages with CD94/NKG2A receptors, contributing to T cell exhaustion [3]. The upregulation of MHC I molecules is also a known mechanism of NK cell evasion. Nevertheless, despite the differences in screened genes in both cell lines Stat1 depletion sensitized tumors to immune pressure in WT mice.

Our single-cell data showed more infiltration of CD8+ T cells and NK cells in B16 tumors compared to 4T1 tumors, and less infiltration of these cells in B16 B6-IC tumors versus B16 NSG-IC and Rag-IC tumors. This indicates that these immune evasion mechanisms were most likely upregulated in B6-selected B16 tumors to evade high CD8 T cell and NK cell immune pressure. Interestingly, when we applied additional immune pressure in our screen through immune checkpoint blockade (ICB), we observed the depletion of different sets of genes affecting tumor fitness. For example, Adar1 depletion affected both 4T1 and B16 cell's fitness under ICB treatment. Adar1 edits endogenous dsRNA, preventing activation of the immune response via the MDA5 pathway. Since ADAR1 is induced by interferon, it helps tumors evade the interferon-mediated immune response triggered by ICB. Therefore, depleting ADAR1 enhances tumor sensitivity to ICB [33].

We further investigated the less explored pathways in both 4T1 and B16 tumors where their depletion sensitized tumors to immune checkpoint blockade (ICB). We identified Rab25, whose depletion enhanced tumor sensitivity to both anti-tumor immunity and ICB in WT mice compared to NSG mice, as well as to CTLA4 inhibition. Our observations revealed that Rab25 plays a role in integrin membrane trafficking and TGF-β regulation, highlighting its potential impact on tumor-immune interactions and response to immunotherapy.

Overall, our in vivo and multi-omics platform successfully tracked the coevolution of tumor and immune cells under various immune pressures. Our findings revealed pathways of immune resistance in 4T1 and B16 cancer mouse models and correlated with clinical data from breast cancer patients treated with ICB. This workflow holds promise for uncovering novel insights into cancer-immune coevolution and resistance, potentially identifying new targets for use either alone or in combination with existing immunotherapies.

1.10. Figures

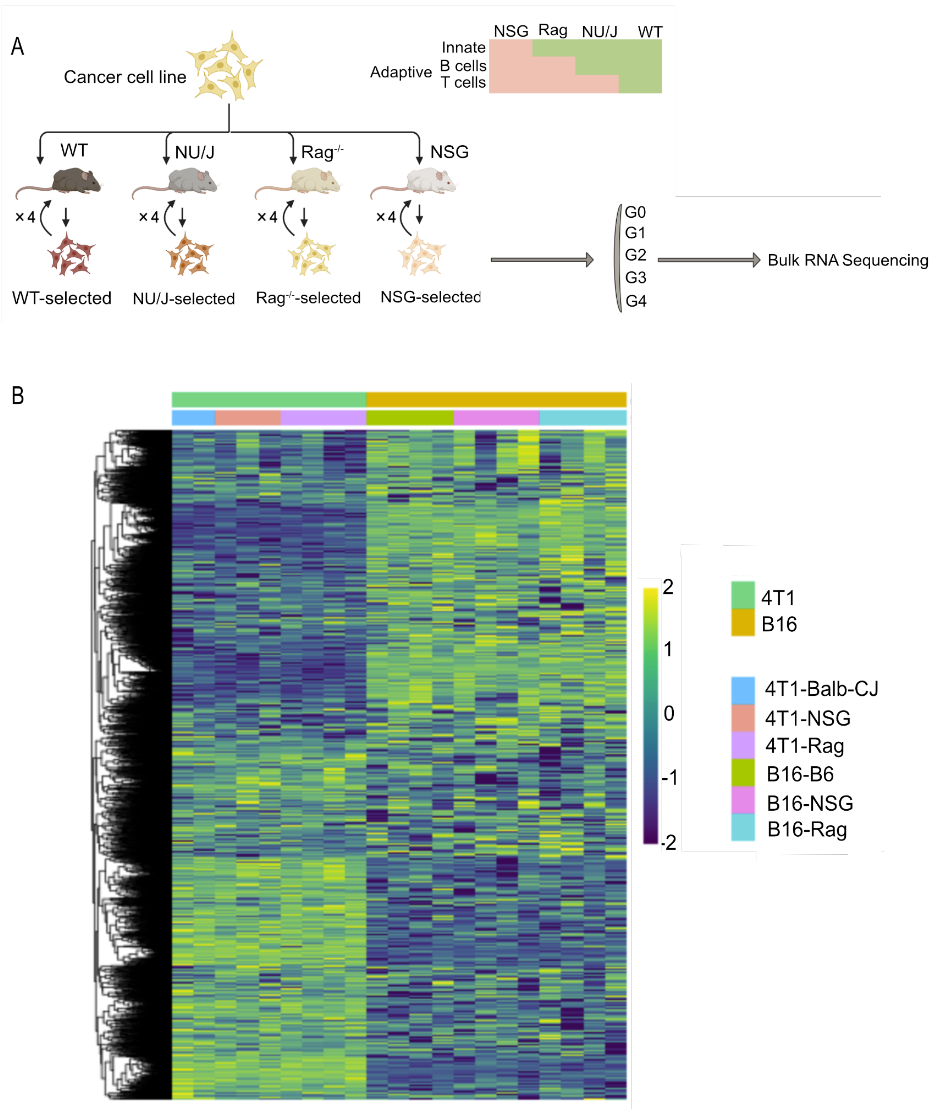


Figure 1.1. Introducing Immunoselection in vivo workflow:

A) A schematic of the in vivo experimental workflow; 4T1 and B16-F10 cells are injected into mice with varying levels of immunocompetency. 4T1 cells were injected to Balb-CJ, Rag-/- , NU/J and NSG mice while B16 cells were injected to B6, Rag -/- and NSG mice for four generations. RNA isolation from cells in each generation was gathered and subjected to RNA sequencing**)** **B)** Heatmap illustrates gene expression comparison of generation 4 of 4T1 and B16 tumors developed in different strains using Deseq2 analysis.

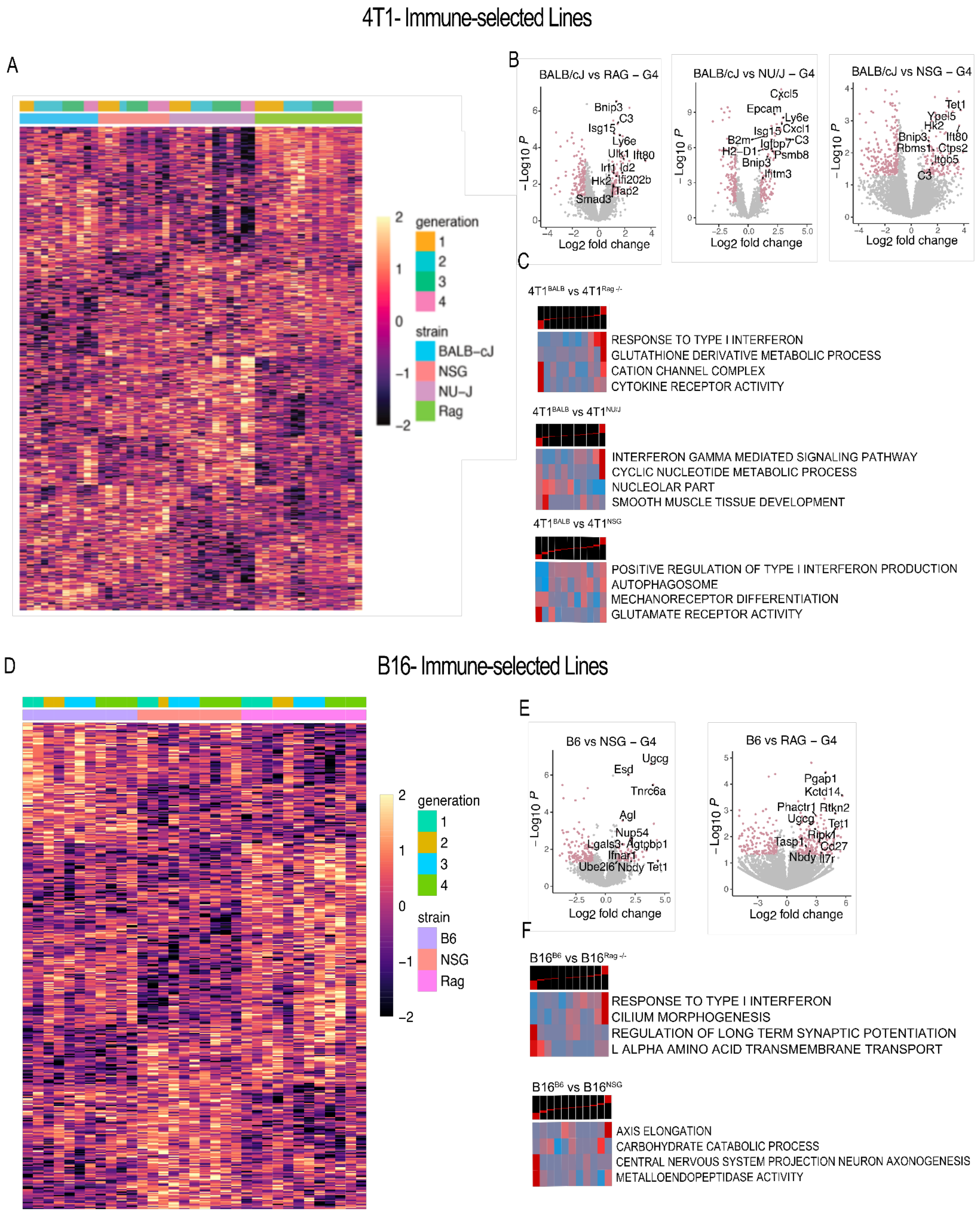


Figure 1.2. Identifying mechanisms of cancer immune co evolution in the immune-selected models

A) Heatmap enrichments of in vivo selected 4T1 cell lines developed in different mouse strains over multiple generations. Genes highlighted are selected based on P-value< 0.05 across various strains using ANOVA test. **B)** Volcano plots comparing Balb-selected vs Rag, NU/J and NSG-selected generation 4 of 4T1 cancer cell lines using DEseq2 analysis. Highlighted in pink are the genes with |Log2FC| >1 and Pv< 0.05. **C)** iPAGE analysis of in vivo selected 4T1 cell lines. (Figure caption continued on the next page)

(Figure caption continued from the previous page) The top panel displays gene expression variances grouped into discrete expression bins. Genes up-regulated in WT-selected tumor samples are positioned on the right, while those with lower expression are on the left. The accompanying heatmap illustrates pathway associations with the expression bins. Red cells signify enrichment of pathway genes within a specific expression bin, while blue cells indicate depletion. Enrichment and depletion levels are determined using log-transformed hypergeometric P values. iPAGE analysis was conducted on genes subjected to DEseq2 Likelihood Ratio Test (LRT) analysis, emphasizing pathways differentially regulated from G0 to G4. **D-F)** Same analysis as in A-C, for B16 cancer cell lines.

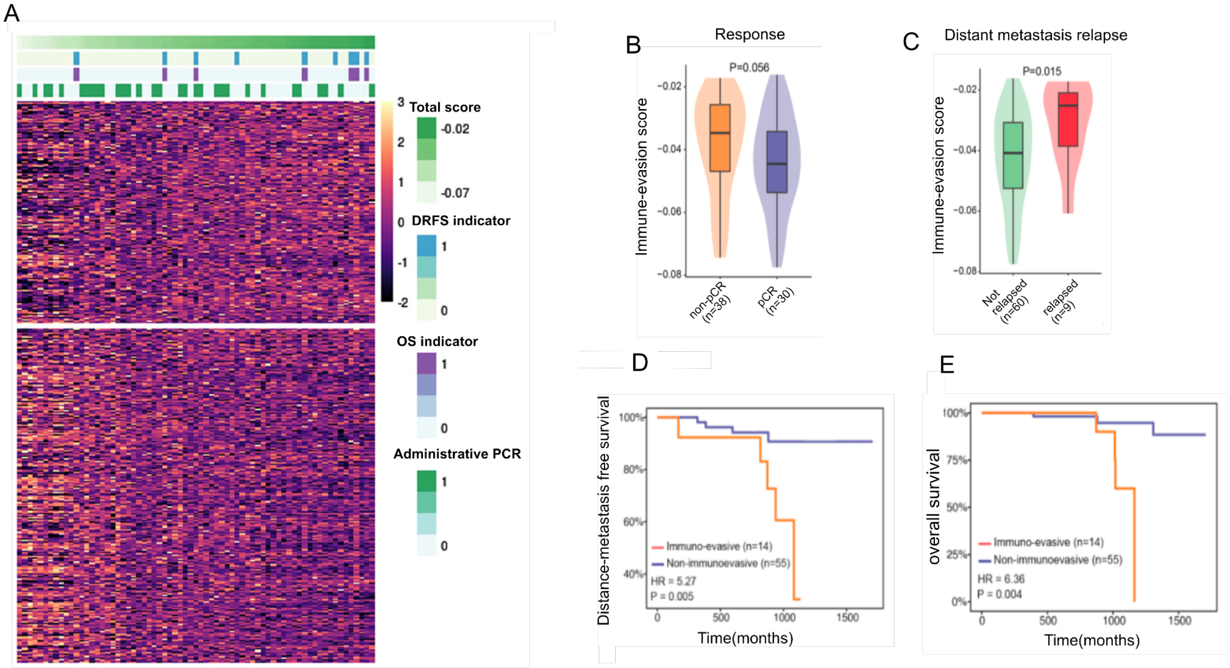


Figure 1.3. Immune evasion score negatively correlates with breast cancer patients' survival outcome

A) Immune evasion markers were selected based on differentially expressed genes in WT vs immune-compromised selected tumors with adjusted Pv<0.5. ISPY-2 trial breast cancer patients undergone anti PD-1 therapy, received an immune evasion score based on the expression of these genes. **B)** Immune evasion score and patients' response to PD-1 treatment, patients were categorized into pCR (partial clinical response) and none-pCR. Immune evasion score is significantly higher in patients who were categorized in none-pCR. **C)** Immune evasion score and patients' relapse status after PD-1 treatment. Immune evasion score is significantly higher in patients who were categorized in none-pCR. Immune evasion score is significantly higher in patients who had relapsed after anti PD-1 treatment. **D)** Immune evasion score and patients' distant metastasis free survival (DMFS). Patients that received higher immune evasion score had DMFS outcome compared to the patient with lower immune-evasion score. **E)** Immune evasion score negatively correlates with patients' survival outcome. Patients that received higher immune evasion scores had worsen survival outcome compared to the patient with lower immune-evasion score.

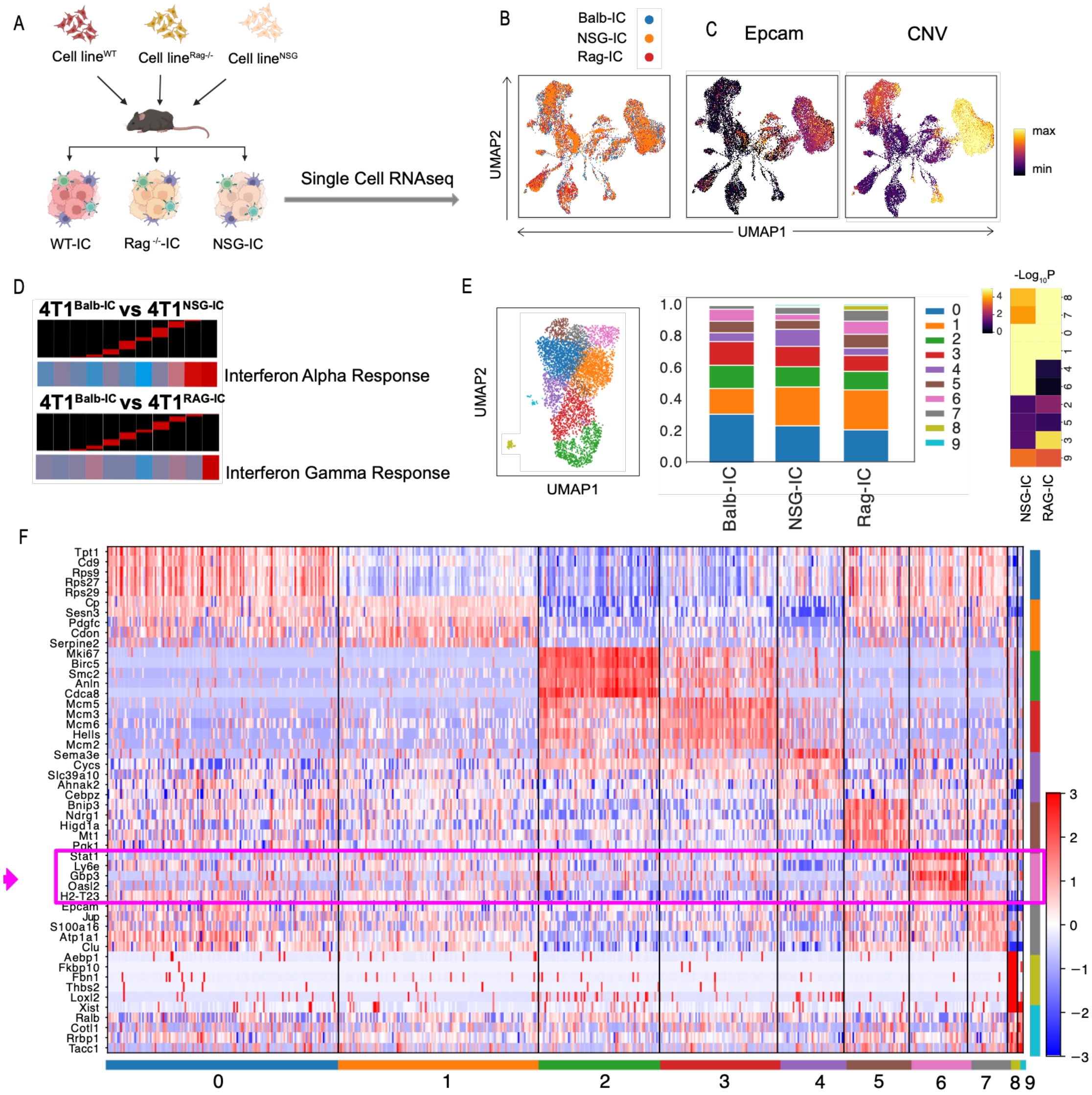


Figure 1.4. Intra tumor heterogeneity of 4T1 Immune-selected lines identifies an immune mediated subcluster

A) A schematic illustrating the workflow for subjecting previously immunoselected tumors to immune pressure. Generation 4 of the selected cell lines were injected into WT mice, and the resulting tumors underwent single-cell RNA sequencing (scRNAseq) for detailed analysis of tumor and tumor microenvironment (TME) clusters. The resulting tumors are labeled as immune-challenged (IC). **B)** UMAP plot overlaying Balb-IC, NSG-IC, and Rag-IC tumors. **C)** UMAP plots highlighting Epcam expression and CNV analysis to emphasize the cancer cell cluster. **D)** iPAGE analysis of the cancer cell cluster among different immune-challenged 4T1 cell lines that were subjected to Deseq2 analysis. **E)** Identification of distinct cancer cell subclusters and their distribution across Balb-IC, NSG-IC, and RAG-IC groups, with a heatmap plot depicting P-values using Fisher's exact test. **F)** Heatmap plot highlighting the top 5 highly variable genes within each cancer cell subcluster.

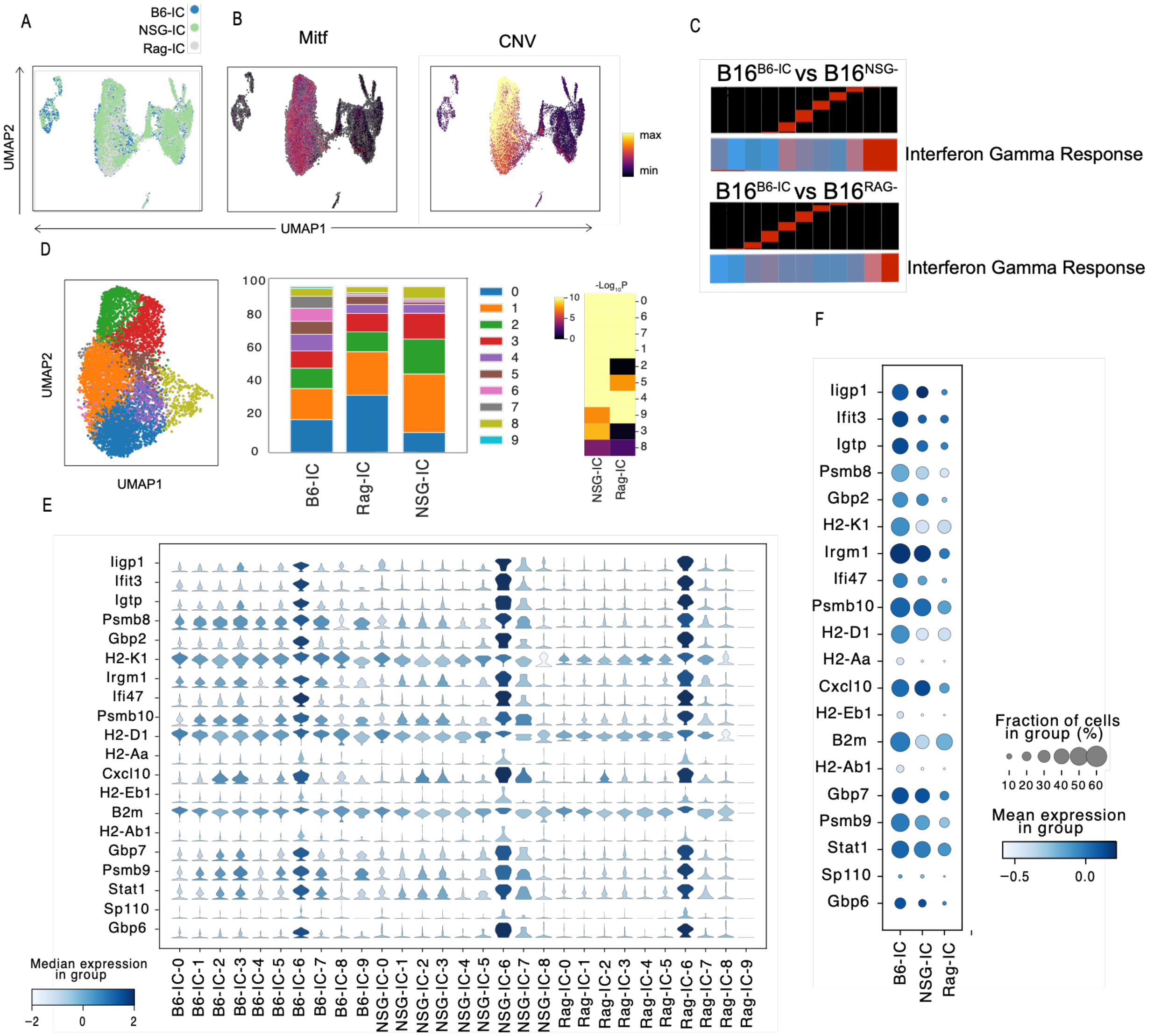


Figure 1.5. Intra tumor heterogeneity of B16 Immune-selected lines identifies an immune mediated subcluster

A) UMAP plot overlaying B6-IC, NSG-IC, and Rag-IC tumors. **B)** UMAP plots highlighting Mitf expression and CNV analysis to highlight the melanoma cancer cell cluster. **C)** iPAGE analysis of the cancer cell cluster among different immune-challenged B16 cell lines that were subjected to Deseq2 analysis. **D)** Identification of distinct cancer cell subclusters and their distribution across B6IC, NSG-IC, and RAG-IC groups, with a heatmap plot depicting P-values using Fisher's exact test. **E)** Stacked volcano plot highlighting the top 20 genes in cluster 6 with median expression across all different cancer cell subclusters in various IC groups. **F)** Dot plot emphasizing the top 20 genes in cluster 6 with median expression across all IC groups collectively.

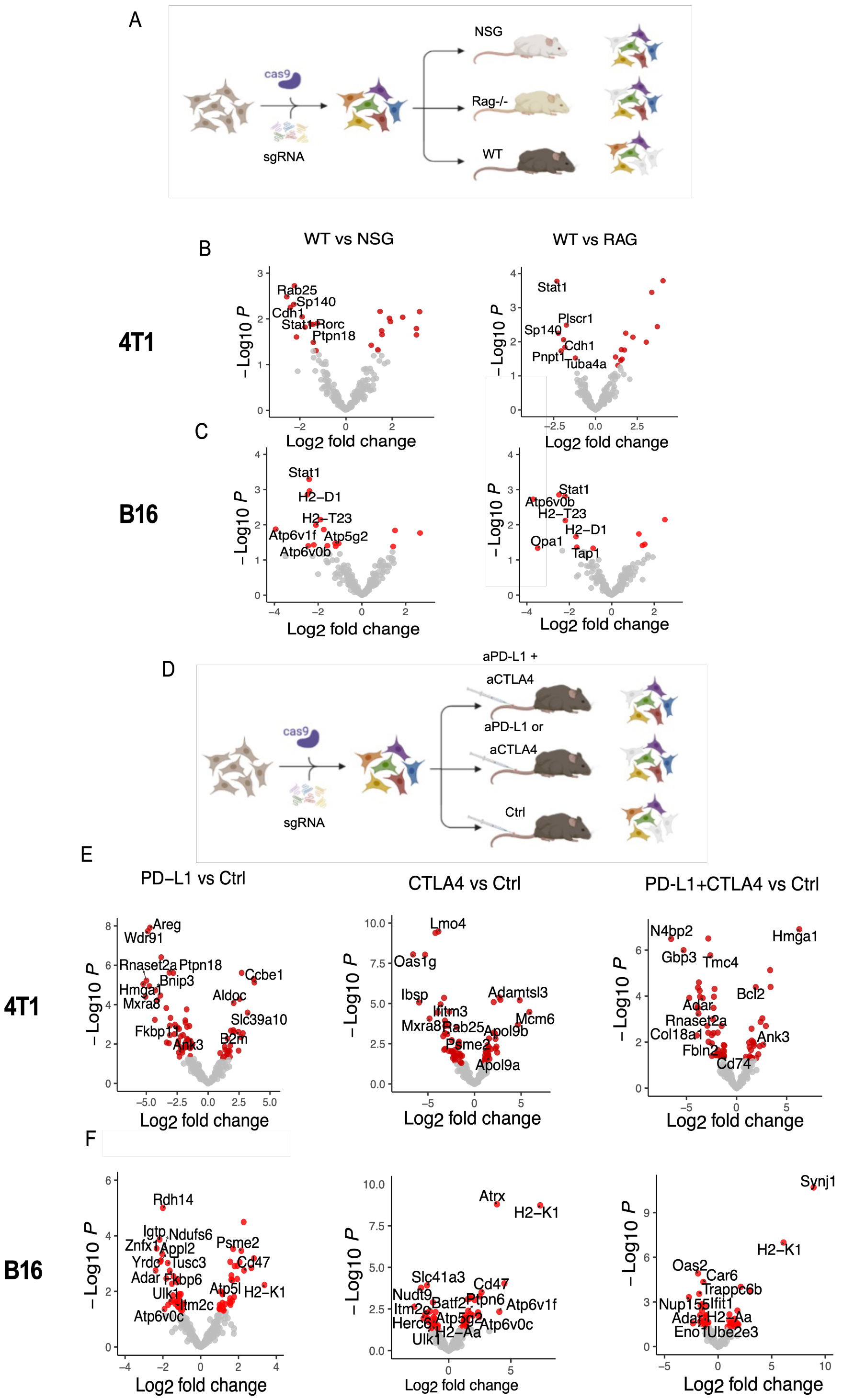


Figure 1.6. In Vivo CRISPRi screens identify immune and ICB resistance pathways

A) Our in vivo CRISPRi screen workflow. We generated 4T1-CI and B16-CI cell line harboring a library comprising approximately 200 genes that were upregulated in Balb and Balb-IC or B6 and B6-IC tumors. (Figure caption continued on the next page)

(Figure caption continued from the previous page) These cells were then injected into NSG, RAG, Balb/cJ, or B6 mice. The resulting tumors underwent sgRNA sequencing for analysis of sgRNA enrichment in the different conditions. **B-C)** Volcano plots show sgRNA enrichment and depletion in 4T1-CI and B16-CI tumors respectively in WT vs NSG/Rag mice. **D)** Injection of 4T1-CI or B16-CI library in WT mice undergoing PD-L1, CTLA4 or combination of PD-L1 and CTLA4 vs their corresponding Igg control. **E-F)** Volcano plots show sgRNA enrichment and depletion in 4T1-CI and B16-CI ICB or Igg treated tumors in WT mice.

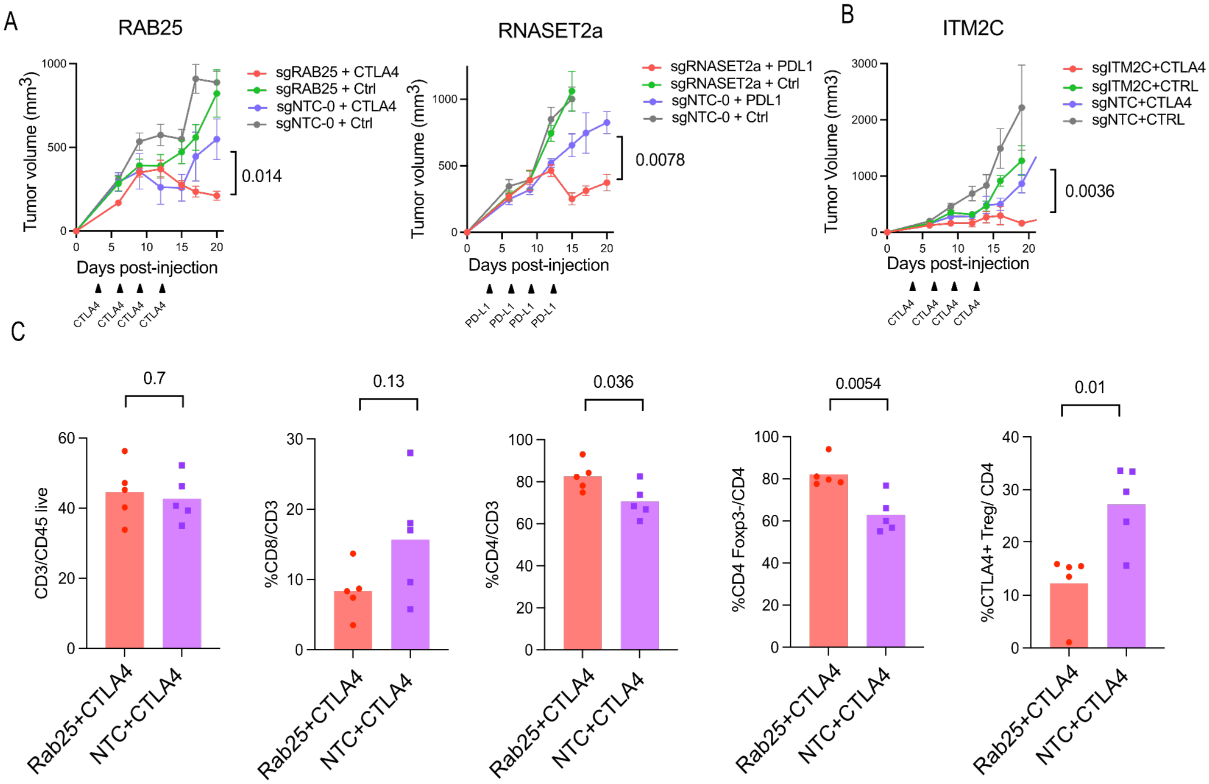


Figure 1.7. Tumor measurements of Rab25, Rnaset2a and ITM2C depleted 4T1 cells confirms sensitization to ICB

A-B) in vivo tumor measurements of 4T1 Rab25-KD, 4T1 Rnaset2a-KD and B16-Itm2c-KD vs their corresponding NTC with or without ICB treatment P values calculated using multiple unpaired t tests. **C)** flow cytometry analysis of T cells of 4T1- RAB25 KD vs 4T1-NTC tumors undergone CTLA4 treatment. P values calculated using unpaired t tests.

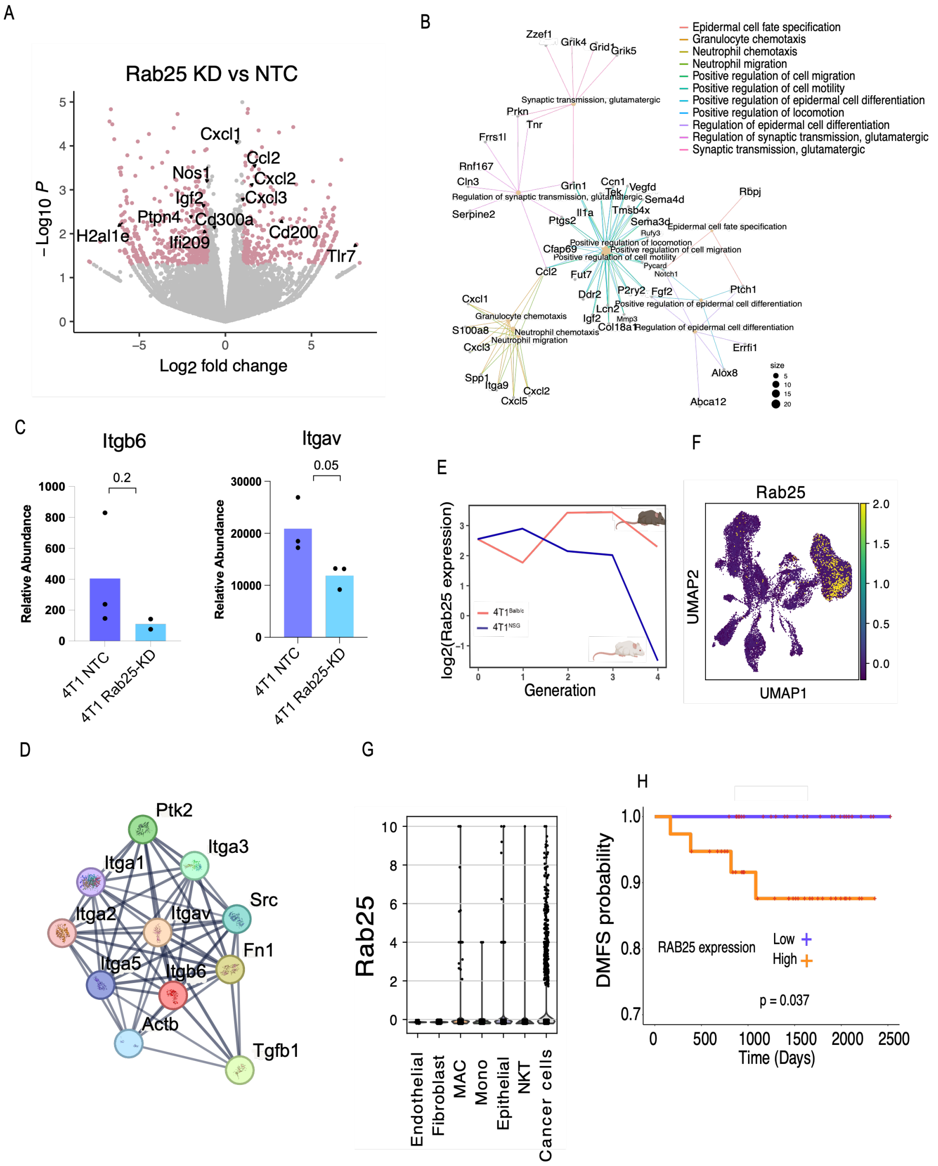


Figure 1.8. Rab25 regulates anti-tumor immunity through integrin membrane trafficking

A) Volcano plot visualizing DEGs in 4T1-RAB25 KD vs 4T1-NTC cell lines. **B)** Network analysis of upregulated genes in 4T1-RAB25 KD cell line (FC>1, Pv<0.05). **C)** relative abundance of integrin protein in mass spec data enriched for membrane proteins, P value calculated using unpaired t-test. **D)** protein network analysis of Itgb6 with n=10 genes. **E)** Rab25 expression in bulk RNAseq data from Balb and NU/J mice over multiple generations. **F-G)** UMAP and violin plots showing Rab25 expression in different immune challenged scRNAseq clusters. **H)** RAB25 expression in ISPY2 trial patients undergone PD1 therapy and their survival outcome.

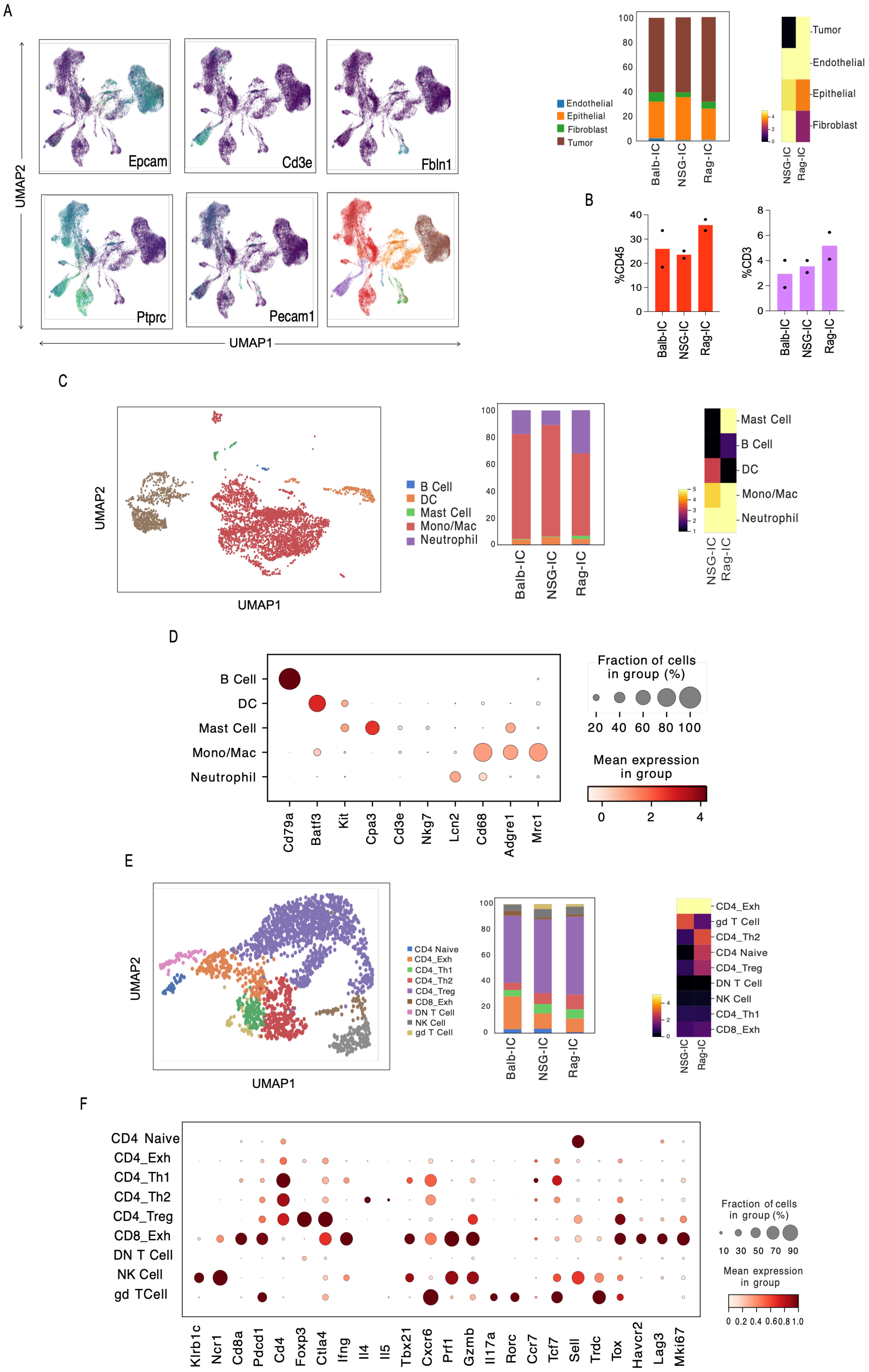


Figure 1.9. Tumor microenvironment analysis of 4T1 immuno-selected tumors

A) UMAP plot illustrating 4T1-IC clusters and their cell type-specific markers, with emphasis on immune cells, epithelial cells, endothelial cells, and fibroblasts. Bar plot displaying the distribution of each cluster across different IC groups. Heatmaps present p-values calculated using Fisher's exact test. (Figure caption continued on the next page)

(Figure caption continued from the previous page). **B)** Distribution of immune and NK/T clusters among groups determined via flow cytometry (n=2). **C)** UMAP plots depict the immune subclusters and their distribution across IC conditions, as indicated by the bar plot. Heatmaps present p-values calculated using Fisher's exact test. **D)** Dot plot highlighting distinct gene signatures for identifying different immune subcluster cell types. **E)** UMAP plot illustrating NK/T subclusters and their distribution among IC conditions, as indicated by the bar plot. P-values calculated using Fisher's exact test are depicted in the heatmap. **F)** Dot plot highlighting diverse gene signatures for identifying various NK/Tsubcluster cell types.

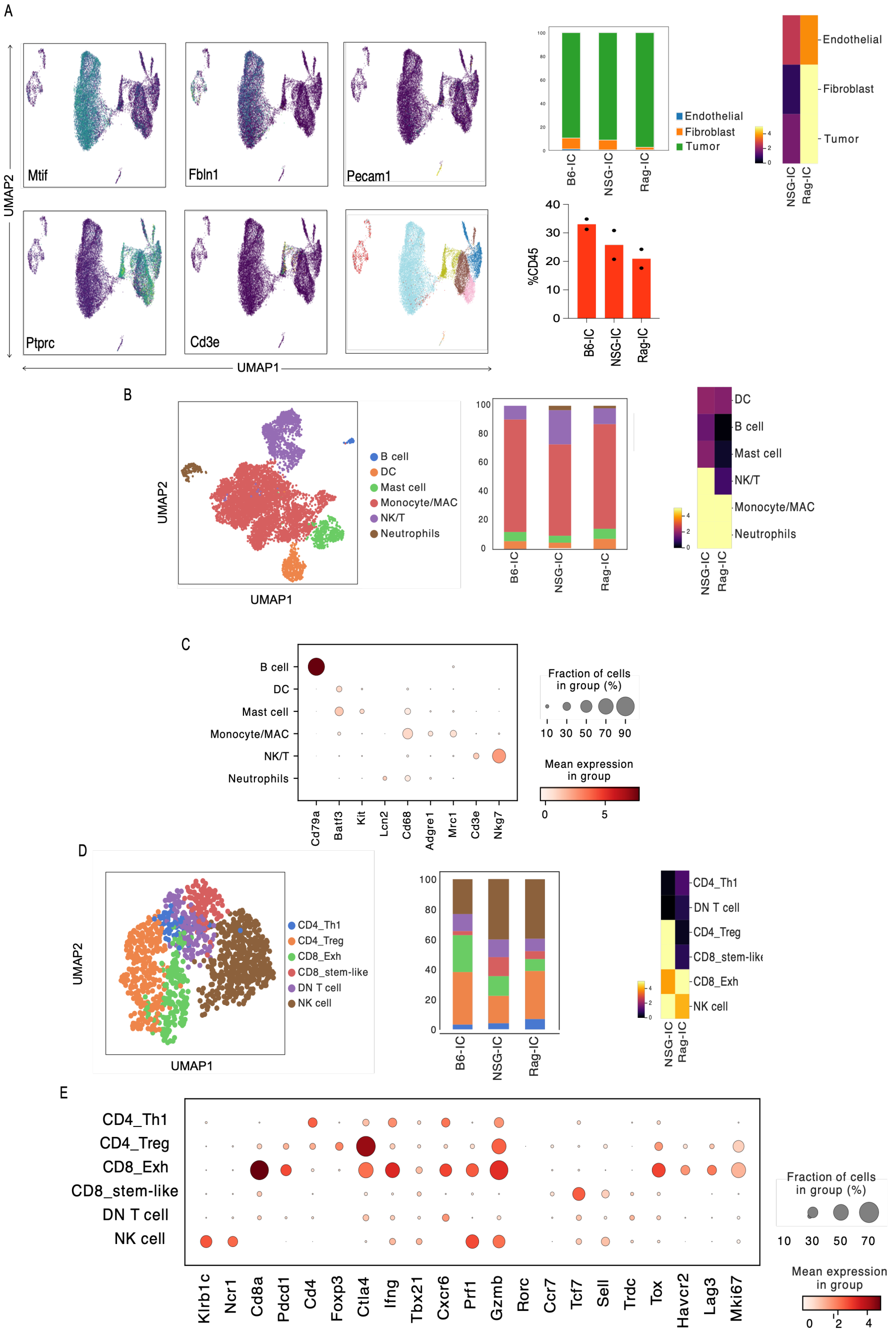


Figure 1.10. Tumor microenvironment analysis of B16 immuno-selected tumors reveals an immunosuppressive phenotype in B6-IC samples

A) UMAP plot illustrating B16-IC clusters and their cell type-specific markers, with emphasis on immune cells, epithelial cells, endothelial cells, and fibroblasts. Bar plot displaying the distribution of each cluster across different IC groups. Heatmaps present p-values calculated using Fisher's exact test. Separate bar plot illustrating the distribution of immune cells among groups determined via flow cytometry (n=2).
(Figure caption continued on the next page)

(Figure caption continued from the previous page) **B)** UMAP plots depict the immune subclusters and their distribution across IC conditions, as indicated by the bar plot. Heatmaps present p-values calculated using Fisher's exact test.

C) Dot plot highlighting distinct gene signatures for identifying different immune subcluster cell types. **D)** UMAP plot illustrating NK/T subclusters and their distribution among IC conditions, as indicated by the bar plot. P-values calculated using Fisher's exact test are depicted in the heatmap. **E)** Dot plot highlighting diverse gene signatures for identifying various NK/T subcluster cell types.

1.11. Materials And Methods:

Cell Line Generation

Gene knockdowns were performed by first transducing 4T1 or B16 cells with dCas9-KRAB construct via lentiviral delivery of: pHR-UCOE-EF1a-dCas9-HAxNLS-XTEN80-Zim3-p2a-GFP. 4T1- dCas9 or B16-dCas9 expressing cells were then sorted by FACS isolation of GFP-positive cells. Guide RNA sequences for CRISPRi-mediated gene knockdown were cloned into pCRISPRia-v2 (Addgene #84832) via BstXI-BlpI sites. After transduction with sgRNA lentivirus, cells were selected with 2 μg/mL puromycin (Gibco). Knockdown of target genes was assessed by reverse transcription of total RNA to cDNA (Maxima H Minus RT, Thermo), then using sequence specific primers along with PerfeCTa SYBR Green SuperMix (QuantaBio) per the manufacturer's instruction. HPRT was used as an endogenous control.

Mouse Experiments

All mouse experiments were performed under the supervision and approval of the Institutional Animal Care and Uses Committee (IACUC) and the University of California, San Francisco (UCSF) (protocol no. AN179718-03F). For immunoselection rounds, 8–10-week-old age-matched female NSG (005557), Rag -/- (003145 or 002216), Balb/cJ or C57BL/6J (000651 or 000664), and NU/J (002019) mice were injected with 500K 4T1 or B16 cells bilaterally. 4T1 cells were injected into mammary fat pads while B16 cells were injected subcutaneously. For in vivo CRISPRi screen 1M cells were injected on each side. Tumors were grown in animals and monitored for 2-3 weeks, or until they reached a size of less than 1.5 cm³, before dissection. All mice were purchased from Jackson Laboratory.

Mouse Treatments

200 ug PD-L1(cat# BP0101) or its Igg Ctrl (cat# BP0090) and 100 ug CTLA4 (cat# BP0032) or its corresponding Igg ctrl (cat# BP0091) were administered to each mice on days 3,6,9 and 12.

Bulk RNAseq Library Preparation And Analysis

The tumor cells were subjected to digestion with collagenase, DNase, and Liberase at 37°C for a duration ranging from 30 minutes to an hour. Following digestion, the resulting mixture was filtered through 70 μm filters to achieve a single-cell suspension. These cells were subsequently cultured in vitro for two to three passages before proceeding to library preparation. RNA extraction was carried out using the Quick-RNA Microprep Kit from Zymo Research, and library preparation utilized the Lexogen QuantSeq Kit. Quality control of the samples was performed with a Tapestation, and sequencing was conducted on a Novaseq SP100. The sequences were processed with Cutadapt to eliminate short and low-quality reads, which were then aligned to the mm10 reference genome using Salmon. Differential expression analysis was conducted with the DESeq2 package.

Single Cell RNAseq

For 4T1-IC tumors, live cells and immune T cells were enriched before mixing and processing on the 10x platform. In contrast, B16-IC tumors were enriched for both live and immune cells. Biological replicates were barcoded, and single-cell RNA-seq libraries were prepared using the Chromium Next GEM Single Cell 3′ Kit v3.1 from 10X Genomics. The prepared libraries were then sequenced on an Illumina NovaSeq sequencer at the UCSF Center for Advanced Technology.

Raw Data Processing

CellRanger v.3.0 (10X Genomics) was used for cell barcode filtering, read alignment, unique molecular identifier counting and to generate a digital gene expression matrix from raw FASTQ files. Reads were aligned to the mouse reference genome mm10 using CellRanger-provided annotations for gene features. Reads were assigned to cells based on their cell barcodes; barcodes that did not appear in the 10X Genomics 3M barcode allow list were removed.

Barcode Demultiplexing And Assignment

Cells were assigned to cell lines of origin by quantifying the relative proportion of detected genetic barcodes. Unique molecular counts for each barcode were determined from barcode-containing reads; a Gaussian kernel density estimation was fitted to the frequency of each barcode across all cells. The interpeak minima of the resulting bimodal distributions were set as the minimum threshold for barcode assignment. Cells were assigned a barcode if the frequency of that barcode exceeded its associated threshold and was tenfold more frequent than the second most frequently occurring barcode. Cells assigned multiple barcodes were designated as doublets and removed.

Single-Cell Data Preprocessing And Visualization

All preprocessing steps were performed using Scanpy. Cells were excluded if they expressed fewer than 200 genes, more than 8,000 genes, or if their mitochondrial gene expression exceeded 10% of the total. Additionally, genes expressed in fewer than three cells were removed. Gene expression counts were normalized to 10,000 counts per cell, then log-transformed and scaled to unit variance across cells. Highly variable genes were identified, and a lower-dimensional representation of the cells was created using principal component analysis (PCA) with this reduced feature set. The Uniform Manifold Approximation and Projection (UMAP) technique, implemented via scanpy.tl.umap with standard settings, was applied to the minimum number of principal components necessary to explain the observed variance. Clusters were then generated using the Leiden algorithm as implemented in scanpy.tl.leiden.

Tumor Genomic DNA Extraction And Library Preparation

Tumors were harvested 2-3 weeks post-injection and processed using Quick-DNA midiprep plus kit (Zymo Research Cat. #D4075) according to the protocol. For each processed tumor, genomic DNA was digested in 5ug scale, 100 µL volume reactions amplified with 5' forward primer: AATGATACGGCGACCACCGAGATCTACACCGCGGTCTGTATCCCTTGGAGAACCACCT, and 3' indexing reverse primer: CAAGCAGAAGACGGCATACGAGANNNNNGCGGCCGGCTGTTTCCAGCTTAGCTCTTAAA. The PCR reactions were then cleaned up using Zymo DNA Clean & Concentrator-25 kit (Zymo Research Cat. #D4033). Samples were pooled and sequenced on a lane of Novaseq X at the UCSF Center for Advanced Technology (CAT).

Screen Cloning

Screen was done using combinatorial sgRNA library cloning as described [43,] the design of each oligo was as follows: PCRadaptor--BstXI--sgRNA_A--BsmBI_2x--sgRNA_B—PCRadaptor

In Vivo Screens

All pool infections were conducted with transduction rates of 15-20%. Transduced cells were selected in puromycin and were maintained in culture at ~1,000× library coverage at all times following infection. 1M cells were injected to two mice bilaterally for each screen condition, maintaining a coverage >1000x. An in vitro culture of the cells was also kept for the same population doubling time as the in vivo screens. Cells were injected to Balb/cj, B6, NSG, or Rag KO mice.For the immunotherapy arm, 4T1 or B16 cells were injected to Balb/cj or B6 mice respectively. On day 3,6,9 and 12 post injection, mice received PD-L1 (200ug/mice), CTLA4(100ug/mice), combination of both or the corresponding Igg Ctrl. ICB drugs were purchased from BioXCell.

CHAPTER 2: REVEALING THE GRAMMAR OF SMALL RNA SECRETION USING INTERPRETABLE MACHONE LEARNING

2.1. Abstract

Small non-coding RNAs can be secreted through a variety of mechanisms, including exosomal sorting, in small extracellular vesicles, and within lipoprotein complexes[44,45]. However, the mechanisms that govern their sorting and secretion are still not well understood. In this study, we present ExoGRU, a machine learning model that predicts small RNA secretion probabilities from primary RNA sequence. We experimentally validated the performance of this model through ExoGRU-guided mutagenesis and synthetic RNA sequence analysis, and confirmed that primary RNA sequence is a major determinant in small RNA secretion. Additionally, we used ExoGRU to reveal *cis* and *trans* factors that underlie small RNA secretion, including known and novel RNA-binding proteins, e.g., YBX1, HNRNPA2B1, and RBM24. We also developed a novel technique called exoCLIP, which reveals the RNA interactome of RBPs within the cell-free space. We used exoCLIP to reveal the RNA interactome of HNRNPA2B1 and RBM24 in extracellular vesicles. Together, our results demonstrate the power of machine learning in revealing novel biological mechanisms. In addition to providing deeper insight into complex processes such as small RNA secretion, this knowledge can be leveraged in therapeutic and synthetic biology applications.

2.2. Introduction

Small non-coding RNAs play a variety of regulatory functions in the cell, including regulation of mRNA stability and protein synthesis[46,47]. However, some small RNAs also reside in the extracellular space, packaged within extracellular vesicles or lipoprotein complexes, for example, where they are thought to play roles in cellular communication[44,48,49]. Many recent studies have focused on the role of these secreted small RNAs as potential biomarkers in various diseases, particularly cancer[50–52]. RNA secretion, however, is not a random process. While some studies have focused on identifying the various mechanisms through which small RNAs are secreted[53,54], our knowledge of the underlying regulatory programs that govern extracellular sorting remains incomplete.

To reveal the *cis*-regulatory grammar that underlies small RNA secretion, we developed ExoGRU, a deep-learning model for predicting secretion probabilities of small RNAs based on their primary sequence. In addition to the commonly used machine learning performance metrics, we also used two independent experimental approaches to validate the veracity of our model. We used ExoGRU to (i) identify mutations that abrogate the secretion of known cell-free small RNAs, and (ii) predict high confidence sets of synthetic sequences that are secreted or retained. Having confirmed the accuracy of ExoGRU using these experimental strategies, we interrogated the model to reveal the cis-regulation RNA secretion grammar that it has learned. In addition to recapitulating known RNA binding proteins involved in small RNA sorting, such as YBX1, we also discovered and validated RBM24 as a novel RNA secretory factor. We also developed exoCLIP, a variation of CLIP-seq[55], that reveals RBP-RNA interactions in the cell-free space using

UV cross-linking immunoprecipitation followed by high-throughput sequencing. Application of exoCLIP to RBM24 and HNRNPA2B1, another factor that was nominated by our model and previously implicated in RNA secretion, further confirmed their direct interactions with target small RNAs in extracellular vesicles.

Our results collectively show the significance of machine learning in uncovering previously unknown biological mechanisms. In addition to capturing the sequence features that mark small RNAs for secretion, our approach provides readily testable hypotheses around the key *trans* factors involved. This not only deepens our understanding of an intricate biological process, but also has practical implications for the design of artificial cell-free RNA species in synthetic biology applications.

2.3. ExoGRU, A Computational Model For Accurate Prediction Of Small RNA Secretion

To learn the small RNA secretory grammar, we first aggregated, curated, and labeled a large compendium of small RNA datasets in the extracellular (EC) or the intracellular (IC) compartment. These datasets, along with their 'EC' versus 'IC' labels, were obtained from three distinct sources: (i) a dataset of intracellular and extracellular small RNAs (between 18 and 50nt) we had previously generated across eight cell line models[50], (ii) the Extracellular RNA Communication Consortium Atlas (exRNA Atlas)[56] dataset, and (iii) The Cancer Genome Atlas (TCGA) small RNA sequencing data[57]. Given that the cell-free RNA content is not correlated with the abundance of small RNAs in the cell, we hypothesized that a *cis*-regulatory grammar serves as a localization signal for small RNA sorting into extracellular space. First, to search for EC-associated RNA sequence and structural features, we compiled the primary sequence, k-mer frequencies (k=1,2,3,4,5,7), *in silico* folding free energy, and predicted secondary structures as input features to train our model (**Figure 2.1A**). Starting with simpler models, we trained linear SVM, Gaussian kernel SVM, and random forests as classifiers. The poor performance of these models (Maximum AUC ROC: 0.71) motivated us to train more complex models with increased learning capacity. We tested various neural network architectures, starting with shallow convolutional neural networks (CNN) and recurrent neural networks (RNN), as well as DeepBind, a previously developed CNN model[58]. Upon hyperparameter tuning, we observed an increase in performance upon switching to a Gated recurrent units (GRUs)-based deep recurrent neural network architecture (**Figure 2.1B**). As shown in **Figure S2.1A**, we benchmarked our GRU model, which we named "ExoGRU" (**Figure 2.1C**), against several existing machine learning and deep learning models. **Figure 2.1D** shows the performance of ExoGRU, evaluated on the held-out test set, in which we achieved an area under receiver operating characteristic (AUROC) of 0.95 and an area under the precision recall curve (AUPRC) of 0.8, respectively. At 83% specificity, the sensitivity of ExoGRU was 91% (see the confusion matrix in **Figure S2.1B**). We also sought to assess the contribution of each input feature to the performance of ExoGRU. From our initial list of features described above, we observed that the primary sequence alone is sufficient to effectively distinguish IC sequences from EC. Furthermore, we conducted a comparative evaluation between our ExoGRU model and several established RNA localization prediction models. Notably, many of these existing models were primarily designed and trained for long non-coding RNAs (lncRNAs), which inherently differ from the shorter small RNAs (smRNAs) that we focus on in our study. Nevertheless, we conducted an extensive analysis of our model's quality metrics in comparison to some of the existing models that accept short RNAs as input. As shown in **Figure S2.1C**, the results revealed significantly superior performance with exoGRU.

2.4. Experimental Verification Of ExoGRU Predictions

To further evaluate the performance of our model, we sought to focus on small RNAs whose status is predicted by ExoGRU with high confidence, i.e., focusing on high confidence true-positives and negatives. For this, we used ExoGRU to select those sequences with the highest and lowest secretion probabilities and labeled them as ECX (high confidence EC) and ICX (high confidence IC) (**Figure S2.1D-E**). Secretion probabilities are computed from the sigmoid-transformed output of the ExoGRU's predictions. The ECX group consists of accurately predicted EC sequences with a secretion probability exceeding 95%, while the ICX group consists of true IC sequences with a secretion probability below 5%. Therefore, both the ECX and ICX groups, by definition, exclude any falsely-predicted sequences. Furthermore, we assessed whether the predictive power of our model was consistent across broad small RNA classes and biotypes. **Figure S2.1F** displays similarly strong performance metrics for miRNAs (microRNAs), snoRNAs (small nucleolar RNAs), scRNAs (small cytoplasmic RNAs), and tRNAs (transfer RNAS), indicating that ExoGRU is capable of accurately predicting small RNA secretion across all these classes.

We next implemented a variety of approaches to experimentally verify the ability of ExoGRU to capture the small RNA secretory grammar among these sequences. First, we generated an exogenously expressed small RNA library composed of two different sets of sequences: high confidence secreted small RNAs (ECX), and mutated variants of ECX

(MUT). The latter set of sequences was generated by randomly mutating ECX small RNAs, in one or two positions, so that ExoGRU no longer classified them as secreted RNAs. We cloned this library, containing both ECX and MUT sequences, in a lentiviral construct downstream of a U6 promoter (pLKO.1 backbone)[59]. We then transduced the MDA-MB-231 breast cancer cell line, which was among the lines used in our original dataset[50,56,57]. We isolated small RNAs from extracellular vesicles (EV), conditioned media (CM) and intracellular (IC) fractions of this library and performed small RNA sequencing across all samples in biological replicates. We then aligned the resulting reads to the reference library to assess the abundance of each small RNA in the extracellular and intracellular space. It should be noted that expressing small RNAs via an exogenous construct may result in RNA species that (i) are mis-localized and therefore rapidly degraded, and (ii) lack the endogenous molecular context they rely on for successful secretion. Of the 400 pairs tested, in 55 cases, both the ECX and MUT pairs were stably expressed and therefore successfully captured by our assay. In order to assign a secretion probability to each small RNA, we compared its abundance in the extracellular fractions (EV or CM) to intracellular RNA (IC). We observed that the resulting 'enrichment scores' were significantly higher for extracellular small RNAs (ECX) compared to their mutated counterparts (MUT), in both CM and EV fractions (**Figure 2.2A**). We observed a large fraction (93%) of exogenously expressed ECX sequences were indeed secreted; and more importantly, slight modifications to these sequences, guided by ExoGRU, resulted in a substantial and significant drop in their secretion potential. To assess the concordance between experimental measurements and ExoGRU predictions, we used a ROC curve to measure the association between experimental and ExoGRU labels at every classification threshold across the CM enrichment score (**Figure 2.2B**) and EV enrichment score **(Figure S2.2A)**. We used the threshold resulting in a specificity of 0.75 to make “EC” and “IC” calls based on the experimental CM enrichment score. We used the resulting experimental classes to generate a confusion matrix against the ExoGRU labels and calculate performance metrics (**Figure 2.2C**). We also performed a similar analysis for the EV fraction, presented in **Figure S2.2B**, by calculating an experimental EV-enrichment score. Our observations in the EV fraction were similar to the conditioned media, albeit with a lower performance (70% accuracy versus 82% in CM). This was not unexpected since EV purification often suffers from technical variation and the recovered RNA levels are substantially lower.

2.5. The Ability Of ExoGRU To Generalize Its Predictions To Synthetic Sequences

We next sought to determine whether ExoGRU can be used for generation, as opposed to mere classification, of synthetic small RNA sequences that are secreted effectively. Furthermore, we sought to assess whether the patterns learned by ExoGRU based on natural small RNAs is sufficiently generalizable to predict secretion probability of synthetic sequences. To these ends, we randomly generated RNA sequences with an average length of 20nt and dinucleotide frequencies matching those observed in the endogenous small RNAs. We then used ExoGRU to estimate their probability of secretion and selected ~400 sequences that were classified as ‘EC’ (labeled REX for randomly generated high confidence EC) and a similar number that were classified as ‘IC’ (labeled RIX for randomly generated high confidence IC). We synthesized REX/RIX sequences and cloned them similar to the above. Finally, we transduced this library into MDA-MB-231 cells and profiled small RNAs from the conditioned media (CM) and extracellular vesicles (EV). If a given randomly generated sequence was observed in the extracellular fraction, it was given the ‘EC’ label, otherwise, it was labeled as IC. In **Figure 2.2D** and **S2.2C**, we have provided the resulting contingency table comparing the experimental and computational labels for CM and EV fractions respectively. The accuracy of ExoGRU class predictions for these synthetic sequences in the CM fraction was 73%, with 82% sensitivity and 59% specificity. We also used a $\chi 2$ test to calculate a *p*-value for the observed counts (P=3.6e-8).

We were intrigued by the ability of ExoGRU to generalize well to previously unseen sequences and to effectively identify entirely synthetic sequences that are efficiently secreted. Therefore, to independently verify the patterns observed for the REX and RIX sets in our sequencing data, we selected eight REX sequences (REX1 through REX8) and three RIX sequences (RIX1 through RIX3). We cloned these under a Pol III promoter (the same pLKO.1 backbone as the library) and generated MDA-MB-231 cell lines for each construct individually. After isolating small RNAs from EV and IC fractions in biological replicates, we performed quantitative RT-PCR to compare the enrichment of each sequence in the extracellular fraction. We used both the abundance and enrichment of small RNAs in the EV and CM fractions as our selection criteria. We used miR-16, which is abundantly secreted, as an endogenous control in this assay and both IC and EV or CM values were first normalized to miR-16. REX1 and REX5 small RNAs, which were significantly enriched in the extracellular fraction based on small RNA sequencing data (corrected p-values 0.033 and P=0.049 respectively), were further validated as EC-associated small RNAs using targeted RT-qPCR (**Figure 2.2E** and **S2.2D**). Finally, we also tested the expression and secretion of REX1 and REX5 in MDA-MB-231 cells when cloned under a CMV promoter in the BdLV backbone60. To do so we used self-cleaving ribozymes61 to express our REX/RIX sequences under this Pol II promoter. In this case as well, we observed a close to a 100-fold enrichment of REX1 and REX5 in the EV fraction (**Figure 2.2F**). Together, our results validate the performance and utility of ExoGRU both as a predictive model that captures the small RNA secretory grammar and also as a generative model that can nominate synthetic small RNAs that are effectively secreted.

2.6. Gaining Insights Into The RNA Secretory Mechanisms By Dissecting The Grammar Learned By ExoGRU

ExoGRU effectively captures the probability of secretion from the primary RNA sequence alone, which implies the presence of an underlying shared sequence grammar that governs this process. *Cis*-regulatory elements often mediate interactions with master regulators, such as RNA binding protein, to influence the RNA life cycle. In fact, several RNA binding proteins have already been shown to play a direct role in RNA sorting into exosomes[44]. In order to systematically explore the role of RBPs in small RNA sorting and secretion, we first focused on applying motif discovery methods to the ECX and ICX sequences to find highly discriminative and class-specific motifs. We used three separate motif finding strategies, namely MEME[62], Homer[63], and FIRE[64]. We identified multiple sequence motifs that were enriched specifically in the ECX sequences. In parallel, we also used CLIP-seq data from the RNA ENCODE project[65] to identify RBPs whose binding sites are enriched among the secreted RNAs. Using signal and peak-calling results of each RBP, and genome coordinates of ECX and ICX sequences, we sought to identify RBPs that are enriched for interactions with the ECX sequences. We applied the

Mann–Whitney statistical test to detect such significantly greater overall signal values among the ECX and ICX regions. In contrast to the motif analysis, ENCODE's eCLIP data resulted in few if any leads. This is not surprising since CLIP data originates from longer RNAs that are nuclease treated into shorter crosslinked fragments. As a result, the much stronger signal from longer RNAs largely masks bonafide small RNA-RBP interactions. In fact, CLIP analysis for small RNA binding has been reported for only a handful of RBPs, notably AGO2[66] and YBX1[67]. Therefore, for the purpose of this study, we focused our downstream analyses on RBPs with enriched bindings sites (**Figure S2.3A**).

Among the RBP motifs enriched in ECX small RNAs, we focused on YBX1, HNRNPA2B1, and RBM24 binding sites since their associated RBPs are also found within the extracellular space[68]. As shown in **Figure 2.3A**, the known motifs for these RBPs were significantly enriched among cell-free small RNAs, even when controlled for length and dinucleotide content. Re-identification of YBX1 through this approach serves as a validation of our strategy given that it is known to be a major factor in microRNA and small RNA sorting into the exosomal compartment[69]. Similarly, while not as well-characterized, HNRNPA2B1 has also been previously implicated in microRNA sorting[70]. RBM24, on the other hand, does not have a canonical role in RNA secretion; however, it is known to be present within exosomes[71]. To gain deeper insights into these sequence features used by exoGRU, we implemented a signal ablation strategy to investigate the influence of masking the identified motifs on the model's predictions. Specifically, we collected approximately 5,000 sequences from the intracellular (IC) and extracellular (EC) datasets that contained matches to our three specified RBP motifs. We subsequently conducted a comparative analysis of the model's mean secretion probabilities before and after masking or completely removing these enriched motifs linked to the proteins of interest. Notably, this analysis revealed a substantial reduction in the model's secretion probabilities for all three selected motifs (**Figure S2.3B**). Among the previously labeled ECX sequences, more than 94% of them are no longer classified as high-confidence extracellular (ECX) when the YBX1 motif (CCUGGC) is masked, with an average secretion probability drop from 0.97 to 0.52 (**Figure S2.3B**). Additionally, For RBM24 motif (GAGUC) more than 77% of ECX sequences are no longer predicted as ECX (Average secretion probability drop from 0.97 to 0.81). Also, for HNRNPA2B1 motif ([ACU]AG[GU][GU]) more than 67% of previously ECX labeled sequences are no longer ECX (Average secretion probability drop from 0.97 to 0.76; **Figure S2.3B**). These findings were consistent with those from the application of saliency maps and DeepLIFT[72] to sequences containing the specified motifs, reinforcing the crucial role played by these identified motifs in shaping the model's predictions.

To further explore the role of HNRNPA2B1 and RBM24 in small RNA secretion, we used CRISPR-interference to knockdown these RBPs and measure their consequences on the cell-free RNA content. We achieved a 77% knockdown for HNRNPA2B1 and 88% for RBM24 in MDA-MB-231 cells using lentiviral transduction as described in methods. We then isolated RNA from EV, CM and IC compartments for small RNA sequencing. As shown in **Figure 2.3B-C**, silencing HNRNPA2B1 and RBM24 resulted in a significant reduction in the abundance of small RNAs that contained their binding sites in both the extracellular vesicles (EV) and conditioned media (CM) fractions. This observation confirms the involvement of these RNA-binding proteins in RNA sorting and secretion. In addition, to further demonstrate the specificity of HNRNPA2B1 and RBM24 for their targets, we grouped the EV enrichment values of small RNAs based on their matches to HNRNPA2B1 and RBM24 motifs, respectively. **Figure S2.3C** demonstrates that EV enrichment of small RNAs carrying HNRNPA2B1 and RBM24 motifs were significantly decreased upon knockdown of HNRNPA2B1 and RBM24 respectively. Notably, this decrease was specific to their cognate motifs.

We next sought to confirm that, as previously claimed, HNRNPA2B1 sorts small RNAs it binds into exosomes. For this, we took advantage of UV-crosslinking co-immunoprecipitation followed by sequencing. CLIP-seq often includes a nuclease digestion step to footprint RBP binding sites across the transcriptome; however, by omitting this step, the small RNA targets bound by an RBP of interest can be profiled instead. We and others have previously used this approach for other RNA-binding proteins, such as AGO2[66] and YBX1[67]. Visualization of radiolabeled RNA crosslinked to HNRNPA2B1 on a denaturing gel revealed a faint but visible band at the correct size range (**Figure S2.3D**). We extracted these HNRNPA2B1-bound RNAs and performed high-throughput sequencing. Motif analysis of the identified binding site showed a strong and highly significant enrichment of the HNRNPA2B1 motif among the bound small RNAs (**Figure S2.3E**), which serves as a technical quality control. Finally, we asked whether these HNRNPA2B1-bound small RNAs were among those depleted from the exosomal space upon HNRNPA2B1 knockdown. Consistently, we observed a marked reduction in the secretion of these RNA, with a higher statistical significance compared to the HNRNPA2B1 motif analysis (**Figure S2.3F**). Together with the prior reports, our results show that HNRNPA2B1 binding to small RNAs is required for their effective secretion.

2.7. HNRNPA2B1 And RBM24 ExoCLIP Shows Enrichment Of EC Predicted Sequences

The presence of RNA-binding proteins HNRNPA2B1 and RBM24 in extracellular vesicles along with their putative small RNA targets strongly suggests direct interactions within the exosomal space. However, direct evidence of RNA binding and the identity of their target RNAs remained lacking. To tackle this problem, we developed a novel approach for capturing the specific RNA molecules that a given RBP interacts with in the exosomal space. This approach, which we have named exoCLIP, is similar to CLIP-seq but uses UV treatment of conditioned media to crosslink RBP-RNA complexes in the cell-free fraction (**Figure 2.4A**). Using exoCLIP, we sought to demonstrate a direct interaction between HNRNPA2B1 and RBM24 and their target small RNAs. In the case of HNRNPA2B1, we tested both the A2 and B1 isoforms. We transduced MDA-MB-231 cells with FLAG-tagged copies of HNRNPA2, HNRNPB1, and RBM24, respectively. We then performed exoCLIP-seq for each line using FLAG co-immunoprecipitation. We used CLIP Toolkit[73] to call peaks for each of the RBPs using two strategies; one based on sequence coverage or signal, and the other based on crosslinking induced mutations (CIMs). Both strategies yielded between hundreds and thousands of RNA targets, a fraction of which mapped to annotated small RNAs (**Figure S2.4A**). These results indicate that HNRNPA2B1 and RBM24 indeed bind their RNA targets directly in the cell-free space. Interestingly, while we observed some correlation between the HNRNPA2 and HNRNPB1 isoforms, there were also many isoform-specific binding sites for these RBPs (**Figure S2.4B**). In **Figure 2.4B**, we have included examples of small RNAs, in this case tRNA fragments, that are bound by each RBP as evidenced by the exoCLIP signal and the presence of crosslinking induced deletions (CIDs). Since we had selected HNRNPA2B1 and RBM24 based on our analysis of high-confidence predictions for EC and IC RNAs from ExoGRU, we expected these predictions to match the exoCLIP results as well. To assess this possibility, we measured the enrichment of bound small RNAs from each dataset among the ExoGRU predicted EC vs IC small RNAs. As shown in **Figure 2.4C**, we observed a significant over-representation of EC small RNAs that are directly bound by HNRNPA2B1 and RBM24.

2.8. Discussion

Extracellular small RNAs play a key role in intercellular communications and regulation of various biological processes[44,48,49,74,75]. Identifying these specific RNA molecules and understanding their mechanisms of action has led to the discovery of different disease associated biomarkers and therapeutic targets[50–52,74,76–78]. However, our understanding of how these RNA molecules are sorted and delivered into the extracellular space is still limited.

Multiple studies have identified different RBPs responsible for RNA secretion into the extracellular space[69,70]. However, the full mechanisms underlying smRNA delivery are still largely unknown. A recent study comparing intracellular vs. extracellular miRNA profiles found multiple “EXOmotifs” and “CELLmotifs” on miRNA responsible for their secretion from or retention in the cells, suggesting that there are various different motifs and RBPs involved in this process[53]. While the study provided valuable information on miRNA distribution in metabolically important cells, we aimed to further explore the mechanisms behind small RNA sorting in cancer cells using machine learning tools and novel molecular biology approaches.

To further decipher the principles of small RNA delivery to the extracellular space, we asked three specific questions: (1) which RNA sequences are selected and secreted; (2) can we develop a computational model that learns the sequence grammar that underlies RNA secretion; and (3) using this model, can we learn the molecular mechanisms that drive this selection? To tackle these questions, we developed ExoGRU, a deep recurrent neural network, to predict the secretion probability for any small RNA given the primary sequence. We rigorously verified our model's ability to capture the small RNA secretory grammar by testing the impact of ExoGRU-guided targeted mutations on the secretion of endogenous small RNAs. We found the RNA primary sequence to be sufficient to discriminate between the intra- and extracellular small RNAs.

Additionally, we used exoGRU to reveal the regulatory grammar captured by the model. Using motif discovery methods and CLIP-seq data combined with high-confidence ExoGRU predictions, we identified several RBPs that preferentially bind to secreted RNAs and are associated with the RNA sorting process. In addition to recapitulating the known involvement of YBX1, we also demonstrated the role of RBM24 and HNRNPA2B1 in RNA secretion through CRISPR-interference and CLIP-seq. We described exoCLIP, a novel approach to capture direct RBP-RNA interactions in cell free media. Using this method, we successfully characterized RBM24, HNRNPA2 and HNRNPB1 RNA targets in the extracellular space. Our exoCLIP-seq data also aligned with ExoGRU predictions as we saw enrichment of EC associated sequences in these data. Overall, our results demonstrate the performance and utility of ExoGRU as a predictive model that captures the small RNA secretory grammar and provides insights into the role of RBPs in small RNA sorting and secretion.

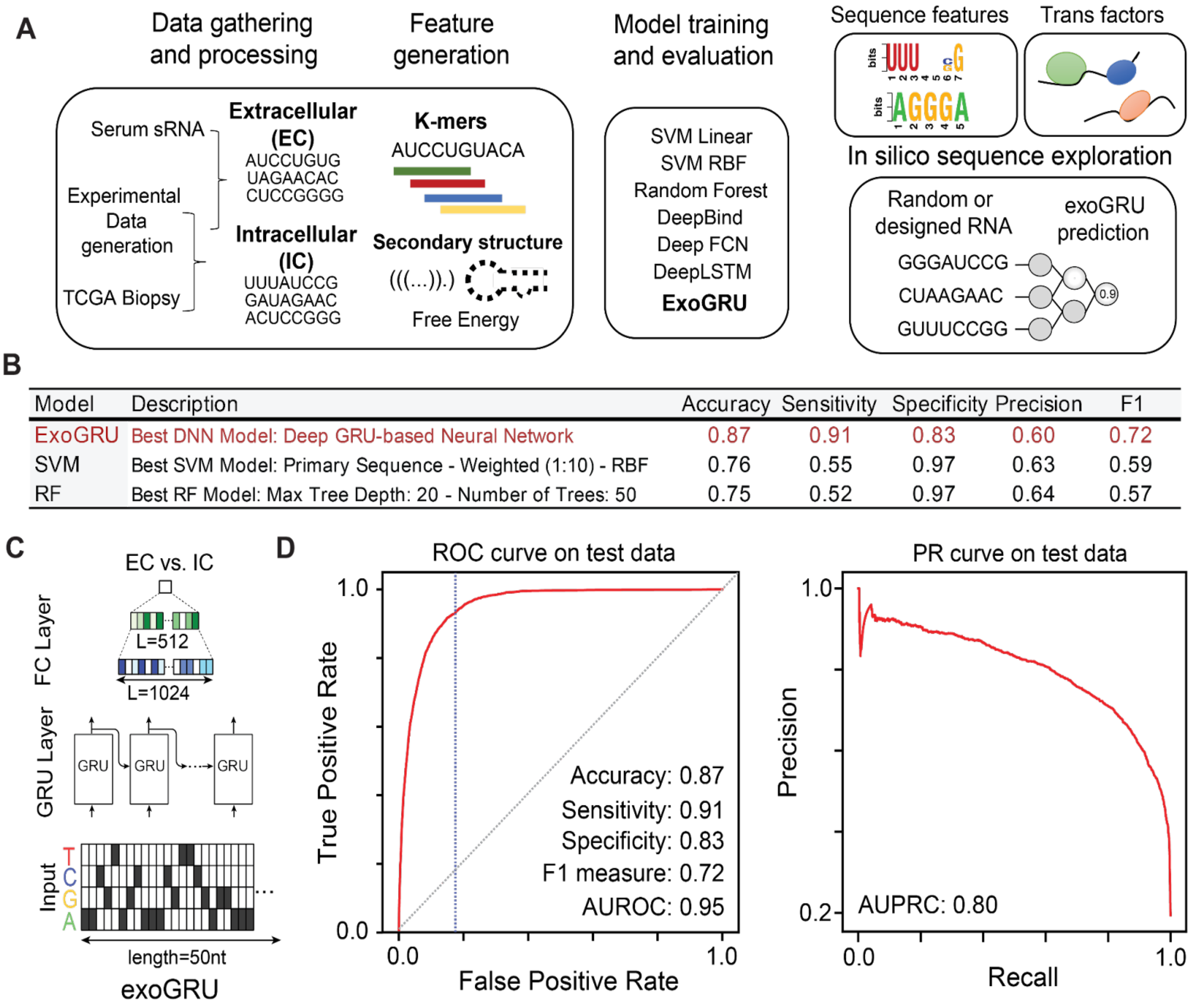
Last but not least, we showed ExoGRU's prediction ability is generalizable to synthetic sequences. This was demonstrated through sequencing and quantitative PCR analysis of randomly-generated but ExoGRU-scored libraries of EC and IC sequences (REX/RIX). The validation process further confirmed the accuracy of the predictions made by the ExoGRU model. Using this feature of ExoGRU we will be able to design fully engineered and efficiently secreted sequences that can be used as biomarkers as well as having further applications in synthetic biology.

2.9. Limitations Of Study

One of the challenges in the biological validation of ExoGRU's findings stems from the expression of small RNA through an exogenous construct. This approach may yield RNA species that (i) experience mis-localization and subsequent rapid degradation, and (ii) lack the crucial endogenous molecular context required for successful secretion. Consequently, not all sequences confidently identified by ExoGRU will be expressed correctly and captured in the extracellular domain. Moreover, the experimental isolation process and sequencing threshold may not efficiently capture lowly abundant secreted RNA, leading to mislabeling these RNA species as intracellular.

The small RNA composition is significantly variable across diverse cell types. Our model was trained on small RNA derived from breast cancer or normal breast tissue, with validation exclusively performed on the MDA-MB-231 cell line. While certain discovered sequences may find expression and validation in other cell types, it is crucial to acknowledge the potential limitations. The model's ability to reproduce similar results may be compromised, emphasizing the need for retraining on new datasets that align more closely with the specific context.

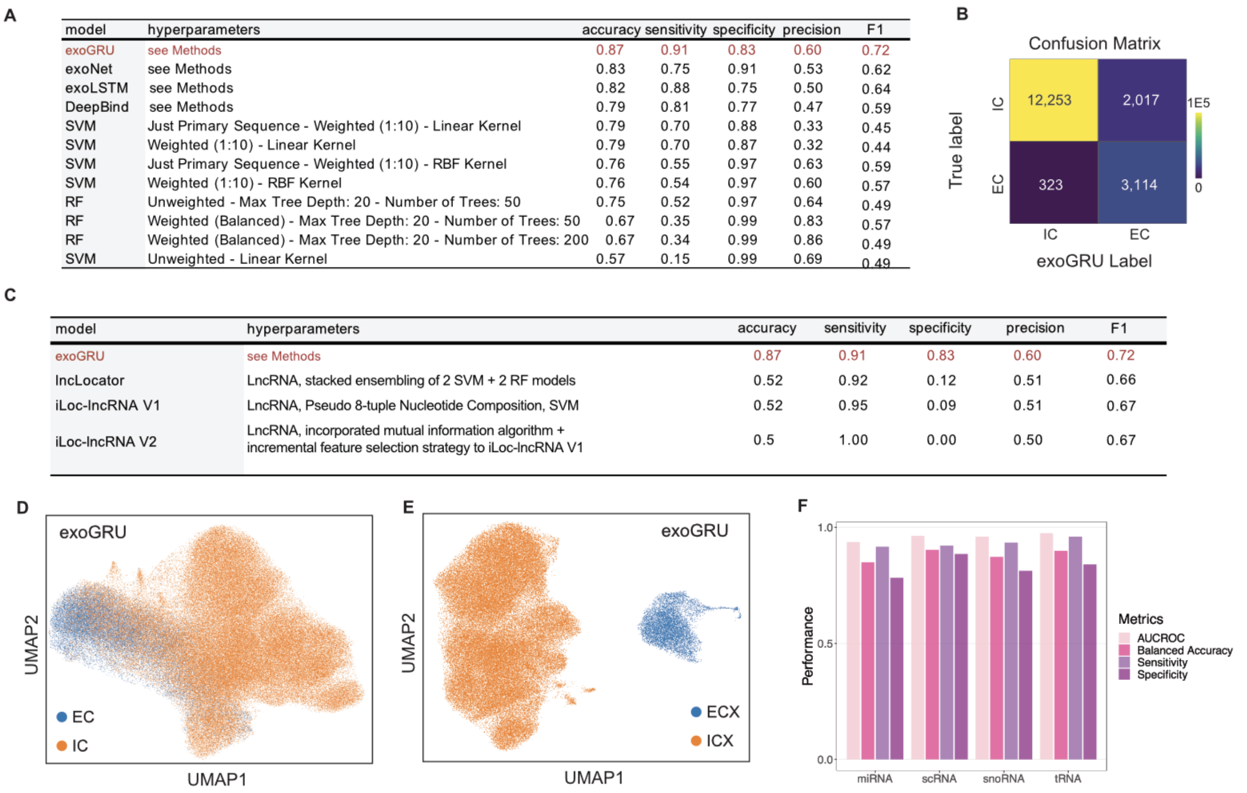
2.10. Figures:



Model	Description	Accuracy	Sensitivity	Specificity	Precision	F1
ExoGRU	Best DNN Model: Deep GRU-based Neural Network	0.87	0.91	0.83	0.60	0.72
SVM	Best SVM Model: Primary Sequence - Weighted (1:10) - RBF	0.76	0.55	0.97	0.63	0.59
RF	Best RF Model: Max Tree Depth: 20 - Number of Trees: 50	0.75	0.52	0.97	0.64	0.57

Figure 2.1. Predicting small RNA secretion from RNA sequence and structural features

(A-B) An overview of our strategy in this study: we used in-house and publicly available data to curate a dataset of intracellular and cell-free small RNA species. Following extensive feature engineering and evaluating various modeling strategies, we selected the best machine learning models for prediction of small RNA secretion. We observed that ExoGRU, a recurrent neural network model, outperforms other models in this task. We then performed feature attribution scoring and model dissection to dissect the cis-regulatory grammar captured by ExoGRU. **(C)** The architecture of ExoGRU following hyperparameter optimization. **(D)** Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the ExoGRU model for the held-out test set. Positive samples are the extracellular (EC) sequences and negative samples are the intracellular (IC) ones. The performance metrics of this model are also listed.



A

model	hyperparameters	accuracy	sensitivity	specificity	precision	F1
exoGRU	see Methods	0.87	0.91	0.83	0.60	0.72
exoNet	see Methods	0.83	0.75	0.91	0.53	0.62
exoLSTM	see Methods	0.82	0.88	0.75	0.50	0.64
DeepBind	see Methods	0.79	0.81	0.77	0.47	0.59
SVM	Just Primary Sequence - Weighted (1:10) - Linear Kernel	0.79	0.70	0.88	0.33	0.45
SVM	Weighted (1:10) - Linear Kernel	0.79	0.70	0.87	0.32	0.44
SVM	Just Primary Sequence - Weighted (1:10) - RBF Kernel	0.76	0.55	0.97	0.63	0.59
SVM	Weighted (1:10) - RBF Kernel	0.76	0.54	0.97	0.60	0.57
RF	Unweighted - Max Tree Depth: 20 - Number of Trees: 50	0.75	0.52	0.97	0.64	0.49
RF	Weighted (Balanced) - Max Tree Depth: 20 - Number of Trees: 50	0.67	0.35	0.99	0.83	0.57
RF	Weighted (Balanced) - Max Tree Depth: 20 - Number of Trees: 200	0.67	0.34	0.99	0.86	0.49
SVM	Unweighted - Linear Kernel	0.57	0.15	0.99	0.69	0.49

C

model	hyperparameters	accuracy	sensitivity	specificity	precision	F1
exoGRU	see Methods	0.87	0.91	0.83	0.60	0.72
lncLocator	LncRNA, stacked ensembling of 2 SVM + 2 RF models	0.52	0.92	0.12	0.51	0.66
iLoc-lncRNA V1	LncRNA, Pseudo 8-tuple Nucleotide Composition, SVM	0.52	0.95	0.09	0.51	0.67
iLoc-lncRNA V2	LncRNA, incorporated mutual information algorithm + incremental feature selection strategy to iLoc-lncRNA V1	0.5	1.00	0.00	0.50	0.67

Supplemental Figure 2.1. ExoGRU confusion matrix and embedding visualization

A) Table compares various quality metrics of ExoGRU's performance against different learning models tested**.** In most categories ExoGRU performs better than all the other models. **B)** The confusion matrix illustrates the comparison between ExoGRU-predicted EC and IC labels and their corresponding true labels from the datasets. A total of 3437 sequences with a true EC label were tested, of which approximately 90% were correctly identified as EC by ExoGRU. Similarly, 14270 sequences with a true IC label were tested, with ExoGRU correctly labeling approximately 86% of them as EC. **C)** Comparison of ExoGRU model with other existing models for prediction of RNA subcellular localization. Quality metrics are also listed for ExoGRU and all other existing models **D)** UMAP projection was used to visualize the 64-dimensional embedding of EC vs IC. EC labeled sequences are those with > 0.5 secretion probability and IC sequences have < 0.5 secretion probability as predicted by ExoGRU. **E)** UMAP projection shows the 64-dimensional embedding of high confidence ECX and ICX calls. ECX labeled sequences are those with > 0.95 secretion probability and ICX sequences are those with < 0.05 secretion probability as predicted by ExoGRU. **F)** Comprehensive analysis of ExoGRU's performance in predicting subcellular localization across various small RNA subtypes including miRNA, scRNA, snoRNA, tRNA.

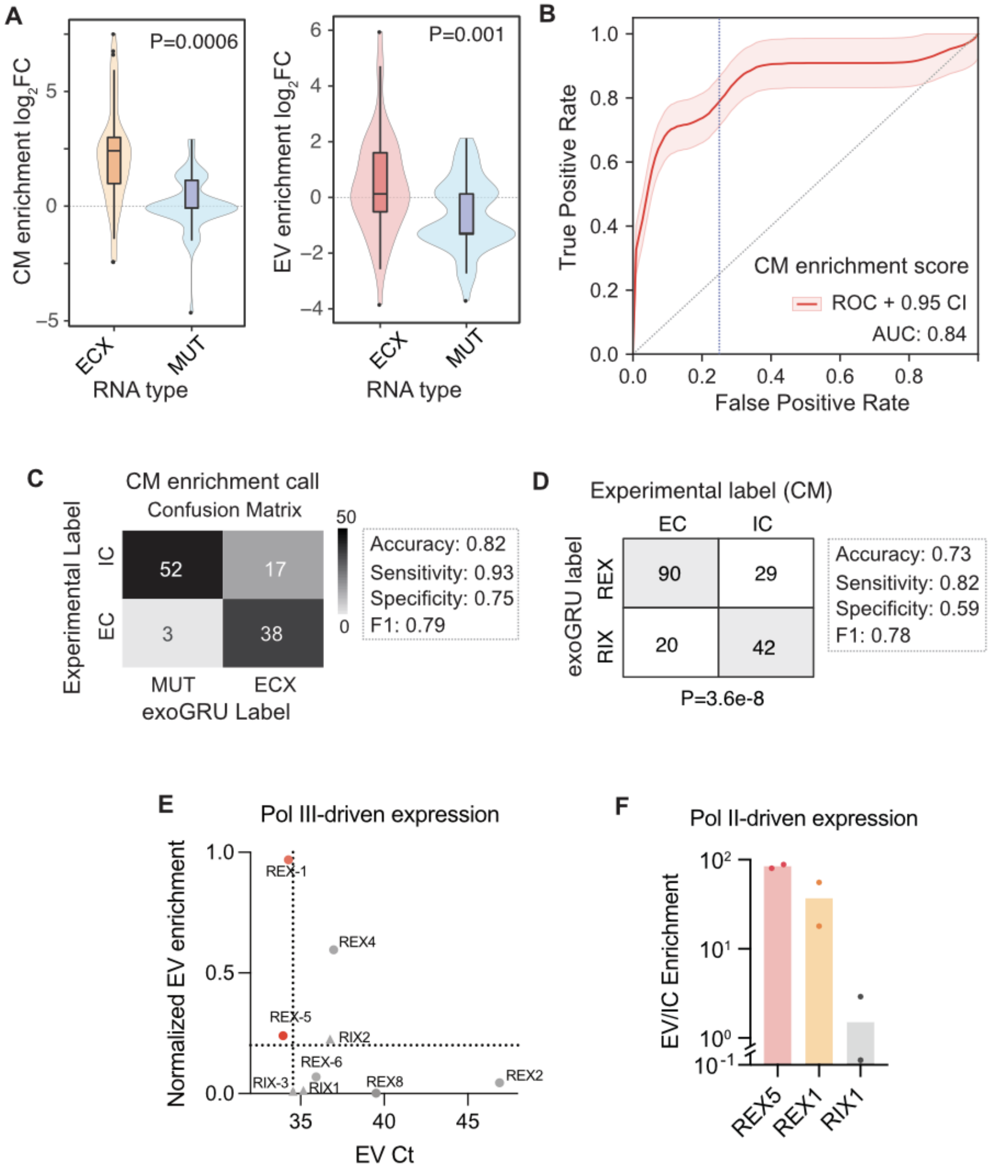
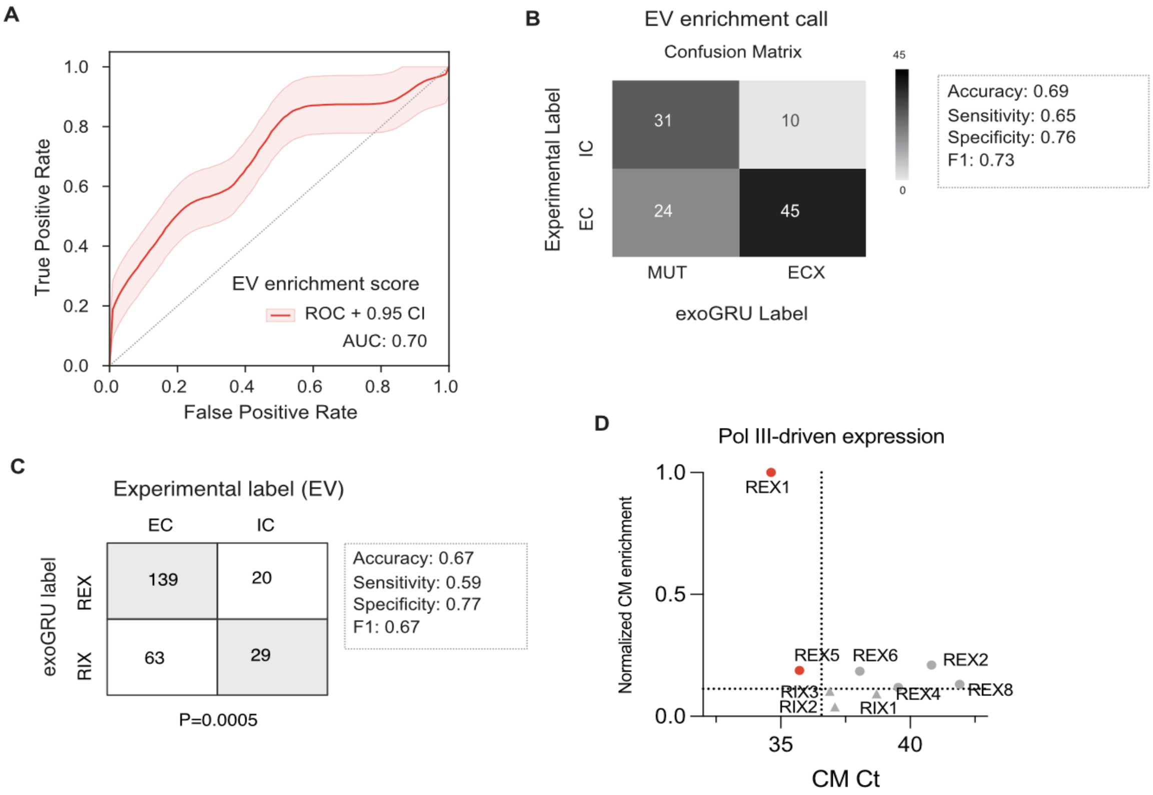


Figure 2.2. Experimental validations of ExoGRU predictions

A) Enrichment scores of ECX vs muted ECX smRNA in CM fractions and EV fractions are shown as Log2 fold change of small RNA abundances in the EV or CM fraction relative to the IC fraction. Total of 55 ECX and 55 matched mutated ECX (MUT) sequences were successfully expressed and used for this analysis. P values are 0.0006 and 0.001 for CM and EV enrichments respectively, calculated using Wilcoxon signed rank test. **B)** ROC curve generated using ECX and MUT experimental CM enrichment score and ExoGRU's localization predictions to measure the association between the experimental vs ExoGRU labels at every classification threshold. The smoothened ROC curve was generated by performing 1000 bootstraps. **C)** EC and IC labels were assigned (Figure caption continued on the next page)

(Figure caption continued from the previous page) to sequences from conditioned media (CM enrichment) using a specificity threshold of 0.75. These experimental labels were subsequently employed to construct a confusion matrix for the classification of ECX and MUT sequences. Performance metrics are provided for this classification. **D)** The presented contingency table illustrates the experimental distribution of ExoGRU generated REX and RIX sequences in CM. The ExoGRU class predictions for these synthetic sequences achieved an accuracy of 73%, with 82% sensitivity and 59% specificity. A χ2 test was applied to calculate a p-value for the observed counts (P=3.6e-8). **E)** Ct Values and Normalized EV Enrichment of REX and RIX Sequences. All sequences were cloned under a RNA polymerase III promoter, and their expression in EV was initially normalized against mir-16. Subsequently, the values were then corrected by their abundance in the IC fraction. The thresholds on Ct and EV enrichment axes (shown as dotted lines) are set as one standard deviation from the average of these values for RIX RNAs. REX-1 and REX-5, highlighted in red, satisfy both constraints (based on their z-scores relative to RIX sequences), with a combined Fisher's P of 1e-11 and 1e-2, respectively. **F)** Independent validation of EV enrichment for REX1, REX5 and RIX1 sequences expressed under RNA polymerase II promoter. The qPCR analysis was conducted in a manner similar to that depicted in Figure 2E.



Supplemental Figure 2.2 ROC curve and confusion matrix for ExoGRU predictions and its experimental validations

A) ROC curve generated using ECX and MUT experimental EV enrichment score and ExoGRU's localization predictions to measure the association between the experimental vs ExoGRU labels at every classification threshold. The smoothened ROC curve was generated by performing 1000 bootstraps. **B)** Using ROC curve analysis, EC and IC labels were assigned to sequences from conditioned media (EV enrichment) using a specificity threshold of 0.75. These experimental labels were subsequently employed to construct a confusion matrix for the classification of ECX and MUT sequences. Performance metrics are provided for this classification. **C)** The presented contingency table illustrates the experimental distribution of ExoGRU generated REX and RIX sequences in EV. The ExoGRU class predictions for these synthetic sequences achieved an accuracy of 67%, with 59% sensitivity and 77% specificity. A χ2 test was applied to calculate a p-value for the observed counts (P=0.0005). **D)** Ct Values and Normalized CM Enrichment of REX and RIX Sequences in the CM fraction. All sequences were cloned under a RNA polymerase III promoter, and their expression in CM was initially normalized against mir-16. Subsequently, the values were further normalized against the corresponding expression of the sequences in IC. REX-1 and REX-5 have significantly lower Ct's than RIX sequences, and they show significantly higher EV enrichment relative to RIX controls, resulting in a combined Fisher's P values of P<1e-100 and P=1e-4 respectively.

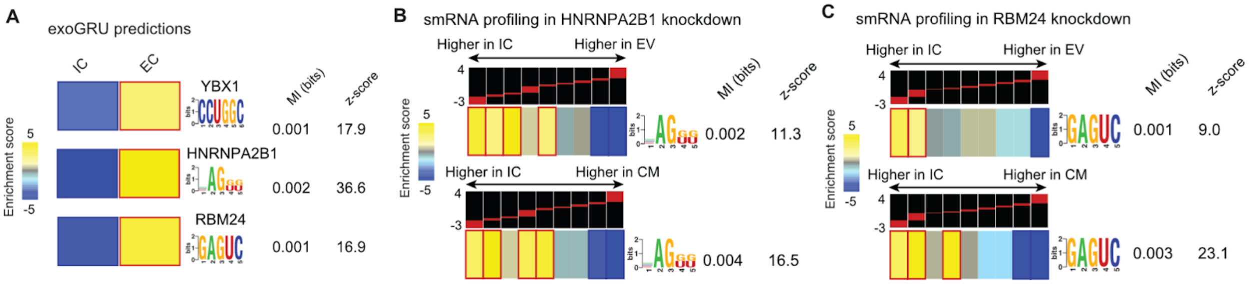
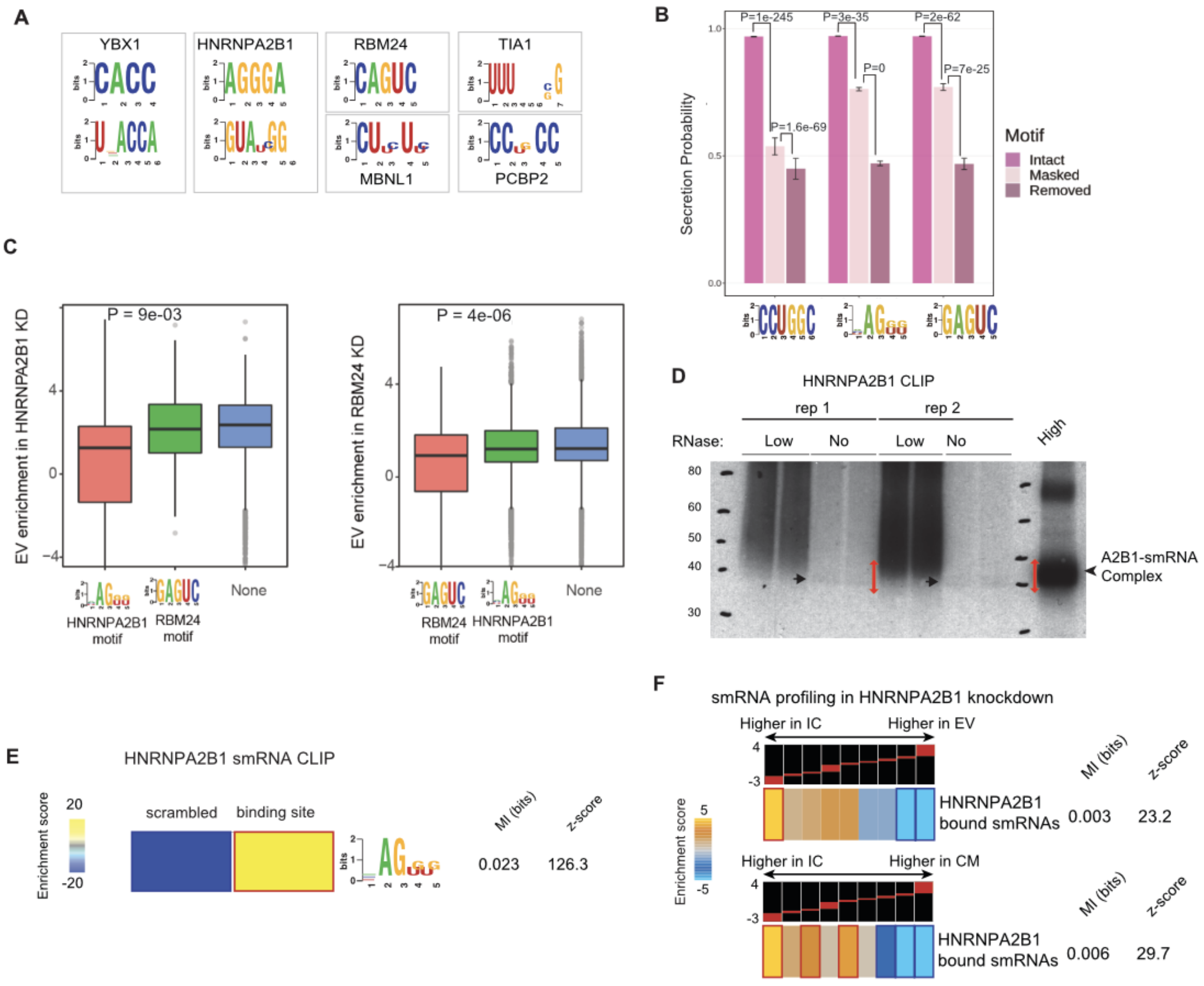


Figure 2.3. Use of ExoGRU in dissecting RNA secretory mechanisms

A) As predicted by exoGRU, YBX1, HNRNPA2B1 and RBM24 motifs are enriched in EC. Each RNA structural motif is shown (far right) along with its pattern of enrichment/depletion across the range of RBPs' expression (far left). In the heatmap representation, a gold entry marks the enrichment of the given motif in its corresponding expression bin (measured by log-transformed hypergeometric P-values), while a light-blue entry indicates motif depletion in the bin. Statistically significant enrichments and depletions are marked with red and dark-blue borders, respectively. Also shown are the mutual information (MI) values and their associated Z scores[64]. Each MI value is used to calculate a z score, which is the number of standard deviations of the actual MI relative to MIs calculated for randomly shuffled expression profiles. Also shown are the mutual information values and their associated *z*-scores measuring the association between motif presence and absence and extracellular enrichment. **B)** Heatmap showing enrichment score of smRNAs containing HNRNPA2B1 motifs in IC, EV, and CM upon decreasing HNRNPA2B1 expression. The log-fold enrichment values were divided into nine equally populated bins, and the enrichment and depletion patterns across the bins were depicted as described in (A). Red and blue borders mark highly significant motif enrichments and depletions, respectively. From left to right, we show the motif names and their sequence information ('motif', in the form of an alphanumeric plot), their associated mutual information values (MI), and their z score **C)** Similar heatmaps showing enrichment score of smRNAs containing RBM24 motifs in IC, EV and CM upon decreasing RBM24 expression.



Supplemental Figure 2.3. ExoGRU RBP motif discovery and CLIPseq to identify small RNA secretion mechanisms

A) RNA binding protein motifs found by exoGRU to be enriched in ECX sequences and their corresponding RBPs. **B)** Comparative analysis of the model's secretion probability before and after masking (or deleting) enriched RBP motifs in the dataset, including 80 sequences with the CCUGGC motif, 304 sequences with the GAGUC motif, and 1888 sequences with the [ACU]AG[GU][GU] pattern. Each bar in the graph illustrates the average secretion probability before and after the masking or deletion of motifs, with the standard error of the mean (SEM) represented by the error bars atop each bar. Statistical significance was assessed using a Wilcoxon test, comparing the secretion probabilities before and after motif masking. To mask the matched motifs, we set the corresponding inputs to 0 across one-hot encoded channel; to remove the motif, we simply deleted that portion of the input sequence. **C)** The first graph illustrates the enrichment of small RNAs containing HNRNPA2B1 motifs, RBM24 motifs, or neither in extracellular vesicles isolated from (Figure caption continued on the next page)

(Figure caption continued from the previous page) MDA cells with reduced expression of HNRNPA2B1 protein. A notable reduction in the enrichment of small RNAs containing HNRNPA2B1 motifs is observed in comparison to those containing RBM24 motifs (p-value = 9e-03). The second graph depicts a similar analysis but in MDA samples with depleted RBM24 protein. A significant decrease in the enrichment of small RNAs containing RBM24 motifs is observed compared to those containing HNRNPA2B1 motifs (p-value = 4e-06). P-values are calculated using Mann-Whitney U test. The Y axis represents the ratio of EV/IC of these motifs in the KD vs CTRL cell line. **D)** Image of radiolabeled RNA bound to HNRNPA2B1 via CLIP. Protein-RNA complexes were treated with no, low and high RNase after crosslinking. The black arrows point to the faint but visible small RNA band in the no RNAse lane. To retrieve the small RNAs bound by HNRNPA2B1, we excised the membrane in the 40-50kDa range for the no-RNase lanes, which corresponds to the artificially created small RNA-HNRNPA2B1 complex in the high-RNase lane (this range is marked by a double-ended red arrows). **E)** Heatmap illustrates the enrichment of HNRNPA2B1 motifs within HNRNPA2B1 binding sites identified through CLIP-seq, compared to scrambled sequences (with di-nucleotide frequency held constant). Red and bolded borders show statistically significant enrichments, as determined by a hypergeometric test (corrected $P < 0.05$). MI value and associated *z*-score are shown. **F)** Heatmap showing pattern of enrichment or depletion of HNRNPA2B1-bound small RNA sequences (captured by CLIP-seq) in EV and CM fractions upon HNRNPA2B1 knockdown. The panels with black bins show how the sequences are partitioned into equally populated bins based on their EV and CM enrichment measures, going from left (lowly expressed in EV/CM) to right (highly expressed in EV/CM). In the heatmap representation, a gold entry marks the enrichment of the HNRNPA2B1-bound small RNA in its corresponding EV or CM expression bins (measured by log-transformed hypergeometric *P*-values), while a light-blue entry indicates HNRNPA2B1-bound small RNA depletion in the bin. Red and blue borders mark highly significant motif enrichments and depletions, respectively. MI value and associated *z*-score are shown.

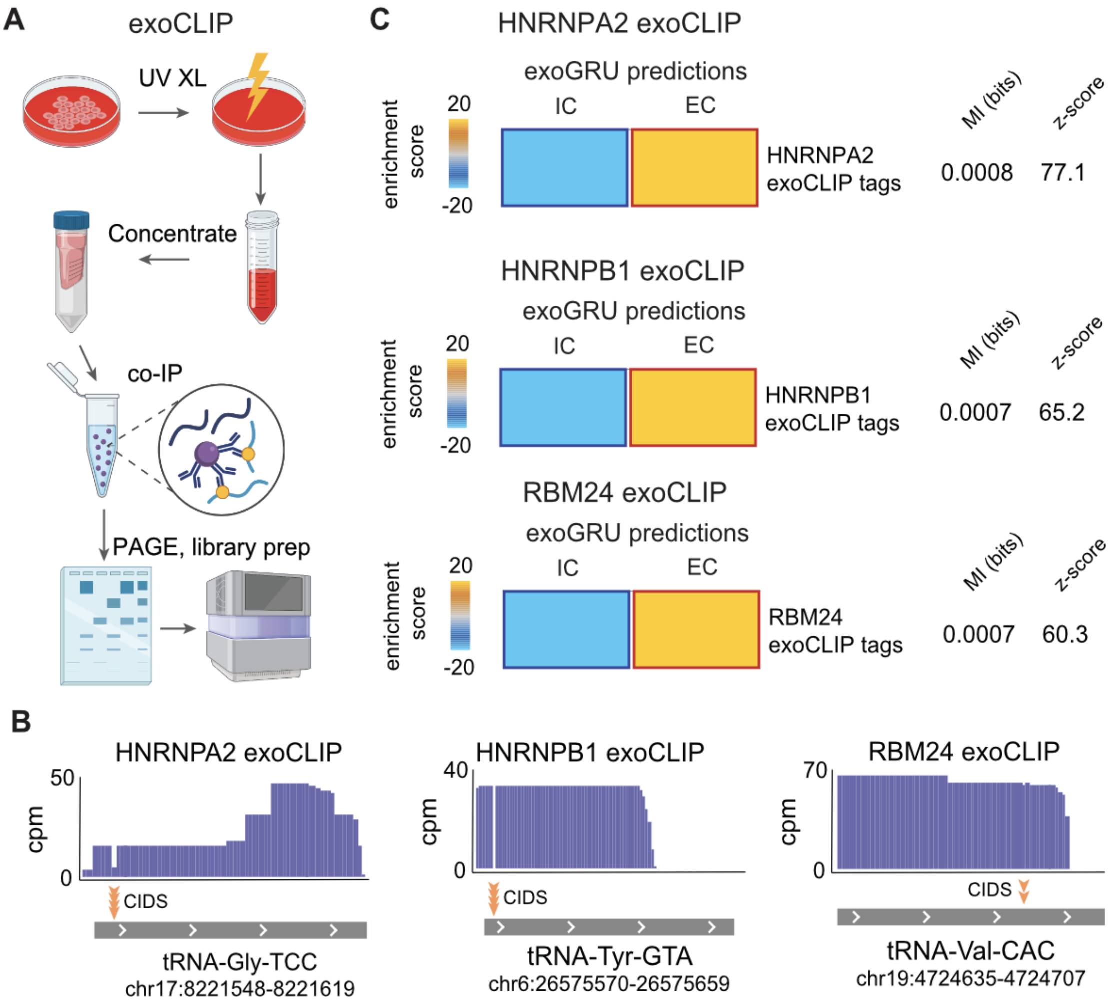
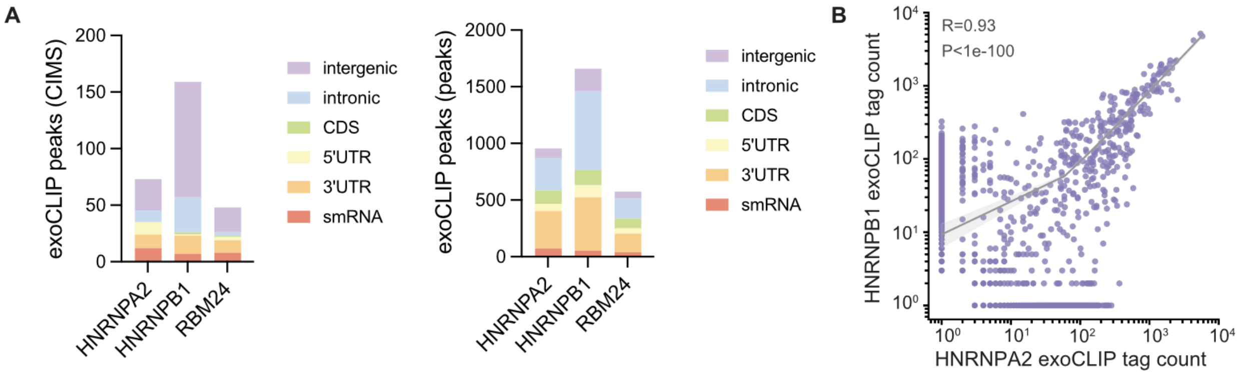


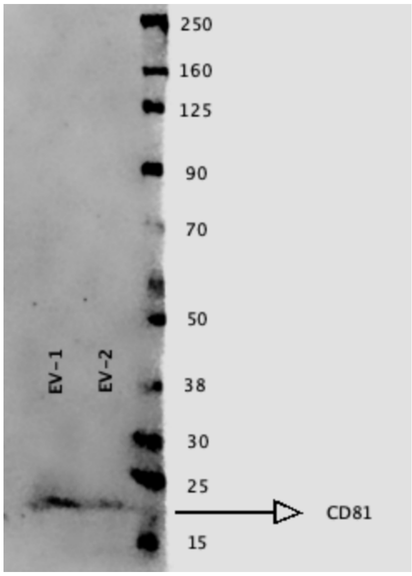
Figure 2.4. Applying exoCLIP to look at the enrichment of HNRNPA2B1 and RBM24 bound smRNA sequences in cell free media

A) Overview of exoCLIP workflow: UV treatment of conditioned media to cross link RBP-RNA complexes and using co-IP to pull down the RBP-RNA complexes of interest followed by RNA library preparation and sequencing. **B)** Examples of tRNA fragments that are associated with HNRNPA2, HNRNPB1, and RBM24 proteins, as extracted from exoCLIP data. The position of crosslinking induced deletions (CIDs) are also highlighted in each case by the yellow arrows. In total, the HNRNPA2 exoCLIP yielded 34 unique reads, with 23 of them exhibiting CIDs at a statistically significant level (P value = 0). The HNRNPB1 and RBM24 exoCLIPs each resulted in 88 unique reads, where 87 reads from HNRNPB1 and 2 reads from RBM24 showed CIDs (P value = 0) P-values are calculated by the CTK package[73]. **C)** Heatmaps illustrate enrichment levels of ExoGRU predicted EC and IC smRNAs in smRNA targets extracted from HNRNPA2, HNRNPB1 and RBM24 exoCLIP. Red and bolded borders show statistically significant enrichments, as determined by a hypergeometric test (corrected P < 0.05).
MI value and associated *z*-score are shown.



Supplemental Figure 2.4. Annotated exoCLIP sequencing results

A) Use of CLIP Toolkit to analyze the HNRNPA2, HNRNPB1 and RBM24 exoCLIP results. For each RNA-binding protein (RBP), two distinct methods were applied to identify peaks. The first plot shows crosslinking induced mutations (CIMs) while the second plot relies on the peak signal or sequence coverage. These plots reveal a substantial abundance of RNA targets associated with each RBP, and these RNA targets were further categorized into various small RNA (smRNA) subtypes.**B)** Scatter plot indicates a high correlation between the distribution of HNRNPA2 and HNRNPB1 tags extracted from exoCLIP data (R=0.93). The plot also shows specific binding sites for these two isoforms.



Supplemental Figure 2.5. Validation of PEG precipitation method in EV isolation

A western blot image illustrating the presence of CD81 protein, a common exosomal marker in EV samples isolated from MDA-MB-231 conditioned media using the PEG precipitation method.

2.11. Materials And Methods

Data Availability

The sequencing data is available in the Gene Expression Omnibus database (GEO: GSE230012).

Cell Culture

All cells were cultured in a 37°C 5% CO2 humidified incubator. The MDA-MB-231 (ATCC HTB-26) breast cancer cell line, and 293T cells (ATCC CRL-3216) were cultured in DMEM high-glucose medium supplemented with 10% FBS, penicillin, streptomycin, and amphotericin B.

All the lentiviral constructs were co-transfected with pCMV-dR8.91 and pMD2.G plasmids using TransIT-Lenti (Mirus) into 293T cells, following manufacturer's protocol. Virus was harvested 48 hours post-transfection and passed through a 0.45 μm filter, and added to target cells 24 hours after they were seeded.

MDA-MB-231 Cells With RBP Knockdowns

Gene knockdowns were performed by first transducing MDA-MB-231 with dCas9-KRAB construct via lentiviral delivery of: pHR-UCOE-EF1a-dCas9-HAxNLS-XTEN80-KRAB-p2a-mCherry. MDA- dCas9-KRAB expressing cells were then sorted by FACS isolation of mCherry-positive cells. Guide RNA sequences for CRISPRi-mediated gene knockdown were cloned into pCRISPRia-v2 (Addgene #84832)[79] via BstXI-BlpI sites (see Supplementary table 1 for sgRNA sequences). After transduction with sgRNA lentivirus, MDA-MB-231 cells were selected with 2 μg/mL puromycin (Gibco). Knockdown of target genes was assessed by reverse transcription of total RNA to cDNA (Maxima H Minus RT, Thermo), then using sequence specific primers along with PerfeCTa SYBR Green

SuperMix (QuantaBio) per the manufacturer's instruction. HPRT was used as an endogenous control (see Supplementary table 1 for primer sequences).

MDA-MB-231 Cells Overexpressing Flag-Tagged RBPs

For generation of flag tagged RBP cell lines, we cloned gblocks containing RBM24, HNRNPA2 or HNRNPB1 and the flag sequences into pLX302-EF1a plasmid via PacI-NheI sites (Supplementary table 2 shows the gblock sequences). Plasmids were delivered to MDA-MB-231 by lentiviral transduction as described above. Expression of RBP-FLAG was assessed using western blot.

MDA-MB-231 Cells Expressing ECX or MUT Sequences Under Pol III Promoter

For expressing ECX/ MUT sequences under U6 promoter we cloned ~400 ECX/ECX_MUT sequence pairs into pLKO.1 plasmid using AgeI and EcoRI sites, and transduced the MDA-MB-231 by lentiviral transduction as described above.

MDA-MB-231 Cells Expressing REX Or RIX Sequences Under Pol III Promoter

For expressing REX1-8 / RIX1-3 sequences under U6 promoter we cloned oligos in Supplementary table 3 into pLKO.1 plasmid using AgeI and EcoRI sites, and transduced the MDA-MB-231 by lentiviral transduction as described above.

MDA-MB-231 Cells Expressing REX Or RIX Sequences Under Pol II Promoter

For cloning REX1-3 and RIX1 sequences under the CMV promoter, we first cloned the ribozyme-small RNA-ribozyme (HH/HDV) cassette[61] into BdLV_Puro_mCherry using PacI and MluI site. We then digested the vector using AsiSI and cloned gblocks containing the sequence of interest (Supplementary table 4) using Gibson assembly. Plasmids were delivered to MDA-MB-231 by lentiviral transduction as previously described.

RT-qPCR For REX/RIX Expression

3.5 ul of isolated RNA was polyA tailed by adding 0.5ul 10X polyA polymerase buffer, 0.5ul 10mM ATP, 0.25ul polyA polymerase (NEB),0.25 ul H2O and incubating at 37°C for 10 minutes. 2.5 ul polyA tailed RNA was then reverse transcribed by adding 0.25ul 10mM dNTPs, 0.1ul 100uM dT T7 primer, 5X RT buffer, 0.15 ul RNAseOUT, 0.25 ul Maxima H Minus RT and 0.75 H2O by incubating at 50C for 15 minutes followed by 85C for 5 minutes. QPCR was done using PerfeCTa SYBR Green SuperMix, T7 primer, and miRNA specific primer as listed on Supplementary table 5. Mir16 primer was used as an endogenous control. To select for EC enriched REX smRNAs, we used both the abundance and enrichment of small RNAs in the EV and CM fractions as a selection criteria. The criteria for Ct and log-fold EV and CM enrichment values were set to be one standard deviation below and above the respective averages of these values for the RIX sequences.

Generation Of SmRNA Libraries

ECX/MUT and REX/RIX oligo pools were ordered from Twist Biosciences. Both oligo pools were separately cloned into pLKO.1-puro plasmid using AgeI and EcorI sites and were transformed into MegaX electrocompetent cells with about 1000X coverage. The smRNA libraries were then transfected to MDA-MB-231 cells using lentivirus as described previously. We maintained a 1000x coverage through the transduction process.

RNA Isolation From Conditioned Media (CM) and Extracellular Vesicles (EV)

MDA cells were seeded in 10 cm or 15 cm plates. The next day the media was removed, and cells were washed with 1X PBS. Cells were then incubated in media prepared with exosome depleted FBS (cat# A2720801) in standard cell culture conditions for 48 hours.

After 48 hours, media was collected and spun down at 500g and passed through 0.4 um to remove any cells. For RNA isolation from conditioned media (CM), we took 1 ml of cell free media and performed RNA isolation using zymo research *Quick*-cfRNA Serum & Plasma Kit (cat# R1059).

The rest of the media was used for exosome isolation. We took advantage of an EV enrichment method using polyethylene glycol (PEG) as outlined in patent# EP2839278B1[80]. by adding Polyethylene Glycol 10000 (PEG, HR2-607) to 10 % final and overnight incubation at 4C. The next day PEG/media mixture was spun down at 3000g at 4C for 1 hour. We then removed the supernatant and proceeded to Zymo Research Quick-RNA Microprep Kit (cat#R1051) for RNA isolation from the EV pellets observed at the bottom of the tube. To confirm the efficacy of our EV isolation through the PEG precipitation method, we present a western blot image of CD81, a well-established exosomal marker, detected in EVs isolated from MDA-MB-231 conditioned media (Figure S2.5). The blot was stained with 1:2000 anti CD81 and 1:10000 IRDye® 800CW Goat anti-Mouse IgG Secondary Antibody and visualized using Licor Odyssey XF.

To delve further into whether the RNA captured in the CM fraction is actively secreted through mechanisms involving lipoprotein complexes, rather than being a result of passive mechanisms like cell death, we conducted a repeated experiment as shown in Figure 2A. This time, we divided the media into two conditions: one with RNAse treatment and one without RNAse treatment. We then extracted RNA from the conditioned media and performed small RNA sequencing. Our analysis revealed that there were no significant differences in RNA sequences (correlation coefficient, R=0.81) between the two treatment conditions. This observation suggests that the RNA sequences in our ECX sequences are mostly secreted through EVs and are thus protected by the RNAse treatment, as they remain relatively unaffected by the enzymatic degradation.

RNA Isolation From Cells

Total RNA for RNA-seq and RT-qPCR was isolated using the Zymo Research Quick-RNA Microprep Kit (cat#R1051) with in-column DNase treatment per the manufacturer's protocol.

ExoCLIP

ExoCLIP of flag tagged RBM24, HNRNPA2, HNRNPB1 MDA-MB-231 cells was done by seeding 12M cells divided in four 15 cm cell culture plates for each cell line in DMEM media as described above. After 24 hours, the media was changed to DMEM with exosome free FBS. 48 hours after the media change, the conditioned media was collected and transferred to 50 ml falcon tubes and spun once at 500g and once at 2000g for 10 min at 4C to clarify the media from any cells. Clarified media was then transferred to 15 cm plates for crosslinking at 200 mJ/cm2 254 nm UV. After the first UV exposure we swirled the media and repeated the crosslinking step for a second time. Crosslinked clarified media was transferred to the centricon plus-70 filter 10K MWCO (millipore sigma UFC701008) and concentrated according to the manufacturer's protocol.

To the concentrated media we added protease inhibitor, SuperaseIN, EDTA, 1M Tris-HCl pH 7.5, and anti-flag magnetic beads (CAT# A36797) and incubated with rotation for 20 hours at 4°C. Beads were magnetized and washed sequentially with cold low salt wash buffer, high salt wash buffer and PNK buffer two times each. This was followed by a PNK mediated dephosphorylation step (2.5ul 10X PNK buffer, 2ul 10X T4 PNK (10unit/ul), 0.5ul SuperaseIN, 20ul H2O) for 20 min at 37C and sequential washes with PNK buffer and high salt wash buffer.

The de phosphorylated RNA-protein complexes were then poly A tailed using yeast PAP, PAP buffer, ATP and SuperaseIn (Jena 600U/ul) at 22 for 5 min. The poly A tailed RNA-Protein complex was then labeled by N3-dUTP, and yeast PAP, PAP buffer and SuperaseIn at 37C for 20 min. Beads were then washed by high salt wash buffer and PBS. The N3-labeled smRNA was stained with 1mM 800cw DBCO at 22C for 30 min. Beads were magnetized and washed with high salt HITS-CLIP WB and PNK buffer respectively, and then resuspended in 20ul of 1X NuPAGE loading buffer + 50mM DTT final concentration diluted in PNK buffer and heated at 75C for 10 min. Beads were placed on the magnet for elution. The eluted RNA protein complexes were then frozen in -80 and later used for WB analysis as described below.

Low Salt Wash Buffer:

1X PBS (TC grade, no Mg++, no Ca++)

1% IGEPAL CA-630

High Salt Wash Buffer:

5X PBS (TC grade, no Mg++, no Ca++)

1% IGEPAL CA-630

1X PNK Buffer**:**

50mM Tris-Cl pH 7.4

10mM MgCl2

1% IGEPAL CA-630

Western Blotting

Eluted RNA-protein complexes from above were run on SDS-PAGE using 4-12% Bis-Tris NuPAGE gels and transferred to protran BA-85 nitrocellulose membrane. The membrane was briefly rinsed in PBS and placed in a sheet protector and imaged with a Licor Odessey instrument.

Protein K Digest And RNA Capture

The RNA-protein complexes imaged as described above appeared as a diffused signal with a modal size of ~15-20kDa above the expected MW of the protein of interest. Average MW of 21 nt long RNA is ~7kDa. Poly(A) tail ~20nt (~6.5kDa), therefore the position of the protein-RNA complex that will generate CLIP tags longer than 20nt is ~14kDa above the expected MW of the protein. HNRNPA2-Flag and HNRNPB1-Flag run at 38 and 39 kDa respectively and RBM24-Flag runs at ~28. Therefore, we cut between 55-85 kDa for HNRNPA2 and HNRNPB1 lanes and 39-70 kDa for RBM24 lane. The MDA only (no flag) lane was cut from 39-85 kDa.

The cut membranes were each transferred to a 1.5 ml Eppendorf tube and treated with 12.5 ul Proteinase K in 200 ul Proteinase K digestion buffer at 55C for 45 min. The samples were then quickly spun down and the 200 ul of supernatant was transferred to a clean Eppendorf tube. Samples were then adjusted for salt by adding 19 ul 5M NaCl and 11 ul H2O per 200 ul sample.

To capture the RNA, we used 30 ul Oligo d(T)25 dynabeads (Invitrogen cat#61002) per IP. Beads were washed 2X with Proteinase K buffer before use. We transferred ~200ul salt-adjusted samples to the beads and incubated at 25C at 300 RPM for 20 min with occasional shaking of 1350 RPM. We then washed the samples/beads 2X with cold high salt wash buffer and 2X with PBS, magnetized and removed the supernatant. RNA was eluted by incubating the beads in 8 ul TE elution buffer at 50C for 5 min. Beads were magnetized and 7.5 ul of eluted RNA was transferred to clean PCR tubes.

Small RNA Library Preparation

Small RNA library preparation for samples taken from exoCLIP was done using Takara Bio SMARTer smRNA-Seq Kit (cat# 635029) with a few modifications. Since our RNA was already poly-A tailed, we skipped this step in the protocol and moved to the cDNA synthesis. We also wanted to incorporate UMI in our cDNA, so we added 2.5ul smRNA mix 1 and 1ul of 10uM dT-UMI RT primer to our 7.5 ul poly A- tailed smRNA and incubated at 75C for 3 min and then placed on ice for 5 min. We then performed reverse transcription as described in the kit's protocol. In the PCR step we also added a 2 ul, 10 uM Universal reverse primer (P7) to the PCR mix and added the 78 ul mix to each cDNA sample. We then added the 2 ul index forward primer to each sample and incubated as described in the protocol. We purified the PCR product using Zymo Research select-a-size MagBead (cat#D4084-50). ExoCLIP sequencing primers and barcodes are listed under supplementary table 6.

Library preparation for RNA isolated from ECX/MUT or REX/RIX transduced MDA-MB-231 cells and small RNA library from HNRNPA2B1 KD and RBM24 KD MDA-MB-231 cells were prepared using an in-house small RNA library preparation. 7.5 ul RNA was polyadenylated using 1ul NEB 10X polyA pol buffer, 1 ul 10 mM ATP, 0.25 ul RiboLock (40 u/ul), 0.25 ul E. coli PolyA pol. 5 u/µl (NEB), and incubated at 16C for 5 min, and then put on ice for maximum 5 min before proceeding to cDNA synthesis. We added 1 ul of RT primer, incubated at 72C for 3 min before putting on ice. We then prepared RT mix on ice using 2 ul 5X RT buffer with DTT (Thermo), 1 ul 10 mM dNTP, 4 ul 5M Betaine, 1 ul Maxima H- RT 200 u/μl (Thermo), 0.25 ul RiboLock 40 u/μl (Thermo), 1 ul 10 μM TSO-UMI primer. We incubated the RT mix at 42C for 30 min and then at 85C for 5 min. The cDNA amplification was carried by using 19 ul cDNA from previous step, 20 ul 5X Phusion HF buffer (Thermo), 2 ul 10 mM dNTP, 2 ul 12 μM Takara Fwd PCR primer, 2 ul 12 μM 12 μM Takara Rev PCR primer, 1 ul Phusion HS II pol. 2 u/μl (Thermo), and 48 ul H2O. We then ran a PCR reaction for cDNA amplification as follows: 30 sec @ 98C - [10 sec @ 98C - 10 sec @ 65C - 5 sec @ 72C]xN cycles (N determined by performing a qPCR). RT, TSO and Takara primers are listed under supplementary table 7. Sample barcodes and indices are listed under supplementary table 8.

PCR reaction was purified through a MN NucleoSpin Gel & PCR Cleanup column (cat#740609) and eluted in 30 μl of water. We ran samples on a 8% TBE gel for 35 mins at 180V and stained the gel with 1X GelGreen in 1X TBE for 2 mins then imaged under Blue light. We cut the fragment of interest (150-200 bp) and placed the gel slices in a 0.5 mL tube with a hole pierced by a 18G needle. Spun the tube in a 1.5 mL tube until the gel was passed through the hole. We then added 400 μl of the DNA gel extraction buffer (10 mM Tris pH 8, 300 mM NaCl, 1 mM EDTA) to the gel, vortexed and froze on dry ice for 30 min, and then thawed overnight on a rotator. Next day we transferred the gel slurry to a Costar filter spin-column and spun at maximum speed until all liquid has passed through. We added 1.5 μl of GlycoBlue and 500 μl of isopropanol to the DNA solution, and put in -80C for 1 hour, then spun at 4C for 30 minutes, air dried for 10 minutes, and resuspended and incubated in 10 μl of 10 mM Tris pH 8 for 10 minutes.

Sequencing And Analysis

Libraries were quantified using Qubit HS dsDNA kit, and also ran on an Agilent bioanalyzer HS DNA Chip or HSD1000 tapestation. All libraries were sequenced as SE65 runs on Illumina Hiseq 4000 at UCSF Center for Advanced Technologies. Reference sequences were collated as fasta files (for ECX and MUT sequences, and RIX and REX sequences separately). UMI-tools[81] (v1.0) and cutadapt[82] (v3.5) were used to extract UMIs and remove linker sequences. BWA[83] (v0.7.17) was used to align reads and then duplicates were removed using the extracted UMIs. Reads mapping to each sequence in the bam file was then counted and DESeq2 (v1.24) was then used to normalize and compare the extracellular fractions to the intracellular fraction and log2 fold-change were reported for each group of sequences.

For the analysis of ECX/MUT experiment in figure 2A, we included 55 ECX RNAs and 55 matched MUT RNAs. For this analysis, we required the RNA to be present in a third of samples in both EV and CM fractions so that the logFC values were meaningful.

HITS-CLIP

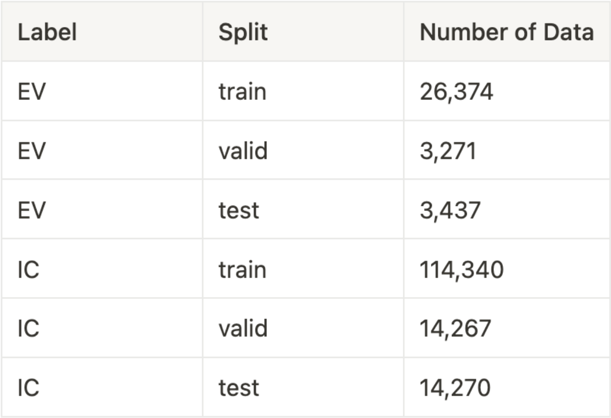
HITS-CLIP for endogenous HNRNPA2B1 was done as described by (Licatalosi et al., 2008)[84] with the modifications previously used for YBX1 small RNA CLIP (Goodarzi et al., 2015)[67]. MDA-MB-231 cells were UV-crosslinked at 400 mJ/cm^2 before cell lysis. Samples with and without RNase treatment were immunoprecipitated with an anti-HNRNPA2B1 antibody (Thermo, PA5-34939) for protein-RNA complexes. RNase treatment was as follows: RNase A (Affymetrix 70194Z, 9,063 units/mg, 4.89 mg/mL); low RNase was 1:2500 and high RNase was 1:50 dilutions, the no RNase was 0 RNase.

Polyphosphatase (Lucigen) was incubated with smRNA samples before ligation and PCR amplification with primers described by (Goodarzi et al., 2015)[67]. Constructed libraries were sequenced on the Illumina HiSeq2000 at the Rockefeller University Genomics Center. The resulting library was then analyzed using the CLIP toolkit (CTK)[73]

Data Acquisition

To train our predictive models, we sourced small-RNA sequences data for intracellular and exosome-specific predictions from three reliable sources: Goodarzi et al. (GSE114366)[50], Extracellular RNA Communication Consortium Atlas (exRNA Atlas), and The Cancer Genome Atlas (TCGA). We first used the GSE114366 data, which was generated to investigate the roles of intracellular and extracellular small-RNAs in breast cancer. We extracted small-RNA-seq data of intracellular small-RNAs (IC) and small-RNAs present in extracellular vesicles (EV) from 8 different breast cancer cell lines. Although the small-RNA selection and secretion machinery may differ among different cell types and states, we assumed that there are some general and common mechanisms that exist in cells. Therefore, we merged all of the IC data, regardless of the cell line, and imported 30,093,690 IC-resided small-RNA sequences to our IC dataset. We also integrated all the EV data and collected 6,127,883 EV-resided small-RNA sequences to our EV dataset. Second, we imported 67,511,039 EV-resided small-RNA sequences from the exRNA Atlas, which were extracted from the serum part of blood cells of 12 samples. Third, we collected 1,8488,703 IC-resided miRNA sequences from the TCGA dataset which were extracted from normal cells of different tissues across the body. We selected miRNA-seq data and not RNA-seq data because our main focus in this research is to investigate the selection and secretion processes of small-RNAs inside a cell. Notably, to avoid data leakage due to sequence similarity, we filtered out redundant (i.e. highly similar) sequences from the aggregated dataset. The table below presents information regarding the total number of samples utilized from each dataset, as well as the percentage of samples held out from each dataset.

Table 2.1. Breakdown of number of samples used from each dataset



Label	Split	Number of Data
EV	train	26,374
EV	valid	3,271
EV	test	3,437
IC	train	114,340
IC	valid	14,267
IC	test	14,270

Data Preprocessing

After integrating data from three distinct sources, we performed several preprocessing steps to clean and organize the data. We first removed sequences that were present in both the IC and EV datasets, assuming that they belonged to the EV, as exported small-RNA sequences can also exist intracellularly. We then eliminated sequences that were less than 18 nts or more than 50 nts, as our focus was on small-RNAs. We also removed any sequences that contained "N" in their primary sequence in order to decrease ambiguity in our dataset. We eliminated duplicated sequences and sequences that were a substring of a bigger one. For example, if we had both ACGU and UACGU sequences in our dataset, we removed the former one. To further clean the data, we used the MEME-Suite's dust tool to mask and delete sequences that carried low-complex regions. The dust tool helps to identify and remove any non-informative regions in the sequences such as repetitive regions. After all of these preprocessing steps, we had 33,083 unique EV small-RNAs and 1,318,795 unique IC small-RNAs that were ready to be used for predictive model training.

Feature Generation

We derived multiple features from IC and EV small-RNA primary sequences to train the classifiers, utilizing the ViennaRNA package to predict RNA secondary structures and free energies for each RNA within our dataset. Two distinct secondary structure representations were generated using ViennaRNA. The first format, created using the RNAfold module, employed a dot-bracket notation. In this notation, nucleotides were represented as either single-stranded, indicated by a dot (.), or double-stranded, denoted by open and closed brackets (e.g., "(" or ")").

The second representation format, referred to as the bulge-graph notation, was created using the forgi module. This representation categorized RNA secondary structures into five distinct types: five-prime (f), three-prime (t), stem (s), interior loop (i), multiloop segment (m), and hairpin loop (h).

To expand the representation of these sequences, we developed an 8-term notation that encapsulated information from both the primary sequence and secondary structures. In this notation, each sequence was represented using the characters {A, C, T, G, a, c, t, g}, where uppercase characters indicated double-stranded nucleotides characterized by bracket secondary structures, and lowercase characters represented single-stranded nucleotides indicating dot secondary structures.

In summary, for each small RNA in our IC and EV dataset, we had four distinct sequence representations: the primary sequence, the dot-bracket secondary sequence, the bulge-graph secondary sequence, and the nucleotide-based secondary sequence. Additionally, we obtained the predicted free energy for each RNA.

These sequences were further analyzed by extracting K-mers (K=1-4, 5, 7) from both the primary and secondary sequences mentioned above. K-mers, or k-grams, represent substrings of length k within any given sequence. For example, for 3-mers, combinations such as "...", "..(", ".(.", "(..", "(.)", "..)", ".).", and "())" were generated for the dot-bracket representation.

To ensure uniformity in our dataset, we normalized the features, including K-mer frequencies and free energy, based on the length of each sequence. Additionally, we included the length of the sequence as a feature.

Given that the majority of our sequences had lengths less than 50 base pairs (bp), sequences exceeding 50 bp in length were truncated, and smaller sequences were padded to achieve a fixed length. This preprocessing step ensured that our data was consistent and ready for input into our model.

Predictive Models

The predictive models examined in this research are divided into two categories: classical machine learning methods and deep learning methods. Classical methods trained/tested in this research include support vector machines (SVMs) and random forests (RFs). On the other hand, deep learning methods include models which are inspired by convolutional neural networks (CNNs) and recurrent neural networks (RNNs). It should be noted that due to the limitations of machine learning models (SVMs and RFs) in handling sequence-based data, all sequence-typed features were removed from the final design matrix for the machine learning experiments.

Support Vector Machines

Support Vector Machines (SVMs) are a family of supervised machine learning algorithms that are commonly used for linear and non-linear classification tasks, as well as for regression tasks. In this study, we conducted an ablation study to evaluate the effectiveness of different input feature spaces and kernel types for SVMs. Specifically, we defined six different training scenarios, which varied in terms of the input feature space and kernel type used.

To evaluate the effectiveness of non-Kmer features, we used two different feature sets: one that included all extracted features, and another that included only the K-mers extracted from the primary sequences. Additionally, we tested both linear and Radial Basis Function (RBF) kernels for the SVMs. Furthermore, due to the inherent class imbalance issue existing in the preprocessed dataset, we tried to mitigate this issue by using weighted SVM by weighting the class parameters inversely proportional to their sample frequencies.

Random Forest

In this study, we employed Random Forest, a powerful tree-based machine learning algorithm, to train on the preprocessed dataset. This choice was made due to the algorithm's ability to handle large feature sets and its robustness to overfitting compared to other existing machine learning models. Similar to the experiments conducted with Support Vector Machines (SVMs), we evaluated the effectiveness of both weighted and unweighted Random Forest models. Additionally, we investigated the impact of different tree population sizes on the performance of the algorithm. Specifically, we tested tree population sizes of 50 and 200 while keeping the tree-depth fixed at 20.

ExoGRU

ExoGRU, as the name suggests, consists of multiple GRU units stacked on top of each other. GRUs introduced a simpler alternative compared to LSTMs. They are able to capture relatively long-term dependencies by utilizing gates in order to control the information flow.

A GRU unit computes the hidden state at time step t as follows:

zt = sigmoid(Wz xt + Uz ht-1 + bz)

rt = sigmoid(Wr xt + Ur ht-1 + br)

h't = tanh(Wh xt + Uh(rt*ht-1)+bh)

ht = (1-zt)ht-1+zth't

Where x_t is the input at time step t, h_{t-1} is the previous hidden state, W_z, U_z, W_r, U_r, W_h, U_h are the weight matrices, b_z, b_r, b_h are the bias terms and sigmoid and tanh are non-linear activation functions. The update gate z_t and reset gate r_t are used to control the flow of information into the hidden state h_t, allowing the network to better handle long-term dependencies.

ExoLSTM

ExoLSTM is also another network we employed in this study. The architecture consists of multiple LSTM units stacked on top of each other. Long Short-Term Memory (LSTM) units are a type of recurrent neural network (RNN) that uses a memory cell to store information over a longer period of time. The memory cell is controlled by gates that determine when to store, update, or discard information in the cell.

An LSTM unit computes the hidden state at time step t as follows:

it = sigmoid(Wi xt + Ui ht-1 + bi)

ft = sigmoid(Wf xt + Uf ht-1 + bf)

ot = sigmoid(Wo xt + Uo ht-1 + bo)

ct = ft * ct-1 + it * tanh(Wc xt + Uc ht-1 + bc)

ht = ot * tanh(ct)

Where x_t is the input at time step t, h_{t-1} is the previous hidden state, c_{t-1} is the previous memory cell state, W_i, U_i, W_f, U_f, W_o, U_o, W_c, U_c are the weight matrices, b_i, b_f, b_o, b_c are the bias terms, and sigmoid and tanh are non-linear activation functions. The input gate i_t, forget gate f_t, output gate o_t and cell state c_t are used to control the flow of information into the hidden state h_t, allowing the network to better handle long-term dependencies.

ExoCNN

ExoCNN is a variant of convolutional neural networks (CNNs) designed to generate predictions from sequences. The architecture of ExoCNN is composed of several layers, including convolution, pooling and fully connected layers, each of which contains tunable weights and biases. One key aspect of the ExoCNN architecture is the use of "conv blocks" as firstly defined in VGG[85] which are composed of multiple consecutive convolution layers followed by a max-pooling operation. In the max-pooling operation, the maximum value is computed for each window of size 2 in the "conv block"’s output matrix, this helps to summarize spatial information into the output while retaining the spatial information. Following the "conv blocks" and max-pooling operations, the output of the last max-pooling operation is flattened and fed to a classifier head with 2-layer fully connected neural network. Similar to the convolution layers, rectified linear activation functions are used in the head. The number of neurons in the hidden layers of ExoCNN's classification head are 1024 and 128 respectively. Finally, the output of the last layer is passed through a sigmoid function which generates a (secretion) probability for the input sequence.

Model Training

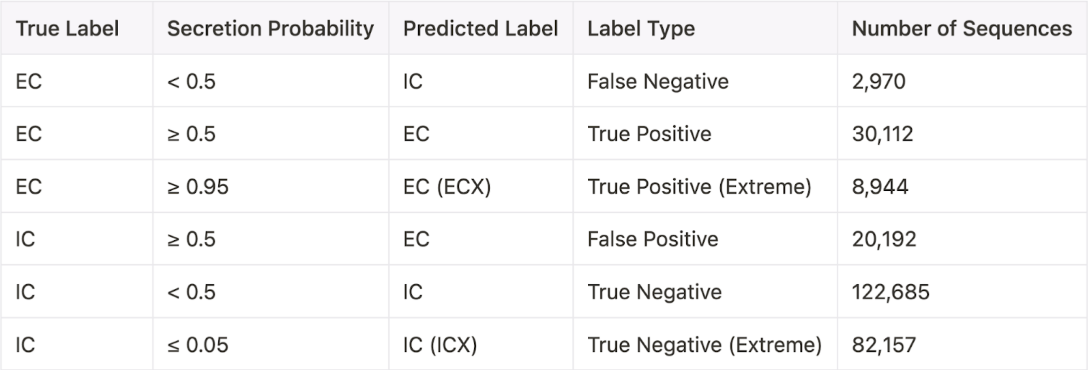
As shown in Table 1, the sample frequency of the IC class is significantly higher than the EV class, which results in a common problem in machine learning known as class imbalance. To address this issue, we downsampled the IC dataset to balance the class frequencies. Additionally, we employed the weighted cross-entropy (WCCE) loss function for training our ExoCNN model in which each class weight is inversely proportional to its sample frequency. The original EV dataset and the downsampled IC dataset were used to train our predictive models. To assess the performance of deep learning models, we performed stratified train/validation/test split with proportions of 0.8, 0.1, and 0.1 on our preprocessed dataset.

We used the Adam optimizer with a learning rate of 0.001 for 100 epochs for all DL models. We employed a batch size of 128 during training. To prevent overfitting, we employed early stopping and learning rate decay techniques during the training process. To initialize the weights of each layer in the network, we used the Xavier initializer. Additionally, we used L1 and L2 regularization techniques with a lambda value of 1e-6 to further prevent overfitting.

Motif Discovery And Enrichment

ExoGRU works within a binary classification framework, each sequence receiving a secretion probability computed from the sigmoid-transformed output of the network's predictions. As demonstrated in Niculescu-Mizil and Caruana et al[86], neural networks trained for binary classification tasks typically yield well-calibrated probabilities, implying that the probabilities generated by ExoGRU serve as reliable estimations of the confidence we place in the model's predictions. Therefore, after training the network with small-RNA sequences, two sets of sequences were identified that the model was highly confident about being secreted or not. Sequences of extracellular vesicles (ECs) are assigned as ECX if the calculated probability exceeds 0.95. Similarly, intracellular (IC) sequences are designated as ICX if the associated probability falls below 0.05 (Table 2).

Table 2.2. Number of smRNA sequences categorized as EC(X) and IC(X) and their corresponding secretion probability as predicted by ExoGRU



True Label	Secretion Probability	Predicted Label	Label Type	Number of Sequences
EC	< 0.5	IC	False Negative	2,970
EC	≥ 0.5	EC	True Positive	30,112
EC	≥ 0.95	EC (ECX)	True Positive (Extreme)	8,944
IC	≥ 0.5	EC	False Positive	20,192
IC	< 0.5	IC	True Negative	122,685
IC	≤ 0.05	IC (ICX)	True Negative (Extreme)	82,157

In order to find motifs more accurately, we removed highly similar sequences from the ECX and ICX sets using the MEME-Suite's purge tool. To find the optimal similarity score threshold, we experimented with different thresholds and checked the number of sequences and removed ones for each threshold. Finally, we used a similarity score threshold of 50. The extreme sequences (ECX and ICX) were clustered based on edit distance and cosine distance. We found motifs based on both unclustered and clustered ECX and ICX, but the results were the same, so we eliminated the clustering step from the analysis pipeline. To perform an exhaustive motif search, we used several motif finders and tested various configurations of the tools. Three motif finding tools were able to discover motifs in the small-RNA sequences: MEME, Homer, and FIRE. We used these tools with three different input sets: ECX only, ECX vs ICX, and ECX vs randomly generated sequences that preserved di-nucleotide frequency. Using these three motif finding tools, three different configurations, and several parameter tuning, we found 10 motifs that were enriched in the ECX sequences. These motifs were related to previously known RNA-binding proteins that are involved in the secretion machinery, and were presented in Figure 4.

With the discovery of the ECX and ICX sequences and the corresponding motifs, we continued our research by identifying secretion-related RNA-binding proteins in two distinct ways. First, we compared the discovered motifs with already known ones in the literature and databases. This allowed us to identify any previously known motifs that were enriched in the ECX sequences and related to known RNA-binding proteins involved in the secretion machinery. Second, we analyzed the eCLIP-seq data of the ENCODE project to identify binding sites of human's RNA-binding proteins. This allowed us to identify any potential RNA-binding proteins that may be involved in the secretion of small-RNAs based on their binding sites in the ECX and ICX sequences.

Motif Comparison

We compared the 10 discovered motifs with known motifs of RNA-binding proteins to detect the proteins that are highly likely to bind to each motif and participate in the secretion machinery. To do this, we used three databases of Ray2013, RBPDB, and ATtRACT, and the MEME-Suite's Tomtom tool to find RNA-binding proteins (RBPs) that significantly bind to our discovered EV-enriched motifs. This motif comparison process gave us 7 proteins that are highly likely to bind to our secretion-related motifs, as shown in Figure 4. As previously mentioned, two of these proteins have already been verified to be involved in the secretion machinery. This comparison process helps us to identify potential players in the secretion process and further investigate them.

RBP Binding Sites Analysis

We aimed to identify RNA binding proteins (RBPs) in the ENCODE database that may have greater interactions with extreme EV sequences, as opposed to IC sequences. Our hypothesis is that these proteins may play a role in the secretion machinery. We filtered out proteins that did not have signals (bigWig file) or peaks (BED file) as their output type and that were not based on the GRCh38 reference genome. This resulted in a final selection of approximately 150 proteins.

To begin, we determined the maximum signal value at each nucleotide position for a specific protein if we have multiple experiments (bigWig files). Next, we extracted signal values for nucleotide positions that overlapped with peak regions, separately for IC and extreme EV sequences. We then used the Mann-Whitney statistical test to compare these two sets of signal values and calculate a p-value to determine if the EV signals were significantly greater than the IC ones.

To obtain comparable signal intensity values and gain a deeper understanding of the interactions between EV extreme sequences and RBPs, we evaluated various scoring methods. In our initial analysis, we obtained signal values (scores) for EV sequences and assigned zero values to regions that did not overlap with peak regions for a specific protein. We also applied this method to IC sequences. This resulted in many zero values in our scores, and the Mann-Whitney test showed a significant sensitivity to the mean in these scenarios.

To address this issue, we modified our approach. Instead of using all peak regions, we applied it to the union of IC and extreme EV regions. This eliminated many zero values from the scores and allowed us to better understand the natural behavior of RBPs. We found that they tend to bind to EV extreme sequences with high signal values and to IC extreme sequences with moderate signal values on average. We then used the Benjamini-Hochberg (BH) method to adjust our p-values and identified proteins with adjusted p-values less than 0.05 as being involved in the EV secretion machinery.

After analyzing the interactions between different proteins and RNA sequences, we also took an intra-protein approach to the problem. To obtain information within each sample (IC vs EV), we used extreme sequences of both IC and EV groups with a secretion probability greater than 0.9 that overlapped with peaked regions. We extracted several features including: the number of EV and IC extreme sequences overlapping with the peaked regions, the total length of each overlapping extremes with peaked regions, the total sum of the signal values for each overlapping extremes, and the mean value of signals for each of the extremes. With this data, we could assess the robustness and reliability of our results as a sanity check and also identify any potential outliers related to the secretion machinery. To do this, we used median absolute deviation (MAD), Z-test, and Percentile rank.

Model Interpretability

A number of strategies have been developed in recent years to help interpret neural network models. For simpler models, such as DeepBind, the convolutional kernels themselves were used to represent features captured by the model[58]. However, for the more complex architectures, including deeper CNNs or RNNs, that are trained on heterogeneous data, customized feature importance analyses are employed. Most famously, DeepLIFT uses a variation of integrated gradient to select those partial sequences across inputs that are most important for the model's prediction and then performs a motif discovery in them, using TF-MoDIsco[72]. In other words, DeepLIFT massively reduces the space in which motif discovery is performed by removing the sequences and parts of sequences that are not informative for the model. Motif discovery can then be effectively performed to identify the features that the model is learning in these sequences. In our case, however, since the sequences are already short for small RNAs, the DeepLIFT scoring is not needed, and we can directly perform motif discovery. However, motif discovery is only performed on sequences that the model is confidently classifying (i.e. ECX). In other words, for these sequences, the model has learned strong features that enable it to make a correct and confident prediction. This approach is very different from performing motif discovery on the initial labels, both in theory and practice. By focusing on the ECX sequences, we are strongly enriching the signal from these sequence features. This is crucial because RNA secretion is a complex process with multiple pathways and many players involved.

To investigate ExoGRU in a more fine-grained way, we extracted approximately 5,000 IC and EC sequences that contain a specific motif associated with RBM24, namely GAGUC. These selected sequences were collectively labeled as "gaguc-intact" and served as the focus of our investigation. We employed DeepLIFT tool to assess the significance of different sequence regions in influencing the model's predictions. Furthermore, we conducted a masking and ablation procedure on the gaguc-intact sequences, called them gaguc-masked and gaguc-removed sequences to show that the model relies on this motif for its prediction. These sets of modified sequences were then used as input to the model, enabling us to examine any changes in the model's predictions. Similarly, We applied the same procedure on the other two motifs (CCUGGC, and [ACU]AG[GU][GU]) as well.

REFERENCES

1. Schreiber, R.D., Old, L.J., and Smyth, M.J. (2011). Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science *331*, 1565–1570. 10.1126/science.1203486.

2. Bruns, C.J., Harbison, M.T., Kuniyasu, H., Eue, I., and Fidler, I.J. (1999). In vivo selection and characterization of metastatic variants from human pancreatic adenocarcinoma by using orthotopic implantation in nude mice. Neoplasia *1*, 50–62. 10.1038/sj.neo.7900005.

3. Dubrot, J., Du, P.P., Lane-Reticker, S.K., Kessler, E.A., Muscato, A.J., Mehta, A., Freeman, S.S., Allen, P.M., Olander, K.E., Ockerman, K.M., et al. (2022). In vivo CRISPR screens reveal the landscape of immune evasion pathways across cancer. Nat. Immunol. *23*, 1495–1506. 10.1038/s41590-022-01315-x.

4. Liu, Y., Hu, Y., Xue, J., Li, J., Yi, J., Bu, J., Zhang, Z., Qiu, P., and Gu, X. (2023). Advances in immunotherapy for triple-negative breast cancer. Mol. Cancer *22*, 145. 10.1186/s12943-023-01850-7.

5. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. 10.1186/s13059-014-0550-8.

6. Vadrevu, S.K., Chintala, N.K., Sharma, S.K., Sharma, P., Cleveland, C., Riediger, L., Manne, S., Fairlie, D.P., Gorczyca, W., Almanza, O., et al. (2014). Complement c5a receptor facilitates cancer metastasis by altering T-cell responses in the metastatic

niche. Cancer Res. *74*, 3454–3465. 10.1158/0008-5472.CAN-14-0157.

7. Medler, T.R., Murugan, D., Horton, W., Kumar, S., Cotechini, T., Forsyth, A.M., Leyshock, P., Leitenberger, J.J., Kulesz-Martin, M., Margolin, A.A., et al. (2018). Complement c5a fosters squamous carcinogenesis and limits T cell response to chemotherapy. Cancer Cell *34*, 561-578.e6. 10.1016/j.ccell.2018.09.003.

8. Nabizadeh, J.A., Manthey, H.D., Steyn, F.J., Chen, W., Widiapradja, A., Md Akhir, F.N., Boyle, G.M., Taylor, S.M., Woodruff, T.M., and Rolfe, B.E. (2016). The complement c3a receptor contributes to melanoma tumorigenesis by inhibiting neutrophil and CD4+ T cell responses. J. Immunol. *196*, 4783–4792. 10.4049/jimmunol.1600210.

9. Zhang, J., and Ney, P.A. (2009). Role of BNIP3 and NIX in cell death, autophagy, and mitophagy. Cell Death Differ. *16*, 939–946. 10.1038/cdd.2009.16.

10. Goodarzi, H., Elemento, O., and Tavazoie, S. (2009). Revealing global regulatory perturbations across human cancers. Mol. Cell *36*, 900–911. 10.1016/j.molcel.2009.11.016.

11. Enjuanes, L., Gorbalenya, A.E., de Groot, R.J., Cowley, J.A., Ziebuhr, J., and Snijder, E.J. (2008). Nidovirales. In Encyclopedia of Virology (Elsevier), pp. 419–430. 10.1016/B978-012374410-4.00775-5.

12. Murray, P.D., McGavern, D.B., Pease, L.R., and Rodriguez, M. (2002). Cellular sources and targets of IFN-gamma-mediated protection against viral demyelination and neurological deficits. Eur. J. Immunol. *32*, 606–615. 10.1002/1521-4141(200203)32:3<606::AID-IMMU606>3.0.CO;2-D.

13. Schömel, N., Hancock, S.E., Gruber, L., Olzomer, E.M., Byrne, F.L., Shah, D., Hoehn, K.L., Turner, N., Grösch, S., Geisslinger, G., et al. (2019). UGCG influences glutamine metabolism of breast cancer cells. Sci. Rep. *9*, 15665. 10.1038/s41598-019-52169-7.

14. Schömel, N., Gruber, L., Alexopoulos, S.J., Trautmann, S., Olzomer, E.M., Byrne, F.L., Hoehn, K.L., Gurke, R., Thomas, D., Ferreirós, N., et al. (2020). UGCG overexpression leads to increased glycolysis and increased oxidative phosphorylation of breast cancer cells. Sci. Rep. *10*, 8182. 10.1038/s41598-020-65182-y.

15. Chang, C.-H., Qiu, J., O'Sullivan, D., Buck, M.D., Noguchi, T., Curtis, J.D., Chen, Q., Gindin, M., Gubin, M.M., van der Windt, G.J.W., et al. (2015). Metabolic competition in the tumor microenvironment is a driver of cancer progression. Cell *162*, 1229–1241. 10.1016/j.cell.2015.08.016.

16. Zheng, S., Li, H., Li, Y., Chen, X., Shen, J., Chen, M., Zhang, C., Wu, J., and Sun, Q. (2023). The emerging role of glycolysis and immune evasion in gastric cancer. Cancer Cell Int. *23*, 317. 10.1186/s12935-023-03169-1.

17. Meissl, K., Macho-Maschler, S., Müller, M., and Strobl, B. (2017). The good and the bad faces of STAT1 in solid tumours. Cytokine *89*, 12–20. 10.1016/j.cyto.2015.11.011.

18. Hix, L.M., Karavitis, J., Khan, M.W., Shi, Y.H., Khazaie, K., and Zhang, M. (2013). Tumor STAT1 transcription factor activity enhances breast tumor growth and immune suppression mediated by myeloid-derived suppressor cells. J. Biol. Chem. *288*,

11676–11688. 10.1074/jbc.M112.441402.

19. Cerezo, M., Guemiri, R., Druillennec, S., Girault, I., Malka-Mahieu, H., Shen, S., Allard, D., Martineau, S., Welsch, C., Agoussi, S., et al. (2018). Translational control of tumor immune escape via the eIF4F-STAT1-PD-L1 axis in melanoma. Nat. Med. *24*, 1877–1886. 10.1038/s41591-018-0217-1.

20. Budhwani, M., Mazzieri, R., and Dolcetti, R. (2018). Plasticity of Type I Interferon-Mediated Responses in Cancer Therapy: From Anti-tumor Immunity to Resistance. Front. Oncol. *8*, 322. 10.3389/fonc.2018.00322.

21. Yi, S., Yan, Y., Jin, M., Bhattacharya, S., Wang, Y., Wu, Y., Yang, L., Gine, E., Clot, G., Chen, L., et al. (2022). Genomic and transcriptomic profiling reveals distinct molecular subsets associated with outcomes in mantle cell lymphoma. The Journal of Clinical Investigation.

22. Wang, T., Yu, Y., Ba, X., Zhang, X., Zhang, N., Wang, G., Bai, B., Li, T., Zhao, J., Zhao, Y., et al. (2023). PTPN18 serves as a potential oncogene for glioblastoma by enhancing immune suppression. Oxid. Med. Cell. Longev. *2023*, 2994316. 10.1155/2023/2994316.

23. Yoon, J., Kim, S., Lee, M., and Kim, Y. (2023). Mitochondrial nucleic acids in innate immunity and beyond. Exp. Mol. Med. *55*, 2508–2518. 10.1038/s12276-023-01121-x.

24. Cai, D., Wang, J., Gao, B., Li, J., Wu, F., Zou, J.X., Xu, J., Jiang, Y., Zou, H., Huang, Z., et al. (2019). RORγ is a targetable master regulator of cholesterol biosynthesis in a cancer subtype. Nat. Commun. *10*, 4621. 10.1038/s41467-019-12529-3.

25. Zhang, L., Xie, B., Qiu, Y., Jing, D., Zhang, J., Duan, Y., Li, Z., Fan, M., He, J., Qiu, Y., et al. (2020). Rab25-Mediated EGFR Recycling Causes Tumor Acquired Radioresistance. iScience *23*, 100997. 10.1016/j.isci.2020.100997.

26. Jeong, B.Y., Cho, K.H., Jeong, K.J., Park, Y.-Y., Kim, J.M., Rha, S.Y., Park, C.G., Mills, G.B., Cheong, J.-H., and Lee, H.Y. (2018). Rab25 augments cancer cell invasiveness through a β1 integrin/EGFR/VEGF-A/Snail signaling axis and expression of fascin. Exp. Mol. Med. *50*, e435. 10.1038/emm.2017.248.

27. Klobučar, M., Sedić, M., Gehrig, P., Grossmann, J., Bilić, M., Kovač-Bilić, L., Pavelić, K., and Kraljević Pavelić, S. (2016). Basement membrane protein ladinin-1 and the MIF-CD44-β1 integrin signaling axis are implicated in laryngeal cancer metastasis. Biochim. Biophys. Acta *1862*, 1938–1954. 10.1016/j.bbadis.2016.07.014.

28. Godwin, T.D., Kelly, S.T., Brew, T.P., Bougen-Zhukov, N.M., Single, A.B., Chen, A., Stylianou, C.E., Harris, L.D., Currie, S.K., Telford, B.J., et al. (2019). E-cadherin-deficient cells have synthetic lethal vulnerabilities in plasma membrane organisation, dynamics and function. Gastric Cancer *22*, 273–286. 10.1007/s10120-018-0859-1.

29. Wu, S.-Y., Fu, T., Jiang, Y.-Z., and Shao, Z.-M. (2020). Natural killer cells in cancer biology and therapy. Mol. Cancer *19*, 120. 10.1186/s12943-020-01238-x.

30. UniProt https://www.uniprot.org/uniprotkb/Q99437/entry.

31. ATP6V1F protein expression summary - The Human Protein Atlas https://www.proteinatlas.org/ENSG00000128524-ATP6V1F.

32. Price, M.J., Patterson, D.G., Scharer, C.D., and Boss, J.M. (2018). Progressive Upregulation of Oxidative Metabolism Facilitates Plasmablast Differentiation to a T-

Independent Antigen. Cell Rep. *23*, 3152–3159. 10.1016/j.celrep.2018.05.053.

33. Ishizuka, J.J., Manguso, R.T., Cheruiyot, C.K., Bi, K., Panda, A., Iracheta-Vellve, A., Miller, B.C., Du, P.P., Yates, K.B., Dubrot, J., et al. (2019). Loss of ADAR1 in tumours overcomes resistance to immune checkpoint blockade. Nature *565*, 43–48. 10.1038/s41586-018-0768-9.

34. Kettwig, M., Ternka, K., Wendland, K., Krüger, D.M., Zampar, S., Schob, C., Franz, J., Aich, A., Winkler, A., Sakib, M.S., et al. (2021). Interferon-driven brain phenotype in a mouse model of RNaseT2 deficient leukoencephalopathy. Nat. Commun. *12*, 6530. 10.1038/s41467-021-26880-x.

35. Gong, Y., Wu, J., Qiang, H., Liu, B., Chi, Z., Chen, T., Yin, B., Peng, X., and Yuan, J. (2008). BRI3 associates with SCG10 and attenuates NGF-induced neurite outgrowth in PC12 cells. BMB Rep. *41*, 287–293. 10.5483/bmbrep.2008.41.4.287.

36. Matsuda, S., Matsuda, Y., and D'Adamio, L. (2009). BRI3 inhibits amyloid precursor protein processing in a mechanistically distinct manner from its homologue dementia gene BRI2. J. Biol. Chem. *284*, 15815–15825. 10.1074/jbc.M109.006403.

37. Boiarsky, R., Haradhvala, N.J., Alberge, J.-B., Sklavenitis-Pistofidis, R., Mouhieddine, T.H., Zavidij, O., Shih, M.-C., Firer, D., Miller, M., El-Khoury, H., et al. (2022). Single cell characterization of myeloma and its precursor conditions reveals transcriptional signatures of early tumorigenesis. Nat. Commun. *13*, 7040. 10.1038/s41467-022-33944-z.

38. Gungabeesoon, J., Gort-Freitas, N.A., Kiss, M., Bolli, E., Messemaker, M., Siwicki, M., Hicham, M., Bill, R., Koch, P., Cianciaruso, C., et al. (2023). A neutrophil

response linked to tumor control in immunotherapy. Cell *186*, 1448-1464.e20. 10.1016/j.cell.2023.02.032.

39. Brown, N.F., and Marshall, J.F. (2019). Integrin-Mediated TGFβ Activation Modulates the Tumour Microenvironment. Cancers (Basel) *11*. 10.3390/cancers11091221.

40. Brummelman, J., Mazza, E.M.C., Alvisi, G., Colombo, F.S., Grilli, A., Mikulak, J., Mavilio, D., Alloisio, M., Ferrari, F., Lopci, E., et al. (2018). High-dimensional single cell analysis identifies stem-like cytotoxic CD8+ T cells infiltrating human tumors. J. Exp. Med. *215*, 2520–2535. 10.1084/jem.20180684.

41. Krishna, S., Lowery, F.J., Copeland, A.R., Bahadiroglu, E., Mukherjee, R., Jia, L., Anibal, J.T., Sachs, A., Adebola, S.O., Gurusamy, D., et al. (2020). Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. Science *370*, 1328–1334. 10.1126/science.abb9847.

42. Cheon, H., Wang, Y., Wightman, S.M., Jackson, M.W., and Stark, G.R. (2023). How cancer cells make and respond to interferon-I. Trends Cancer *9*, 83–92. 10.1016/j.trecan.2022.09.003.

43. Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol. *38*, 954–961. 10.1038/s41587-020-0470-y.

44. O'Brien, K., Breyne, K., Ughetto, S., Laurent, L.C., and Breakefield, X.O. (2020). RNA delivery by extracellular vesicles in mammalian cells and its applications. Nat.

Rev. Mol. Cell Biol. *21*, 585–606. 10.1038/s41580-020-0251-y.

45. Sork, H., Conceicao, M., Corso, G., Nordin, J., Lee, Y.X.F., Krjutskov, K., Orzechowski Westholm, J., Vader, P., Pauwels, M., Vandenbroucke, R.E., et al. (2021). Profiling of Extracellular Small RNAs Highlights a Strong Bias towards Non-Vesicular Secretion. Cells *10*. 10.3390/cells10061543.

46. Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O., and Margalit, H. (2007). Regulation of gene expression by small non-coding RNAs: a quantitative view. Mol. Syst. Biol. *3*, 138. 10.1038/msb4100181.

47. Patil, V.S., Zhou, R., and Rana, T.M. (2014). Gene regulation by non-coding RNAs. Crit. Rev. Biochem. Mol. Biol. *49*, 16–32. 10.3109/10409238.2013.844092.

48. Mittelbrunn, M., and Sánchez-Madrid, F. (2012). Intercellular communication: diverse structures for exchange of genetic information. Nat. Rev. Mol. Cell Biol. *13*, 328–335. 10.1038/nrm3335.

49. Chen, X., Liang, H., Zhang, J., Zen, K., and Zhang, C.-Y. (2012). Secreted microRNAs: a new form of intercellular communication. Trends Cell Biol. *22*, 125–132. 10.1016/j.tcb.2011.12.001.

50. Fish, L., Zhang, S., Yu, J.X., Culbertson, B., Zhou, A.Y., Goga, A., and Goodarzi, H. (2018). Cancer cells exploit an orphan RNA to drive metastatic progression. Nat. Med. *24*, 1743–1751. 10.1038/s41591-018-0230-4.

51. Badowski, C., He, B., and Garmire, L.X. (2022). Blood-derived lncRNAs as biomarkers for cancer diagnosis: the Good, the Bad and the Beauty. NPJ Precis. Oncol. *6*, 40. 10.1038/s41698-022-00283-7.

52. Pardini, B., Sabo, A.A., Birolo, G., and Calin, G.A. (2019). Noncoding rnas in extracellular fluids as cancer biomarkers: the new frontier of liquid biopsies. Cancers (Basel) *11*. 10.3390/cancers11081170.

53. Garcia-Martin, R., Wang, G., Brandão, B.B., Zanotto, T.M., Shah, S., Kumar Patel, S., Schilling, B., and Kahn, C.R. (2022). MicroRNA sequence codes for small extracellular vesicle release and cellular retention. Nature *601*, 446–451. 10.1038/s41586-021-04234-3.

54. Tosar, J.P., Gámbaro, F., Sanguinetti, J., Bonilla, B., Witwer, K.W., and Cayota, A. (2015). Assessment of small RNA sorting into different extracellular fractions revealed by high-throughput sequencing of breast cell lines. Nucleic Acids Res. *43*, 5601–5616. 10.1093/nar/gkv432.

55. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215. 10.1126/science.1090095.

56. Ainsztein, A.M., Brooks, P.J., Dugan, V.G., Ganguly, A., Guo, M., Howcroft, T.K., Kelley, C.A., Kuo, L.S., Labosky, P.A., Lenzi, R., et al. (2015). The NIH extracellular RNA communication consortium. J. Extracell. Vesicles *4*, 27493. 10.3402/jev.v4.27493.

57. Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a shared vision for cancer genomic data. N. Engl. J. Med. *375*, 1109–1112. 10.1056/NEJMp1607591.

58. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the

sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838. 10.1038/nbt.3300.

59. Stewart, S.A., Dykxhoorn, D.M., Palliser, D., Mizuno, H., Yu, E.Y., An, D.S., Sabatini, D.M., Chen, I.S.Y., Hahn, W.C., Sharp, P.A., et al. (2003). Lentivirus-delivered stable gene silencing by RNAi in primary cells. RNA *9*, 493–501. 10.1261/rna.2192803.

60. Amendola, M., Venneri, M.A., Biffi, A., Vigna, E., and Naldini, L. (2005). Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. Nat. Biotechnol. *23*, 108–116. 10.1038/nbt1049.

61. Gao, Y., and Zhao, Y. (2014). Self-processing of ribozyme-flanked RNAs into guide RNAs *in vitro* and *in vivo* for CRISPR-mediated genome editing. J. Integr. Plant Biol *56*, 343–349. 10.1111/jipb.12152.

62. Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

63. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. 10.1016/j.molcel.2010.05.004.

64. Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. Mol. Cell *28*, 337–350. 10.1016/j.molcel.2007.09.027.

65. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719. 10.1038/s41586-020-2077-3.

66. Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486. 10.1038/nature08170.

67. Goodarzi, H., Liu, X., Nguyen, H.C.B., Zhang, S., Fish, L., and Tavazoie, S.F. (2015). Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. Cell *161*, 790–802. 10.1016/j.cell.2015.02.053.

68. Keerthikumar, S., Chisanga, D., Ariyaratne, D., Al Saffar, H., Anand, S., Zhao, K., Samuel, M., Pathan, M., Jois, M., Chilamkurti, N., et al. (2016). ExoCarta: A Web-Based Compendium of Exosomal Cargo. J. Mol. Biol. *428*, 688–692. 10.1016/j.jmb.2015.09.019.

69. Shurtleff, M.J., Temoche-Diaz, M.M., Karfilis, K.V., Ri, S., and Schekman, R. (2016). Y-box protein 1 is required to sort microRNAs into exosomes in cells and in a cell-free reaction. eLife *5*. 10.7554/eLife.19276.

70. Villarroya-Beltri, C., Gutiérrez-Vázquez, C., Sánchez-Cabo, F., Pérez-Hernández, D., Vázquez, J., Martin-Cofreces, N., Martinez-Herrera, D.J., Pascual-Montano, A., Mittelbrunn, M., and Sánchez-Madrid, F. (2013). Sumoylated hnRNPA2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs. Nat. Commun. *4*, 2980. 10.1038/ncomms3980.

71. He, M., Qin, H., Poon, T.C.W., Sze, S.-C., Ding, X., Co, N.N., Ngai, S.-M., Chan, T.-F., and Wong, N. (2015). Hepatocellular carcinoma-derived exosomes promote motility of immortalized hepatocyte through transfer of oncogenic proteins and RNAs. Carcinogenesis *36*, 1008–1018. 10.1093/carcin/bgv081.

72. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. arXiv. 10.48550/arxiv.1704.02685.

73. Shah, A., Qian, Y., Weyn-Vanhentenryck, S.M., and Zhang, C. (2017). CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. Bioinformatics *33*, 566–567. 10.1093/bioinformatics/btw653.

74. Sadik, N., Cruz, L., Gurtner, A., Rodosthenous, R.S., Dusoswa, S.A., Ziegler, O., Van Solinge, T.S., Wei, Z., Salvador-Garicano, A.M., Gyorgy, B., et al. (2018). Extracellular rnas: A new awareness of old perspectives. Methods Mol. Biol. *1740*, 1–15. 10.1007/978-1-4939-7652-2_1.

75. Sohail, A.M., Khawar, M.B., Afzal, A., Hassan, A., Shahzaman, S., and Ali, A. (2022). Multifaceted roles of extracellular RNAs in different diseases. Mil. Med. Res. *9*, 43. 10.1186/s40779-022-00405-z.

76. Pita, T., Feliciano, J.R., and Leitão, J.H. (2020). Extracellular RNAs in Bacterial Infections: From Emerging Key Players on Host-Pathogen Interactions to Exploitable Biomarkers and Therapeutic Targets. Int. J. Mol. Sci. *21*. 10.3390/ijms21249634.

77. Wu, D., Tao, T., Eshraghian, E.A., Lin, P., Li, Z., and Zhu, X. (2022). Extracellular RNA as a kind of communication molecule and emerging cancer biomarker. Front. Oncol. *12*, 960072. 10.3389/fonc.2022.960072.

78. Ilieva, M., and Uchida, S. (2022). Extracellular RNAs as communicators in cardiovascular disease: a narrative review. ExRNA *4*, 14–14. 10.21037/exrna-22-3.

79. Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife *5*. 10.7554/eLife.19760.

80. EP2839278B1 - Methods for exosome isolation - Google Patents https://patents.google.com/patent/EP2839278B1/en.

81. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499. 10.1101/gr.209601.116.

82. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. *17*, 10. 10.14806/ej.17.1.200.

83. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. 10.1093/bioinformatics/btp324.

84. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469. 10.1038/nature07488.

85. Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv. 10.48550/arxiv.1409.1556.

86. Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with

supervised learning. In Proceedings of the 22nd international conference on Machine learning - ICML ’05 (ACM Press), pp. 625–632. 10.1145/1102351.1102430.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Bahar Zirak
DBC46FAD4539455...
Author Signature

5/28/2024
Date