INVESTIGATING THE BIOLOGICAL DETERMINANTS OF EARLY LUNG ADENOCARCINOMA BEHAVIOR THROUGH DATA INTEGRATION

By

Maria Fernanda Senosain Ortega

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Cancer Biology

May 31st, 2022

Nashville, Tennessee

Approved:

Vito Quaranta, Ph.D.

Carlos F. Lopez, Ph.D.

Jonathan M. Irish, Ph.D.

Ken S. Lau, Ph.D.

In memory of my dearest friend, colleague and mentor, Dr. Pierre P. Massion.

”Inspired by patients, driven by science.”

ACKNOWLEDGMENTS

My time at Vanderbilt has been immensely rewarding, exciting and full of gratifying experiences. My journey actually started several years ago, when I was given then opportunity to participate of the Vanderbilt International Summer Research Academy as an undergraduate back in the summer of 2014. My first experience in cancer research was by the hand of Dr. Pierre Massion and his amazing lab, which ignited in me the desire to purse a career in this field. Few years later, I returned to his lab as a graduate student and the rest is history. This wonderful journey would not have been possible without the support and help of a long list of people to whom I will be eternally grateful.

First, I would like to thank my former advisor Dr. Pierre Massion, whose untimely passing left on all of us who knew him a deep void which is slowly but surely being filled by gratitude and hope. Thanks for believing in me more than I ever could, and for supporting my crazy dream of becoming a computational scientist even when I had no idea what the heck I was doing. Thanks for never giving up on me and for always being my greatest cheerleader. Thanks for all our long talks about science, faith and life in general. It hurts not being able to cross this finish line with you physically present, but I know that you are by my side and we will see each other again.

I would also like to thank my other academic advisors, Dr. Vera Pancaldi and Dr. Carlos Lopez. Thanks Vera for welcome me into your lab with open arms. My time with your group was an integral part of my training and I will be forever in debt. Thanks for helping me get to the finish line, for pushing me when I needed it the most and for always being there for me. Thanks Carlos for also welcome me into your lab and becoming my primary advisor upon Pierre's passing. I enjoyed very much my time in the LoLab and I will be forever grateful for having your unconditional support, especially during the last stretch of this marathon. I would be remiss without thanking my other committee members, Dr. Vito Quaranta, Dr. Jonathan Irish and Dr. Ken Lau. To all these amazing scientists that have been part of my training, thanks for providing thoughtful and critical feedback of my research, I have learned so much from you throughout these past years. It has been an honor to be mentored by such talented group of scientists.

To the Massion Lab family, thanks for teaching me how to do translational cancer research as a team, particularly those who were part of my project, I could not have made it without your help. To the LoLab family, thanks for make me feel welcome since day one and for all the insightful feedback and scientific discussions. I am also thankful for my training from the IGP to the Cancer Biology Program, for the support of the BRET office and for all the career development opportunities that have been given to me.

To the friendships I forged while in Nashville, my second home, especially Mehnaaz and John, thanks for making this ride much more fun and enjoyable! To my partner Victor, thanks for your unconditional love and support, for introducing me to the world of computational sciences and for being a role model for me. To my friends and family in Peru, thanks for being by my side all this time and for making feel home a bit closer with your constant love and support. Special thanks to my all time best friends Alondra, Dyanne, Claudia, Brenda and Fania!

To my Mom and Dad, thanks for supporting my education since my early years and for always believing that nothing was impossible for me. I owe you each and every one of my achievements and I am extremely grateful for all the sacrifices you made to get me where I am today. To my siblings Josecito and Analu, my life companions, thanks for bringing me so much light and joy and inspiring me to be a better version of myself. And of course, to my sweet puppies Olivia and Leia, for their endless loving licks. You all have been a constant source of love and happiness.

And last but not least, to God for all the blessings He has given me and the wonderful people He put in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

1.1 Acknowledgements

This chapter is adapted from “Intratumor Heterogeneity in Early Lung Adenocarcinoma” published in Frontiers in Oncology and has been reproduced in line with publisher policies [1].

1.2 Introduction

Over the last decades, several efforts have been made to reduce mortality among lung cancer patients. While advances in diagnostic and therapeutics have occurred, long-term survival rates compared to other cancers have barely improved [2]. Therefore, new approaches are needed. In the context of lung adenocarcinoma (LUAD), this is of great importance due to the high rate of overdiagnosis and lack of accuracy in predicting indolent vs. aggressive behavior of the tumor [3]. In order to better predict disease behavior, it is crucial to understand the cellular and molecular underpinnings of the tumor. Thus, the study of intratumor heterogeneity and its clonal composition has become an attractive strategy to understand tumor progression and behavior [4, 5, 6, 7, 8]. In the recent years emerging single-cell analysis platforms have allowed the deep profiling of the tumor microenvironment (TME), and seem promising approaches for the dissection and of tumor heterogeneity [9].

1.3 Clinical overview of Lung Adenocarcinoma

Adenocarcinoma is a subclass of non-small cell lung cancer, which develops within the glandular cells of smaller airways along the outer edges of the lungs. It is the most common histological type, accounting for about 40% of all lung cancer cases. This type of lung cancer mostly occurs among current or former smokers, however it is also the most prevalent type of lung cancer in non-smokers [2]. Thus, the exposure to environmental carcinogens combined with genetic susceptibility may also play an important role in the development of the disease [10, 11].

The survival rate for lung cancer mostly depends on the stage at the time of diagnosis. On average, the current 5-year survival rate is about 18%, but if detected early it can lead to a better prognosis, with a 5-year survival rate of 54% for localized stage [2]. However, only 15% of all cases are diagnosed on time, while the vast majority (57%) are diagnosed at a late stage [12]. Therefore, screening for lung cancer in high risk individuals is important.

1.4 Computer Tomography-based detection and risk stratification of LUADs

In the past years, numerous randomized trials have assessed the power of lung cancer screening showing that it is possible to detect lung cancer at an early stage in more than 40% of the cases [13, 14]. Furthermore, the 5- and 10-year survival rates among lung cancer patients enrolled in screening programs were close to 90%, which is very reassuring [15]. The largest lung cancer screening trial at the moment, The National Lung Screening Trial (NLST), enrolled 53,452 high risk individuals for lung cancer across 33 U.S. medical centers and reported a 20% relative risk reduction in mortality using low-dose computed tomography (CT) screening compared to chest radiography (CXR) screening [16]. Despite these encouraging statistics, it is worth to mention that 96% of the nodules detected through CT screening were benign. Moreover, confirmed lesions detected through CT screening range from very indolent to severely aggressive cancers. Therefore, screening, which by definition seeks to spot malignant nodules in asymptomatic individuals, bears the inherent feature of overdiagnosis. This phenomenon can be defined as the detection of a cancer that in other circumstances would have not become clinically evident , and represents a serious drawback for lung cancer screening in that it generates unnecessary treatment, morbidity, additional expenses, and anxiety and distress to the patient. A while after the NLST results were published, another study focused on the estimation of overdiagnosis in the NLST, reporting a probability of 18.5% that any lung cancer detected by CT was an overdiagnosis, as well as probabilities of 22.5% for non-small cell lung cancer and 78.9% for adenocarcinoma *in situ* [3]. In that sense, a careful assessment of the images is crucial to ensure a more accurate prognosis. Additionally, the ongoing investigation in the discovery of new biomarkers offers a promising avenue to assist or eventually guide the screening and diagnosis process of high risk individuals.

The current clinical treatment decisions are mostly based on the composition of the lesions on single time point or serial imaging (pure ground glass = indolent, significant or increasing solid component = concern for invasion). However, this practice is subjective and limited by intra-observer and inter-observer variability[17]. In that context, Foley et al.[18, 19] developed and validated an imaging software Computer-Aided Nodule Assessment and Risk Yield (CANARY), which successfully risk stratifies screen-detected lung adenocarcinomas based on clinical disease outcomes. 294 eligible patients diagnosed with LUAD spectrum lesions in the low-dose CT arm of the National Lung Screening Trial were identified retrospectively. The most recent low-dose CT scan before the diagnosis of LUAD was analyzed using CANARY blinded to clinical data. Using unsupervised clustering, nine natural exemplars were identified as basic radiographic features of LUAD nodules. Based on their parametric CANARY signatures, all the LUAD nodules were risk stratified into Good, Intermediate, and Poor, and yielded significantly different survival curves, allowing for noninvasive risk stratification of the nodules into three groups with distinct post-treatment progression-free survival. In a following publication, this group presented a cumulative aggregate of normalized distributions of ordered CANARY exemplars, the Score Indicative of Lung Cancer Aggression (SILA)[20]. The SILA discriminated between indolent and invasive LUAD and, prediction of linear extent of histopathologic tumor invasion was possible. In stage I LUAD, three separate SILA prognosis groups were identified: indolent, intermediate, and poor, with 5-year survival rates of 100%, 79%, 58%, respectively. Cox proportionality hazard modeling predicted a 50% increase in mortality, for a 0.1 unit increase in the SILA over a median follow-up time of 3.6 years. In conclusion, tools like CANARY and SILA could ultimately facilitate individualized management of incidentally or screen-detected LUADs.

1.5 The molecular landscape of Lung Adenocarcinoma

Over the years, genomic alterations occur and accumulate and in some cases those alterations may lead to oncogenesis. The somatic genomic alterations that are involved in cancer development are known as "driver alterations" and the ones that are not are known as "passenger alterations" [21]. LUAD has one of the highest mutational burdens compared to other cancers [22, 23]. Those high rates of somatic alterations and genomic rearrangements include a large load of passenger events per tumor genome, which makes the identification of driver alterations even more challenging [24]. Despite the difficulties, several genomic alterations have been described in the past years, some of which are currently known as canonical driver alterations, and some others that have recently been reported and may be novel driver events [24, 25, 26, 27].

Driver genomic alterations in LUAD are generally associated with events that lead to the constitutive activation of signaling proteins, which commonly occur in oncogenes of the receptor tyrosine kinase (RTK)/RAS/ RAF pathway [28]. In the TCGA study, 62% of the tumors harbored such alterations [26]. *KRAS* driver mutations were reported in 32% of TCGA samples [26]. Along with *HRAS* and *NRAS* (0.9%), the other members of the RAS family, these proteins play an important role in the regulation of signaling pathways that control cell proliferation [29]. Additionally, *KRAS* mutations are highly correlated with poor prognosis in early LUAD [30]. Cancer-associated mutations in *EGFR* were present in 11% of TCGA samples [26]. *EGFR*, as well as other member of the EGFR family the oncogene *HER2* (1.7%), are known to be involved in the regulation of several cellular processes including cell motility, angiogenesis, cell proliferation and apoptosis [31]. Likewise, some *EGFR* mutations are related to an improved prognosis [32]. Another important oncogene is *BRAF*, which works downstream of RAS proteins and has a crucial role in the RAS-MAPK pathway. Driver mutations of this gene were present in 7% of TCGA samples and are not known to be associated with prognosis [26, 33]. *MAP2K1* encodes for a protein that operates downstream of *BRAF* and was found mutated in 0.9% of TCGA samples [26]. *MET* exon 14 skipping is another cancer driver event which results in the loss of a negative regulatory site, and occured in 4.3% of TCGA samples [26]. Gene fusions, were reported for the genes *ROS1*, *ALK* and *RET*, which were altered in 1.7%, 1.3%, 0.9% of TCGA samples, respectively [26, 28, 34].

In addition to the drivers described above, for the 38% of the samples that did not carry a driver oncogene mutation, the TCGA study proposed previously unrecognized driver genes that might be involved in the RTK/RAS/RAF pathway activation [26]. They identified significant amplification events of *HER2* and *MET* in the oncogene-negative samples. Higher *MET* copy number in primary LUAD at the time of diagnosis has been associated with poor prognosis [33]. *NF1*, a tumor suppressor that negatively regulates the *RAS* oncogene, was mutated in 8.3% of the samples [26, 35]. *RIT1* is mutated in 2.2% of LUAD cases, and has been identified as a new oncogene driver as its mutations have been shown to activate MAPK and PI(3)K signaling in NIH3T3 cells [26, 36].

Besides the RTK/RAS/RAF pathway, other relevant somatic genomic alterations have been identified. *TP53* was commonly mutated in 46% of the samples [26]. *PIK3CA*, a crucial positive regulator of the PI(3)K-mTOR pathway, was mutated in 7% of the cases, and *STK11*, a tumor suppressor from the same pathway, was mutated in 17% of the cases [26]. Other mutated tumor suppressors were *KEAP1* (17%), *RB1* (4%), and *CDKN2A* (4%). In a large-scale project that characterized copy-number alterations in LUAD, the most common amplification was found in chromosome 14q13.3, which corresponds to NKX2-1 (TTF1), a transcription factor involved in lung development [25]. The inhibition of this gene led to reduced cell viability and colony formation in LUAD cell lines [25]. This gene was also reported amplified in 14% of TCGA samples [26]. Other significant amplifications in the TCGA study included the telomerase reverse transcriptase *TERT* (18%), and *MDM2* (8%), a negative regulator of p53 [26]. The most significant deletion (19%) was the *CDKN2A* locus, which codes for the proteins p16 and p14arf, two important tumor suppressors and cell cycle regulators of the TP53 pathway [26, 37]. Some of the alterations described above are depicted in Fig. 1.1.

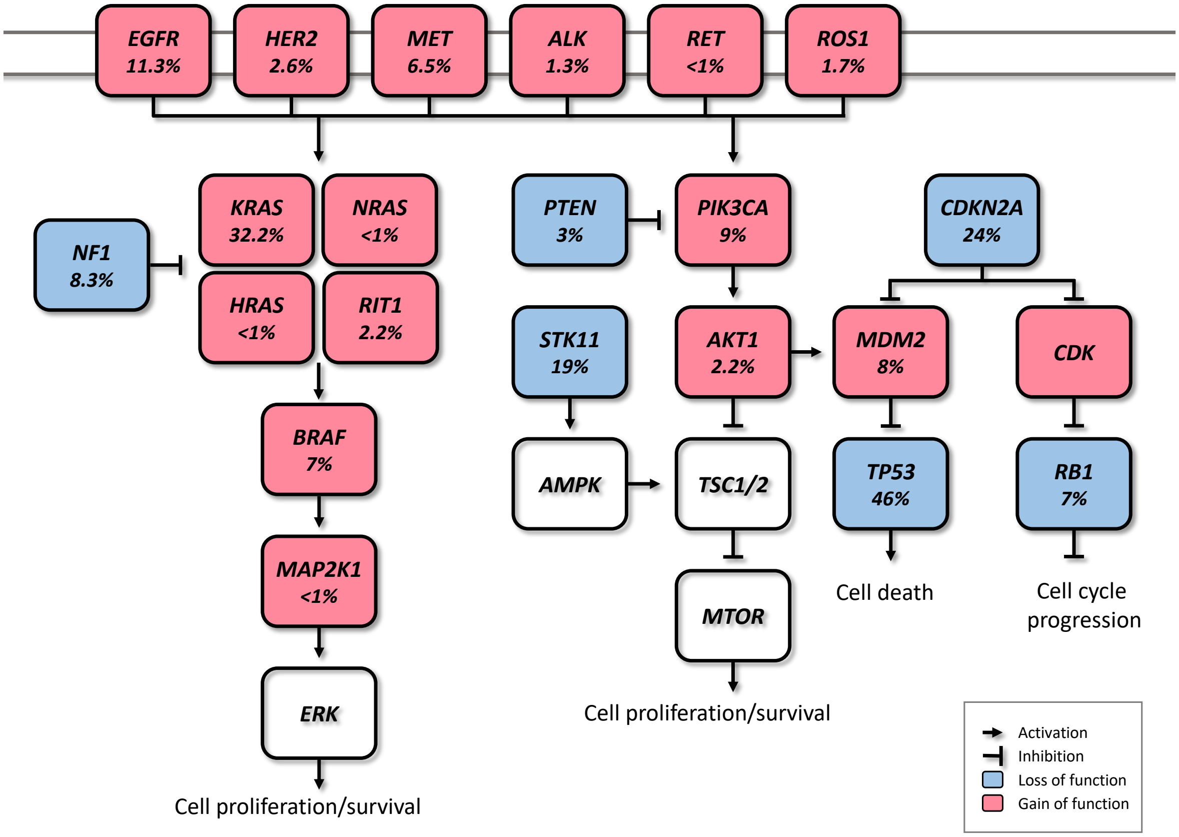


Figure 1.1: Canonical molecular pathways altered in LUAD. Graphical representation of the most mutated pathways in lung adenocarcinoma. The numbers correspond to the percentage of samples that carry that genomic alteration in TCGA.

The understanding of LUAD molecular alterations has significantly impacted patient survival in the past years through the development of targeted therapies. Patients with advanced or metastatic tumors bearing *EGFR* mutations, *EML4-ALK* rearrangement or *ROS1* fusions have benefited from those. Erlotinib, gefitinib and afatinib are some of the drugs currently used to treat patients with *EGFR* exon 19 deletion or exon 21 mutations [38, 39, 40]. Alectinib, ceritinib and crizotinib have shown effectiveness in patients with *ALK* alterations, and the latter is also used in patients with *ROS1* translocation [41, 42, 43, 44]. The advances on genomic phenotyping of LUAD have also benefited the development of immunotherapy. In a healthy individual, the immune checkpoint PD-1 expressed in T cells protects against autoimmunity and inflammation. In cancer, PD-L1 expressed on tumor cells binds to PD-1 resulting in immunosupression and immune evasion. Nivolumab, pembrolizumab and atezolizumab are some of the PD-1/PD-L1 FDA approved inhibitor drugs that have shown improved survival in advanced NSCLC patients compared to standard therapies [45, 46, 47]. Another immunecheckpoint under the radar is CTLA-4. Two clinical trials (NCT02000947, NCT02352948) are currently investigating the effects of a combination therapy of dual checkpoint inhibition using durvalumab and tremelimumab, PD-1 and CTLA-4 inhibitors respectively. However, early results suggest that this strategy did not significantly improved overall survival, although treatment with durvalumab alone provided a significant overall survival improvement. [48, 49]. These and other targeted therapies have been extensively reviewed previously [35, 50, 51].

More recently, the molecular characterization of early LUAD lesions has also provided some insights on tumor behavior. A recent study from our group has characterized 21 adenocarcinoma *in situ* (AIS), 27 minimally invasive adenocarcinoma (MIA) and 54 fully invasive adenocarcinoma using deep targeted genome sequecing [52]. This work uncovered molecular features associated with aggressive early LUAD clinical behavior and disease progression. Most genomic alterations in LUAD were already present in AIS and 21 significantly mutated genes including known drivers such as KRAS, EGFR and TP53 were shared among the three groups, suggesting their step-wise role in malignant transition. APOBEC signature was associated with worse survival compared to DNA mismatch repair signature, and KRAS codon 12 mutations were associated with aggressive tumor behavior. Finally, an ensemble-level progression model using phylogenetic analysis inferred the role of many known alterations in LUAD progression and introduced several new players such as EPPK1, ATM, SMAD4, KMT2C and KMT2D, which deserve to be further investigated. This brings new insights into the distinction between indolent and aggressive tumor behavior and will potentially have future implications in early LUAD clinical management.

1.6 Intratumor heterogeneity and clonal architecture

Intratumor heterogeneity is a highly complex phenomenon and it represents a major challenge in the assessment of cancer, as it acts as a confusing factor resulting in inaccurate diagnosis, prognosis and treatment of the disease [4]. As mentioned before, LUAD is a very heterogeneous disease with one of the highest mutational burdens across different cancer types [22, 23]. Therefore, a comprehensive understanding of the natural history of these tumors is urgently needed.

The study of tumor growth from an evolutionary perspective is not a new approach. In the early 70's, Alfred Knudson proposed that for a particular cell to became cancerous, both alleles of a given tumor suppressor gene must be mutated, also known as the ”two-hit hypothesis” [53]. In 1976, Peter Nowell applied evolutionary models to study tumor progression and treatment failure, and proposed a clonal evolution model in which a tumor arises from a single mutated cell (“clone”) and tumor progression occurs as a result of subsequent alterations, in which fitter and more aggressive clones replace the original clone cells [54]. This linear evolution model was supported mostly by early studies that focused in a single gene rather than in the whole genome , and therefore clonal diversity was underestimated [55]. Advances in new sequencing technologies allowed genome wide sequencing, which have elucidated a more complex clonal structure than previously thought [23].

In the past years, other evolutionary models have derived from applied phylogenetic inference to next-generation sequencing data. In neutral evolution, all driver alterations are thought to be present in the original neoplastic cell and subsequent alterations are neutral, thus it is characterized by the absence of selection and heterogeneity arises from stochastic processes as a byproduct of tumor progression [56]. In punctuated evolution, it is postulated that tumor heterogeneity is generated in the early development of the neoplasia as a punctuated burst, followed by neutral evolution [57, 58]. Branching evolution, also known as the trunk-branch model, is defined by the gradual accumulation of driver mutations in subclonal populations [59]. In this model, the "trunk" of the tumor consists of progenitor clones bearing early somatic alterations that drive tumorigenesis. Those early alterations are potentially ubiquitous events. Conversely, somatic events that occur later are heterogeneous events and are present in the subclones which make up the "branches" of the tumor and are tumor progression drivers.

Multiregion sequencing has been the most successful strategy to investigate intratumor heterogeneity and clonal evolution in LUAD to date [5, 6, 7]. The studies conducted by De Bruin and colleagues, Zhang and colleagues, and most recently Jamal-Hanjani and colleagues, provide evidence suggesting that intratumor heterogeneity and branched evolution might be a universal phenomenon across LUAD (Fig. 1.2). Most known driver alterations [26, 28] were mapped to the trunks of the tumors, which suggests that those canonical alterations occur early in tumor evolution. Truncal driver mutations almost always occurred before genome doubling suggesting a particular role in tumorigenesis. On the other hand, truncal genome doubling events occurred before subclonal diversification but after the acquisition of driver mutations, which suggests that chromosomal instability may be a crucial step that induces copy number alterations followed by a burst of mutational heterogeneity (Fig. 1.2). Furthermore, the association of drug resistance and patient relapse with chromosomal instability [60], supports the hypothesis that the ability of chromosomal instability to generate extensive subclonal divergence could be compromising the effectiveness of therapeutics strategies that target truncal driver mutations due to the overlooked and already present clonal heterogeneity [5]. Besides, data from these studies suggest that certain alterations in non-canonical cancer genes may also drive tumor development and subclonal diversification.

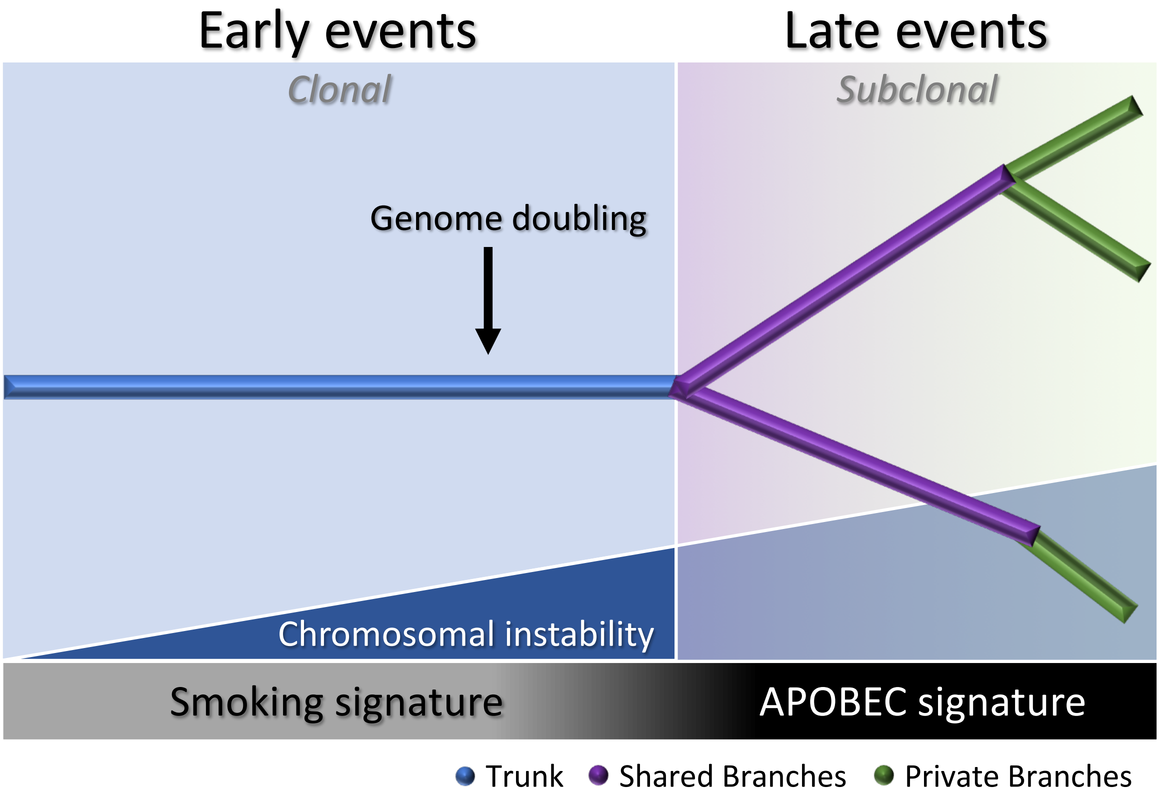


Figure 1.2: Branching process of tumor evolution in LUAD. A tumor is depicted as a tree structure with the trunk representing ubiquitous (clonal) mutations present in all tumor regions (blue); shared branches representing heterogeneous (subclonal) mutations present in some tumor regions (purple), and private branches (also subclonal) representing unique mutations present in one tumor region only (green). The blue right triangle shows how as the chromosomal instability increases, the subclonal diversification is triggered. The bottom bar indicates that the smoking signature is associated with early events whereas the APOBEC signature is associated with late events.

Another important feature of the disease addressed by these groups was the influence of smoking status in the clonal history of the tumors. Smoking signature (signature 4) is characterized by a high proportion of C>A transversions [23]. In these studies, tumors from former and current smokers showed a decrease in the proportion of C>A transversions in subclonal mutations compared to early mutations, which suggests a relative decrease in the mutational burden due to smoking during tumor development [5, 6, 7]. Moreover, the decrease of C>A transversions was followed by an increase in C>T and C>G mutations, which indicates APOBEC cytidine deaminase activity [23]. This suggests that APOBEC mutagenesis may be playing a role in subclonal expansion in these tumors. In addition, a prolonged tumor latency period was reported by two groups [5, 7]. In the study conducted by De Bruin and colleagues, a tumor from a patient that ceased smoking 20 years before surgery bore the smoking signature in more than 30% of truncal mutations, which suggests that these events occurred within a smoking tumorigenic setting more than 20 years ago [5]. Likewise, Jamal-Hanjani and colleagues reported that 7 patients that were former smokers for several years before surgery, presented a smoking mutational signature suggesting tumor latency for several years before clinical manifestation of the disease [7]. Furthermore, Zhang and colleagues and Jamal-Hanjani and colleagues found an association between the proportion of subclonal genomic alterations and recurrence [6, 7]. In the cohort studied by first group, the three patients that relapsed had a significantly higher proportion of subclonal mutations compared to the patients with no relapse, suggesting that the degree of subclonal divergence may be associated with post-surgical relapse [6]. In contrast, the second group did not find a significant association between the proportion of subclonal mutations and disease recurrence in their cohort, but found that patients with a large proportion of copy-number alterations were at higher risk for relapse or death compared to patients with a low proportion [7]. Additionally, this group found that many late driver mutations corresponded to alterations that have been reported in other tumor types, and most of them are involved in genome maintenance processes such as DNA damage response, chromatin remodeling and histone methylation. They hypothesized that late mutations may be responsible for providing advantages to the emerging subclones and enabling the late stages of the disease as they may remove tissue specific constrains on the neoplastic genome [7].

These studies raised the question if single-region biopsy is informative enough to help the health providers make accurate treatment decisions. Intratumor heterogeneity has proven to be an intrinsic phenomenon to LUAD, and it may compromise the ability of a single biopsy to comprehensively and accurately describe the complexity of the disease for an optimal cancer control. In a handful of cases, a large proportion of subclonal events were found in a single region but were absent in other regions of the same tumor, evidencing the limitations of a single-region biopsy in accurately explaining the clonal architecture of the tumor and highlighting the power of multiregion sequencing to better capture the clonality of the tumor which could help to prioritize some drug targets [5, 6, 7]. Nonetheless, in the study conducted by Zhang et al., while they observed that multiregion sequencing is a better strategy to understand intratumor heterogeneity they also provided evidence that demonstrates that an increase in sequencing depth (∼277x to ∼863x) allowed the identification of most of the driver mutations in the tumors studied and many subclonal mutations were detectable in all regions of individual tumors. This suggests that a single biopsy analysis might be sufficient if the sequencing depth is increased [6].

1.7 The tumor microenvironment of Lung Adenocarcinoma

It is known that the immune microenvironment plays a pivotal role in LUAD development, thus it may also shape intratumor heterogeneity. Neoantigen presentation is an important step for cytolytic T cell response and it is guided by the human leukocyte antigen (HLA) class I molecule, which presents intracellular peptides on the cell surface for the T cell receptors to recognize [61]. A person's genome contains up to six different HLA class I alleles encoded by the genes *HLA-A*, *HLA-B* and *HLA-C*. Each HLA allotype presents peptide antigens based on specific anchor residues within the peptide sequence that are required for the peptides to bind. Therefore, loss of heterozygosity (LOH) results in loss of an HLA allotype and thus loss of the ability to bind those peptides that only contain anchor residues able to bind to the lost HLA molecule, hence fewer neoantigens can be presented to T cells. The impairment of tumor neoantigen presentation as a consequence of LOH in HLA class I was recently suggested as a mechanism of immune evasion in NSCLC [62]. In this study, both lung adenocarcinomas and squamous cell carcinomas tumors with HLA LOH presented higher mutational burden compared to tumors without HLA LOH, with a significant increase in subclonal mutations. Furthermore, tumors harboring HLA LOH were enriched in neoantigens predicted to bind the missing HLA alleles and presented high PD-L1 staining on immune cells. This mechanism may facilitate the sub clonal expansion of cells harboring previously antigenic mutations that had become undetectable to the immune system. A following study from the same group, found that the immune microenvironment tends to be highly heterogeneous between and within patients, showing distinct regions with different levels of immune evasion within individual tumors [63]. Additionally, tumors showing high immune infiltration and HLA allelic preservation also presented neoantigen depletion suggesting that immune evasion occurs by HLA LOH or neoantigen suppression. One of the possible mechanisms for the latter is promoter hypermethylation, which explains 23% of the neoantigens included in this study, suggesting that other mechanisms must be in place. Further elucidation of the mechanisms involved in neoantigen-associated immune escape could have important clinical implications in therapy selection and response prediction.

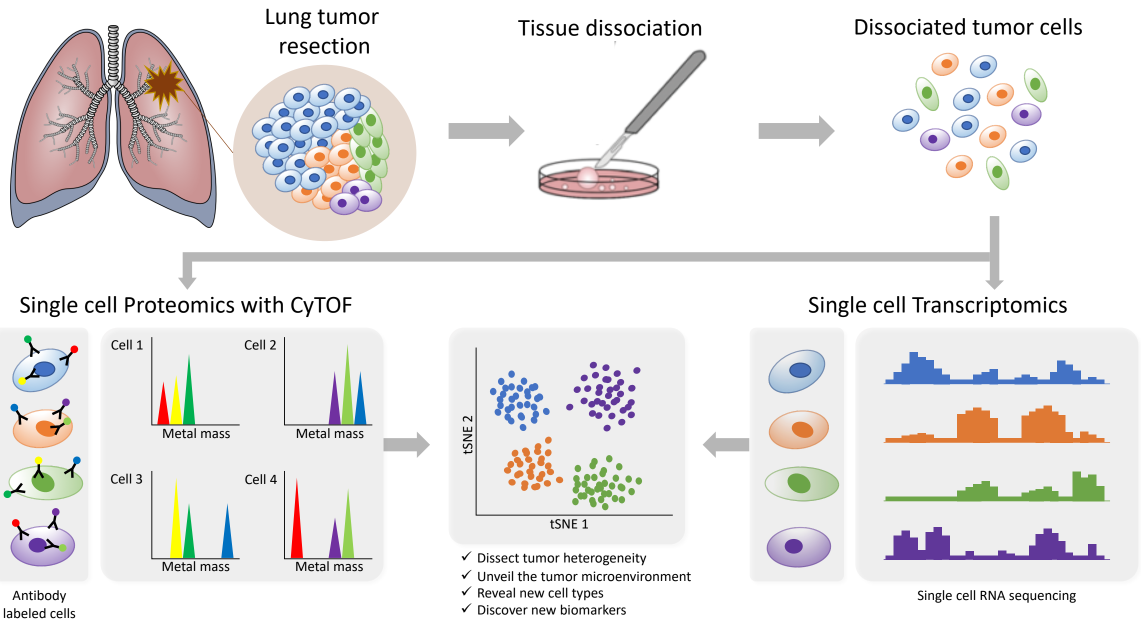


Figure 1.3: Investigating intratumor heterogeneity and the TME with single cell approaches. A lung tumor resection is dissociated into single cell suspension which can be used in different applications. CyTOF uses metal-labeled antibodies to detect a limited number of proteins in the cells. Single cell RNA-Seq reveals the transcriptome of each individual cell. Both can be analyzed through computational strategies to dissect intratumor heterogeneity.

In recent years, more studies focusing on the TME are starting to implement the use of single-cell based technologies, which can elucidate tumor heterogeneity with high resolution by detecting cells individually instead of a bulk signal and yield loads of information (Fig. 1.3). Using single-cell proteomics mass cytometry analysis with paired tumor tissue, normal tissue and peripheral blood, Lavin and colleagues intended to provide an innate immune cell atlas of early LUAD [64]. In this study, early lesions have shown to bear a unique and TNM stage-independent immune signature, with a particular subset of tumor-infiltrating myeloid cells different from normal lung –PPARγ^{hi} macrophages enrichment and CD141+ dendritic cells (DC) depletion)– which could be compromising T cell immunity and may offer a new avenue of intervention in T cell immunotherapies. PPARγ is a transcription factor known to drive an immunosuppressive program [65]. Lymphotoxin beta, inflammatory response inducer, has been previously shown to act on high endothelial venules (HEV) to promote lymphocyte homing to peripheral lymph nodes in vivo [66]. The authors found that the CD141+ DC subset expressed lymphotoxin beta transcripts in lung tumor tissues which suggests that CD141+ DC contribute to tertiary lymphoid structure formation likely through HEV-mediated recruitment of lymphocytes. Therefore, an induced expansion of intratumoral CD141+ DC may serve as a potential anti-tumor immunity strategy. This study highlights the importance of paired analysis to identify tumor-associated immune alterations from normal tissue-imprinting. Other study that also focused on tumor infiltrating myeloid cells (TIM), used single-cell RNA seq to profile a compare TIM populations between mice and humans in the context of NSCLC [67]. Although the goal of this study was to establish similarities between mouse and human TIM expression programs, the comprehensive annotation of the different myeloid populations is an important contribution for future studies on clinical implications of the heterogeneity of these cell types. The authors reported that mouse and human TIM subsets show one-to-one equivalence and that blood myeloid cells poorly reflect TIM states. Due to the overlap of TIM states between patients they assessed the association with patient survival addressing the expression of genes specific to each subpopulation. They identified three conserved subsets of neutrophils, N1 that express canonical neutrophil markers, N2 which are tumor specific and promote tumor growth, and N2 which have a expression signature of type I interferon response. They found that human neutrophil subsets N2 and N5 showed an abundance of marker genes associated with poor survival. Conversely, the marker genes of human DC subset 2, which preferentially interacts with CD4+ T cells, showed positive association with survival. Guo and colleagues also investigated the immune system of NSCLC with single-cell RNA seq but focusing on T cell subpopulations of 14 patients [68]. They identified two new CD8+ T cell pre-exhausted subsets, which together with the presence of highly migratory effector T cells may provide an explanation for positive responses to immunotherapy. When they interrogated LUAD TGCA data with their expression signature, they found that patients mainly clustered into two groups: one enriched in pre-exhausted CD8+ T cells, non-activated Tregs and activated CD4+ T cells, and the other enriched in exhausted T cells and activated Tregs. Patients from group 1 had significantly better prognosis than patients from group 2, therefore T cell composition could be a potential clinical biomarker for LUAD patients. In a different study, Lambrechts and colleagues used single-cell RNA sequencing and reported a comprehensive 52,698-cell catalog of the TME transcriptome of lung cancer samples, most of which were LUAD patients [69]. They identified 52 different stromal subtypes including different populations of cancer-associated fibroblasts, endothelial cells and infiltrating immune cells, some of which were further validated through immunofluorescence. Further analysis of TCGA data indicated that the abundances of some subpopulations and their correlation with patient survival differ between LUAD and squamous cell carcinoma (SCC) and that they were influenced by clinical characteristics such as stage. Low expression of CD8+ T cell cluster 8 marker genes were positively and negatively associated with survival in LUAD patients and SCC, respectively. This cluster represented CD8+ cytotoxic T cells per their high granzyme and IFN expression, and was characterized by high T cell exhaustion marker expression (LAG3). These and other gene expression changes in tumor stroma reveal potential new directions for intervention.

In conclusion, the TME represents an important component of tumor heterogeneity in LUAD and is strongly associated with disease progression and predicted outcome. Although the different flavors of bulk profiling of the tumors are still providing a significant amount of information, it is important to acknowledge that single-cell approaches offer a new level of granularity that are allowing us to deeply dissect and further understand LUAD heterogeneity and its implications in early stages of the disease. Nevertheless, such techniques are highly expensive which currently limits the number of samples per study. A combination of both bulk and single-cell approaches as reported in some of the studies mentioned above may be a suitable alternative to get the most out of the data while state-of-the-art techniques become more affordable through the years.

1.8 Multi-omic Data Integration Strategies and Limitations

As biological data acquisition for some data types becomes increasingly more affordable, the amount of data collected at different molecular levels also increases. One of the main aims of using a multi-omic strategy is to put that wealth of information to good use to better classify biological samples, such as in medical studies aiming to improve patient stratification. Unsupervised data integration can potentially capture complex relationships within data types and reveal groups of samples that otherwise would go unnoticed. Multi-omics data integration can also be done in a supervised way to predict response variables, such as clinical outcomes, or for the identification of biomarkers associated with the response variable[70]. There are multiple tools and methods that have been developed in the previous years to leverage multi-omics data. In a recent review published by Subramanian et al.[71], they described some of these tools grouped base on their approach (similarity, correlation, network, Bayesian, multivariate, fusion) and their applications (disease subtyping, disease insight, biomarker prediction). One of these tools is the Multi-Omics Factor Analysis (MOFA), which is a Bayesian method intended for biomarker prediction[72]. MOFA is an unsupervised method to integrate multi-omics datasets on the same or partially overlapped samples. It infers an interpretable low-dimensional data representation as hidden factors on multiple data modalities using a Bayesian framework that supports both numerical and categorical data. Nevertheless, as MOFA use linear models to represent relationships between data it can fail to capture nonlinear associations between and within modalities. Another very solid and versatile toolkit is mixOmix, which provides a set of supervised and unsupervised multivariate methods for data integration focused on disease subtyping and biomarker prediction[73]. This package offers a variety of methods such as PCA, independent PCA, partial least squares regression (PLS), sparse PLS, canonical correlation analysis (CCA), and PLS discriminant analysis (PLS-DA) to classify or cluster samples. Additionally, their novel DIABLO framework enables the integration of the same biological N samples measured on different omics platforms using sparse PLS-DA to identify highly correlated multiomics signatures to discriminate disease subtypes. Finally, an example of a network method is Similarity Network Fusion (SNF)[74], which as its name states creates an individual network for each data type and then fuses these into a single similarity network using a nonlinear method based on message passing theory. In this process, weak connections disappear with iterations while strong connections are propagated till convergence. This method focuses on disease subtyping.

These tools and many others contribute to the rapidly developing field of multi-omics data integration. However, it is important to consider and address some limitations[70]. One overlooked challenge in data collection of multi-omics studies is the lack of uniformity in methods for missing value imputation and the need for sensitivity analysis to assess the impact of imputation in the downstream analyses. In terms of the integrative analysis itself, some of the limitations include the heterogeneity in signal-to-noise ration among different omics technologies, the poor biological interpretability of multi-omic models, and the need for more biologists trained to use cloud-based services as the datasets are becoming bigger and demand more computational power. Finally, despite the large amount of multi-omics studies and publicly available datasets, the retrieval of multi-omics data is still a problem as most of the times there is a lack of connection of samples across modalities, making this task usually manual when not impossible. Therefore, there is an urgent need for standards for data annotation and storage in multi-omic studies. In conclusion, with the advent of high throughput technologies and those becoming more accessible there has been an increase in the numbers of multi-omics studies in the past years, which is revolutionizing the field of biomedical research and systems biology. However, there are still several challenges need to be addressed or for which solutions are still limited.

1.9 Summary and Dissertation Outline

LUAD is a devastating disease and despite the ongoing research efforts, the overall survival rates have barely improved in the past years. While screening programs have proven to significantly increase the chance of survival in high risk individuals, there is also a high probability of overdiagnosis. Therefore, the molecular determinants of early tumor development behavior need to be further investigated. In the past years, it has become more evident that intratumor heterogeneity profiling of LUAD is the most effective strategy to understand tumor progression. In this context, the rapidly evolving field of single-cell technologies offers a novel set of tools that is unraveling the complexity of LUAD and other cancers with a resolution never reached before. Furthermore, as LUAD is a consequence of complex biological processes, it is necessary to take an integrative approach combining data from different modalities to understand the interrelationships of multiple biological layers and their functions.

In this dissertation, I aim to investigate the biological determinants of early lung adenocarcinoma indolence or aggressiveness. I hypothesize that the integration of biological, clinical and radiomics data of early stage LUAD will improve the discrimination between indolent and aggressive tumors which in turn may offer novel and personalized avenues for intervention. In the next chapters, I will present the methodologies and results of my research studies. In Chapter 2, I will describe in detail the methods used to acquire, process and analyze the data collected from LUAD cell lines and LUAD patients across different data modalities. In Chapter 3, I hypothesize that single-cell proteomic analysis of early stage LUAD will provide new insights into the cellular and molecular determinants of indolent and aggressive tumors. I will report the validation of a LUAD-focused Mass Cytometry antibody panel on LUAD cell lines and present the analysis of a set of ten early stage primary LUADs with indolent and aggressive behaviors showing some valuable insights on immunogenicity of the tumors. In Chapter 4, I will present the results of my investigation of the biological determinants of early lung adenocarcinoma indolence or aggressiveness using radiomics as a surrogate of behavior. The integration of Next Generation Sequencing (NGS) data, proteomics and radiomics features is the central piece of this section and will reveal novel insights that connect tumor biology and clinical characteristics of LUAD. Finally, in Chapter 5 I will summarize the main conclusions of this work and discuss the implications of future directions of this research.

CHAPTER 2

Materials and Methods

2.1 Cell lines and cell culture

Human lung adenocarcinoma cell lines A549, PC9, H23 and Human Burkitt's lymphoma cell line Ramos were obtained from ATCC. H3122 was provided by Dr. Christine Lovly (Vanderbilt University) [75]. Cells were grown in RPMI 1640 medium containing 10% heat-inactivated FBS (Life Technologies, cat# 16140071) and 1X Pen/Strep at 37°C, 100% humidity, and 5% CO2. All cells used were in a low passage number (<5). These cell lines harbor different genetic alterations (Table 2.1).

Cell line	Genetic Alteration
A549	KRAS activating mutation CDKN2A locus deletion
H3122	EML4-ALK variant 1, activating mutation
PC9	EGFR activating mutation TP53 inactivating mutation
H23	TP53 inactivating mutation KRAS activating mutation

Table 2.1: LUAD Cell lines genomic profiles

2.2 Human specimens

PBMCs were obtained from a healthy donor under an Internal Review Board (IRB) approved protocol 030763 and tumor tissues samples were collected from patients undergoing lung resection surgery following an IRB approved protocol 000616 at the Vanderbilt University Medical Center. Informed consent was obtained from all subjects. Samples from Chapter 3 were obtained from 10 lung adenocarcinoma patients, from which 5 were males and 6 were females. The ages from this patients ranged from 58 to 88 with a median of 72 (Table 2.2). Samples from Chapter 4, were obtained from 92 lung adenocarcinoma patients, from which 43 were males and 49 were females. The ages from this patients ranged from 48 to 90 with a median of 66.5 (Table 2.3).

Characteristic		Patients (N=10)
Sex		
	Male	4
	Female	6
Age		
	Median	68.5
	Range	56 - 86
Race		
	Caucasian	10
Smoking Status		
	Smoker	1
	Ex-smoker	9
Family History of Cancer		
	Lung	2
	Other	5
Nodule size (mm)		
	Median	31.5
	Range	9.7 - 61
Pathological Stage		
	Stage 0	1
	Stage IA	2
	Stage IB	1
	Stage IIA	1
	Stage IIB	4
	Stage IIIB	1
Tumor Location		
	RLL	3
	RUL	4
	LLL	3
Risk Stratification (CANARY)		
	LPS	4
	SPS	6

Table 2.2: Summarized patient characteristics for Chapter 3

Characteristic		**Patients (N=92)**
Sex		
	Male	43
	Female	49
Age		
	Median	66.5
	Range	48 - 90
Race		
	African American	6
	Asian	2
	Caucasian	84
Smoking status		
	Never smoked	15
	Smoker	11
	Ex-smoker	66
Family history of cancer		
	Lung	12
	Unknown	29
	Other	51
Nodule size (cm)		
	Median	2.3
	Range	0.8 - 7.3
Pathological Stage		
	Stage 0	1
	Stage IA	43
	Stage IB	11
	Stage IIA	13
	Stage IIB	15
	Stage IIIA	5
	Stage IIIB	2
	Stage IV	2
Predominant histology		
	Acinar	52
	Lepidic	2
	Micropapillary	9
	Mucinous acinar	3
	Papillary	9
	Solid	17
Tumor location		
	LLL	12
	LUL	15
	RLL	16
	RML	1
	RUL	43

SILA score		
	Median	0.625
	Range	0.049 - 0.853
SILA groups		
	Indolent	14
	Intermediate	26
	Aggressive	52

Table 2.3: Summarized patient characteristics for Chapter 4

2.3 Sample collection and processing

All tissue samples were processed within one hour of surgery. Lung tissues were minced, digested with Collagenase and DNase I for one hour at 37°C. Single-cell suspension was filtered (70 um and 40 um) and cryopreserved for long-term storage as previously described [76]. Cell viability was assessed before cryopreservation and after thawing. For bulk analyses, lung tissues were snap froze and stored at -80 °C.

2.4 Patient risk stratification and radiomics assessment

2.4.1 Computer-Aided Nodule Assessment and Risk Yield (CANARY)

We analyzed the chest CT scans of the patients using a Computer-Aided Nodule Assessment and Risk Yield (CANARY) software to differentiate and stratify risk of lung adenocarcinomas [18]. CANARY analysis was performed on the CT images taken within 3 months prior surgery for all patients involved in this study. Semi-automated nodule segmentation using CANARY software detects nine classes of nodule characteristics based on voxel histogram features within the CT images which in turn helps in risk stratification of the nodule. These features are coded as Violet (V), Indigo (I), Blue (B), Green (G), Yellow (Y), Orange (O), Red (R), Cyan (C), and Pink (P). The V, I, R, O class represents solid density voxel. Classes B, C, G represent ground-glass opacity and P and Y classes indicate lepidic and invasive growth. The overall prediction of histopathological tissue invasion helps in a risk stratification of the lesions into Good (G) and Poor (P) risk groups, which we refer in the main paper as LPS and SPS, respectively. Samples were classified as shown (Table 2.2).

2.4.2 Score Indicative of Lung Cancer Aggression (SILA)

SILA is a cumulative aggregate of normalized distributions of above mentioned 9 ordered CANARY exemplars and provides a continuous variable in range of 0 to 1[20]. In addition to discrimination between indolent and invasive adenocarcinoma, it also helps in predicting the degree of invasion, disease-free survival and cancer-related mortality in stage I LUAD on the basis of CT. The continuous scale can be thresholded at multiple levels, if needed. We set two SILA thresholds and categorized three distinct histopathologic and prognostic groups for stage I LUAD. These thresholds were computed by using two approaches: automatic histogram-based multilevel thresholding and pathology-based threshold selection. In the automatic approach, the histogram constructed from the SILA values for stage I LUAD nodules in the cohort is divided into three partitions by using a well-known multilevel thresholding algorithm. Pathology-based SILA thresholds were assigned based on TImax (maximum linear extent of tumor invasion) in stage I LUAD. Three distinct survival groups were discovered: best survival in indolent tumors (AIS and MIA), intermediate survival in tumors with TImax from 6 to 20 mm, and worst survival in tumors with TImax greater than 20 mm. The group with a SILA of 0.338 or lower (SILA at the upper 95% confidence interval [CI] of the indolent group) was defined as the good-prognosis group. The group with a SILA of 0.338 to 0.675 (SILA at the upper 95% CI of the TImax ¼ 15- to 20-mm group) was defined as the intermediate prognosis group, and the group with a SILA of 0.675 or higher was defined as the poor-prognosis group.

2.4.3 HealthMyne©

HealthMyne©platform allows semi-automatic lesion segmentation of the delineated volumes of interest, followed by extraction of radiomic features. The user initializes the lesion segmentation by drawing a long axis on ROI in an axial plane of the multiplanar reconstruction. A 2D segmentation is updated in real-time with interactive feedback of the lesion boundary and 2D segmentations on the other MPR planes are immediately proposed. If the contour on a MPR plane seem unsatisfactory, the user can update the segmentation by either drawing long axes on the other MPR views or using a 2D brush tool. When the segmentation is satisfactory, the user can confirm to initiate the 3D segmentation computation. Based on these initial user interactions, the RPM™ algorithms combined statistical sampling methods together with deep learning strategies in order to delineate the target volume and provide an automatic 3D segmentation. The 3D segmentation is reviewed by scrolling through slices on the MPR views. Interactive editing tools including 2D and 3D brushes can be used to reduce/enlarge or add details to the proposed volume segmentation. As the 3D segmentation is confirmed by the user, the measure of the long and short lesion axes is automatically determined by leveraging the volume delineation. A large number of radiomic features are extracted from the segmented volume. Redundant features or features with high inter/intra-user variability were removed. The radiomic risk score is derived from regression shrinkage and subset selection via LASSO method.

2.5 Mass cytometry

2.5.1 Antibody panel

We have developed a comprehensive antibody panel that comprises a total of 34 antibodies, including markers for cellular lineage (immune cells, epithelial cells, endothelial cells, fibroblasts/mesenchymal cells), cancer markers and signaling pathways. Metal-conjugated antibodies were purchased from Fluidigm and customized conjugations were performed using Maxpar Multi-Metal labeling Kits (Fluidigm) with purified antibodies from different sources (see Table 2.4).

2.5.2 Sample preparation and data acquisition

Cryopreserved samples were thawed and stained with our antibody panel (Table 2.4) as previously described [76]. Cell lines were detached from culture flasks using TrypLE Express (Gibco) and processed following the same protocol. For intracellular staining, cells were permeabilized with methanol. To prevent cell loss, an additional fixation step was added to the protocol after the washing steps of the intracellular staining. We controlled for batch effect using EQ Four Element Calibration Beads (DVS Sciences/Fluidigm). Prior sample acquisition, cells were resuspended in 1X calibration beads in deionized water to reach a concentration of 5×10^5 cells/ml. Cells were filtered using FACS tubes with filter caps (Corning Falcon) and collected using a standard/narrow bore on a Helios CyTOF system at the Mass Cytometry Center of Excellence at Vanderbilt University.

Antigen	Isotope	Level	Clone	Source	Catalog #
EpCAM	141-Pr	Surface	9C4	Fluidigm	3141006B
c-caspase3	142-Nd	Intracellular	D3E9	Fluidigm	3142004A
TP53∗	143-Nd	Intracellular	DO-7	Biolegend	645802
HLA-ABC	144-Nd	Surface	W6/32	Fluidigm	3144017B
CD31	145-Nd	Surface	WM59	Fluidigm	3145004B
Thioredoxin	146-Nd	Intracellular	2G11/TRX	Fluidigm	3146016B
b-CAT	147-Sm	Intracellular	D10A8	Fluidigm	3147005A
HER2	148Nd	Surface	29D8	Fluidigm	3148011A
p-STAT6	149-Sm	Intracellular	18/P-Stat6	Fluidigm	3149004A
p-STAT5	150-Nd	Intracellular	Y694	Fluidigm	3150005A
TTF1∗	151-Eu	Intracellular	D2E8	CST	12373
p-AKT	152-Sm	Intracellular	D9E	Fluidigm	3152005A
ki67∗	153-Eu	Intracellular	ki67	Biolegend	350523
CD45	154-Sm	Surface	HI30	Fluidigm	3154001B
CD56/NCAM	155-Gd	Surface	B159	Fluidigm	3155008B
Vimentin	156-Gd	Intracellular	RV202	Fluidigm	3156023A
p-STAT3	158-Gd	Intracellular	Y705	Fluidigm	3158005A
CD4∗	159-Tb	Surface	RPA T4	Biolegend	300502
MDM2∗	160-Gd	Intracellular	Polyclonal	Abcam	ab38618
Cytokeratin∗	161-Dy	Intracellular	C-11	Abcam	ab7753
MET∗	162-Dy	Surface	L6E7	CST	8741
TP63∗	163-Dy	Intracellular	W15093A	Biolegend	687202
CK7	164-Dy	Intracellular	RCK105	Fluidigm	3164020A
EGFR∗	165-Ho	Surface	AY13	Biolegend	352902
CD44	166-Er	Surface	BJ18	Fluidigm	3166001B
p-ERK	167-Er	Intracellular	D13.14.4E	Fluidigm	3167005A
CD8	168-Er	Surface	RPA-T8	Fluidigm	3168002B
CD24	169-Tm	Surface	ML5	Fluidigm	3169004B
CD3e	170-Yb	Surface	SP34-2	Fluidigm	3170007B
CD11b∗	171-Yb	Surface	ICRF44	Biolegend	301337
p-S6	172-Yb	Intracellular	N7-548	Fluidigm	3172008A
HLA-DR	174-Yb	Surface	L243	Fluidigm	3172008A
CD274/PDL1	175-Lu	Surface	29E.2A3	Fluidigm	3175017B
Histone H3	176-Yb	Intracellular	D1H2	Fluidigm	3176016A

Table 2.4: Mass cytometry antibody panel for lung adenocarcinoma.
*Customized conjugated antibodies.

2.5.2.1 Cell lines

To validate our antibody panel we used four LUAD cell lines (Table 2.1) and PBMCs from a healthy donor. In one experiment, we pooled and stained the 4 cell lines and PBMCs in the same proportions (0.5 million cells per group) and we repeat this experiment. In other experiment, we stained and run the different cell groups separately (1 million cells per group). All cells were stained with the same panel (Table 2.4) and we used Histone H3 expression to identify nucleated intact cells.

2.5.2.2 Human samples

Patient samples were stained and processed in the same fashion as cell lines. For every batch, a control was stained and run on the same day. This control was a mixture of A549 and Ramos cells, 1 million cells of each.

2.5.3 Data preprocessing

Collected events from both validation experiments with cell lines and human samples were processed in the same fashion. Prior to analysis, all mass cytometry FCS files were normalized using the premessa R package (https://github.com/ParkerICI/premessa, version 0.2.4), an R implementation of the MATLAB bead normalization software [77]. Normalized data was initially analyzed in Cytobank [78].

2.5.3.1 Data cleaning: manual

For the first dataset that will be presented in Chapter 3, noise reduction parameters were as follows: cells with Histone H3 < 10 were considered dead and excluded, only cells with an event length 10-70 were considered singlets and included.

2.5.3.2 Data cleaning: automated

For the complete CyTOF dataset that will be presented in Chapter 4, I applied an automated data cleaning strategy, which was deployed in an R package. This tool uses classification models to automatically remove the noise from the data having as input the normalized files of the samples and their batch controls. An initial phase removes debris in two steps: first removes events with no expression of ”mandatory” markers (e.g. His H3 for nucleated cells) and events not expressing at least one of the cell type specific markers; the second step removes debris using a classification model trained on Gaussian Discrimination parameters gating, per Fluidigm recommendations (Fig. 2.1). The final phase removes the beads using another classification model trained on the designated beads channels.

To train the first model, we used a random sample of FCS files from our dataset and proceed with manual gating, labeling and then spliting the dataset into training and test (Fig. 2.2). To train the beads model, we first used a random random sample of FCS files, applied arcsinh tranformation (cofactor=5) and performed an unsupervised detection of the beads using a clustering method, which can be either by k-means or Gaussian Mixture Models (Fig. 2.3). Clustering results will be evaluated and only the files in which the events identified as beads show a coefficient of variation (CV) < 0.05 (”good” files) will be selected to be part of the training and test sets. For both models, since CyTOF experiments usually render a large number of events and we do not need that many events to train a model, labeled events from the initial files are concatenated and we take a random sample from it. Models evaluation and further details can be found in the package website (https://msenosain.github.io/denoisingCTF/index.html.)

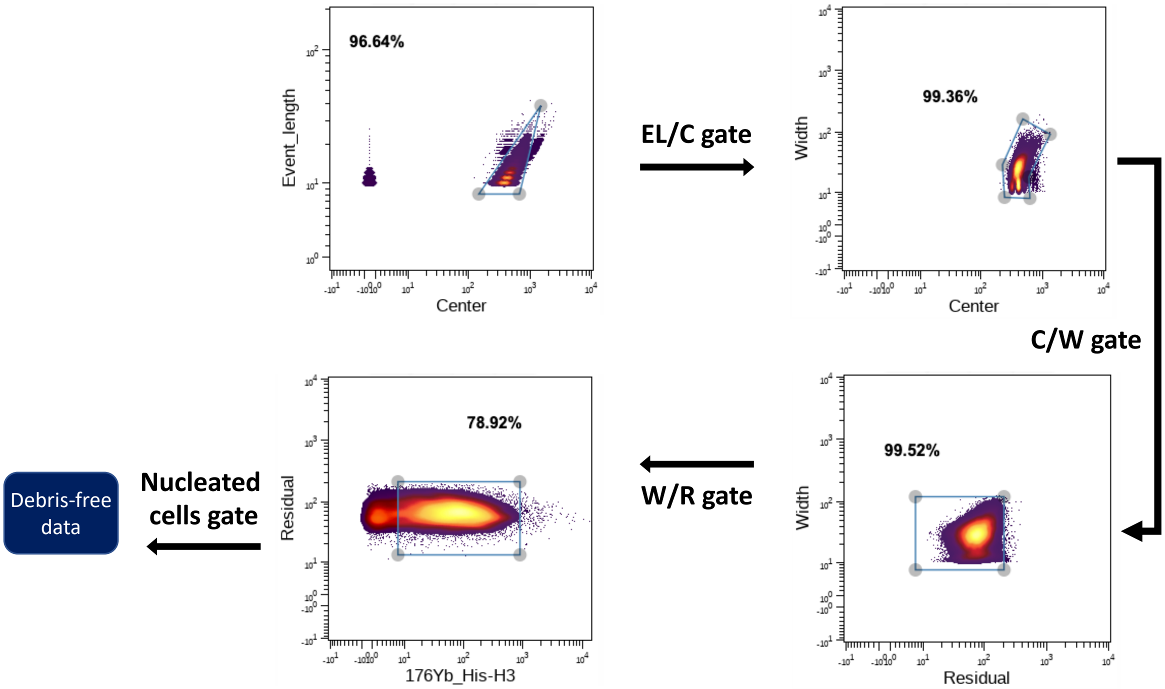


Figure 2.1: Gaussian Parameters-based manual gating example. Based on Fluidigm recommendations

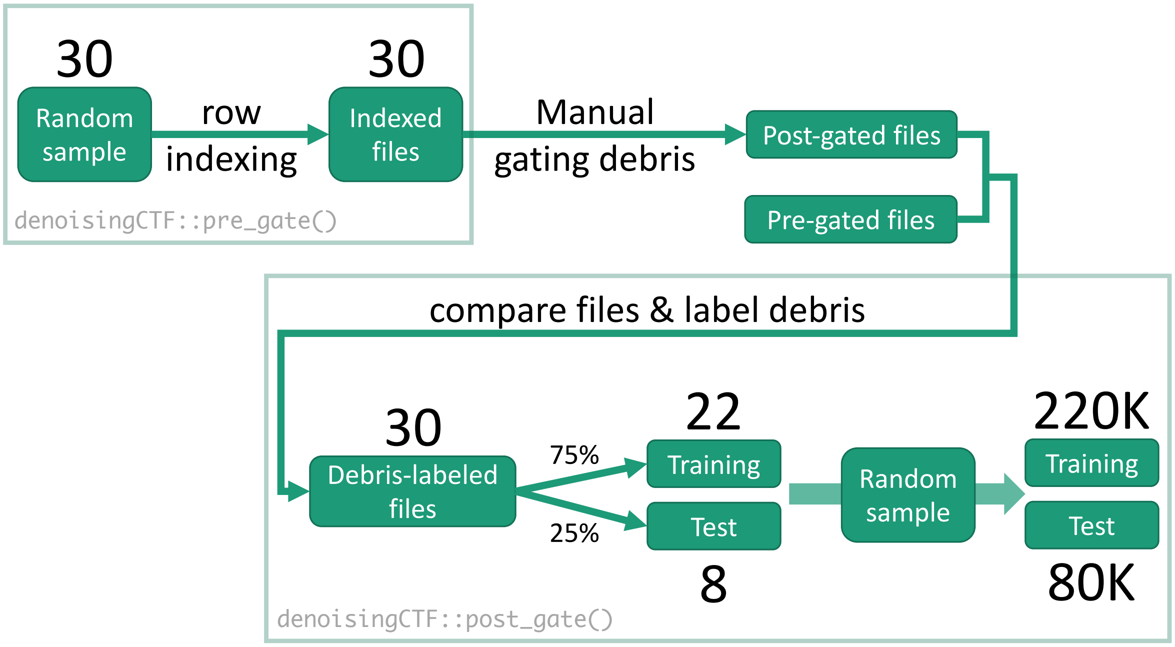


Figure 2.2: Debris model data workflow. Strategy to build the training and test sets for the debris classification model. A total of 220 000 and 80 000 events (i.e. cells, rows) were used for training and test sets, respectively.

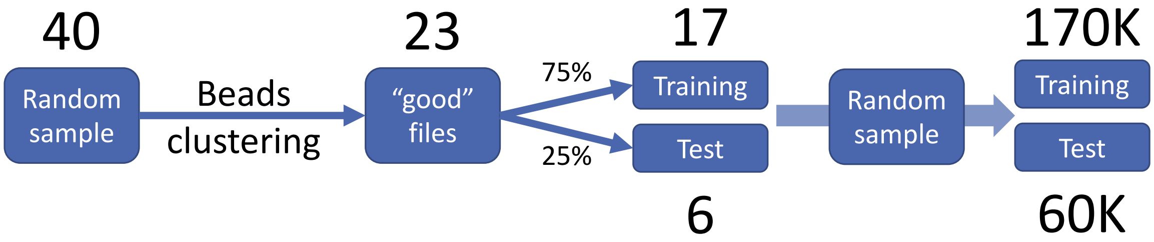


Figure 2.3: Beads model data workflow. Strategy to build the training and test sets for the beads classification model. A total of 170 000 and 60 000 events (i.e. cells, rows) were used for training and test sets, respectively.

2.5.4 Data analysis

2.5.4.1 Cell lines

For data shown in Figs. 3.1-3.2 we used the data acquired for each cell line individually, performed random equal subsampling (15,000 events per sample), and concatenated the files. UMAP plots shown in Figs. 3.1-3.2 were generated in R using all markers of Table 2.4, except for Histone H3. We used k-means for clustering analysis and applied the same markers. To determine the optimal number of clusters *k* to target, we used the 'elbow' criterion, for which the total within-cluster sum of squares was calculated for a range of values of *k* [79]. Clustering was performed with $k = 8$.

2.5.4.2 Human samples

For Chapter 3, to determine cellular identity, we performed k-means using markers that identify main cellular populations (EpCAM, CD31, CD45, vimentin, cytokeratin and cytokeratin7). We targeted for a large number of clusters (k=10) to allow for more granularity and prevent rare cell populations from being engulfed into dominant clusters. These were annotated based on protein expression and clusters with similar characteristics were merged. Final cell types were annotated as epithelial cancer cells, endothelial cells, mesenchymal cells and immune cells. Epithelial cancer cells were defined as EpCAM+/cytokeratin+/cytokeratin7+, endothelial cells as CD45-/CD31+, mesenchymal cells as vimentin+/CD45-/CD31-/EpCAM-/cytokeratin-/cytokeratin7- and immune cells as CD45+. We performed a second clustering round for immune cells only(k=10) using immune cell markers CD8, CD24, CD3, CD11b, CD56 and HLA-DR. Cluster were annotated into myeloid cells (CD45+ /CD3-/CD11b+), cytotoxic T cells (CD45+/CD3+/CD8+), helper T cells (CD45+/CD3+/CD4+) and other immune as the remaining CD45+ cells. Fig. 3.3A is a representation of the annotated cell types of the 10 tumors using the same markers from the two clustering rounds to generate the UMAP plots, for which we obtained a random sample without replacement for a total of 4000 events per sample. Epithelial cancer cells from each entire sample were subseted and clustered using k-means (k = 10) and the following markers: EpCAM, c-casp3, TP53, HLA-DR, HLA-ABC, CD31, thioredoxin, *beta*-catenin, HER2, p-STAT3, p-STAT5, p-STAT6, TTF1, p-AKT, Ki67, CD56, vimentin, MDM2, cytokeratin, MET, TP63, CK7, EGFR, CD44, p-ERK, CD24, p-S6, PDL1. Fig. 3.4A is a representation of the clusters of the 10 tumors using the same markers from the previous clustering to generate the UMAP plots, with random sampling without replacement for for a total of 2000 events per sample. For Chapter 4, to determine cellular identity, we performed k-means using markers that identify main cellular populations (EpCAM, CD31, CD45, vimentin, cytokeratin and cytokeratin7). The optimal number of clusters was determined by calculating the Within Cluster Sum of Squares (WSS) for different k values, plotting k vs WSS and choosing the k in which we see a pronounced bend or ”elbow” (k=10). The clusters were annotated based on protein expression and clusters with similar characteristics were merged. Final cell types were annotated as Epithelial cancer cells (EpCAM+/cytokeratin+/cytokeratin7+), Endothelial cells(CD45-/CD31+), Fibroblasts/Mesenchymal cells (vimentin+/CD45-/CD31-/EpCAM-/cytokeratin-/cytokeratin7-) and Immune cells (CD45+). We performed a second clustering round for immune cells only using immune cell markers CD8, CD4, CD3, CD11b, and CD56. Clusters were annotated into Myeloid cells (CD45+/CD3-/CD11b+), CD8+ T cells (CD45+/CD3+/CD8+), CD4+ T cells (CD45+/CD3+/CD4+), Double negative T cells (CD45+/CD3+/CD4-/CD8-) and Other immune as the remaining CD45+ cells. Each identified cell subset, including the non-immune cells, underwent an independent round of clustering using the protein markers showed in their corresponding heatmap (Fig. S2-S9, panel C). We then calculated the percentage of each subset per patient and compared cluster frequencies between groups using non-parametric test Wilcoxon rank-sum (Fig. S2-S9, panel D). For each cell type we calculated the Spearman correlation between protein markers (Fig. S2-S9, panel E). We also calculated the Spearman correlation of the proportion of cell type clusters among the patients (Fig. S10). Finally, we calculated the bulk median protein per patient and compared patients between groups using non-parametric test Wilcoxon rank-sum

(Fig. S11).

2.6 Multiplex immunofluorescence validation of CyTOF data

2.6.1 Tissue microarray

TMA was generated from lung tissue blocks from patients with LPS and SPS lung adenocarcinoma. Two tissue cores were used to represent one patient. First, specific cases were selected to match samples, analyzed by CytOF, next, every core was evaluated by pathologist to ensure tissue quality (no massive areas with necrosis, stroma, large vessels; no processing artefacts).

2.6.2 Staining

TMA paraffin blocks were cut into 5 μm sections. Hematoxylin Eosin staining was used for visual evaluation of morphology to ensure comparable tissue samples were used for analysis. Multiplexed Immunofluorescent (mxIF) stain was performed with following antibodies: anti-PanCK, Clone AE1/AE3 (Invitrogen); anti-CD45, Clone HI30 (Biolegend); anti-CD3 (Agilent Inc., Dako); anti-HLADR, Clone SPM288 (Novus Biologicals LLC.). Multistep mxIF staining was perform, where after blocking, in a first step tissue was incubated with mouse anti-CD45 antibodies, followed by Fab fragment anti-mouse-Cy3 (Jackson ImmunoResearch). Tissue was washed well to remove unbound antibodies, blocked with mouse IgG and incubated with directly conjugated mouse PanCK-FITC, HLADR-Cy7 and rabbit anti-CD3 antibodies. Next, after washing, CD3 was detected in additional step with anti-rabbit-Cy5 (Thermo Fisher Scientific) antibodies. Nuclei were stained with DAPI (Thermo Fisher Scientific). Slides were coverslip with prolong gold (Invitrogen) and dried overnight. Whole slide imaging was performed on Aperio Versa 200 (Leica) scanner.

2.6.3 Single cell analysis

To perform single cell analysis of multiplexed fluorescent stained images, image analysis pipeline was built in KNIME (Knime.com) analytical platform (KNIME 4.1.2 with integrated image processing and analysis extensions) [80, 81]. DAPI-stained images were used to generate nuclear masks using deep learning algorithm [82]. Cell segmentation was generated by circular outgrow of nuclear masks. Single cell features were extracted by aligning nuclear or cell masks to specific fluorescent stain images. Geometrical, statistical, and texture features were extracted for each segmented cell. For cell classifications, training set of positive and negative cells was annotated. These annotations along with extracted from each specific stain features, were used for machine learning where XG boost AI models were generated for each marker. These models were applied to whole data set and resulting probabilities with $p \geq 0.9$ cutoff were used for initial binary cell classification: "PanCK+ or PanCK-" "CD45+ or CD45-" "CD3+ or CD3-". Cell classification using combination of binary markers yielded following cell classes: "Epithelial/Tumor cells" (PanCK+CD45-CD3-), "T-cells" (CD3+CD45+PanCK-), "Immune (none-T) cells" (CD45+CD3-PanCK-), "Other cells" (CD45-CD3-PanCK-). Quantitative data from single cell features (such as X, Y coordinates, HLA-DR expression and etc.) was used for correlation and spatial analysis. Continuous scale of fluorescent signal was used to quantify HLA-DR expression on tumor cells. For this, signal intensities normalized to DAPI (sums fluorescent signals) were used. Total cell number and specific class cell number per image were quantified and percent calculations were made. Correlation between HLA-DR expression on Tumor cells and T cell number was determined by Spearman's rank-order correlation test. In neighborhoods of 100 micrometers diameter for each (processing) Tumor cell, HLA-DR median signal intensity on neighboring Tumor cells and number of T cells were calculated in Python and used as inputs for correlation analysis. Spatial analysis was performed in KNIME by calculation of distances from each T cell to nearest 1st and 2nd Tumor cell using similarity search node.

2.7 TCGA LUAD data set

Fragments Per Kilobase of transcript per Million (FPKM) normalized read counts of RNA-Seq from LUAD patients and matching clinical data were downloaded from National Cancer

Institute Genomic Data Commons Data Portal (https://portal.gdc. cancer.gov/projects/TCGA-LUAD).

2.8 Cell type enrichment analysis with xCell

Using TCGA data, we selected patients with disease stage between I and III. After applying log transformation ($log_2(FPKM+1)$) we computed the quantiles of expression of MHC-II related genes. Patients were labeled as ”low” if the expression of the gene in question was below the first quantile (25%) and ”high” if it was higher than the third quantile (75%). Cell type enrichment analysis results for TCGA data were downloaded from the xCell website (https://xcell.ucsf.edu/) and patient groups were compared.

2.9 Whole Exome Sequencing

2.9.1 Sample preparation and data acquisition

DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) following the kit protocol. A quantitation and integrity assessment were completed using the whole genomic DNA. An aliquot of each sample was analyzed on the Agilent TapeStation and quantitated using a Picogreen assay. The samples were normalized and plated using the BioMek FX liquid handler. Libraries were prepared using 12-50 ng of DNA and the Twist Biosciences library preparation kit (P/N 104207) per manufacturer’s instructions. Libraries were then captured using the Twist Comprehensive Exome panel (P/N 102031). Individual libraries were assessed for quality using the Agilent 2100 Bioanalyzer and quantified with a Qubit Fluorometer. The adapter ligated material was evaluated using qPCR prior to normalization and pooling for sequencing on the QuantStudio 12K Flex. The libraries were sequenced using the NovaSeq 6000 instrument with 150 bp paired end reads. RTA (version 2.4.11; Illumina) was used for base calling and data QC was completed using MultiQC v1.7. Each sample was analyzed using the DRAGEN Enrichment Pipeline v3.7.5 to calculate alignment and capture metrics.

2.9.2 Data preprocessing

Sequence data from genomic DNA were aligned to the reference human genome (GRCh38) by BWA aligner[83]. For quality Control purpose, multiple stages of quality control (QC) on sequencing data were carried out. Raw data QC was performed by FastQC[84] and QC3[85]. Alignment QC and Variants QC were performed using QC3[85]. GATK software 4.1.8.1 was used for somatic single nucleotide variants (SNVs), short insertion and deletion variant (INDELs), and somatic CNV calling[86]. Briefly, the reads pre-processing (RealignerTargetCreator, IndelRealigner, BaseRecalibrator) was performed as described in GATK Best Practices Workflows[86]. Then MuTect2 [87] was used for somatic mutation (SNVs and INDELs) calling and GATK was used for somatic CNV calling. All the identified variants were annotated by ANNOVAR to gene and transcript level[88]. All variants outside the target regions or synonymous variants were removed. Then all the variants were annotated to public database including dbSNP[89], Exome Aggregation Consortium (ExAC)[90], NHLBI GO Exome Sequencing Project (ESP) and COSMIC[91]. To remove possible germline mutations, variants reported in dbSNp or ExAC or ESP with minor allele frequency in normal population larger than 1% were removed.

2.9.3 Data analysis

The resulting processed file (Mutation Annotation Format, MAF), was analyzed using the R package maftools[92]. We used the Oncoplot to visualize the top 25 mutated genes, and the Forest plot to compare Indolent + Intermediate tumors versus Aggressive and identify the significantly mutated genes (Fig. S12A,C). Finally, we calculated the Spearman correlation between the SILA score and the logarithm base 10 of the mutational load (number of mutations per patients) (Fig. S12B).

2.10 Bulk RNA Sequencing

2.10.1 Sample preparation and data acquisition

RNA was extracted using the RNeasy Plus Mini Kit (Qiagen) following the kit protocol. RNASeq libraries were prepared using 300 ng of RNA and the NEBNext Ultra II Directional RNA Library Prep kit (NEB, Cat: E7760L). Fragmentation, cDNA synthesis, end repair/dA-tailing, adaptor ligation and PCR enrichment were performed per manufacturer's instructions. Individual libraries were assessed for quality using the Agilent 2100 Bioanalyzer and quantified with a Qubit Fluorometer. The adapter ligated material was evaluated using qPCR prior to normalization and pooling for sequencing. The libraries were sequenced using the NovaSeq 6000 with 150 bp paired end reads. RTA (version 2.4.11; Illumina) was used for base calling and data QC was completed using MultiQC v1.7 by the Vanderbilt Technologies for Advanced Genomics (VANTAGE) core (Vanderbilt University, Nashville, TN).

2.10.2 Data preprocessing

Quality control (QC) analysis was performed on all sequencing reads using FastQC package developed by the Babraham Institute bioinformatics group. Reads with poor quality were trimmed and adapter sequences were removed by cutadapt g. Reads were then aligned to human genome (hg38) using STAR[93] and quantified by featureCounts[94]. Alignment quality was checked by QC3[85]. Any RNA-Seq experiment with poor quality was removed.

2.10.3 Data analysis

Starting from the raw counts, we removed low variance genes and filtered out genes from chromosomes X and Y. We used the package DESeq2 to perform differential gene expression analysis[95] and the package fgsea for the Gene Set Enrichment Analyses[96] with the Molecular Signature Database (MSigDB) hallmark gene set collection[97] and the REACTOME database[98]. The transcription factor activity was inferred using the VIPER package[99]. Individual pathways scores per patient sample were obtained using the Gene

Set Variation Analysis (GSVA) tool[100]. Liberzon2015Gillespie2022

2.11 Single Cell RNA Sequencing

2.11.1 Sample preparation and data acquisition

After dead cell removal with MACS Dead Cell Removal Kit, (Miltenyi Biotec, Germany), cells (5,000-10,000 cells per sample) were submitted for processing using the 10X Genomics platform. Libraries were prepared using P/N 1000006, 1000080, and 1000020 following the manufacturer's protocol. The libraries were sequenced using the NovaSeq 6000 with 150 bp paired end reads. RTA (version 2.4.11; Illumina) was used for base calling and analysis was completed using 10X Genomics Cell Ranger software v4.0.0.

2.11.2 Data preprocessing

We used 10x Genomics Cell Ranger 4.0.0 software to obtain the feature barcode matrices per sample. For further preprocessing steps we used the scanpy tool[101]. For more details see https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html.

2.11.3 Data analysis

We first computed a principal component analysis to reduce the dimensionality of the data and then computed a neighborhood graph on the first 40 principal components. We then used the Leiden graph-clustering method[102] and obtained 25 clusters which then were annotated into 7 major cell types: B cells, Cancer cells, Endothelial cells, Mural cells, Myeloid cells and T cells. We calculated the cell type proportions for each patient and compared indolent versus aggressive tumors using non-parametric test Wilcoxon rank-sum. Each cell type underwent an additional clustering step and again cluster proportions between groups were compared. To better understand the identity of the clusters we used the split violin visualization from scanpy, and showed the top 30 marker genes for each cluster when compared to the rest.

2.12 Data integration

For the data integration effort, we selected only the features that were significantly associated with tumor behavior. From the CyTOF dataset, we included the cell type cluster proportions and the bulk protein expression per patient. In the latter, for a protein marker to be considered, the median of at least one patient group (indolent, intermediate or aggressive) should be above 1.44, which in raw values (before the arcsinh transform) correspond to 10 ”pushes” which is the default lower limit of the Helios^{TM}[103]. From the RNA-Seq data set, we selected all the pathways with adjusted p value ¡ 0.05, and a normalized enrichment score (NES) ¿ 1.5. We then used the GSVA package to calculate individual expression scores of these pathways for each patient. For the HealthMyne radiomics features dataset, we performed a Spearman pairwise correlation against the SILA score and selected only those significantly correlated (adjusted p value ¡ 0.05). Only patients with complete data were selected, all the matrices concatenated and the features were scaled and centered. Patients and features were clustered independently using k means (k=4, by elbow method as described in the CyTOF methods section). Cluster IDs for each patient and feature can be found in Tables S11-12. To visualize the feature interactions we computed a similarity matrix and also performed a PCA and plotted the first two components for both features and patients.

2.13 Statistical analysis

For correlation analysis we used Spearman’s rank correlation test and adjusted p-values for multiple hypothesis using the Benjamini & Hochberg method [104]. Comparison of categorical variables was performed using the Mann-Whitney U test. Survival curves were generated using the Kaplan-Meier method, and statistically significant differences were analyzed with the log rank test. All statistical tests were two-sided and p values less than 0.05 were considered statistically significant. The analyses were performed in R 4.0.3 and Python 3.

2.14 Code Availability

All the code used to analyze the data and generate the visualizations and tables can be accessed at https://github.com/msenosain/TMA36_data-analysis.

CHAPTER 3

HLA-DR cancer cells expression correlates with T cell infiltration and is enriched in lung adenocarcinoma with indolent behavior

3.1 Acknowledgements

This chapter is adapted from “HLA-DR cancer cells expression correlates with T cell infiltration and is enriched in lung adenocarcinoma with indolent behavior” published in Scientific Reports and has been reproduced in line with publisher policies. [105]

3.2 Abstract

Lung adenocarcinoma (LUAD) is a heterogeneous group of tumors associated with different survival rates, even when detected at an early stage. To investigate whether CyTOF identifies cellular and molecular predictors of tumor behavior. We developed and validated a CyTOF panel of 34 antibodies in four LUAD cell lines and PBMC. We tested our panel in a set of 10 LUADs, classified into long- (LPS) (n=4) and short-predicted survival (SPS) (n=6) based on radiomics features. We identified cellular subpopulations of epithelial cancer cells (ECC) and their microenvironment and validated our results by multiplex immunofluorescence (mIF) applied to a tissue microarray (TMA) of LPS and SPS LUADs. The antibody panel captured the phenotypical differences in LUAD cell lines and PBMC. LPS LUADs had a higher proportion of immune cells. ECC clusters (ECCc) were identified and uncovered two LUAD groups. ECCc with high HLA-DR expression were correlated with CD4+ and CD8+ T cells, with LPS samples being enriched for those clusters. We confirmed a positive correlation between HLA-DR expression on ECC and T cell number by mIF staining on TMA slides. Spatial analysis demonstrated shorter distances from T cells to the nearest ECC in LPS. In conclusion, our results demonstrate a distinctive cellular profile of ECC and their microenvironment in LUAD. We showed that HLA-DR expression in ECC is correlated with T cell infiltration, and that a set of LUADs with high abundance of HLA-DR+ ECCc

and T cells is enriched in LPS samples. This suggests new insights into the role of antigen presenting tumor cells in tumorigenesis.

3.3 Introduction

Recently, the National Lung Screening Trial (NLST) reported a 20% relative mortality risk reduction using low-dose computed tomography (CT) over chest X-ray screening [16]. However, lung tumors detected through CT screening range from indolent to aggressive. Aggressive lung cancers have doubling times of 50 to 150 days, yet CT screening has been shown to detect slow growing tumors with doubling times of 400 days or more [106]. Lung cancer screening bears the inherent risk of overdiagnosis in up to 18% of tumors [107]. Recent efforts in radiomics have been reported to predict this phenomenon, however its biological determinants remain unknown [19, 108, 109].

LUAD is a highly heterogeneous disease. Assuming that subpopulations may be responsible for a particular behavior, these may be rare and difficult to detect at an early stage with standard bulk analyses [5, 6, 7]. Until recently, the molecular profiling of tumors has been based on an average phenotype of hundreds of thousands of cells, including neoplastic cells and cells of the tumor microenvironment (TME). Although this approach has proven to be useful in many applications, there is a significant loss of information, particularly affecting the detection of rare cell subsets that could be responsible for cancer initiation, plasticity and recurrence. Emerging single-cell technologies can overcome such limitation, providing high resolution information essential for a better understanding of the tumor cellular composition [9]. Among those, mass cytometry is a rapidly evolving technology capable of measuring the expression of $\sim$40 proteins on individual cells using antibodies labeled with heavy metal isotopes [110]. To date, some studies have investigated LUAD from a single-cell perspective [64, 67, 68, 69, 111, 112], however the molecular determinants of early LUAD behavior as for why some tumors progress faster than others remain unknown.

Here, we hypothesized that single-cell proteomic analysis of early stage adenocarcinoma of the lung will provide new insights into the cellular and molecular determinants of indolent and aggressive tumors which in turn may offer novel and personalized avenues for intervention. We developed a comprehensive mass cytometry antibody panel that will allow us to investigate LUAD behavior, which includes markers for cellular lineage, tumor cell markers and signaling pathways. To this end, we have validated our panel using LUAD cell lines and PBMC and we present the analysis of a set of ten early stage primary LUADs of the lung with indolent and aggressive behaviors showing some valuable insights on immunogenicity of the tumors.

3.4 Results

3.4.1 LUAD mass cytometry antibody panel captures the cellular diversity between LUAD cell lines and PBMC

To validate our mass cytometry panel, we used a combination of LUAD cell lines that harbor different mutations and therefore have different protein expression patterns (Table 2.1). We also included PBMC from a healthy donor in the mix to mimic the immune cells that can be found in a tumor. All cells were pooled in the same proportion, stained and run through the CyTOF machine as a single sample. Additionally, cells were run separately to confirm our findings. Protein expression by cell line was consistent across replicates (see Appendix A Fig. S1-S7). Dimensionality reduction algorithm UMAP [113] allowed us to visualize the multiple parameters measured in a two dimensional map (Fig. 3.1A-B). Our panel captured phenotypic differences among the cell lines and PBMC in the parameter space, visualized as independent islands in the UMAP plot (Fig. 3.1A). Epithelial markers EpCAM, pan-cytokeratin and cytokeratin 7 were positive in LUAD cell lines, but not always expressed on the same cells (Fig. 3.1B). Receptor tyrosine kinases EGFR and MET were highly expressed in all LUAD cell lines as expected. Cell line H3122 was positive for TTF1 as previously reported [75], and cell lines PC9 and H23 which harbor inactivating TP53 mutations expressed high levels of the latter (Table 2.1). A549 expressed high levels of CD24. Human PBMC were all CD45 positive and divided into three major islands: CD3+ CD4+ (T helper cells), CD3+ CD8+ (cytotoxic T cells), and CD3- CD11b+ cells (myeloid cells). Additionally, basal kinase activity as represented by phosphorylation of ERK, S6, STAT5 and, in lesser degree, AKT was detected mostly in LUAD cell lines, reflecting the constitutive activation of these pathways (Fig. 3.1C).

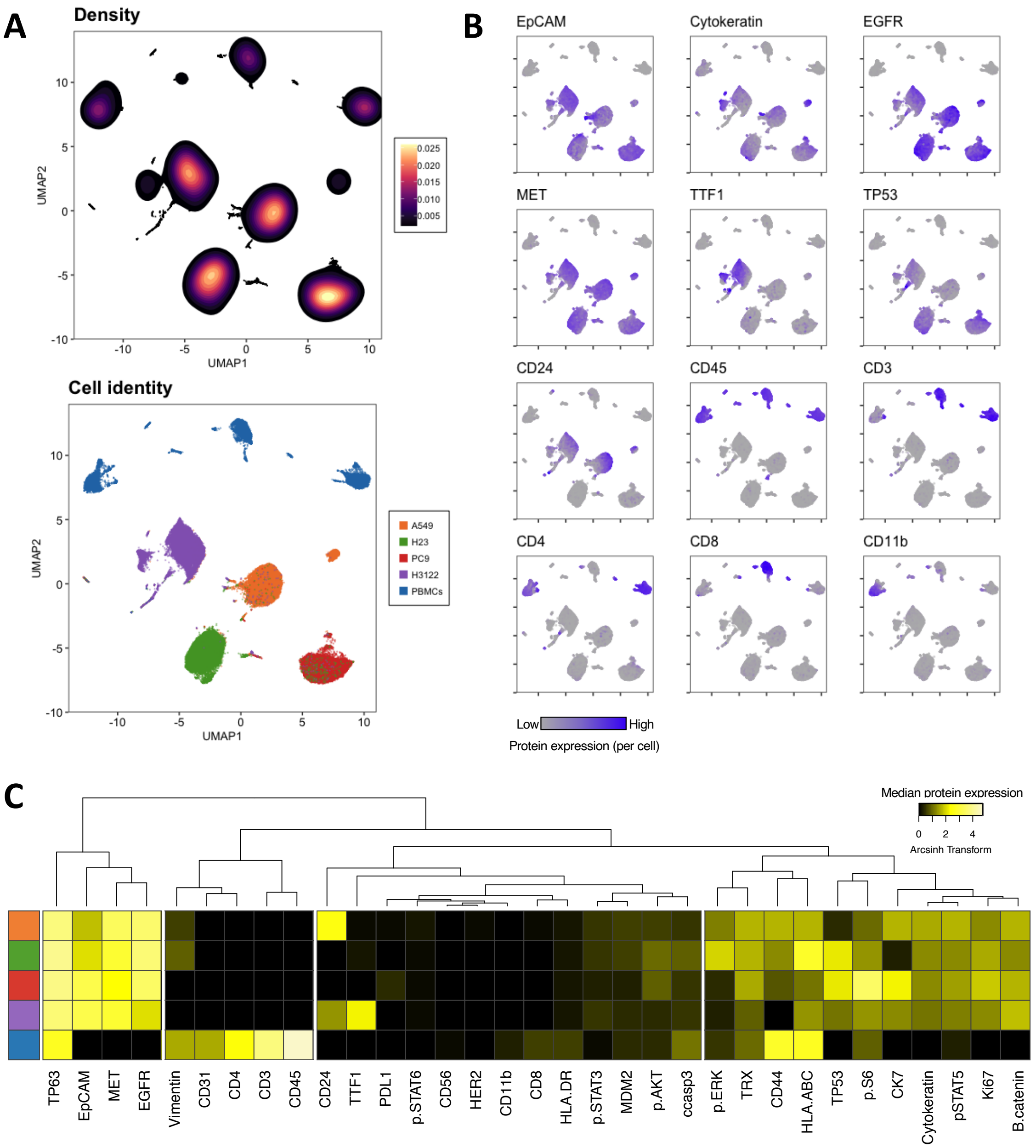


Figure 3.1: Mass cytometry panel and unsupervised computational analysis capture cellular diversity in LUAD cell lines and PBMC. (A) Density (above) and cell identity (below) UMAP representations show separation of the cellular populations based on single-cell protein expression. (B) UMAP plots correspond to the same cells from (A) showing single cell expression of the labeled protein. (C) Heatmap shows median protein expression of arcsinh transformed values (cofactor = 5) for each protein on each cell population. Colors on the left represent the cellular populations and match those represented in (A).

To test if our clustering strategy was successful in identifying the different cell types in the mix, we determined the optimal number of clusters and studied their composition. To determine the optimal number of clusters k to target with k-means clustering, we used the ‘elbow’ criterion, for which the total within-cluster sum of squares was calculated for a range of values of k [79]. Clustering was performed with $k = 8$. The resulting clusters represented with high accuracy the different cell types present in the mix (Fig. 3.2). Cluster 2 was 94.6% composed by H23 cells, cluster 3 was 97.4% composed by A549 cells; cluster 5 was 86% composed by H3122 cells and cluster 7 was 90% composed of PC9 cells. For the immune clusters, clusters 4, 6 and 8 were 100% composed by PBMC. Based on their protein expression, these could be annotated as CD11b+ monocytes, CD8+ T cells and CD4+ T cells, respectively. Finally, cluster 1 is a mix of cells dominated by A549 and H3122 cells, driven by a high pan-cytokeratin and cytokeratin 7 expression. Altogether, these results show that our mass cytometry antibody panel can successfully identify different cancer subsets as well as some immune populations.

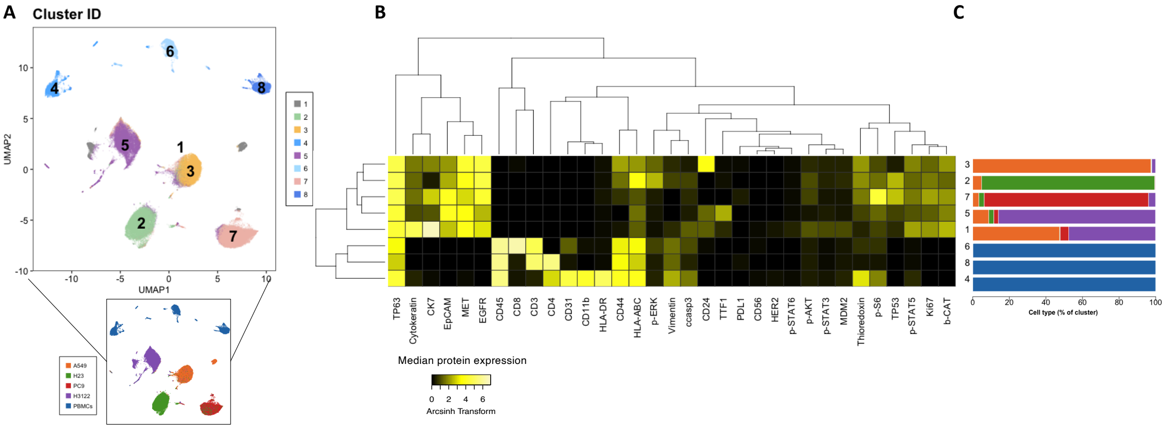


Figure 3.2: Clustering analysis of LUAD cell lines and PBMC. (A) UMAP plot is the same as in Fig. 1 but colors represent 8 clusters obtained with k-means. (B) Heatmap shows median protein expression of arcsinh transformed values (cofactor = 5) for each protein on each cluster. (C) Stacked barplots represent cluster composition (percentage per cell type). Colors match those represented in (A)(bottom).

3.4.2 Mass cytometry analysis identifies main cell types in LUADs and captures differences between tumors with long and short predicted survival

LUADs human samples characterized by different predicted behavior classified into long- (LPS) (n = 4) and short-predicted survival (SPS) (n = 6) were stained with our antibody panel (Table 2.4, see Appendix A Fig. S8). We identified the major cell types (ECC, endothelial, mesenchymal and immune cells) based on the expression of protein markers (Fig. 3.3B). EpCAM+/pan-cytokeratin+/cytokeratin 7+ cells were annotated as ECC; CD31+/CD45- cells were annotated as endothelial cells; vimentin+ CD31- CD45- and negative for epithelial markers cells were annotated as mesenchymal cells. All CD45+ cells and negative for epithelial markers were annotated as immune cells. The latter were further classified into T helper cells (CD3+/CD4+/CD8-), cytotoxic T cells (CD3+/CD8+/CD4-), myeloid cells (CD11b+/CD3-) and the remainder CD45+ cells were annotated as ”Other immune”. While the number of cells acquired varied between samples, we included all cells collected for each tumor in the analysis and used the cell type relative abundances (i.e. percentages) for comparisons.

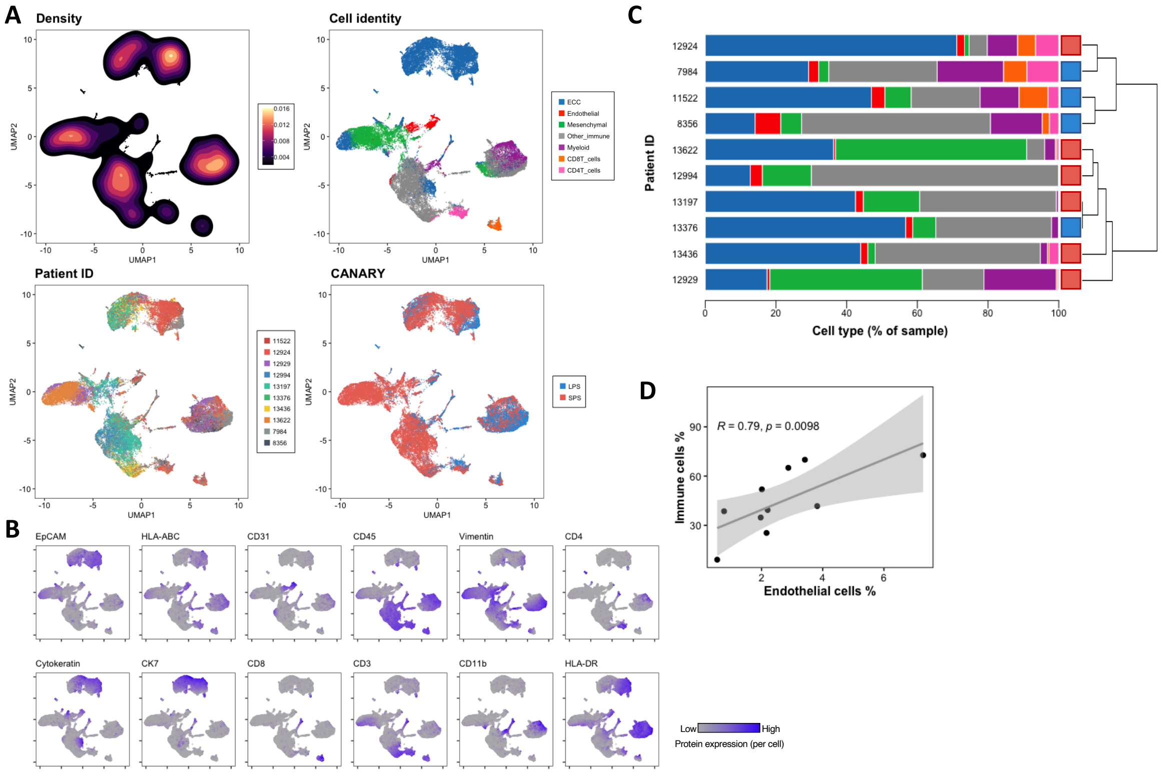


Figure 3.3: Mass cytometry antibody panel distinguishes epithelial and non-epithelial cell types in 10 early LUADs. (A) UMAP plots of a random sample of 4000 cells per patient colored by Density, Cell identity, Patient ID and CANARY prediction. Seven cell types were identified based on k-means clustering and marker expression profiles. Patient CANARY risk stratification is represented as a light blue for long-predicted survival (LPS) and dark blue for and short-predicted survival (SPS). (B) UMAP plots correspond to the same cells from (A) showing single cell expression of selected labeled protein. (C) Stacked barplots with cell type percentage per patient. Colors match those in (A) Cell identity plot. Dendrogram was calculated from a patient-patient Spearman correlation matrix. (D) Spearman correlation analysis of the relative abundance of immune cells vs. endothelial cells.

Fig. 3.3A is a representation of an equal sampling of annotated cell types of the 10 tumors using dimensionality reduction algorithm UMAP [113]. Cell types separated based on their marker expression (Fig. 3.3A, Cell identity). Additionally, events (i.e. cells) did not cluster by sample but were mixed among the different islands in the plot (Fig. 3.3A, Patient ID). We further investigated the distribution of these cell types across the 10 tumors by performing hierarchical clustering on the correlation matrix based on the subpopulations relative abundances (Fig. 3.3C). Samples clustered in two main groups, one enriched in T cells and myeloid cells and one with lower to no abundance of those cell types and higher abundance of mesenchymal cells on average. The first group of samples was composed by 3 LPS samples (7984, 11522, 8356) and one SPS sample (12924). The other group of samples was mainly composed of SPS samples (13622, 12994, 13197, 13436, 12929) and one LPS sample(13376) (see Appendix A Table S1). Additionally, we found a statistically significant positive correlation between endothelial cells and immune cells in the LUAD samples (Fig. 3.3D, see Appendix A Fig. S9). When LPS and SPS tumor samples were compared, we found that LPS had a higher median percentage of endothelial cells and immune subtypes, whereas SPS samples had a higher median percentage of fibroblasts/mesenchymal cells (see Appendix A Fig. S10) We compared LPS vs SPS protein expression by cell types (see Appendix A Fig. S10-S16). We found a tendency towards a higher expression of HLA-DR and HLA-ABC in endothelial cells from LPS tumors (see Appendix A Fig. S11). In epithelial and mesenchymal cells there was higher HLA-DR expression in LPS compared to SPS tumors, with the latter cell type showing a significant difference (p=0.038) (see Appendix A Fig. S12-S13). The immune cells as a whole also showed a tendency towards higher HLA-DR expression in LPS tumors (see Appendix A Fig. S14). CD8+ T cells showed a significantly higher expression of HLA-ABC in LPS tumors (p=0.032) (see Appendix A Fig. S15). CD4+ T cells showed a tendency towards higher expression of activation marker CD44 in LPS tumors (see Appendix A Fig. S16). Finally, myeloid cells presented a tendency towards higher expression of HLA-ABC and HLA-DR in LPS tumors (see Appendix

A Fig. S17). To confirm that the HLA-DR higher expression in most cell types of LPS tumors was not due to an artifact of the antibody, we assessed the expression of this protein in our batch control cell lines A549 and Ramos (see Appendix A Fig. S18). Results were consistent across batches, with A549 showing minimal expression of HLA-DR and Ramos showing high expression of the protein in question as expected.

Based on these results, we conclude that our mass cytometry antibody panel enables the identification of major cell types in LUADs, allowing for comparison across tumors of different predicted behavior. We found that our set of samples divided in two main groups based on their cellular composition, one enriched on T cells (LPS predominant) and one depleted on T cells (SPS predominant). Additionally, we found a tendency towards a higher HLA-DR expression in LPS samples, suggesting an immunogenic profile on these tumors.

3.4.3 Unsupervised analysis of ECC reveals HLA-DR+ subsets associated with T cell infiltration

Because distinct subpopulations of malignant cells have been associated with disease outcome [7], we tested whether our antibody panel detects different subsets of ECC and whether LPS or SPS tumors are particularly enriched for any subset. We computationally extracted the ECC of each tumor from the pool of cells (Fig. 3.3).

We used $k = 10$ to achieve more granularity and dig deeper into the differences of the ECC. Fig. 3.4A is an equal-sampling representation of the 10 ECCc of the 10 LUAD samples using dimensionality reduction algorithm UMAP [113]. ECCc separated based on their protein expression (Fig. 3.4A-B, Cluster ID) and cells did not grouped by sample but were mixed among the different islands in the plot (Fig. 3.4A, Patient ID). We then assessed the sample ECCc composition across the 10 tumors by hierarchical clustering on the correlation matrix based on the cluster relative abundances as described above (Fig. 3.4C). A first set of samples with very similar profile composed by 3 LPS samples(7984, 11522, 8356) and one

SPS sample (12924) were enriched in clusters 7, 8 and 9, which have a high expression of HLA-DR, TTF1, *beta*-catenin, and all three epithelial markers EpCAM, pan cytokeratin and cytokeratin 7. This group of LUADs is composed by the same patients that clustered together in Fig. 3.3C as well. Another set of LUADs composed by 3 SPS samples (13436, 13197, 12994) and one LPS sample (13376) were enriched in clusters 1, 3 and 6, which are HLA-DR and TTF1 negative. Within this group, SPS samples 13197 and 12994 were also enriched in cluster 4, which is also HLA-DR and TTF1 negative and has high vimentin expression. A last set of 2 SPS samples (13622, 12929) were enriched in clusters 5 and 10, which present high expression of vimentin, MDM2 and p-STAT3, and are negative for HLA-DR, TTF1 and *beta*-catenin. When we assessed the correlation of these epithelial clusters with the other cell types in the TME, we found that 3 clusters were significantly correlated with some immune subsets (Fig. 3.4D, see Appendix A Fig. S19). Epithelial cancer clusters 7, 8 and 9 were significantly correlated with CD4+ (r=0.96, p<2.2e-16; r=0.9, p<0.001; r=0.78, p=0.012) and CD8+ T cells (r=0.95, p<2.2e-16; r=0.89, p=0.0014; r=0.76, p=0.016). Interestingly, these specific clusters as described above, are characterized by high HLA-DR, TTF1 and *beta*-catenin, among which the former has been associated with an immunogenic profile and favorable prognosis in several cancers [114, 115].

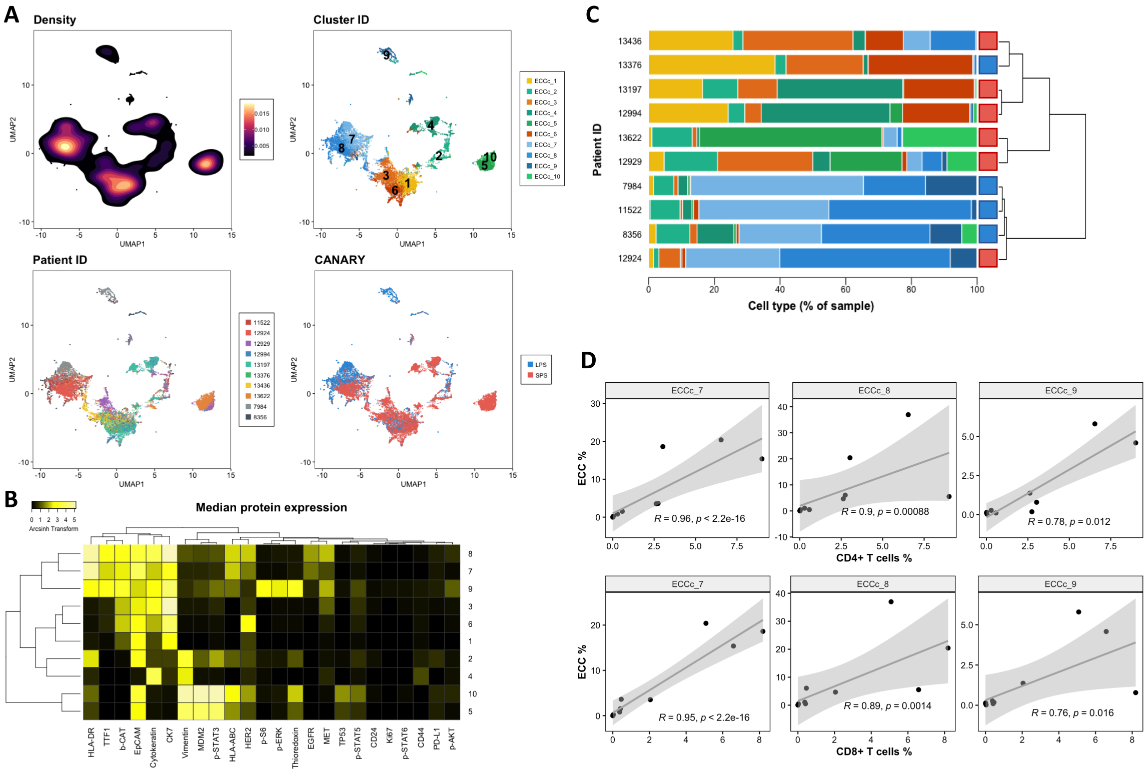


Figure 3.4: Unsupervised analysis of ECC reveals intra- and inter-tumor heterogeneity. (A) UMAP plots of a random sample of 2000 ECC per patient colored by Density, Cell identity, Patient ID and CANARY prediction. Ten clusters were obtained based on k-means clustering. Patient CANARY risk stratification is represented as a light blue for long-predicted survival (LPS) and dark blue for and short-predicted survival (SPS). (B) Heatmap shows median protein expression of arcsinh transformed values (cofactor = 5) for each protein on each ECCc. (C) Stacked barplots with ECCc percentage per patient. Colors match those in (A). Dendrogram was calculated from a patient-patient Spearman correlation matrix. (D) Spearman correlation analysis of the relative abundance of ECCc 7, 8 and 9 vs CD4+ and CD8+ T cells, respectively.

Thus, our results show that this mass cytometry antibody panel allows the detection of subpopulations of malignant epithelial cells. Based on the cellular subsets described here, we found a high degree of intra- and inter-tumor heterogeneity. Furthermore, a significant positive correlation of HLA-DR+ ECCc with T cell infiltration and the enrichment of HLA-DR+ ECCc predominantly in LPS tumors suggests the occurrence of an immunogenic process that may be associated with a more favorable outcome.

3.4.4 Validation with mIF suggests immunogenic profile in LSP tumors and RNA-Seq-based cell type enrichment analysis of independent cohort supports findings

To validate our mass cytometry results and to gain insights into the spatial distribution of cellular interactions, we used mIF staining of TMA sections of LUAD. We generated a TMA from lung tissue blocks from patients with LPS and SPS LUAD, using two tissue cores per patient. Cases were selected to match samples analyzed by CyTOF and every tissue core was evaluated by a pathologist to ensure tissue quality (no areas of necrosis, predominant stroma or large vessels. With the exception of one patient sample (ID 7984) which stained cores were excluded due to a significant loss of material during staining, all CyTOF samples were included in this analysis along with some extra to increase statistical power. Fluorescent staining was performed for PanCK, CD45, CD3, HLA-DR, DAPI. Slides were scanned and images were extracted. Cell nuclei were segmented using deep learning algorithm (cellpose.org) [82] and were further processed in KNIME analytical platform where cell segmentation, feature extraction and cell classification were performed [80]. Using a combination of binary markers we annotated the following cell types: “ECC/Tumor cells” (PanCK+CD45-CD3-), “T-cells” (CD3+CD45+PanCK-), “Immune (none-T) cells” (CD45+CD3-PanCK-), “Other cells” (CD45-CD3-PanCK-). Quantitative data from single cell features (such as X, Y coordinates, HLA-DR expression and etc.) was used for correlation and spatial analysis (Fig.3.5A-C). We computed the correlation between HLA-DR expression on tumor cells and T cell number by Spearman’s rank-order correlation test. For this, in neighborhoods of

100 micrometers diameter for each (processing) tumor cell, HLA-DR median signal intensity on neighboring tumor cells and number of T cells were calculated and used as inputs for correlation analysis. We found a significant positive correlation of HLA-DR expression in tumor cells and T cell number (r=0.25, p=2.2e-5), confirming our previous findings (Fig.3.5B, Fig.3.4D). Next, spatial analysis was performed in KNIME by calculation of distances from each T cell to nearest 1st and 2nd tumor cell. T cells in LPS tumors showed a shorter distance to the first tumor cell compared to SPS tumors (Fig.3.5C, see Appendix A Fig. S20), demonstrating that LPS tumors are more immunogenic than SPS tumors. These results support our CyTOF findings and further demonstrate by spatial analysis that LPS tumor cells are in closer proximity with T cells compared to SPS tumors, suggesting that the HLA-DR and T cell infiltration play an important role in the indolent behavior of these tumors.

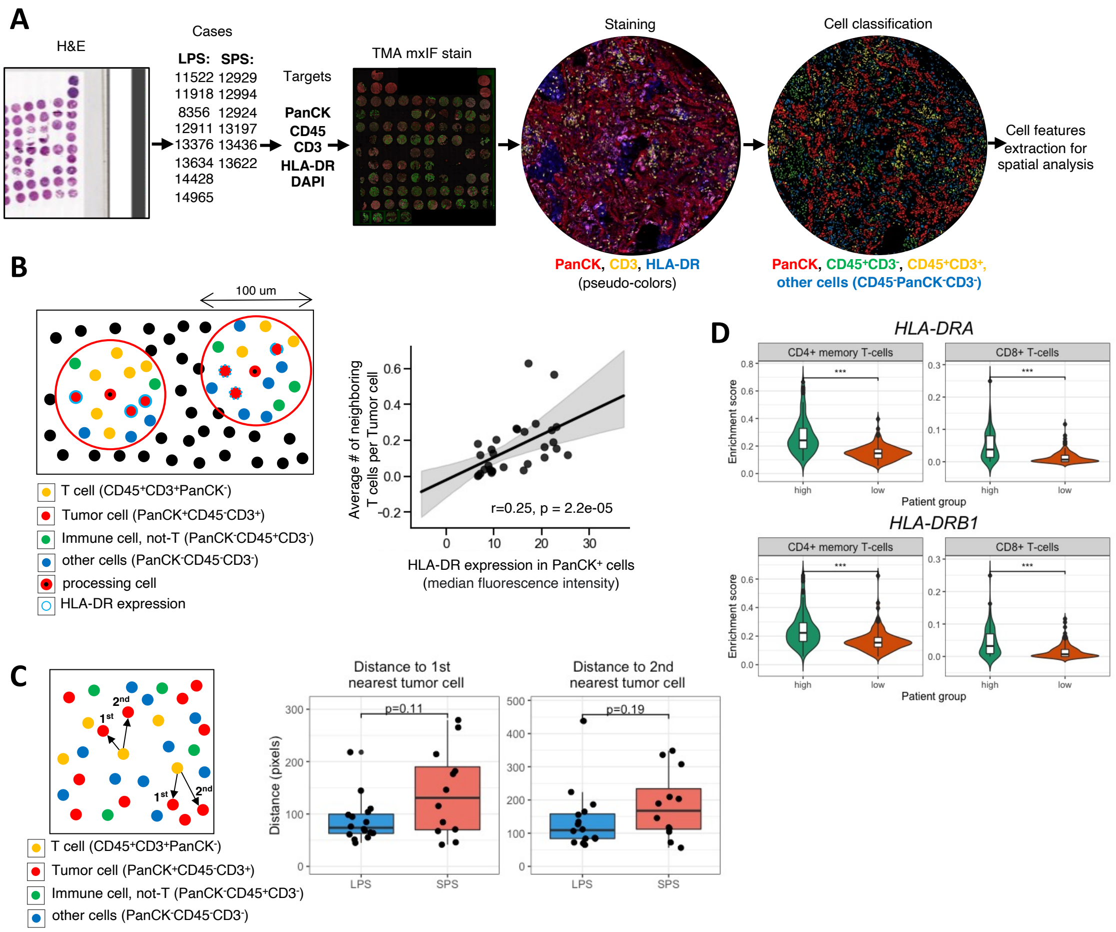


Figure 3.5: Validation by mIF on matching samples and cell enrichment analysis on RNA-Seq data from TCGA (A) Experiment design. TMA was generated from lung tissue blocks from patients with LPS and SPS lung adenocarcinoma. Two tissue cores were used to represent one patient. Fluorescent staining was performed for PanCK, CD45, CD3, HLA-DR, DAPI. Slides were scanned and images were extracted. Cell nuclei were segmented using deep learning algorithm (cellpose.org) and were further processed in KNIME analytical platform. Cell classification using combination of binary markers yielded following cell classes: "ECC/Tumor cells" (PanCK+CD45-CD3-), "T-cells" (CD3+CD45+PanCK-), "Immune (none-T) cells" (CD45+CD3-PanCK-), "Other cells" (CD45-CD3-PanCK-). (B) Correlation between HLA-DR expression on Tumor cells and T cell number was determined by Spearman's rank-order correlation test. For this, in neighborhoods of 100 micrometers diameter for each (processing) Tumor cell, HLA-DR median fluorescence intensity in Tumor cells and average number of neighboring T cells per sample were calculated and used as inputs. (C) Spatial analysis was performed in KNIME by calculation of distances from each T cell to nearest 1st and 2nd Tumor cell. (D) Cell enrichment analysis on LUAD RNA-Seq data from TCGA using xCell, comparing enrichment of CD4+ memory T cells and CD8+ T cells between patients with high (n=120) vs. low (n=120) gene expression of *HLA-DRA* and *HLA-DRB1*. Significance was assessed by Mann-Whitney U test (*** = pvalue <0.001).

Finally, acknowledging the limited sample size of our study we decided to further validate our results using the LUAD cohort from The Cancer Genome Atlas Research Network (TCGA). In a recent study, Ma and colleagues used the same cohort and found that the top pathways associated with better prognosis were enriched for immune cell signaling-related pathways, and that MHC-II genes were among the common genes shared by these pathways[116]. When performing survival analysis they found that up-regulation of MCH-II genes was significantly associated with an improved overall survival rate. Taking these results into account, we decided to take a step further and performed cell type enrichment analysis on the same RNA-Seq data using xCell, a gene signatures-based method robustly trained and validated that identifies immune and stroma cell types[117]. When comparing samples with high vs low expression of MHC-II-related genes we found that those with high expression had significantly higher enrichment scores for multiple T cell subtypes such as CD4+ memory T cells and CD8+ T cells (Fig.3.5D, see Appendix A Table S3). Altogether, these results provide an additional validation to our findings and highlighting the potential role of HLA-DR in tumor behavior and prognosis of LUAD.

3.5 Discussion

Predicting behavior of early detected LUAD presents a major challenge to patients and their providers. In this study, we presented the development and validation of a mass cytometry antibody panel that aims to further our understanding of the biological determinants of early LUAD behavior and thus improve the discrimination between indolent and aggressive tumors. First, we tested our panel in LUAD cell lines and PBMC and showed that dimensionality reduction and unsupervised clustering algorithms performed optimally. We were able to accurately capture the cellular diversity between and within different cell types. Second, when we tested our panel on ten primary LUAD we saw that the relative abundance of endothelial cells is positively correlated with immune cell infiltration. LUADs with LPS had a higher proportion of endothelial and immune cells, whereas a group of LUADs predicted to have SPS had higher proportion of mesenchymal cells. Third, when considering the ECC compartment, samples showed high inter- and intra-tumor heterogeneity and HLA-DR+ subpopulations were positively correlated with T cell infiltration. Specifically, a group of four samples that clustered together by cell type abundance in Fig. 3.3C which presented a high percentage of CD8+ and CD4+ T cells and myeloid cells, also clustered together based on their ECCc profile (Fig. 3.4) which was enriched in HLA-DR+ cells. Three of these samples were LPS tumors classified as stage IA or 0 cancers with small nodule size based on their CT scans (Table 2.1), and their histology is mostly lepidic which is associated with a favorable prognosis [118] (see Appendix A Table S2). Conversely, the one LPS sample that deviated from this profile is a stage IB cancer, presents a bigger nodule size compared to the other LPS samples and has a predominant lepidic pattern but it also has a micropapillary component which is typically associated with a worse prognosis [118]. Finally, we validated our CyTOF findings by immunofluorescence and spatial analysis, in which we confirmed that the T cell abundance was positively correlated with HLA-DR expression in pan-cytokeratin+ cells and that T cells in LPS samples were closer to the first tumor cell in the space compared to SPS samples (Fig. 3.5).

The hypothesis that intra-tumor heterogeneity is associated with disease progression is not novel per se [119]. However, most studies in LUADs are based on bulk tissue analysis, which provides an average phenotype affecting the detection of rare subsets and overlooking the contribution of the TME. Single-cell technologies can overcome such limitation, providing high resolution information. Recently, the development and improvement of tissue dissociation protocols have made possible the application of single cell analysis to solid tumors [76]. A recent study using mass cytometry investigated the TME of LUAD focusing on the innate immune component [64]. The authors focused on comparing blood to normal and cancer tissues, for which the latter had a higher T cell content and they identified changes in tumor infiltrating myeloid cell subpopulations that could impair anti-tumor T cell immunity. Association with clinical outcome was not reported, however. Another study used single-cell RNA Seq and obtained a deep profile of lung cancer samples, most of which were LUAD patients, focusing on the TME and highlighting its heterogeneity and importance in tumor development [69]. Additional analysis of TCGA data showed that the abundances of some subpopulations and their correlation with patient survival differ between LUAD and squamous cell carcinoma and that they were influenced by clinical characteristics such as stage. An important component of the immune response in tumor biology is played by the interaction of the major histocompatibility complex molecules class I and II. MHC-I has been widely studied in cancer and there are some pivotal publications dedicated to LUAD specifically [62, 63]. In contrast, the role of MHC-II or HLA-DR in LUAD is less well understood. HLA-DR is constitutively expressed in antigen presenting cells but its expression can be induced in other tissues under, such as tumor cells, under inflammatory conditions [115]. Their main role is antigen presentation to CD4+ T cells, which when activated support CD8+ T cell activation and generation of memory T cells. Furthermore, tumor specific HLA-DR expression is associated with favorable outcomes in cancer patients [115]. In a recent study, Johnson and colleagues addressed the effect of HLA-DR expression in cancer cells on T cell recruitment and anti-PD1 therapy response using non-small cell lung cancer murine models [120]. They found that HLA-DR expression in cancer cells correlated with response to anti-PD1 therapy and showed by mechanistic experiments that overexpression of CIITA, a master regulator of the MHC-II pathway, in anti-PD1 resistant cells resulted in HLA-DR expression and increased T cell infiltration, whereas loss of CIITA in anti-PD1 responsive cells resulted in reduced HLA-DR expression and decreased T cell infiltration. In our data we found a strong association between HLA-DR expression in ECC and T cell abundance, mainly in LPS tumors. In addition, we found by spatial analysis an increased proximity of T cells to tumor cells in LPS tumors, suggesting that an immunogenic process could be responsible for the indolent behavior. How HLA-DR expressing ECC and closely related T cell infiltration in space contribute to the behavior of early LUAD remains to be studied.

Our results prove mass cytometry as a suitable tool to dissect LUAD biology at the single cell level and to investigate the interplay between the TME and the epithelial compartment [114, 121, 122]. Our work also has limitations. In this preliminary study, we are including a limited number of tumors per group (LPS, SPS) and we present these results as a proof of concept for the use of mass cytometry as a relatively novel application in LUAD research. Results will be further validated in a larger cohort which is part of an ongoing study. Additionally, with this analysis we are limited to a fixed number of proteins compared to single cell RNA Seq in which thousands of transcripts can be analyzed. Yet, the latter carries the uncertainty that missing data could be non-expressed genes or non-detected genes, and for that mass cytometry data is more reliable. Additionally, protein expression of tumors presents high variability, and normal lung tissue control is not always available. We are also limited by the amount of tissue that we could collect and by the overrepresentation of SPS LUADs as we are biased towards larger lesions. As for clinical limitations, the aggressiveness and indolence of LUADs are confounded by the heterogeneous treatments patients undergo and we do not know the true natural history of early LUAD. Finally, is important to consider that CANARY is not a perfect tool, and that other predictors should be consider in the future.

The difficulty in predicting behavior of early detected LUAD presents a major challenge to patients and their providers. These preliminary results of mass cytometry in early LUAD suggest a distinct cellular profile among LPS vs SPS tumors, implying an important role for T cell infiltration linked to HLA-DR expression. Future work will refine these results, integrate data from other platforms (i.e. radiomics, transcriptomics, genomics, etc.) and determine whether the combination of ECC subpopulations with specific subpopulations of cells in the TME predicts tumor behavior. We postulate that ultimately this work will allow us to better predict tumor behavior and integrate this evidence to improve current management of early LUADs.

CHAPTER 4

Multi-omics profiling of early lung adenocarcinoma reveals an association between radiomics features and tumor biology

4.1 Abstract

Lung adenocarcinoma (LUAD) is a heterogeneous group of tumors associated with different survival rates, even when detected at an early stage. Here, we aim to investigate the biological determinants of early LUAD indolence or aggressiveness using radiomics as a surrogate of behavior. We present a set of 92 LUAD patients with data collected across different methodologies. Patients were risk-stratified using the Computed Tomography–Based Score Indicative of Lung Cancer Aggression (SILA) tool (continuous score, 0=least aggressive, 1= most aggressive). We grouped the patients as indolent ($x <= 0.4$, n=14), intermediate ($0.4 > x <= 0.6$, n=27) and aggressive ($0.6 > x <= 1$, n=52). Using CyTOF we identified subpopulations characterized by high HLA-DR expression that were associated with indolent behavior. In the RNA-Seq dataset, pathways related to immune response were associated with indolent behavior, while pathways associated with cell cycle and proliferation were associated with aggressive behavior. We used HealthMyne (HM) software to extract radiomics features from the CT scans of the patients and computed pairwise correlation with SILA to select significant variables. When we integrated these datasets we identified four feature signatures and four patient clusters that were associated with survival. Using single cell RNA-Seq, we found that indolent tumors had significantly more T cells and less B cells than aggressive tumors, and that the latter had a higher abundance of regulatory T cells and T helpers. In conclusion, we found a bridge between radiomics and tumor biology which could improve the discrimination between indolent and aggressive ADC tumors and in turn may offer novel and personalized avenues for intervention.

4.2 Introduction

Lung cancer has the highest mortality rate among cancers worldwide, causing more deaths than breast, cervical, prostate and colorectal cancers, which have established population-based screening programs[123]. The 5-year survival rate for these patients is only 15%, mainly because 70% of them are diagnosed at a late stage[124]. Among lung cancer subtypes, lung adenocarcinoma (LUAD) still remains the more frequent[125]. In the past years, the NLST trial and more recently the NELSON trial have shown that lung cancer mortality is significantly reduced in individuals who undergo low-dose and volume CT screening, respectively[16, 126]. However, in both cases the overdiagnosis rate for a follow-up of 10 years is relatively high, 18.5% and 19.9% respectively. Additionally, LUAD is a heterogeneous disease both clinically and biologically. The recent advances in single cell technologies have allowed researchers to dissect the cellular heterogeneity of the tumor and learn more about the tumor microenvironment (TME) and its role in tumorigenesis, tumor development, progression and metastasis[64, 67, 69, 127]. On the other hand, advances in imaging technologies, specifically in the radiomics field, have allowed for the development of new tools to aid diagnosis and prognosis of these tumors[18, 20, 109, 128, 129, 130, 131]. Despite these research efforts, the biological determinants for the difference in tumor behavior remain obscure even though these have a direct implication in the efficacy and cost-effectiveness of lung cancer screening, particularly when considering the risks of overdiagnosis and overtreatment[3, 132]. In a recent publication, we showed that using a single cell technology we could dissect some of the main cell types of LUAD and found that the protein expression of MHC-II was associated with indolent behavior and increased T cell infiltration[105]. Here, we investigate the biological determinants of early lung adenocarcinoma indolence or aggressiveness using radiomics as a surrogate of behavior. We hypothesize that integration of biological, clinical and radiomics data of early stage LUAD will improve the discrimination between indolent and aggressive tumors which in turn may offer novel and personalized avenues for intervention. To this end, we generated a unique and comprehensive multi-omics dataset and an integrative analytical strategy that provides a deep prolifiling of tumor biology of LUAD in association with noninvasive CT-based risk stratification, granting a link between a widely used medical tool and the biology of the tumor.

4.3 Results

4.3.1 Multi-omic profiling of LUAD tumors using radiomics as a surrogate of behavior

To characterize the biological landscape of lung adenocarcinoma in association with their radiomics-based predicted behavior (i.e. indolent vs aggressive), we designed a multi-omic profiling study of surgically resected primary tumors. We present a comprehensive set of 92 lung adenocarcinoma patients who were treatment naive at the time of surgery and were representative of the lung adenocarcinoma distribution across age, sex, mutational status, and smoking status (Table 2.3, see Appendix B Table S1). Additionally, over 90% of the cohort is composed by early stage tumors.

Data was collected across different methodologies (Fig.4.1, see Appendix B Table S2). Surgically resected specimens (one per patient) were split and processed as: single cell suspension for CyTOF and single cell RNA-Seq, and fresh frozen tissue for RNA seq and whole exome seq (WES). Although data collection at every level was not possible for all specimens, close to 60% of the patients have data collected for CyTOF, RNA-Seq, and radiomics, allowing data integration (Fig.4.1A).

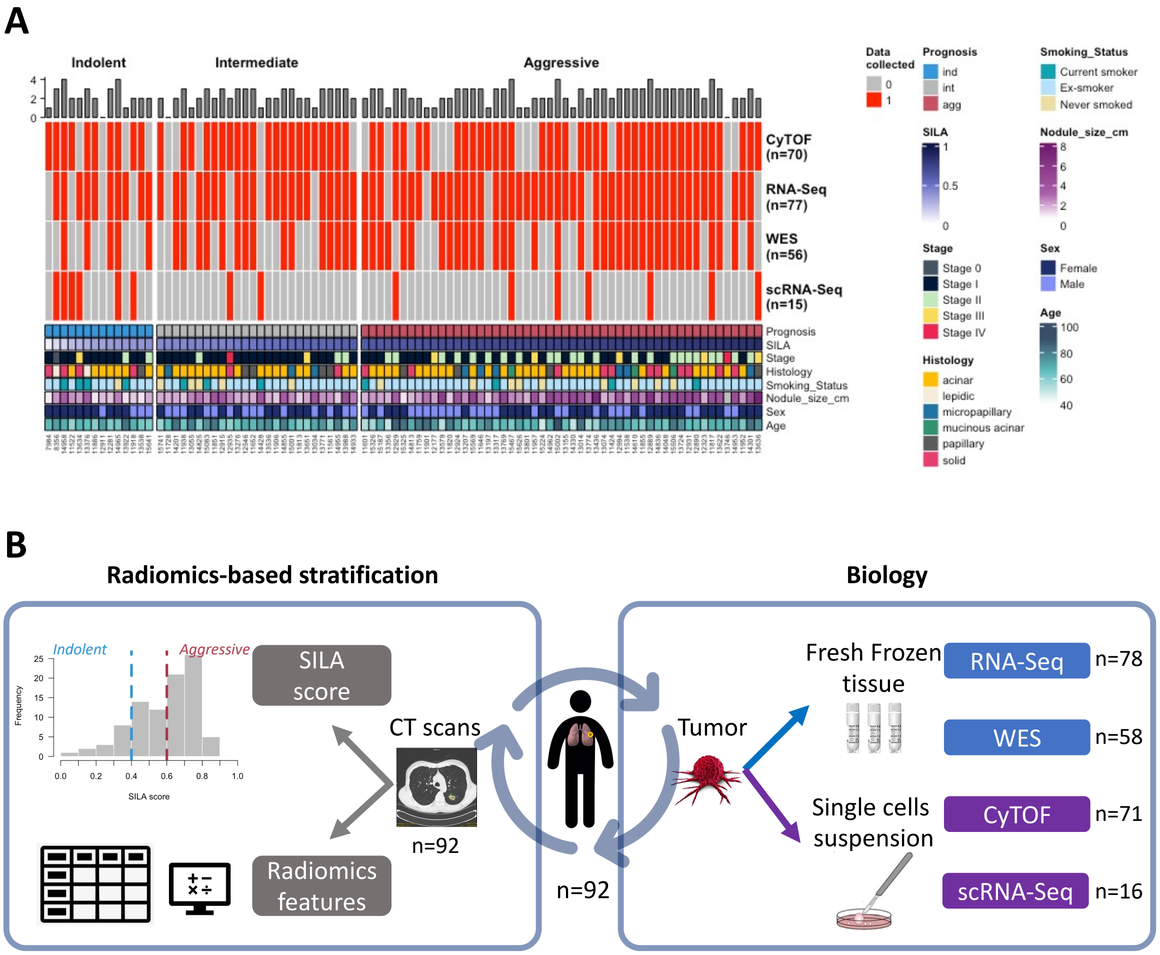


Figure 4.1: Summary of LUAD datasets and study workflow. (A) Heatmap showing the datasets included in this study (rows) by patient (columns) where red means data has been collected for that specific patient and gray that it has not. The bottom annotation show some clinical characteristics of the patient cohort. (B) Study workflow. For each of the 92 patients, tumor nodules from CT scans were analyzed to obtain SILA score and radiomics features (left), and for some of them biological data was collected from surgically resected tumors (right).

In addition to the clinical data, chest CT scans for each patient were analyzed and radiomics features were extracted with the HealthMyne software[128]. To risk-stratify the patients we used the Computed Tomography-Based Score Indicative of Lung Cancer Aggression (SILA) which analyses the CT scans of the patients and outputs a continuous score that ranges between 0 and 1, 0 being the least aggressive and 1 the most aggressive. This score has been validated to accurately correlate with histopathologic assessment, providing a scoring system to noninvasively predict the degree of histologic tumor invasion in LUAD [20]. We then grouped these into indolent (0-0.4), intermediate (>0.4-0.6), and aggressive (<0.6 - 1) (Fig.4.1B left).

4.3.2 LUADs of predicted indolent behavior are enriched in HLA-DR protein expression

LUADs human samples characterized by different predicted behavior classified into indolent (n = 10), intermediate (n = 21), and aggressive (n = 39) were stained with our previously validated antibody panel[105]. We identified the major cell types (epithelial cancer cells (ECC), endothelial cells, mesenchymal cells and immune cells) based on the expression of protein markers (Fig. 4.2A). EpCAM+/pan cytokeratin+/cytokeratin 7+ cells were annotated as ECC; CD31+/CD45- cells were annotated as endothelial cells; vimentin+/CD31-/CD45- and negative for epithelial markers cells were annotated as mesenchymal cells. All CD45+ cells were annotated as immune cells. The latter were further classified into CD4+ T cells (CD3+/CD4+/CD8-), CD8+ T cells (CD3+/CD8+/CD4-), double negative T cells (CD3+/CD8-/CD4-), myeloid cells (CD11b+/CD3-) and the remainder CD45+ cells were annotated as ”Other Immune”. The relative abundance (frequencies) of these main cell types were not significantly different between patient groups (see Appendix B Fig. S1).

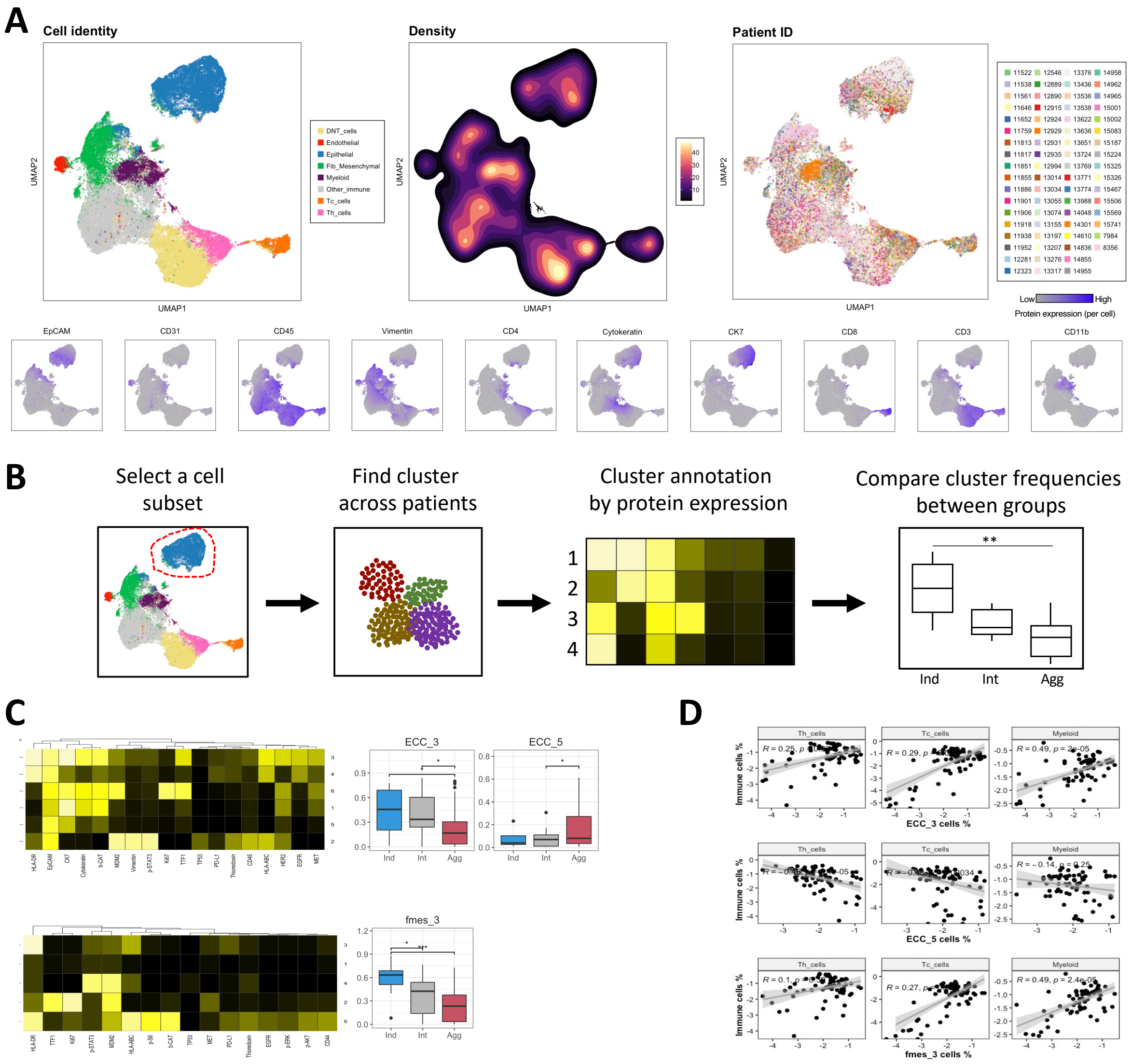


Figure 4.2: CyTOF analysis of LUAD samples reveal subsets associated with HLA-DR protein expression (A)UMAP representation colored by cell type (epithelial cancer cells (blue), endothelial cells (red), fibroblasts/mesenchymal cells (green), CD8+ T cells (orange), CD4+ T cells (pink), double negative T cells (yellow), myeloid cells (purple) and other immune cells (grey)), by density, by patient ID, and by protein expression of lineage markers (bottom). (B) Analysis workflow of the clustering by cell subset. (C) Heatmap of median protein expression per protein marker per cluster (left) and differential abundance analysis (right) for ECC (top) and fibroblast/mesenchymal cells (bottom). Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (D) Spearman correlation analysis of the relative abundance of ECC3, 5 and Fmes 3 vs CD4+, CD8+ T cells, and myeloid cells respectively.

Each subset individually went through an additional clustering step. Clusters were annotated by protein expression and then their frequencies within individual patient samples were compared between groups (Fig. 4.2B). In the ECC compartment, from a total of 6 clusters ECC cluster 3 (ECC3) relative abundance was significantly higher in patients with predicted indolent and intermediate behavior compared to aggressive (Fig. 4.2C, see Appendix B Fig. S2). ECC3 is characterized by a high expression of HLA-DR, pan-cytokeratin, cytokeratin 7 (CK7), *beta*-catenin and TTF1, as opposed to ECC4 which is other ECC cluster that expresses HLA-DR but lacks expression of the former. ECC5 and ECC2 were significantly higher in aggressive LUAD compared to intermediate, however the latter was mainly composed by two tumors only. The former lacked expression of every other marker except for EpCAM and CK7, whereas the latter presented high expression of EpCAM, vimentin, MDM2 and p-STAT3. ECC6 was the only cluster expressing the proliferation marker Ki67, with aggressive tumors having a slightly higher median compared to the other groups. In terms of protein co-expression, HLA-DR, HLA-ABC and EpCAM protein expression were highly correlated (r>0.45), and PD-L1 expression was also correlated with the first two (r>0.4) (see Appendix B Fig. S2E). Another group of highly correlated proteins were pan-cytokeratin, CK7 and *beta*-catenin, as well as the pairs of MET and EGFR, and TTF1 and Ki67 (r>0.45, see Appendix B Fig. S2E). In the fibroblasts/mesenchymal cells compartment cluster 3 (Fmes3) relative abundance was significantly higher in patients with predicted indolent and intermediate behavior compared to aggressive (Fig. 4.2C, see Appendix B Fig. S4). Fmes3 presented the highest expression of HLA-DR among the 5 clusters and also had a moderate expression of HLA-ABC. This cell type also presented a subset engaged in proliferation (Fmes2) with high expression of Ki67, TTF1 and MDM2 (see Appendix B Fig. S2C). In the protein co-expression analysis, Ki67 and TTF1 showed the highest correlation (r=0.65), followed by p-STAT3 and MDM2 (r=0.47), and HLA-DR and PD-L1 (r=0.45) (see Appendix B Fig. S4E). HLA-DR and HLA-ABC correlation was also significant but not as high as in the cancer cells (r=0.38). Although our CyTOF panel did not include sufficient markers to further annotate the identified immune cell types, we also performed reclustering on these with the aim of undercover some degree of heterogeneity if present (e.g. proliferative vs non proliferative) (see Appendix B Fig. S5-9). OIC cluster 4 (OIC4) was significantly enriched in patients with predicted indolent behavior compared to aggressive, and it was characterized by a high HLA-DR, HLA-ABC and vimentin expression (see Appendix B Fig. S9). OIC2 was significantly enriched in aggressive compared to indolent tumors, and it was characterized for the lack of expression of most markers and a moderate to low vimentin expression. Furthermore, the expression of HLA-DR and HLA-ABC was highly correlated (r=0.61), as was the expression of Ki67, TTF1 and MDM2 (r>0.45). Additionally, as HLA-DR (an isotype of MHC-II) is known to be involved in antigen presentation, we wanted to see if the relative abundance of the above mentioned subsets were significantly correlated with enrichment or depletion of CD8+ and CD4+ T cells and myeloid cells (Fig. 4.2D, see Appendix B Fig. S10). Indeed, ECC3, fmes3 and OIC4, clusters enriched in indolent tumors, were positively correlated with CD8+ and CD4+ T cells and myeloid cells, whereas ECC5 and OIC2, clusters enriched in aggressive tumors were negatively correlated with CD8+ and CD4+ T cells and myeloid cells. Finally, we calculated the median ”bulk” protein expression for each protein per sample (see Appendix B Fig. S11). We found that bulk HLA-DR protein expression is significantly higher in indolent and intermediate tumors compared to aggressive. Altogether, these results validate our previous findings[105], showing that HLA-DR expression in cancer cells and now also in fibroblasts/mesenchymal cells correlates with T cell and myeloid cell enrichment and that these cells are particularly abundant in LUADs with indolent behavior, calling for a potentially immunogenic environment and therefore a more favorable prognosis.

4.3.3 Transcriptomic profiles of lung ADCs are associated with proliferation, immune response and extracellular matrix organization

Fresh frozen tissue from a set of 77 LUADs human samples characterized by different predicted behavior (indolent n=10, intermediate n=21, aggressive n=46) was processed and the RNA was extracted and sequenced. A subset of those were also used to obtain whole exome sequence (WES) (indolent n=5, intermediate n=15, aggressive n=36) for genomic analysis. The mutational landscape of our LUAD cohort was very similar to what is expected for this cancer type [1, 26], with *KRAS* being the top mutated gene (41%) followed by *RYR2* (34%) and *MUC16* (32%) (see Appendix B Fig. S12A). *TP53* (27%) and *EGFR* (21%) were also among the top 15 mutated genes, and the latter was exclusive from *KRAS* alterations, as expected. We computed the mutational load for all 56 samples and found that it was mildly but significantly correlated with the SILA score (r=0.27, p=0.04), suggesting that genomic instability increases with the degree of predicted aggressiveness of the tumor (see Appendix B Fig. S12B). To perform a clinical enrichment analysis of the mutations, we opted for combining Indolent and Intermediate tumors, as the former was too small to compare on its own. Among the top significantly enriched tumors in Aggressive samples versus the Indolent+Intermediate group were *CTNND2*, *CACNA1E*, *SORCS1*, *PRDM9*, *NPAP1*, *APOB* and *ADAMTS12* (see Appendix B Fig. S12C).

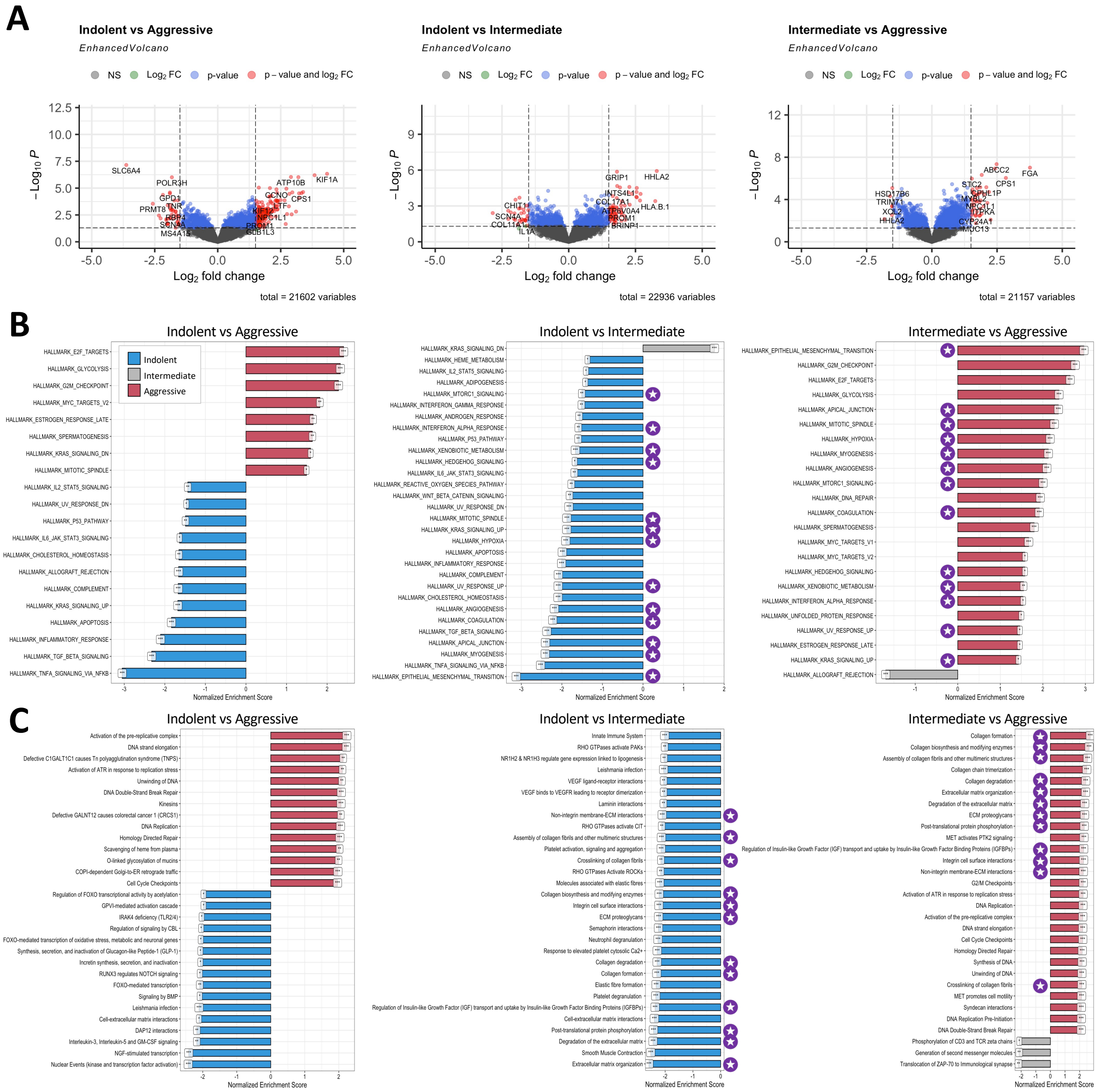


Figure 4.3: Transcriptomic analysis of LUAD highlights profiles associated with risk stratification (A) Volcano plots for indolent vs aggressive, indolent vs intermediate, and intermediate vs aggressive tumors, showing differentially expressed genes by fold change (FC) and p value. Cutoffs are log2FC> |1.5| and pval<0.05 (B) Gene Set Enrichment Analysis with Hallmark and (C) REACTOME databases for indolent vs aggressive, indolent vs intermediate, and intermediate vs aggressive tumors. Purple icon indicates pathways upregulated in both indolent and aggreSsive tumors when compared to intermediate.

We then performed differential gene expression analysis on the RNA-Seq data (Fig. 4.3A, see Appendix B Table S3). When comparing indolent vs aggressive, among the top dysregulated genes were *SLC6A4*, *KIF1A*, *HMGA2*, *ATP10B*, *POLR3H*, *GRIP1*, and *INTS4L1*. When comparing indolent vs intermediate, some of the top dysregulated genes were *HHLA2*, *GRIP1*, *DLGAP1-AS5*, *INTS4L1*, *PKHD1*, and *IGHV4-61*. When comparing intermediate vs aggressive, the top dysregulated genes were *ABCC2*, *FGA*, *B4GALNT1*, *MEGF10*, *CPS1*, and *STC2*. A detailed list of the differentially expressed genes (DEG) is presented in Table S3 (see Appendix B). Furthermore, gene set enrichment analysis (GSEA) of the differentially expressed genes was performed to understand their biological functions in the patient groups with different predicted behavior using the Hallmark[97] and REACTOME[98] databases (Fig. 4.3B-C, see Appendix B Tables S5-7). When comparing aggressive vs indolent or aggressive vs intermediate, pathways associated with proliferation and cell cycle were up-regulated, such as G2M Checkpoint, E2F targets, DNA replication and elongation, etc. This suggests that the tumors predicted to be aggressive, share a strong proliferative signal compared to tumors with lower SILA scores. On the other hand, when comparing indolent vs aggressive or indolent vs intermediate, pathways related with immune response were up-regulated, such as Inflammatory response, Complement, TGF-$beta$ signaling, TNF$alpha$ signaling via NFkB, Leishmania infection, IL-3, IL5 and GM-CSF, Innate immune system, etc. Eventhough we saw the ”Allograft rejection” pathway (a pathway associated with the expression of MHC classes I and II genes) present when comparing indolent or intermediate vs aggressive tumors, the pathways ”Antigen processing-Cross presentation” and ”MHC class II antigen presentation” were only up-regulated in aggressive when compared to intermediate, suggesting that the high HLA-DR protein expression we previously saw associated with indolent tumors (Fig. 4.2C-D) might be a consequence of an inflammatory microenvironment rather than the cause of inflammation by antigen presentation. Interestingly, when comparing either aggressive or indolent vs intermediate, pathways related to structural components such as extracellular matrix (ECM) organization, collagen formation or degradation, epithelial mesenchymal transition (EMT), angiogenesis, hypoxia, among others, were up-regulated. A detailed list of the dysregulated pathways is presented in Tables S5-7 (see Appendix B). Finally, we used the VIPER algorithm to infer transcription factor (TF) activity from gene expression data in the compared groups (see Appendix B Table S4). When comparing indolent vs aggressive gene expression, the *FOXO1* and *SPI1* regulons were down-regulated in aggressive tumors; when comparing indolent vs intermediate, the *HIF1A* and *SPI1* regulons were down-regulated in intermediate tumors; and when comparing intermediate vs aggressive, the *FOXM1* and *HIF1A* regulons were up-regulated in aggressive tumors (see Appendix B Table S4 for more details). We see once again a pattern shared by indolent and aggressive tumors, this time the activation of the *HIF1A* regulon, which correlates well with the structural pathways up-regulated in these patients.

4.3.4 Data integration reveals an association between radiomics features and tumor biology

A fundamental part of this study is the use of computer extracted quantitative features from the chest CT scans of the LUAD patients, also known as radiomics. We first used SILA to obtain a score predictive of tumor aggressiveness and risk-stratify our cohort (Fig. 4.1. However, we are also interested in dissecting these images at a more granular level. Using the HealthMyne picture archiving and communication system (www.healthmyne.com) lung nodules were segmented from CT scans for feature extraction. We obtained 300+ features, and then we filtered those that were significantly correlated with the SILA score. We then ended up with 61 features, and only 5 of them were negatively correlated with the SILA score (i.e. features associated with good prognosis) (see Appendix B Table S8). Percentage of ground glass opacity was one of the them, whereas solid percentage was positively correlated with SILA.

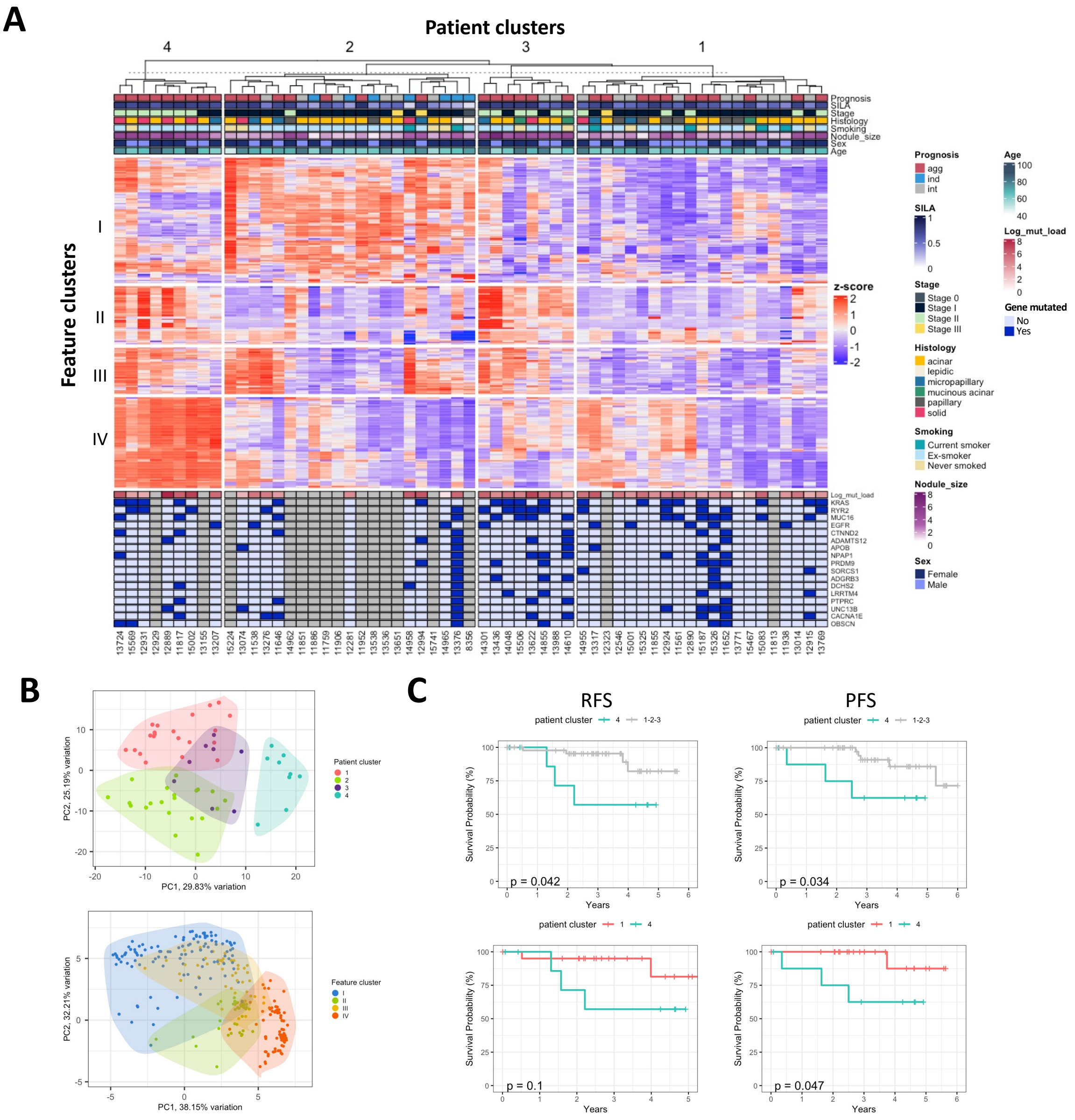


Figure 4.4: Data integration reveals an association between radiomics features and tumor biology. (A)Heatmap showing the z-score per patient (columns) per feature (rows) split by clusters. Top annotation shows some clinical characteristics and bottom annotation shows mutated genes. (B) Principal component analysis of patients (top) and features (bottom) colored by cluster. (C) Recurrence Free Survival (RFS) (left) and Progression Free Survival (PFS) of patients from cluster 4 vs 1,2,3 (top) and 4 vs 1 (bottom).

To this end, we have obtained several features at different biological and clinical levels that are significantly associated with the SILA score and therefore with the predicted level of aggressiveness of the tumors. Using those results as our feature selection strategy, we integrated a total of 301 features from the CyTOF, RNA-Seq and radiomics datasets on 59 patients with complete data across all modalities (Fig. 4.4, see Appendix B Tables S9-10). From the RNA-Seq dataset we used the significantly dysregulated pathways from the gene set enrichment analysis (Fig. 4.3B-C, see Appendix B Tables S5-7) to avoid redundancy. We used the Gene Set Variation Analysis (GSVA) algorithm to compute individual pathway scores for each patient sample. Features were scaled, centered and then clustered, resulting in four feature clusters (I - IV) (Fig. 4.4A, see Chapter 2 for details). Feature cluster I (FI) included all CyTOF features that were significantly enriched in indolent tumors (HLA-DR+ subpopulations), and bulk HLA-DR protein expression. It also included the five radiomics features that were positively correlated with SILA score such as percent GGO, root mean square and surface area to volume ratio (see Appendix B Table S8 for definitions). From the gene expression data a variety of pathways fell here: pathways associated with immune response, antigen presentation, cytokine cascades, etc; pathways associated with tumor initiation and growth signals such as NOTCH1 and MYC but also pathways associated with tumor suppression such as TP53 and PTEN signaling; and finally pathways associated with apoptosis, hypoxia and reactive oxygen species (ROS). All of these features together suggested a scenario in which the tumors were initiating or attempting growth but opposing signals were fighting back to prevent proliferation and the immune response could either be the cause or the consequence of this process. Feature cluster II (FII) included mostly radiomics features positively correlated with SILA score, a CyTOF subpopulation (ECC5) enriched in aggressive tumors, and the pathways ”O-linked glycosylation of mucins” and ”KRAS signaling down”. Feature cluster III (FIII) included the radiomic feature ”GLCM homogeneity” and then pathways associated with structural components such as collagen degradation and formation, ECM organization, angiogenesis, cell motility and EMT. Finally, cluster IV (FIV) was composed by pathways associated with cell proliferation, mytosis, DNA replication and cell cycle. When we performed a PCA on the features clusters and plotted the first two components (>70% of variance explained) we observed that FI and FIV showed almost no overlap, whereas FIII mostly overlapped with FI, and FII overlapped mostly with FIV (Fig. 4.4B bottom). To better understand those overlapping features, we generated similarity matrix (see Appendix B see Appendix B Fig. S13). These results show that there is an almost exclusive expression of either features from FI or FIV, and that some radiomics features from FII behave very similarly to features from FIV. This suggests a potential of using radiomics features to predict the degree of proliferative activity of the tumor. We then clustered the patients to find groups with similar feature characteristics and we found four clusters (1 - 4) (Fig. 4.4A). Patient cluster 1 (P1) was expressing low levels of most of the features clusters, except for a subset of it that were expressing moderate levels of FIV. Patient cluster 2 (P2), was a group of patients with moderate to high levels of FII and low levels of FIV, and a subset of them presented high levels of FIII. Patient cluster 3 (P3) presented moderate levels of FII and FIII and low levels of F1 and FIV. Finally, patient cluster 4 (P4) was characterized for a high level of FIV, moderate levels of FII and FIII, and low FI. When we performed a PCA on the patient clusters and plotted the first two components (<55% of variance explained), we observed that clusters P1, P2 and P4 were fairly different from each other, while P3 overlapped with P1 and P2 (Fig. 4.4B top). Lastly, when we assessed the recurrence (RFS) and progression free survival (PFS) of the patient clusters, we found that patients from P4 had the worst prognosis when compared with the other three clusters and also, but with reduced significance, when compared to P1 alone. Altogether, these results demonstrated the feasibility of integrating data from different modalities to obtain insights on the tumor biology which can be linked to clinical features.

4.3.5 In depth profiling of the LUAD tumor microenvironment by single cell RNA-Seq analysis

In an effort to better understand the microenvironment of tumors with different predicted behavior, we performed single cell RNA Sequencing of 15 tumors (indolent n=6 of which 3 were P2, intermediate n=2, aggressive n=7 of which 1 was P1 and 4 P4) (Fig. 4.5). After quality filtering (see Chapter 2), we obtained 44867 cells. Out of these, 14795 cells (%33) came from indolent tumors, 7107 cells (%16) from intermediate tumors, and 22974 (%51) from aggressive tumors. After gene normalization and filtering, we applied PCA on 1871 highly variable genes, and performed a graph-based clustering[102] to classify the cells into groups of similar gene expression. We annotated those clusters and identified 7 major cell types: B cells, T cells, myeloid cells, endothelial cells, cancer cells, mural cells and fibroblasts (Fig. 4.5A-B, Fig S14). Aggressive tumors were significantly enriched in B cells, while indolent tumors showed a significantly higher proportion of T cells (see Appendix B Fig. S14C), and we see a similar pattern for patients from P4 and P2, respectively (see Appendix B Fig. S14B).

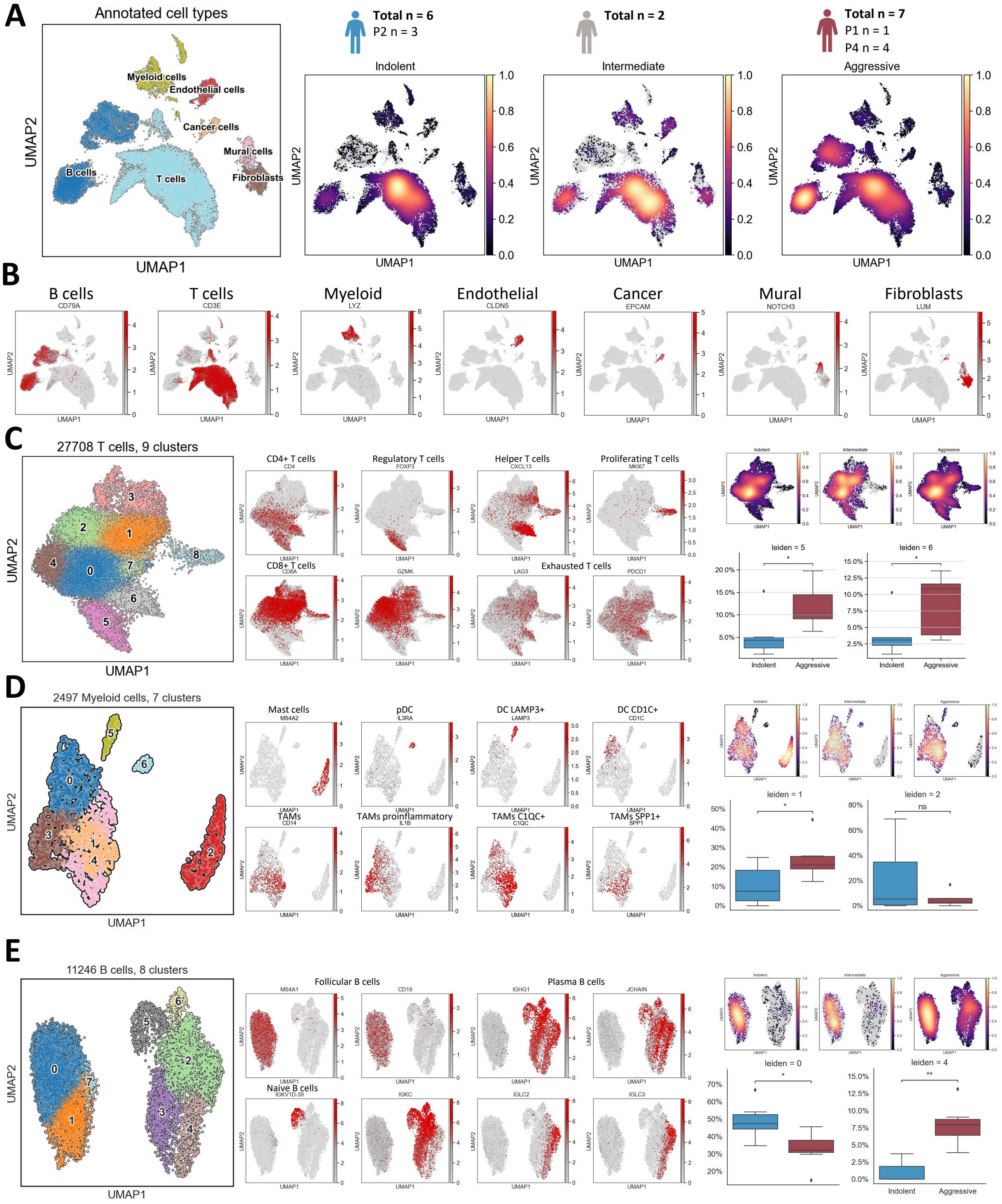


Figure 4.5: Profiling of LUAD tumor microenvironment by single cell RNA-Seq analysis. (A) UMAP representation colored by cell type using all cells (left) and by density grouped by risk group (right). (B) UMAP representation colored by gene expression of top lineage gene markers for each main cell type. (C) Reclustering analysis for T cells, (D) myeloid cells, and (E) B cells. UMAP representation colored by cluster, followed by UMAP representation colored by gene expression of some subset representative markers. On the far right we have UMAP by density grouped by risk group (top) and differential abundance analysis (bottom). Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001.

We then performed an additional clustering step to find subclusters within each of these main cell types (Fig. 4.5C-E, see Appendix B Fig. S15-21). In the T cell group we obtained 9 clusters (Fig. 4.5C, see Appendix B Fig. S15). Clusters 0,4,5,6 and 7 were identified as CD4+ T cells and clusters 1,2 and 3 were identified as CD8+ T cells. Clusters 5 and 6 were significantly enriched in aggressive tumors compared to indolent. Cluster 5 showed high FOXP3 expression which is characteristic of regulatory T cells, whereas cluster 6 showed high expression of CXCL13, a chemokine expressed by helper T cells. Numerous CD8+T cells also expressed GZMA, GZMB, GZMK and GNLY, which encode the cytotoxic molecules granzymes A, B and K and granulosyn, respectively. In addition to granzymes an other cytotoxic molecules, cluster 3 also expressed FCGR3A, a gene that encodes CD16, which presumably indicates that these are NKT cells (see Appendix B Fig. S15B). Cluster 8 corresponded to proliferating T cells, both CD8+ and CD4+. A fair amount of cells, particularly those in cluster 6 were expressing LAG3 and PDCD1, markers of T cell exhaustion. When we look at the samples classified by the data integration clusters from Fig. 4.4, patients from P4 and P2 followed similar patterns as aggressive and indolent, respectively, while the patient from P1 behaved like the indolent group but with less concentration of cytotoxic T cells (Fig S15A). In the myeloid cell compartment we found 7 clusters, from which clusters 1, 3 and 4 were tumor associated macrophages (TAM) expressing genes such as HLA-DRB1 and CD14 Fig. 4.5D, see Appendix B Fig. S16). Cluster 3 was enriched in proinflammatory TAM markers such as IL1B, while clusters 4 and 1 expressed C1QC and SPP1 genes. Clusters 0, 5 and 6 were dendritic cells (DC), with 0 being CDC1+ DCs, 5 being LAMP3+ DCs and 6 being plasmacytoid DCs expressing IL3A. Finally, cells from cluster 2 were identified as mast cells for their unique expression of MS4A2. Aggressive tumors as well as P4 tumors were enriched in cluster 1, while the mast cell subset (cluster 2) was dominated by one particular indolent tumor (11522) (see Appendix B Fig. S16A,C-D). In the B cell compartment we found 8 clusters, from which clusters 0, 1 and 7 corresponded to follicular B cells, given their expression of MS4A1 and CD19 and HLA-DR related genes

(Fig. 4.5E, Fig S17B). Cluster 5 was identified as naïve B cells, and clusters 2, 3, 4 and 6 were plasma B cells. Indolent tumors, but no P2 tumors, were enriched in cluster 0, and aggressive tumors were enriched in cluster 4. Tumors from P2 had little to no fraction of B cells in general, while tumors from P1 and P4 behaved similarly to each other and also were similar to aggressive tumors (see Appendix B Fig. S17A). Mural cells are composed by 6 clusters, from which clusters 0, 1, 2, 4 and 5 are characterized by the expression of some collagen genes, NOTCH3, ACTA2, PDGFRB which are commonly expressed in smooth muscle cells (SMC), and cluster 3 is characterized by the expression of KLF4 and MGP, genes associated with mesenchymal cells and regulation of SMC. Indolent and P2 are slightly enriched in cluster 3 cells while aggressive tumors appear to be enriched in cluster 0 cells (see Appendix B Fig. S18). In the fibroblasts compartment we found 7 clusters, from which both indolent and aggressive tumors were enriched in clusters 1 and 3, which were characterized for the expression of various collagen genes including COL1A1 and COL1A2, and intermediate tumors were enriched in cluster 2, characterized by the expression of some MFAP4, A2M, LIMCH1, among others (see Appendix B Fig. S19). In terms of the data integration patient groups, P2 and P4 were also enriched in clusters 1 and 3. In the endothelial compartment we found 7 clusters, however, the majority of these cells come from patients 14428 (intermediate) and 13634 (indolent) (see Appendix B Fig. S20). Finally, in the cancer cell compartment, indolent tumors present very few cells, intermediate tumors were enriched in cluster 1, and aggressive tumors were enriched in cluster 0 (see Appendix B Fig. S21). Cells from cluster 1 were characterized for the expression of some HLA-DR related genes, as well as lung-specific markers SFTPB and MUC1. Cells from cluster 0 expressed THE the collagen III gene COL3A1 and MIF, a gene that encodes the macrophage migration inhibitory factor. To recapitulate some of the main findings of this section, indolent tumors show higher percentage of T cells compared to aggressive tumors, but aggressive tumors are significantly enriched in regulatory and helper T cells. Aggressive tumors show a higher percentage of B cells compared to indolent tumors, which can be explained by a lack of plasma B cells in the latter. Aggressive tumors also show a higher percentage of CD14+/C1QC+/SPP1+/IL1B- TAMs. Indolent tumors also present an enrichment in mesenchymal mural cells, while aggressive tumors seem to be enriched in SMC-like cells, which correlates well with a more solid tumor component. The interesting finding from the RNA-Seq dataset in which both indolent and aggressive tumors appear to share an up-regulated signature for structural cellular pathways could be explained by looking at the fibroblasts compartment, in which tumors from both groups have an enrichment in fibroblasts with high expression of several collagen genes. In summary, these results give us a deeper understanding of the cellular subsets in LUAD and their transcriptomic profiles which help us to better understand the biological differences between indolent and aggressive tumors.

4.4 Discussion

Understanding the biology of lung adenocarcinomas in the context of tumor behavior is crucial to improve the current clinical standards of diagnosis and treatment, particularly in early stages of the disease. In this study, we presented a comprehensive set of early stage LUAD patients risk-stratified into predicted indolent, intermediate or aggressive behavior groups based on radiomics, with data collected across different biological layers. First, we used our previously validated CyTOF panel [105] to assess the difference between indolent and aggressive tumors at the proteomic level4.2. We found that indolent tumors were significantly enriched in a subset of cancer cells and a subset of fibroblast/mesenchymal cells characterized by high HLA-DR protein expression, compared to aggressive tumors and that these subsets were positively correlated with CD8+ T cells, CD4+ T cells and Myeloid cell abundance. HLA-DR bulk protein expression was also significantly higher in indolent vs aggressive tumors. We previously showed that HLA-DR expression was enriched in indolent tumors and that it was correlated with an increased abundance of T cells [105]. In the present study, we were able to confirm those CyTOF results in a bigger cohort and the other data modalities also suggested an increased immune response in indolent tumors compared to aggressive.

While MHC-II expression is usually restricted to antigen presenting cells (APC), it has been shown that its expression can also be induced in non-APCs in response to an inflammatory microenvironment and there is evidence of MHC-II molecule expression in cancer cells associated with good prognosis in various cancer types such as melanoma, breast cancer and esophageal cancer[114, 115, 122, 133, 134]. In a recent study[120], the authors assessed the effect of cancer cell-specific MHC-II expression in LUAD on T cell recruitment to tumors and response to anti-PD-1 therapy in murine models. They found that loss of CIITA, a master regulator of the MHC-II pathway, decreased MHC-II expression in cancer cells and turned the cells anti-PD-1 resistant. This effect was associated with reduced levels of Th1 cytokines, reduced T cell infiltration and macrophage recruitment, and increased B cell abundance. The opposite occurred with enforced expression of CIITA. They validated these results in surgically resected human LUADs, showing that MHC-II expression improved survival and positively correlated with T cell expression. These results align well with our findings, and highlight the potential of MHC-II expression in cancer cells as an independent biomarker of sensitivity to checkpoint inhibitors. In our single cell RNA-Seq data we found that indolent tumors were enriched in T cells, but aggressive tumors were enriched in T regs and T helpers specifically 4.5. Also aggressive tumors were enriched in B cells and indolent tumors mostly lacked plasma B cells. The influence of plasma B cells in NSCLC, has been mostly studied in the context of immunotherapies or adjuvant chemotherapies, in which cases it has been associated with improved prognosis[135, 136]. However is important to note that most of these tumors are late stage or metastatic. We then investigated the difference in gene expression between tumors of different predicted behavior4.3. When comparing indolent vs aggressive, the serotonin transporter *SLC6A4* was the top downregulated gene. It has been reported to be overexpressed in normal lung compared to LUAD and its deregulation has been associated with tobacco consumption [137, 138]. *KIF1A* and *HMGA2* were some of the top upregulated genes in aggressive tumors, the first one has been associated with drug resistance in breast cancer [139, 140] and the latter was reported to be associated with reduced overall survival in

LUAD patients, positively regulating lung cancer proliferation, progression and metastasis [141, 142]. In the Gene Set Analysis, when comparing aggressive vs indolent or intermediate, pathways associated with proliferation and cell cycle were up-regulated, and when comparing indolent vs aggressive or intermediate, pathways related with immune response were up-regulated. Although we found that the Hallmark pathway Allograft rejection[97], a gene set that includes MHC-I and II related genes as well as granzymes and cytokines such as INFG, was up-regulated in indolent tumors, pathways related with antigen presentation were not, suggesting that the high HLA-DR protein expression we saw in indolent tumors might be a consequence of an inflammatory microenvironment rather than the cause of inflammation by antigen presentation. An unexpected finding appeared when we compared either aggressive or indolent vs intermediate. Patients from both extremes shared up-regulation of pathways related to structural functions such as extracellular matrix organization, collagen formation and degradation, EMT, etc. These patients also presented an increased inferred activity of the HIF-1 alpha transcription factor, which is a master regulator of cellular and systemic homeostatic response to hypoxia[143, 144]. One possible explanation is the dual effect of some of these actors. For example, HIF-1 alpha may promote both tumorigenesis and apoptosis under different circumstances [145]. The authors claim that most of the conflicting data can be explained by the different cutoffs used to define high HIF-1 alpha expression. They analyzed the expression of HIF-1 alpha in NSCLC by immunohistochemistry, defining as low cutoff the median staining (¿5%) and as high cutoff ¿60%, and found that when using the latter an association with poor prognosis was significant. In a recent study of ours[146] using the same LUAD patient samples we described in Chapter 3[105], we found, by multiplex immunofluorescence, that indolent and aggressive tumors did not show significant different in neither the amount of collagen fibers or the average length of fibers. However, when we performed spatial analysis we found that tumor cells from the indolent group were co-localized with an increased number of immune cells. Additionally, tumor cells from aggressive LUADs were co-localized with lower number of collagen fibers and these fibers generally had smaller length, which may indicate involvement of these cells in the processes of collagen degradation and ECM remodeling. It is known that increased collagen deposition also increases the stiffness of the tumor and this has been associated with poor prognosis in several cancer types[147]. Some *in vitro* studies show that T cells migrate slower through collagen gels of high density compared to low density[148, 149]. Other *in vitro* studies have also demonstrated that T cells preferentially migrate along the collagen fibers, indicating that the collagen orientation could control the migration of T cells[150]. The overexpression of these signatures in our cohort could also suggest that both tumor types have the potential for metastasis but indolent tumors have other tools to counteract these while aggressive tumors have tools to support them. Additionally, when we looked into the fibroblasts compartment in our single cell RNA-Seq data (see Appendix B Fig. S19), we see similarities between indolent and aggressive tumors, however in the mural cells compartment aggressive tumors appear to have higher density of smooth-muscle-like cells which show high collagen expression compared to other cells in this subset (Fig S.18). We also see a higher number of T regs and T helpers in aggressive tumors, which has been associated with an stiffer microenvironment[151]. In that study, collagen led to an increase in the CD4:CD8 ratio among the infiltrating T cells and the CD4+ T cells were skewed toward a Th2 phenotype. We then integrated biological and radiomics features that were significantly associated predicted tumor behavior4.4. We found 4 main feature signatures:(I) immune response, growth initiation signals, and tumor suppression; (II) radiomics features positively correlated with SILA; (III) ECM organization and other structural components; (IV) proliferation and cell cycle. I and IV were strongly negatively correlated, and some features from II such as percentage of solid component were positively correlated with IV, while percent of GGO was positively correlated with I. Multiple radiomics studies and tools have focused on prediction of invasiveness, and association of solid or glass ground opacity (GGO) component with outcome. Our results are in agreement with the literature in that tumors with increased GGO percent show improved prognosis whereas tumors with higher solid percentage are associated with poor survival[129, 130, 131, 152]. However, there is no study in LUAD at the moment that has demonstrated correlation between radiomics features and specific and detailed biological signatures such as cell cycle, proliferation, DNA replication, mitosis, immune response, etc. We demonstrated a strong positive correlation between features associated with solid components and proliferation signatures, and these were also strongly but negatively correlated with immune response (Fig. 4.4). Similarly, GGO and other radiomics features negatively correlated with SILA showed an opposite relationship. This is a unique and unprecedented finding that connects a tool widely use in the clinic with biological insights of the tumor.

Our results show a unique and previously unseen potential bridge between tumor biology and the developing field of radiomics. However, our work also has its limitations. In the clinic, there is fewer patients that come with indolent tumors compared to aggressive ones, therefore our cohort has a reduced number of these samples which limits the study of intra-patient heterogeneity in this subset and introduces some degree of bias as we have an overrepresentation of aggressive tumors. In the same line, aggressive tumors are, for the most part, bigger than indolent tumors, which inherently influences the total number of cells and thus our ability to capture intracellular heterogeneity. These tissues are also less affected by cell loss during tissue processing. As for clinical limitations, the approach to define the aggressiveness or indolence of LUAD is still at the discretion of the researcher as there is no gold standard. The behavior of LUADs are confounded by the heterogeneous treatments patients undergo and we do not know the true natural history of early LUAD, as prospective studies to simply observe the natural history of the tumor without intervention would be unethical. In this study, all patients had resection of their primary lung nodule and an accompanying CT scan of that nodule obtained few weeks or days before surgery. We decided to use SILA, a CT-based tool that predicts the degree of histologic tissue invasion and patient survival specifically design for LUAD. We acknowledge that this, as any other predictive tool, is not flawless but it has been thoroughly validated[20]. Finally, each data set that we presented in this study has its own limitations and its own biases. For instance, the CyTOF dataset is limited to a fixed number of proteins compared to single cell RNA Seq in which thousands of transcripts can be analyzed. Yet, the latter is affected by sparsity and the cost limits the number of samples and number of cells to be sequenced. Additionally, both datasets require the tumor to be processed to obtain single cells, introducing an additional component of perturbation to the system and incidentally selecting for some cell types. The RNA-Seq and WES technologies are much more affordable, thus we can sequence more samples but can only interpret the results as a bulk. Despite these limitations, the strength of this study is to have all those datasets together to fill in the missing pieces. Although we present unique findings in each dataset, we were also able to find a common thread and results that complement each other.

In conclusion, we presented a unique and comprehensive collection of datasets in LUAD from which we were able to elucidate previously unknown insights on the biology of the tumors related to their predicted behavior, and data integration provided an evident and unprecedented link between tumor biology and radiomics. We also showed the important role of the TME, both in the immune compartment and the stromal compartment, in defining the indolence or aggressiveness of the tumors. Finally, experimental and mechanistic validations are needed to further understand these relationships. This is a rich data collection with huge potential that could be further explored in the future to answer multiple other research questions regarding LUAD. We believe that this work contributes to the knowledge and characterization of LUAD tumor biology in relation with its indolence or aggressiveness and further research can potentially integrate this evidence into clinical settings to improve current management of early LUADs.

CHAPTER 5

Discussion and Future Directions

5.1 Summary

In Chapter 1, I introduced the clinical and biological current knowledge on LUAD, and highlighted one of the unanswered questions remaining in the field: How can we better predict the disease behavior? The data presented in this dissertation begins to address that question by dissecting the biology of LUAD tumors of opposite behavior. Clinically, it is known that screening dramatically reduces lung cancer mortality[16, 126], but we also know that there is a significant percentage of overdiagnosis which can potentially translate into overtreatment[3]. This is of especial interest because a large number of LUAD patients detected at an early stage are senior of have other, which puts them at a higher risk during invasive procedures. We could reduce the number of patients that undergo those procedures if we knew how to identify potentially inconsequential lung cancers from aggressive ones, and therefore improve patient care. In an attempt to do that, several radiological tools have been developed in recent years[18, 20]. However, the link between the clinical diagnosis and the biological understanding of the disease is still very limited. In this chapter, I also outlined some of the main basic science research findings that have improved our understanding on LUAD biology, such as TCGA[26]. As research technologies developed, we have been able to dig deeper into the systemic processes of LUAD. Three pivotal studies brought our attention into intratumor heterogeneity and clonal architecture[5, 6, 7], suggesting that it is a universal phenomenon across LUAD and that it might be associated with survival and drug resistance. Then, single cell technologies allowed us to learn more about the tumor microenvironment and how it interacts with the cancer cells influencing tumorigenesis, tumor development and tumor progression[64, 67, 68, 69]. I closed the chapter on the importance of multi-omics data integration for the advancement of LUAD research and also stating some of its limitations.

In Chapter 2, I provided a detailed description of the materials and methods used for this dissertation.

In Chapter 3, I presented the validation of what became one of my main tools to dissect LUAD biology, a customized CyTOF antibody panel focused on LUAD oncogenic markers. I used LUAD cell lines and PBMCs to validate the panel in a controlled dataset. Our antibody panel captured the heterogeneity between and within cell lines. Then, I tested the panel in a sample of 10 LUADs, 4 being indolent and 6 aggressive tumors. I was able to identify main cell types such as epithelial cancer cells, endothelial cells, fibroblasts/mesenchymal cells, CD8+ T cells, CD4+ T cells, myeloid cells, and other unclassified immune cells. I further dissected the cancer cell compartment and found a subset of them characterized by high HLA-DR expression and were enriched in indolent tumors. Interestingly, the abundance of these subsets were positively correlated with CD8+ and CD4+ T cell abundance. These results were then validated by multiplex immunofluorescence, which confirmed a positive correlation of HLA-DR expression in cancer cells and T cell number. The spatial analysis also showed shorter distances from T cells to the nearest cancer cell in indolent tumors. These preliminary results proved our CyTOF antibody panel as a reliable tool to dissect intratumor heterogeneity.

Finally, in Chapter 4 I presented a comprehensive study that involved the use of radiomics, our previously validated CyTOF panel, WES, RNA-Seq, and single cell RNA-Seq and the integration of some of those to provide a deeper understanding of LUAD biology with respect to their radiomics-based predicted indolence or aggressiveness. The CyTOF results were in agreement with our previous findings presented in Chapter 3, HLA-DR expression associated with indolent behavior and with the abundance of T cells, but this time shown in a larger cohort. The transcriptomic analysis followed this line, showing that pathways associated with immune response were enriched in indolent tumors, while pathways associated with cell cycle and proliferation were enriched in aggressive tumors. As part of the data integration effort, I found that some radiomics features were correlated with immune response and some with cell proliferation, and those two were, for the most part, mutually exclusive. The single cell RNA-Seq data provided more detailed insights, such as the enrichment of T regs and Plasma B cells in aggressive tumors, and indolent tumors having more T cells overall.

5.2 Future Directions

5.2.1 Further validation using an independent/larger cohort

One of the most immediate things that remains to be done is to validate these results in an independent cohort. Due to the uniqueness and complexity of this data collection, and in particular in the data integration step, is quite challenging to find something similar in a public repository. At the bare minimum, we need chest CT images taken within 3 months prior surgery, and some high-throughput biological data collected on them, preferably RNA-Seq. As most of CT images in LUAD-related studies are acquired as part of routine care and not as part of a controlled research study, such as the TCGA collection, one of the biggest issues that can introduce confounding effects is the heterogeneity of these images in terms of scanner modalities, manufacturers and acquisition protocols. A potential option is to use one of our research group's previous LUAD cohorts for which we have CT scans and tissue microarray from which biological data could be obtained. Another possibility, is to use TCGA image repository, The Cancer Imaging Archive (TCIA), to retrieve CT scans and matching RNA-Seq data. However, as mentioned before one of the biggest issues is the heterogeneity of the images and also that these are at least a couple of decades old, which is a big gap in terms of imaging technology advances. Furthermore, in our study most of the clinical data was not noticeably associated with the patient clusters we found, perhaps in a larger cohort one can see differences (e.g. smokers vs non smokers). Along the same line, there is need for a more diverse cohort, as the majority of the patients in our cohort are Caucasian. Finally, as mentioned in the previous chapter, the datasets collected for this study can be further explored and this represents a potential opportunity to apply for a research grant that can address similar research questions in a bigger cohort, and the main results of this dissertations could be used as strong preliminary data.

5.2.2 The role of MHC-II in LUAD tumorigenesis and tumor progression

The high HLA-DR protein expression in indolent tumors was one of the main findings of this dissertation. Initially shown in a small cohort and then in a larger one, it was demonstrated that HLA-DR protein expression was negatively correlated with the SILA score, and that specific subsets of cancer cells and fibroblasts/mesenchymal cells also expressed these and their abundance was positively correlated with T cell and myeloid cell abundance. However, the functional and mechanistic role of MHC-II in the indolence of LUAD tumors is unclear. It has been reported that the expression of MHC-II and related pathway components is associated with improved prognosis in many other cancers [114, 115, 120, 122, 133, 134], but correlative associations in human tumors do not establish causality, therefore *in vitro* and *in vivo* experiments are necessary to understand the role of MHC-II. This is of particular interest for immunotherapy research, as response biomarkers are still not well established. Tumor specific MHC-II may play a role in CD4+ T cell stimulation, although different depending on the subset its function could be pro- and anti-tumor. For instance, Th1 cells secrete activating cytokines, whereas regulatory T cells have an immunosuppressive effect, playing a central role in tumor immune evasion. However, since most of the literature suggests that tumor specific MHC-II expression is associated with favorable prognosis this could suggest that it is somehow failing to activate T regs. Another intriguing avenue regarding the mechanistic function of MHC-II in tumor cells, is the origin of its expression and the status of its regulatory elements during cancer. The expression of MHC-II and its related machinery is driven by the transcriptional master regulator class II transactivator (CIITA)[115]. Promoters I and III drive constitutive expression of MHC-II in dendritic and B cells, respectively. Promoter IV is inducible by INFγ stimulation in various cell types, and it depends on JAK/-

STAT signaling. The transcription factor INF regulatory factor-1 (IRF-1) is also induced by INFγ and its loss also impairs INFγ-mediated CIITA induction[153]. The inducible expression of HLA-DR may also be regulated by retinoblastoma (Rb) protein [154]. It has been reported that some cells can induce CIITA expression with INFγ stimulation without producing functional MHC-II at the cell surface, and in instances where Rb function is lost as result of mutation, the defect can be rescued by reconstitution of functional Rb protein. This suggests that MHC-II expression at cell surface can be also regulated at the post-CIITA level. Additionally, in a breast cancer study MHC-II suppression by RAS/MAPK activation was reported[155]. This is in line with our findings, although not exclusively related with HLA-DR expression, where pathways associated with proliferation and cell cycle were negatively correlsted with immune response. All this considered, it would be interesting to understand what mechanisms of regulation are influencing the expression of HLA-DR in indolent tumors or which are inhibiting it in aggressive tumors.

5.2.3 The role of the extracellular matrix and stromal cells in LUAD behavior

One of the most surprising findings in this dissertation was the similarities in pathway expression of indolent and aggressive tumors when compared to intermediate, which were associated with stromal components. This was also reflected in the single cell RNA-Seq data in the fibroblasts subset, but in the mural cell compartment aggressive tumors were enriched in smooth-muscle-like cells while indolent tumors were enriched in mesenchymal-like cells, possibly pericytes. In a previous publication[146], we found that there was no difference in the amount of collagen fibers between indolent and aggressive tumors, but indolent tumors showed longer fibers. Understanding the role of these components is crucial, as it is known that collagen and ECM remodeling has an important role in cancer development[147]. An increased stiffness has been associated with poor outcome in other cancers, and it might support tumor progression, vascularization, and metastasis[147, 156]. The type of collagen may also affect tumor behavior differently, and although we did not find differences in overall density it would be interesting to explore if these tumors have a different composition of collagen fibers. Additionally, ECM components, such as collagen, have been reported to directly or indirectly influence T cell migration, phenotype and function. *In vitro* studies have also shown that T cells preferentially migrate along the collagen fibers, thus collagen orientation could also control the migration of T cells[150]. Therefore, collagen orientation is another interesting avenue to explore.

5.2.4 The study of LUAD as a system and advancement in multi-omics data integration strategies

LUAD, as other cancers and medical conditions in general, is a disease that must be studied as a system. Although we see that drugs treating specific actors in a pathway can have initial good results in some patients, the disease usually comes back and then the drug is no longer effective. One of the reasons is intra-tumor heterogeneity, meaning heterogeneity in cancer cell populations but also in TME cell types and stromal components. Thus, system approaches are needed and so is the development of multi-omics data integration strategies. In the data integration section presented in Chapter 4 (Fig. 4.4), we saw that even though the main topic of Feature cluster 1 (F1) was immune response, there was also a decent amount of other pathways associated with tumor initiation as well as tumor suppression. One can speculate that tumors enriched in F1 (P2), most of them indolent or intermediate, are in a stage in which the tumor is actively sending growth signals but the mechanisms of tumor suppression and immune response are still functional and fighting the tumor back. A systems approach to study these tumors and their mechanisms *in vitro* or *in vivo* would be ideal to better understand the picture, however recreating the TME is still a challenging task. 3D cancer models are important step towards that goal. In a recent study, the authors developed and validated a 3D lung cancer model in fibrin gel to investigate the angiogenic potential of cancer cells and its responses to hypoxia and therapeutics [157]. Another research group developed a similar model, which they call microphysiologic 3D tumor model with vascularized properties, to assess the effectiveness of ROR1-CAR T cells in lung and breast cancer[158]. They showed that ROR1-CAR T cells penetrated deep into tumor tissue and eliminated multiple layers of tumor cells located above and below the basal membrane. These two studies, however, use established cancer cell lines, thus 3D models or 2D culturing of tumor derived cells is still an unsolved challenge. Regarding data analytics, we need to keep developing and improving multi-omics data integration strategies in cancer research. In a recent perspective article by Tarazona and colleagues[70], the authors highlight some of the neglected challenges in multi-omics studies going from data collection, through data integration, to community distribution. The authors suggest, among other things, that we must improve our awareness on the differences of the methods we aim to integrate and think about how our missing data imputation strategies may affect the integrative analysis results. However, one of the issues that caught my attention the most was the need for standardization of multi-omics studies data distribution. Even though the amount of multi-omics studies have significantly increased in the past years, our way to distribute the data is still highly heterogeneous, calling for better sample annotation across modalities, more detailed data acquisition descriptions, and a unified storage strategy to allow widely use of data available to the public.

5.3 Concluding Remarks

In conclusion, this dissertation provided a comprehensive and deep profiling of LUAD indolence and aggressiveness at the biological bulk and single cell levels, as well as at the clinical and radiomics levels. This is a hypothesis generating study that has uncovered several potential future research avenues. It has also highlighted the importance and power of data integration to improve our systemic understanding of LUAD and to help reduce the gap between basic science research and clinical practice. Ultimately, I hope that my scientific findings contribute to the advancement of cancer research and directly or indirectly impact LUAD patient lives for the best.

References

[1] Senosain, M.-F. & Massion, P. P. Intratumor Heterogeneity in Early Lung Adenocarcinoma. *Frontiers in Oncology* **10**, 1–9 (2020). URL https://www.frontiersin.org/article/10.3389/fonc.2020.00349/full.

[2] Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA: a cancer journal for clinicians* **66**, 7–30 (2016). URL http://arxiv.org/abs/gr-qc/9809069http://dx.doi.org/10.1080/01422419908228843http://www.tandfonline.com/doi/abs/10.1080/01422419908228843http://www.ncbi.nlm.nih.gov/pubmed/26742998http://doi.wiley.com/10.3322/caac.21332. 9809069.

[3] Patz, E. F. *et al.* Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer. *JAMA Internal Medicine* **174**, 269 (2014). URL http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2013.12738.

[4] Diaz-Cano, S. J. Tumor heterogeneity: Mechanisms and bases for a reliable application of molecular marker design (2012). URL www.mdpi.com/journal/ijms.

[5] de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes deines lung cancer evolution. *Science* **346** (2014).

[6] Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delinetated by multiregion sequencing. *Science* **346**, 256–259 (2014).

[7] Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **376**, 2109–2121 (2017). URL http://www.nejm.org/doi/10.1056/NEJMoa1616288.

[8] Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017). NIHMS150003.

[9] Ortega, M. A. *et al.* Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clinical and Translational Medicine* **6** (2017). URL https://doi.org/10.1186/s40169-017-0177-y.

[10] Gibelin, C. & Couraud, S. Somatic alterations in lung cancer: Do environmental factors matter? *Lung Cancer* **100**, 45–52 (2016). URL http://dx.doi.org/10.1016/j.lungcan.2016.07.015http://linkinghub.elsevier.com/retrieve/pii/S0169500216304135.

[11] Rivera, G. A. & Wakelee, H. *Lung Cancer in Never Smokers*, 43–57 (Springer International Publishing, Cham, 2016). URL https://doi.org/10.1007/978-3-319-24223-1_3.

[12] Torre, L. A., Siegel, R. L. & Jemal, A. *Lung Cancer Statistics*, 1–19 (Springer International Publishing, Cham, 2016). URL https://doi.org/10.1007/978-3-319-24223-1_1.

[13] van Klaveren, R. J. *et al.* Management of Lung Nodules Detected by Volume CT Scanning. *New England Journal of Medicine* **361**, 2221–2229 (2009). URL http://www.nejm.org/doi/abs/10.1056/NEJMoa0906085.

[14] Blanchon, T. *et al.* Baseline results of the Depiscan study: A French randomized pilot trial of lung cancer screening comparing low dose CT scan (LDCT) and chest X-ray (CXR). *Lung Cancer* **58**, 50–58 (2007).

[15] Vinet, L. & Zhedanov, A. Survival of Patients with Stage I Lung Cancer Detected on CT Screening. *New England Journal of Medicine* **355**, 1763–1771 (2006). URL http://www.nejm.org/doi/abs/10.1056/NEJMoa0904327http://www.ncbi.nlm.nih.gov/pubmed/20573919http://arxiv.org/abs/1011.1669http://dx.doi.org/10.1088/1751-8113/44/8/085201http://www.nejm.org/doi/abs/10.1056/NEJMoa060476. 1011.1669.

[16] National Lung Screening Trial Research Team *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine* **365**, 395–409 (2011). URL http://www.nejm.org/doi/10.1056/NEJMoa1102873http://www.ncbi.nlm.nih.gov/pubmed/21714641http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4356534http://www.ncbi.nlm.nih.gov/pubmed/19038878{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/21714641{%}5Cnht. 15334406.

[17] Naidich, D. P. *et al.* Recommendations for the management of subsolid pulmonary nodules detected at CT: A statement from the Fleischner Society. *Radiology* **266**, 304–317 (2013).

[18] Maldonado, F. *et al.* Noninvasive computed tomography-based risk stratification of lung adenocarcinomas in the national lung screening trial. *American Journal of Respiratory and Critical Care Medicine* **192**, 737–744 (2015).

[19] Foley, F. *et al.* Computer-Aided Nodule Assessment and Risk Yield Risk Management of Adenocarcinoma: The Future of Imaging? *Seminars in Thoracic and Cardiovascular Surgery* **28**, 120–126 (2016). URL http://dx.doi.org/10.1053/j.semtcvs.2015.12.015.

[20] Varghese, C. *et al.* Computed Tomography–Based Score Indicative of Lung Cancer Aggression (SILA) Predicts the Degree of Histologic Tissue Invasion and Patient Survival in Lung Adenocarcinoma Spectrum. *Journal of Thoracic Oncology* **14**, 1419–1429 (2019). URL https://doi.org/10.1016/j.jtho.2019.04.022.

[21] Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009). URL http://dx.doi.org/10.1038/nature07943. arXiv:1108.1502v2.

[22] Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013). URL http://dx.doi.org/10.1038/nature12634. arXiv:1011.1669v3.

[23] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). URL https://www.nature.com/nature/journal/v500/n7463/pdf/nature12477.pdfhttp://www.nature.com/articles/nature12477. NIHMS150003.

[24] Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).

[25] Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007). URL http://www.nature.com/doifinder/10.1038/nature06358.

[26] Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014). URL http://www.nature.com/doifinder/10.1038/nature13385. NIHMS150003.

[27] Berger, A. H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* **30**, 214–228 (2016). URL https://doi.org/10.1016/j.ccell.2016.06.022. 15334406.

[28] Pao, W. & Hutchinson, K. E. Chipping away at the lung cancer genome. *Nature Medicine* **18**, 349–351 (2012). URL http://dx.doi.org/10.1038/nm.2697.

[29] Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nature reviews. Cancer* **3**, 11–22 (2003). URL http://www.nature.com/doifinder/10.1038/nrc969http://www.ncbi.nlm.nih.gov/pubmed/12509763.

[30] Kadota, K. *et al.* KRAS Mutation Is a Significant Prognostic Factor in Early-stage Lung Adenocarcinoma. *The American Journal of Surgical Pathology* **40**, 1579–1590 (2016). URL http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage{&}an=00000478-201612000-00002.

[31] Wieduwilt, M. J. & Moasser, M. M. The epidermal growth factor receptor family: Biology driving targeted therapeutics. *Cellular and Molecular Life Sciences* **65**, 1566–1584 (2008). URL http://link.springer.com/10.1007/s00018-008-7440-8.

[32] Liu, W.-s. *et al.* Prognostic value of epidermal growth factor receptor mutations in resected lung adenocarcinomas. *Medical Oncology* **31**, 771 (2014). URL http://link.springer.com/10.1007/s12032-013-0771-9.

[33] Sholl, L. M. Biomarkers in Lung Adenocarcinoma: A Decade of Progress. *Archives of Pathology & Laboratory Medicine* **139**, 469–480 (2015). URL http://www.archivesofpathology.org/doi/10.5858/arpa.2014-0128-RA.

[34] Bergethon, K. *et al.* ROS1 rearrangements define a unique molecular class of lung cancers. *Journal of Clinical Oncology* **30**, 863–870 (2012).

[35] Chalela, R. *et al.* Lung adenocarcinoma: From molecular basis to genome-guided therapy and immunotherapy. *Journal of Thoracic Disease* **9**, 2142–2158 (2017).

[36] Berger, A. H. *et al.* Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* **33**, 4418–4423 (2014). URL http://dx.doi.org/10.1038/onc.2013.581.

[37] De Snoo, F. A. & Hayward, N. K. Cutaneous melanoma susceptibility and progression genes. *Cancer Letters* **230**, 153–186 (2005).

[38] Wu, Y. L. *et al.* First-line erlotinib versus gemcitabine/cisplatin in patients with advanced EGFR mutation-positive non-small-cell lung cancer: Analyses from the phase III, randomized, open-label, ENSURE study. *Annals of Oncology* **26**, 1883–1889 (2015).

[39] Mitsudomi, T. *et al.* Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *The Lancet Oncology* **11**, 121–128 (2010). URL http://dx.doi.org/10.1016/S1470-2045(09)70364-X.

[40] Sequist, L. V. *et al.* Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutations. *Journal of Clinical Oncology* **31**, 3327–3334 (2013).

[41] Shaw, A. T. *et al.* Alectinib in ALK-positive, crizotinib-resistant, non-small-cell lung cancer: A single-group, multicentre, phase 2 trial. *The Lancet Oncology* **17**, 234–242 (2016). URL http://dx.doi.org/10.1016/S1470-2045(15)00488-X.

[42] Shaw, A. T. *et al.* Ceritinib in ALK-rearranged non-small-cell lung cancer. *New England Journal of Medicine* **370**, 1189–1197 (2014).

[43] Chuang, J. C. & Neal, J. W. Crizotinib as first line therapy for advanced ALK-positive non-small cell lung cancers. *Translational Lung Cancer Research* **4**, 639–641 (2015).

[44] Shaw, A. T. *et al.* Crizotinib in ROS1-rearranged non-small-cell lung cancer. *New England Journal of Medicine* **371**, 1963–1971 (2014).

[45] Borghaei, H. *et al.* Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *New England Journal of Medicine* **373**, 1627–1639 (2015).

[46] Herbst, R. S. *et al.* Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial. *The Lancet* **387**, 1540–1550 (2016).

[47] Rittmeyer, A. *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *The Lancet* **389**, 255–265 (2017).

[48] Garon, E. B. *et al.* Safety and activity of durvalumab + tremelimumab in immunotherapy (imt)-pretreated advanced nsclc patients. *Journal of Clinical Oncology* **36**, 9041–9041 (2018). URL https://doi.org/10.1200/JCO.2018.36.15_suppl.9041. https://doi.org/10.1200/JCO.2018.36.15_suppl.9041.

[49] Kowalski, D. M. *et al.* Arctic: durvalumab tremelimumab and durvalumab monotherapy vs soc in 3l advanced nsclc treatment. *ESMO 2018 Congress* (2018). URL https://oncologypro.esmo.org/Meeting-Resources/ESMO-2018-Congress/ARCTIC-durvalumab-tremelimumab-and-durvalumab-monotherapy-vs-SoC-in-_3L-advanced-NSCLC-treatment.

[50] Parums, D. V. Current status of targeted therapy in non-small cell lung cancer. *Drugs of today (Barcelona, Spain : 1998)* **50**, 503–25 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25101332.

[51] Hirsch, F. R. *et al.* Lung cancer: current therapies and new targeted treatments. *The Lancet* **389**, 299–311 (2017). URL http://dx.doi.org/10.1016/S0140-6736(16)30958-8.

[52] Qian, J. *et al.* Genomic Underpinnings of Tumor Behavior in in situ and Early Lung Adenocarcinoma. *American Journal of Respiratory and Critical Care Medicine* rccm.201902–0294OC (2019). URL https://browzine.com/articles/358537148https://www.atsjournals.org/doi/10.1164/rccm.201902-0294OC.

[53] Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* **68**, 820–823 (1971). URL http://www.pnas.org/cgi/doi/10.1073/pnas.68.4.820.

[54] Nowell PC. & Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976). URL http://www.ncbi.nlm.nih.gov/pubmed/959840.

[55] Swanton, C. Intratumor heterogeneity: Evolution through space and time. *Cancer Research* **72**, 4875–4882 (2012).

[56] Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48**, 238–244 (2016). URL http://dx.doi.org/10.1038/ng.3489. 15334406.

[57] Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013). URL http://dx.doi.org/10.1016/j.cell.2013.03.021.

[58] Sottoriva, A. *et al.* A big bang model of human colorectal tumor growth. *Nature Genetics* **47**, 209–216 (2015). URL http://dx.doi.org/10.1038/ng.3214. 15334406.

[59] Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883–892 (2012). URL http://www.nejm.org/doi/abs/10.1056/NEJMoa1113205.

[60] Lee, A. J. *et al.* Chromosomal instability confers intrinsic multidrug resistance. *Cancer Research* **71**, 1858–1870 (2011).

[61] Cruz-Tapias, P., Castiblanco, J. & Anaya, J.-M. Major histocompatibility complex: Antigen processing and presentation (2013). URL https://www.ncbi.nlm.nih.gov/books/NBK459467/.

[62] McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259–1271.e11 (2017).

[63] Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019). URL http://www.nature.com/articles/s41586-019-1032-7.

[64] Yonit Lavin, A. *et al.* Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses Comparing single tumor cells with adjacent normal tissue and blood from patients with lung adenocarcinoma charts early changes in tumor immunity and provides insights to g. *Cell* **169**, 750–757.e15 (2017). URL http://dx.doi.org/10.1016/j.cell.2017.04.014.

[65] Reddy, R. C. Immunomodulatory role of PPAR-γ in alveolar macrophages (2008). URL http://jim.bmj.com/lookup/doi/10.2310/JIM.0b013e3181659972.

[66] Moussion, C. & Girard, J. P. Dendritic cells control lymphocyte entry to lymph nodes through high endothelial venules. *Nature* **479**, 542–546 (2011). URL http://dx.doi.org/10.1038/nature10540.

[67] Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317–1334.e10 (2019). URL https://doi.org/10.1016/j.immuni.2019.03.009https://linkinghub.elsevier.com/retrieve/pii/S1074761319301268.

[68] Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine* **24**, 978–985 (2018). URL http://dx.doi.org/10.1038/s41591-018-0045-3http://www.nature.com/articles/s41591-018-0045-3.

[69] Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine* **24**, 1277–1289 (2018). URL http://dx.doi.org/10.1038/s41591-018-0096-5http://www.nature.com/articles/s41591-018-0096-5.

[70] Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science* **1**, 395–402 (2021).

[71] Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights* **14**, 117793221989905 (2020). URL http://journals.sagepub.com/doi/10.1177/1177932219899051.

[72] Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14**, 1–13 (2018).

[73] Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Computational Biology* **13**, 1–19 (2017).

[74] Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333–337 (2014).

[75] Lovly, C. M. *et al.* Rationale for co-targeting IGF-1R and ALK in ALK fusion-positive lung cancer. *Nature Medicine* **20**, 1027–1034 (2014).

[76] Leelatian, N. *et al.* Single cell analysis of human tissues and solid tumors with mass cytometry. *Cytometry Part B - Clinical Cytometry* **92**, 68–78 (2017). 15334406.

[77] Finck, R. *et al.* Normalization of mass cytometry data with bead standards. *Cytometry Part A* **83 A**, 483–494 (2013). NIHMS150003.

[78] Kotecha, N., Krutzik, P. O. & Irish, J. M. Web-based analysis and publication of flow cytometry experiments. *Current Protocols in Cytometry* 1–24 (2010). NIHMS150003.

[79] Ketchen, D. & Shook, C. The Application of Cluster Analysis in Strategic Management Research : An Analysis and Critique Author (s): David J . Ketchen , Jr . and Christopher L . Shook Published by : Wiley Stable URL : http://www.jstor.org/stable/2486927 Accessed : 06-06-2016 06. *Strategic management journal* **17**, 441–458 (1996).

[80] Dietz, C. *et al.* Integration of the ImageJ Ecosystem in KNIME Analytics Platform. *Frontiers in Computer Science* **2**, 1–17 (2020).

[81] Vasiukov, G. *et al.* Myeloid Cell-Derived TGFβ Signaling Regulates ECM Deposition in Mammary Carcinoma via Adenosine-Dependent Mechanisms. *Cancer research* **80**, 2628–2638 (2020). URL http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-19-3954http://www.ncbi.nlm.nih.gov/pubmed/32312837http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7299805.

[82] Stringer, C., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *bioRxiv preprint bioRxiv:2020.02.02.931238* (2020). URL https://www.biorxiv.org/content/early/2020/02/03/2020.02.02.931238. https://www.biorxiv.org/content/early/2020/02/03/2020.02.02.931238.full.pdf.

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

[84] Fastqc (2015). URL https://qubeshub.org/resources/fastqc.

[85] Guo, Y. *et al.* Multi-perspective quality control of Illumina exome sequencing data using QC3. *Genomics* **103**, 323–328 (2014). URL http://dx.doi.org/10.1016/j.ygeno.2014.03.006.

[86] Van der Auwera, G. & O'Connor, B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, Incorporated, 2020). URL https://books.google.com/books?id=wwiCswEACAAJ.

[87] Benjamin, D. *et al.* Calling somatic snvs and indels with mutect2. *bioRxiv* (2019). URL https://www.biorxiv.org/content/early/2019/12/02/861054. https://www.biorxiv.org/content/early/2019/12/02/861054.full.pdf.

[88] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164–e164 (2010). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603.

[89] Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311 (2001). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.1.308.

[90] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). URL http://www.nature.com/articles/nature19057.

[91] Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research* **45**, D777–D783 (2017). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1121.

[92] Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research* **28**, 1747–1756 (2018).

[93] Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

[94] Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). 1305.3347.

[95] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 1–21 (2014).

[96] Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. *bioRxiv* (2019).

[97] Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* **1**, 417–425 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26771021http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4707969.

[98] Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687–D692 (2022). URL https://academic.oup.com/nar/article/50/D1/D687/6426058.

[99] Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics* **48**, 838–847 (2016).

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

[101] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018). URL https://onlinelibrary.wiley.com/doi/10.1111/1462-2920.13787https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1382-0http://www.ncbi.nlm.nih.gov/pubmed/28474475http://www.ncbi.nlm.nih.gov/pubmed/29409532http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5802054.

[102] Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 1–12 (2019). 1810.08473.

[103] Olsen, L. R., Leipold, M. D., Pedersen, C. B. & Maecker, H. T. The anatomy of single cell mass cytometry data. *Cytometry Part A* **95**, 156–172 (2019).

[104] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 Published by :. *Journal of the Royal Statistical Society. Series B* **57**, 289–300 (1995). 95/57289.

[105] Senosain, M. F. *et al.* HLA-DR cancer cells expression correlates with T cell infiltration and is enriched in lung adenocarcinoma with indolent behavior. *Scientific Reports* **11**, 1–13 (2021). URL https://doi.org/10.1038/s41598-021-93807-3.

[106] Midthun, D. E. Early detection of lung cancer. *F1000Research* **5**, 739 (2016). URL http://f1000research.com/articles/5-739/v1.

[107] Patz, E. F. *et al.* Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer. *JAMA Internal Medicine* **174**, 269 (2014). URL http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2013.12738.

[108] Mimae, T. *et al.* What are the radiologic findings predictive of indolent lung adenocarcinoma? *Japanese Journal of Clinical Oncology* **45**, 367–372 (2015).

[109] She, Y. *et al.* The predictive value of CT-based radiomics in differentiating indolent from invasive lung adenocarcinoma in patients with pulmonary nodules. *European Radiology* 1–8 (2018).

[110] Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780–791 (2016). URL http://dx.doi.org/10.1016/j.cell.2016.04.019. 15334406.

[111] Min, J. W. *et al.* Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell RNA-seq. *PLoS ONE* **10**, 1–17 (2015).

[112] Kashima, Y. *et al.* Combinatory use of distinct single-cell RNA-seq analytical platforms reveals the heterogeneous transcriptome response. *Scientific Reports* **8**, 1–16 (2018). URL http://dx.doi.org/10.1038/s41598-018-21161-y.

[113] McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018). URL https://doi.org/10.21105/joss.00861.

[114] Johnson, D. B. *et al.* Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nature Communications* **7**, 1–10 (2016). URL http://dx.doi.org/10.1038/ncomms10582.

[115] Axelrod, M. L., Cook, R. S., Johnson, D. B. & Balko, J. M. Biological consequences of MHC-II expression by tumor cells in cancer. *Clinical Cancer Research* **25**, 2392–2402 (2019).

[116] Ma, K.-Y. *et al.* Single-cell RNA sequencing of lung adenocarcinoma reveals heterogeneity of immune response–related genes. *JCI Insight* **4** (2019). URL https://insight.jci.org/articles/view/121387.

[117] Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 220 (2017). URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1349-1.

[118] Zombori, T. *et al.* The more the micropapillary pattern in stage I lung adenocarcinoma, the worse the prognosis—a retrospective study on digitalized slides. *Virchows Archiv* **472**, 949–958 (2018).

[119] Jamal-Hanjani, M., Quezada, S. A., Larkin, J. & Swanton, C. Translational implications of tumor heterogeneity. *Clinical Cancer Research* **21**, 1258–1266 (2015).

[120] Johnson, A. M. *et al.* Cancer Cell–Intrinsic Expression of MHC Class II Regulates the Immune Microenvironment and Response to Anti–PD-1 Therapy in Lung Adenocarcinoma. *The Journal of Immunology* **204**, 2295–2307 (2020).

[121] He, Y. *et al.* MHC class II expression in lung cancer. *Lung Cancer* **112**, 75–80 (2017). URL http://dx.doi.org/10.1016/j.lungcan.2017.07.030.

[122] Park, I. A. *et al.* Expression of the MHC class II in triple-negative breast cancer is associated with tumor-infiltrating lymphocytes and interferon signaling. *PLoS ONE* **12**, 1–14 (2017).

[123] Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**, 394–424 (2018). URL http://doi.wiley.com/10.3322/caac.21492.

[124] Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians* **68**, 7–30 (2018).

[125] Barta, J. A., Powell, C. A. & Wisnivesky, J. P. Global epidemiology of lung cancer. *Annals of Global Health* **85**, 1–16 (2019).

[126] de Koning, H. J. *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine* **382**, 503–513 (2020).

[127] Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications* **11** (2020).

[128] Balagurunathan, Y., Schabath, M. B., Wang, H., Liu, Y. & Gillies, R. J. Quantitative Imaging features Improve Discrimination of Malignancy in Pulmonary nodules. *Scientific Reports* **9**, 1–14 (2019). URL http://dx.doi.org/10.1038/s41598-019-44562-z.

[129] Fan, L. *et al.* Radiomics signature: a biomarker for the preoperative discrimination of lung invasive adenocarcinoma manifesting as a ground-glass nodule. *European Radiology* **29**, 889–897 (2019).

[130] Wu, G. *et al.* Diagnosis of invasive lung adenocarcinoma based on chest CT radiomic features of part-solid pulmonary nodules: A multicenter study. *Radiology* **297**, 451–458 (2020).

[131] He, B. *et al.* A machine learning-based prediction of the micropapillary/solid growth pattern in invasive lung adenocarcinoma with radiomics. *Translational Lung Cancer Research* **10**, 955–964 (2021).

[132] Thalanayar, P. M., Altintas, N., Weissfeld, J. L., Fuhrman, C. R. & Wilson, D. O. Indolent, potentially inconsequential lung cancers in the Pittsburgh Lung Screening Study. *Annals of the American Thoracic Society* **12**, 1193–1196 (2015).

[133] Forero, A. *et al.* Expression of the MHC Class II Pathway in Triple-Negative Breast Cancer Tumor Cells Is Associated with a Good Prognosis and Infiltrating Lymphocytes. *Cancer Immunology Research* **4**, 390–399 (2016). URL http://cancerimmunolres.aacrjournals.org/lookup/doi/10.1158/2326-6066.CIR-15-0243.

[134] Dunne, M. R. *et al.* HLA-DR expression in tumor epithelium is an independent prognostic indicator in esophageal adenocarcinoma patients. *Cancer Immunology, Immunotherapy* **66**, 841–850 (2017).

[135] Lohr, M. *et al.* The prognostic relevance of tumour-infiltrating plasma cells and immunoglobulin kappa C indicates an important role of the humoral immune response in non-small cell lung cancer. *Cancer Letters* **333**, 222–228 (2013). URL http://dx.doi.org/10.1016/j.canlet.2013.01.036.

[136] Patil, N. S. *et al.* Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. *Cancer Cell* 289–300 (2022).

[137] Chen, B., Gao, S., Ji, C. & Song, G. Integrated analysis reveals candidate genes and transcription factors in lung adenocarcinoma. *Molecular Medicine Reports* **16**, 8371–8379 (2017).

[138] Ishii, T., Wakabayashi, R., Kurosaki, H., Gemma, A. & Kida, K. Association of serotonin transporter gene variation with smoking, chronic obstructive pulmonary disease, and its depressive symptoms. *Journal of Human Genetics* **56**, 41–46 (2011).

[139] De, S., Cipriano, R., Jackson, M. W. & Stark, G. R. Overexpression of kinesins mediates docetaxel resistance in breast cancer cells. *Cancer Research* **69**, 8035–8042 (2009).

[140] Guerrero-Preston, R. *et al.* Differential promoter methylation of kinesin family member 1a in plasma is associated with breast cancer and DNA repair capacity. *Oncology Reports* **32**, 505–512 (2014).

[141] Ahlemann, M. *et al.* Overexpression Facilitates mTOR-dependent Growth Transformation. *Molecular Carcinogenesis* **967**, 957–967 (2006).

[142] Gao, X. *et al.* HMGA2 regulates lung cancer proliferation and metastasis. *Thoracic Cancer* **8**, 501–510 (2017).

[143] Wang, G. L., Jiang, B. H., Rue, E. A. & Semenza, G. L. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 5510–5514 (1995).

[144] Iyer, N. V. *et al.* Cellular and developmental control of O2 homeostasis by hypoxia-inducible factor 1α. *Genes and Development* **12**, 149–162 (1998).

[145] Swinson, D. E. *et al.* Hypoxia-inducible factor-1α in non small cell lung cancer: Relation to growth factor, protease and apoptosis pathways. *International Journal of Cancer* **111**, 43–50 (2004).

[146] Vasiukov, G. *et al.* Integrated Cells and Collagen Fibers Spatial Image Analysis (2021). URL https://www.frontiersin.org/article/10.3389/fbinf.2021.758775.

[147] Rømer, A. M. A., Thorseth, M. L. & Madsen, D. H. Immune Modulatory Properties of Collagen in Cancer. *Frontiers in Immunology* **12**, 1–15 (2021).

[148] Wolf, K. *et al.* Physical limits of cell migration: Control by ECM space and nuclear deformation and tuning by proteolysis and traction force. *Journal of Cell Biology* **201**, 1069–1084 (2013).

[149] Sadjadi, Z., Zhao, R., Hoth, M., Qu, B. & Rieger, H. Migration of Cytotoxic T Lymphocytes in 3D Collagen Matrices. *Biophysical Journal* **119**, 2141–2152 (2020). URL https://doi.org/10.1016/j.bpj.2020.10.020. 2001.05331.

[150] Pruitt, H. C. *et al.* Collagen fiber structure guides 3D motility of cytotoxic T lymphocytes. *Matrix biology : journal of the International Society for Matrix Biology* **85-86**, 147–159 (2020). URL https://doi.org/10.1016/j.matbio.2019.02.003http://www.ncbi.nlm.nih.gov/pubmed/30776427http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6697628.

[151] Sadtler, K. *et al.* Developing a pro-regenerative biomaterial scaffold microenvironment requires T helper 2 cells. *Science* **352**, 366–370 (2016).

[152] Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **5** (2014).

[153] Muhlethaler-Mottet, A., Berardino, W. D., Otten, L. A. & Mach, B. Activation of the MHC class II transactivator CIITA by interferon-γ requires cooperative interaction between Stat1 and USF-1. *Immunity* **8**, 157–166 (1998).

[154] Lu, Y., Tschickardt, M. E., Schmidt, B. J. & Blanck, G. IFN-γ inducibility of class II transactivator is specifically lacking in human tumour lines: Relevance to retinoblastoma protein rescue of IFN-γ inducibility of the HLA class II genes. *Immunology and Cell Biology* **75**, 325–332 (1997).

[155] Loi, S. *et al.* RAS/MAPK activation is associated with reduced tumor-infiltrating lymphocytes in triple-negative breast cancer: Therapeutic cooperation between MEK and PD-1/PD-L1 immune checkpoint inhibitors. *Clinical Cancer Research* **22**, 1499–1509 (2016).

[156] Fang, M., Yuan, J., Peng, C. & Li, Y. Collagen as a double-edged sword in tumor progression. *Tumor Biology* **35**, 2871–2882 (2014).

[157] Kniebs, C. *et al.* Establishment of a Pre-vascularized 3D Lung Cancer Model in Fibrin Gel—Influence of Hypoxia and Cancer-Specific Therapeutics. *Frontiers in Bioengineering and Biotechnology* **9**, 1–11 (2021).

[158] Wallstabe, L. *et al.* ROR1-CAR T cells are effective against lung and breast cancer in advanced microphysiologic 3D tumor models. *JCI Insight* **4** (2019).

APPENDIX

Appendix A

Supplementary material for Chapter 3. This section is adapted from the Online Supplementary Information of “HLA-DR cancer cells expression correlates with T cell infiltration and is enriched in lung adenocarcinoma with indolent behavior” published in Scientific Reports and has been reproduced in line with publisher policies[105].

Patient ID	CANARY	Batch ID	Batch #	Events*
7984	LPS	32618	4	16787
8356	LPS	32118	2	7471
11522	LPS	32418	3	8194
12924	SPS	32618	4	255991
12929	SPS	32418	3	48359
12994	SPS	32118	2	104147
13197	SPS	32418	3	51198
13376	LPS	31618	1	653176
13436	SPS	31618	1	32681
13622	SPS	32618	4	501184

Table S1. CyTOF Sample batches.

**Number of events (cells) after pre-processing.*

Pt ID	CANARY	Age at collection	Sex	Race	Smoking Status	Age Started	Age Quit	Pack Years	Family History Cancer Type	CT Nodule Size (mm)	CT Nodule Location	8th Edition Path Stage	Biological data
7984	LPS	66	Female	Caucasian	Ex-smoker	15	69	50	Unknown	9.7	RLL	Stage IA1	CyTOF
8356	LPS	72	Female	Caucasian	Ex-smoker	18	60	37	Pancreatic	23.4	LLL	Stage 0	Both
11522	LPS	57	Female	Caucasian	Ex-smoker	16	57	20.5	Unknown	28	RUL	Stage IA3	Both
12924	SPS	70	Male	Caucasian	Ex-smoker	17	70	53	Melanoma Skin Cancer	31	LLL	Stage IIB	Both
12929	SPS	86	Female	Caucasian	Ex-smoker	16	41	37.5	Unknown	37	LLL	Stage IIB	Both
12994	SPS	76	Male	Caucasian	Ex-smoker	20	44	24	Brain	60	RUL	Stage IIIB	Both
13197	SPS	78	Male	Caucasian	Ex-smoker	20	55	35	Lung Cancer	13	RLL	Stage IIB	Both
13376	LPS	64	Female	Caucasian	Current smoker	16	N/A	20	Bladder	41	RUL	Stage IB	Both
13436	SPS	56	Male	Caucasian	Ex-smoker	21	51	45	Gynecological Cancer	61	RLL	Stage IIB	Both
13622	SPS	67	Female	Caucasian	Ex-smoker	12	67	110	Lung Cancer	32	RUL	Stage IIA	Both
11918	LPS	68	Male	African American	Ex-smoker	18	43	25	Gastrointestinal Cancer	22	RUL	Stage IA1	MxIF
12911	LPS	72	Male	African American	Ex-smoker	31	61	15	Unknown	12	LUL	Stage IA2	MxIF
13634	LPS	67	Female	Caucasian	Current Smoker	13	N/A	N/A	Other	N/A	RUL	Stage IIIB	MxIF
14428	LPS	73	Male	Caucasian	Current Smoker	N/A	N/A	45	Gynecological Cancer	38	RUL	Stage IA2	MxIF
14965	LPS	62	Female	Caucasian	Never smoker	N/A	N/A	N/A	Other	N/A	LUL	Stage IA3	MxIF

Table S2. Detailed patient clinical characteristics

Pt ID	CANARY	Solid	Acinar	Lepidic	Mucinous	Micropapillary
7984	LPS		+++	+		
8356	LPS			+++		
11522	LPS			+++		
12924	SPS	+	+	+		
12929	SPS	+	+++			
12994	SPS		+++			+
13197	SPS		+++			+
13376	LPS		+	+++		+
13436	SPS		+++	+	+++	+
13622	SPS	+				

Table S3. Histologic subtypes of ADC

Gene	Cell type	Size.high	Size.low	Median.High	Median.Low	p.value	p.adjusted
HLA.DRA	CD4+ memory T-cells	120	120	0.2415	0.1475	1.05E-18	3.93E-18
HLA.DRA	CD4+ naive T-cells	120	120	0.10225	0.02488	6.61E-24	4.96E-23
HLA.DRA	CD8+ naive T-cells	120	120	0.0067875	0.005175	0.16879963	0.19781206
HLA.DRA	CD8+ T-cells	120	120	0.037765	0.006478	2.23E-13	5.08E-13
HLA.DRA	CD8+ Tcm	120	120	0.05113	0.008943	6.63E-26	8.29E-25
HLA.DRB5	CD4+ memory T-cells	120	120	0.2002	0.15705	5.77E-05	7.46E-05
HLA.DRB5	CD4+ naive T-cells	120	120	0.08357	0.028725	5.51E-17	1.53E-16
HLA.DRB5	CD8+ naive T-cells	120	120	0.0053235	0.005488	0.93774746	0.96609699
HLA.DRB5	CD8+ T-cells	120	120	0.022435	0.0100045	7.07E-05	8.99E-05
HLA.DRB5	CD8+ Tcm	120	120	0.034265	0.010875	6.01E-12	1.13E-11
HLA.DRB6	CD4+ memory T-cells	120	120	0.21425	0.16055	6.07E-08	9.10E-08
HLA.DRB6	CD4+ naive T-cells	120	120	0.08662	0.026945	1.04E-12	2.10E-12
HLA.DRB6	CD8+ naive T-cells	120	120	0.006645	0.005613	0.65471749	0.70148302
HLA.DRB6	CD8+ T-cells	120	120	0.03306	0.011785	7.97E-07	1.13E-06
HLA.DRB6	CD8+ Tcm	120	120	0.04045	0.01496	4.07E-13	8.98E-13
HLA.DRB1	CD4+ memory T-cells	120	120	0.2232	0.1545	1.01E-09	1.69E-09
HLA.DRB1	CD4+ naive T-cells	120	120	0.100035	0.02494	6.49E-23	4.06E-22
HLA.DRB1	CD8+ naive T-cells	120	120	0.005806	0.005212	0.23544515	0.27166748
HLA.DRB1	CD8+ T-cells	120	120	0.032355	0.006944	3.71E-09	5.80E-09
HLA.DRB1	CD8+ Tcm	120	120	0.04474	0.0102135	1.12E-17	3.49E-17
HLA.DQA1	CD4+ memory T-cells	119	120	0.2328	0.1486	7.12E-13	1.48E-12
HLA.DQA1	CD4+ naive T-cells	119	120	0.09937	0.025295	2.78E-24	2.60E-23
HLA.DQA1	CD8+ naive T-cells	119	120	0.005452	0.005777	0.97313001	0.97313001
HLA.DQA1	CD8+ T-cells	119	120	0.02962	0.0100045	1.17E-07	1.73E-07
HLA.DQA1	CD8+ Tcm	119	120	0.04678	0.0092275	1.13E-20	5.28E-20
HLA.DQB1	CD4+ memory T-cells	118	120	0.2117	0.1659	2.71E-05	3.57E-05
HLA.DQB1	CD4+ naive T-cells	118	120	0.09492	0.02308	5.31E-22	3.06E-21
HLA.DQB1	CD8+ naive T-cells	118	120	0.005436	0.0053665	0.91751113	0.96609699
HLA.DQB1	CD8+ T-cells	118	120	0.02703	0.00977	3.78E-06	5.16E-06
HLA.DQB1	CD8+ Tcm	118	120	0.03978	0.01308	7.20E-14	1.74E-13
HLA.DQA2	CD4+ memory T-cells	120	120	0.2308	0.149	1.15E-13	2.70E-13
HLA.DQA2	CD4+ naive T-cells	120	120	0.08491	0.02863	2.99E-17	8.98E-17
HLA.DQA2	CD8+ naive T-cells	120	120	0.0055865	0.0053505	0.96662672	0.97313001
HLA.DQA2	CD8+ T-cells	120	120	0.031205	0.009417	3.61E-09	5.76E-09
HLA.DQA2	CD8+ Tcm	120	120	0.047205	0.014475	8.83E-18	2.88E-17
HLA.DQB2	CD4+ memory T-cells	117	120	0.1855	0.15905	0.00865005	0.01081256
HLA.DQB2	CD4+ naive T-cells	117	120	0.09023	0.02578	1.32E-18	4.70E-18
HLA.DQB2	CD8+ naive T-cells	117	120	0.005838	0.004884	0.25157037	0.28160863
HLA.DQB2	CD8+ T-cells	117	120	0.02652	0.0076725	5.58E-06	7.48E-06
HLA.DQB2	CD8+ Tcm	117	120	0.03031	0.011095	1.41E-09	2.31E-09
HLA.DOB	CD4+ memory T-cells	120	120	0.2475	0.1467	8.62E-18	2.88E-17
HLA.DOB	CD4+ naive T-cells	120	120	0.1095	0.02623	3.79E-25	4.06E-24
HLA.DOB	CD8+ naive T-cells	120	120	0.006561	0.005043	0.11101426	0.13215983
HLA.DOB	CD8+ T-cells	120	120	0.041405	0.0056895	2.12E-19	8.39E-19
HLA.DOB	CD8+ Tcm	120	120	0.056605	0.0078265	3.51E-27	6.57E-26
HLA.DMB	CD4+ memory T-cells	119	120	0.2486	0.1475	7.56E-20	3.15E-19
HLA.DMB	CD4+ naive T-cells	119	120	0.1043	0.024645	1.82E-21	9.73E-21
HLA.DMB	CD8+ naive T-cells	119	120	0.006214	0.0054535	0.52646757	0.58066276
HLA.DMB	CD8+ T-cells	119	120	0.04049	0.007085	6.30E-13	1.35E-12
HLA.DMB	CD8+ Tcm	119	120	0.05465	0.008451	1.66E-26	2.48E-25
HLA.DMA	CD4+ memory T-cells	120	120	0.2063	0.15705	2.67E-06	3.71E-06
HLA.DMA	CD4+ naive T-cells	120	120	0.090215	0.02454	3.55E-23	2.42E-22
HLA.DMA	CD8+ naive T-cells	120	120	0.0062675	0.0054115	0.24704118	0.28072861
HLA.DMA	CD8+ T-cells	120	120	0.029275	0.0072795	3.88E-09	5.94E-09
HLA.DMA	CD8+ Tcm	120	120	0.033825	0.01038	5.33E-17	1.53E-16
HLA.DOA	CD4+ memory T-cells	120	119	0.2183	0.1504	2.64E-11	4.61E-11
HLA.DOA	CD4+ naive T-cells	120	119	0.10365	0.02488	1.27E-28	4.75E-27
HLA.DOA	CD8+ naive T-cells	120	119	0.005298	0.006384	0.05667027	0.06855275
HLA.DOA	CD8+ T-cells	120	119	0.0355	0.009776	3.64E-10	6.21E-10
HLA.DOA	CD8+ Tcm	120	119	0.041985	0.009255	2.91E-20	1.28E-19
HLA.DPA1	CD4+ memory T-cells	119	120	0.233	0.14855	4.20E-15	1.05E-14
HLA.DPA1	CD4+ naive T-cells	119	120	0.1043	0.022955	1.48E-27	3.70E-26
HLA.DPA1	CD8+ naive T-cells	119	120	0.005774	0.0053435	0.9403344	0.96609699
HLA.DPA1	CD8+ T-cells	119	120	0.03645	0.007038	1.13E-11	2.07E-11
HLA.DPA1	CD8+ Tcm	119	120	0.05308	0.007936	4.79E-24	3.99E-23
HLA.DPB1	CD4+ memory T-cells	120	120	0.23745	0.15055	2.00E-12	3.94E-12
HLA.DPB1	CD4+ naive T-cells	120	120	0.1132	0.024285	1.18E-29	8.82E-28
HLA.DPB1	CD8+ naive T-cells	120	120	0.006382	0.006102	0.64400735	0.70000798
HLA.DPB1	CD8+ T-cells	120	120	0.040475	0.0073185	3.08E-12	5.92E-12
HLA.DPB1	CD8+ Tcm	120	120	0.052265	0.0092275	6.63E-21	3.31E-20
HLA.DPB2	CD4+ memory T-cells	120	119	0.23275	0.1523	1.74E-11	3.11E-11
HLA.DPB2	CD4+ naive T-cells	120	119	0.089785	0.03013	6.64E-16	1.72E-15
HLA.DPB2	CD8+ naive T-cells	120	119	0.007628	0.005168	0.0457618	0.05626451
HLA.DPB2	CD8+ T-cells	120	119	0.03291	0.009776	3.98E-07	5.74E-07
HLA.DPB2	CD8+ Tcm	120	119	0.048415	0.01312	5.91E-16	1.58E-15

Table S4. Summary of cell type enrichment analysis on ADC TCGA using xCell.

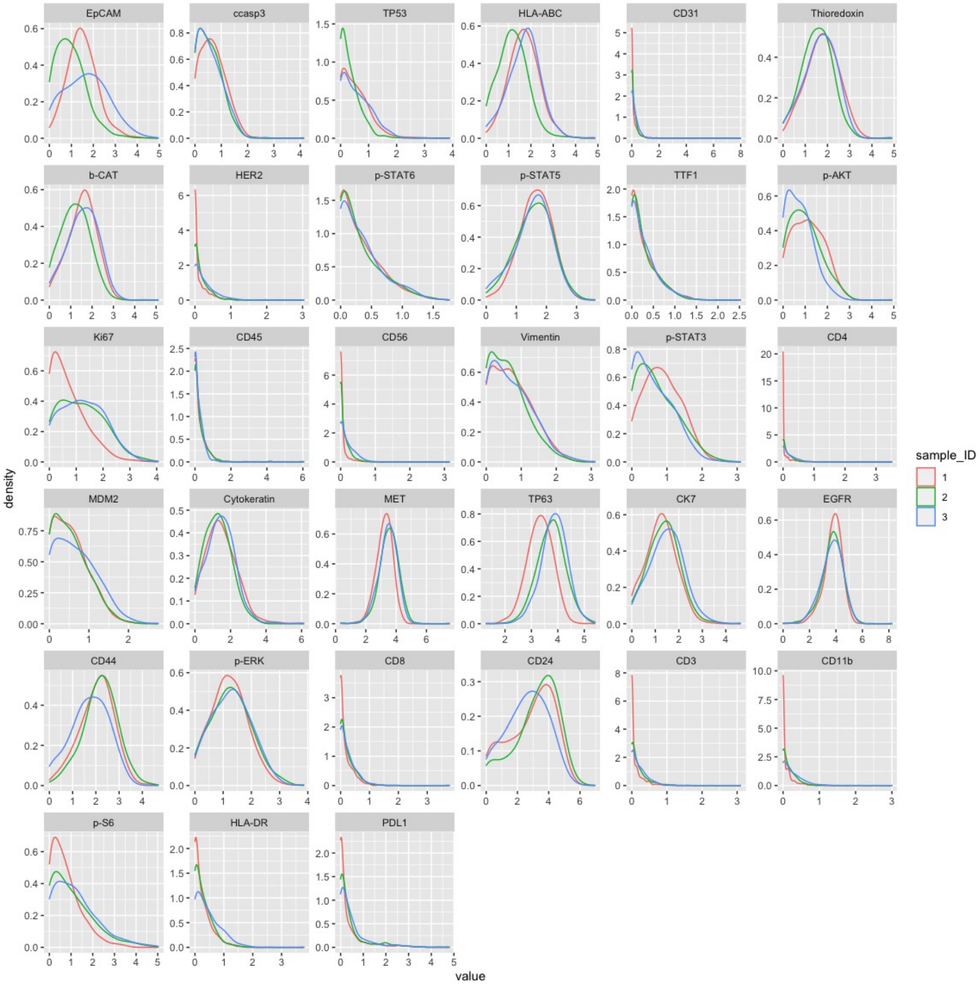


Figure S1. A549 protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for A549 cell line.

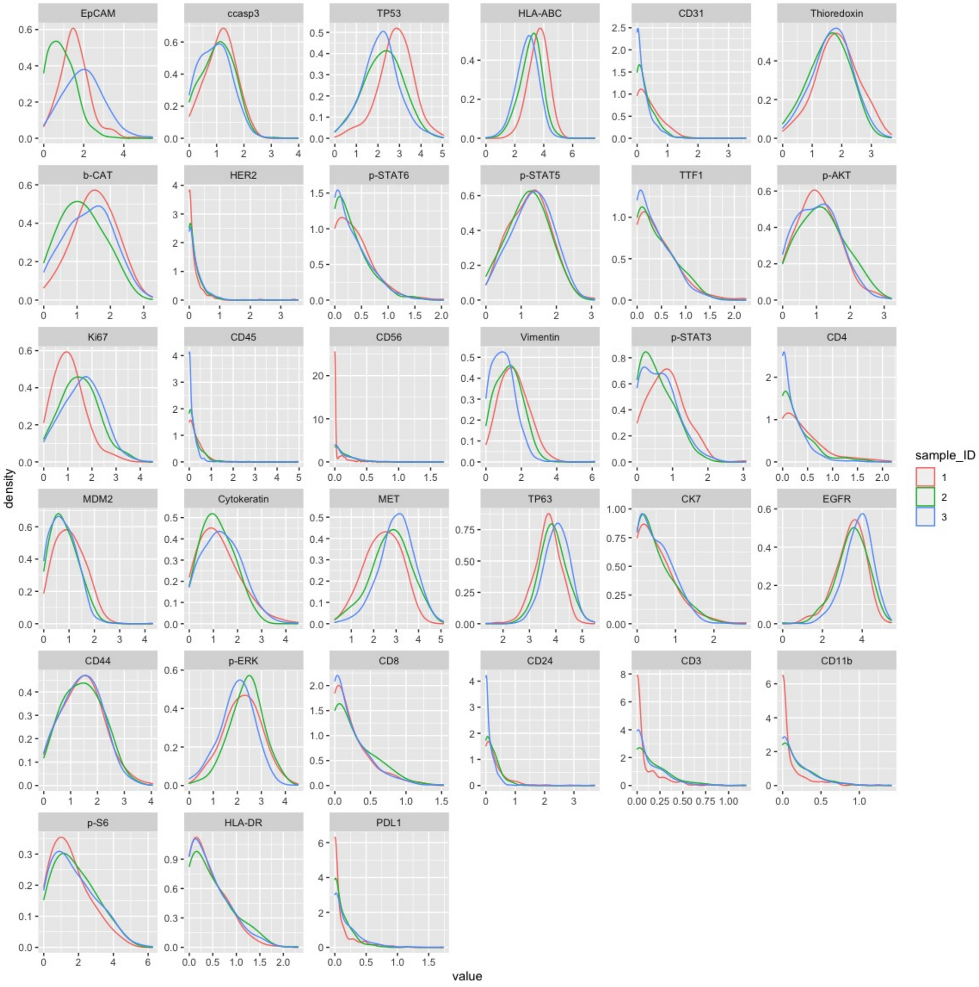


Figure S2. H23 protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for H23 cell line.

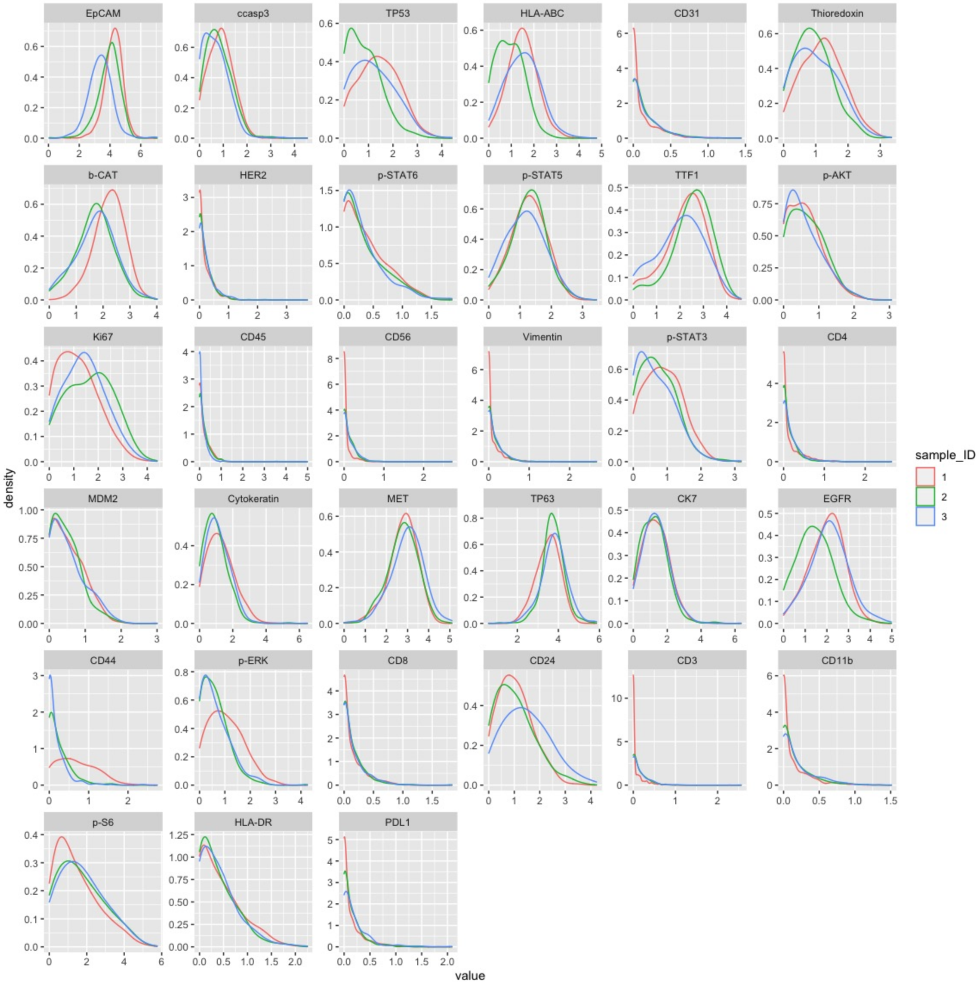


Figure S3. H3122 protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for H3122 cell line.

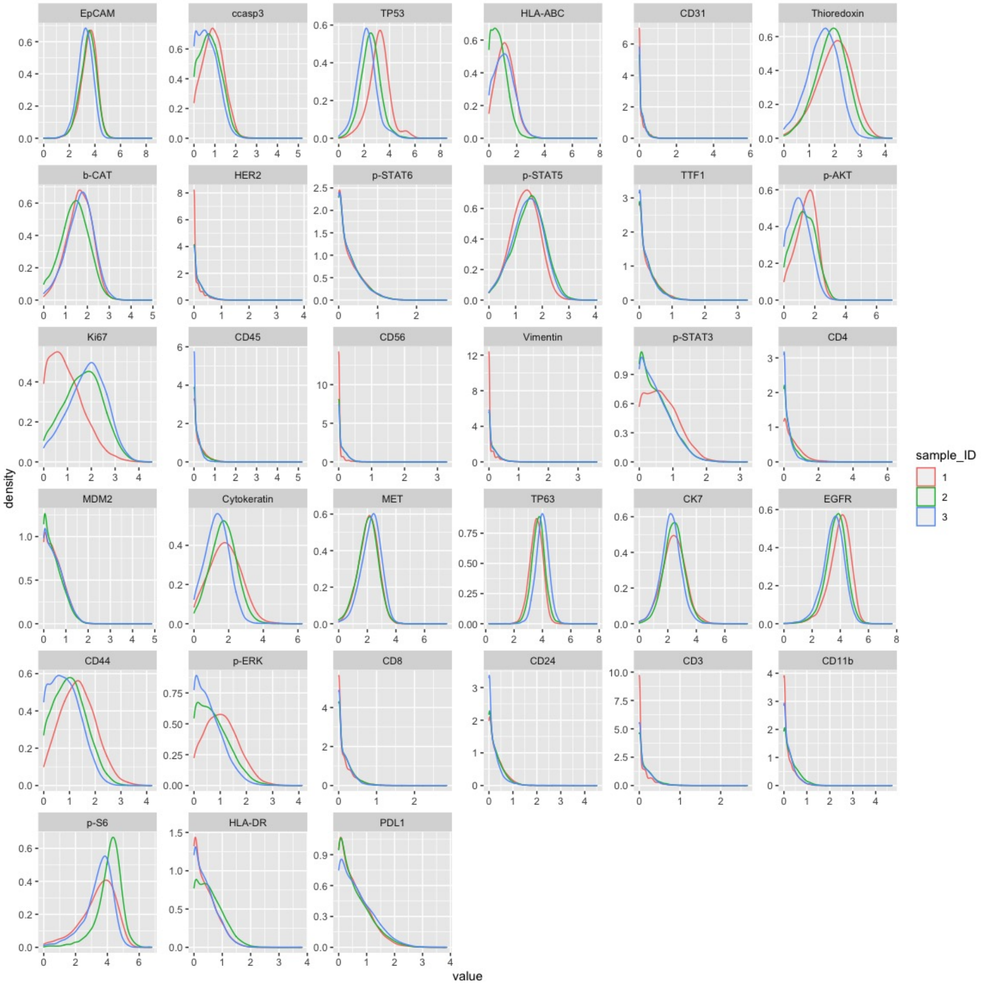


Figure S4. PC9 protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for PC9 cell line.

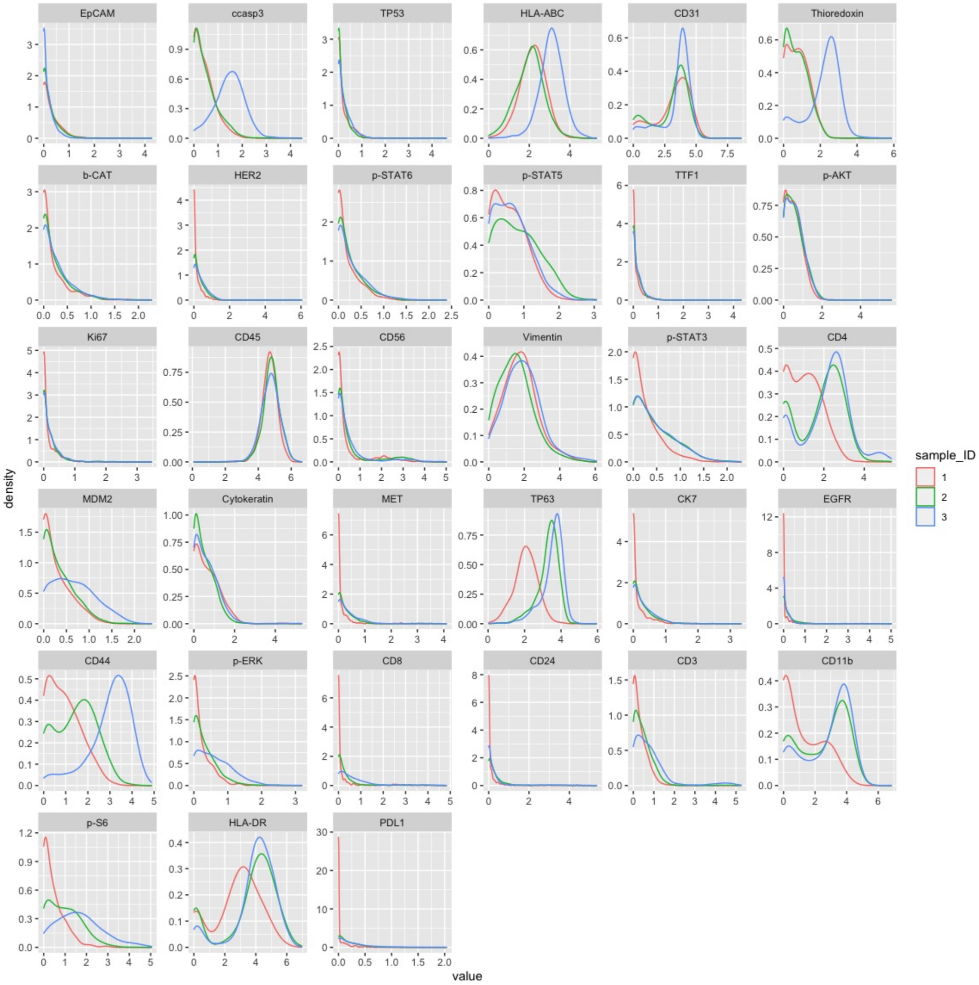


Figure S5. Monocytes protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for monocytes from the PBMC sample.

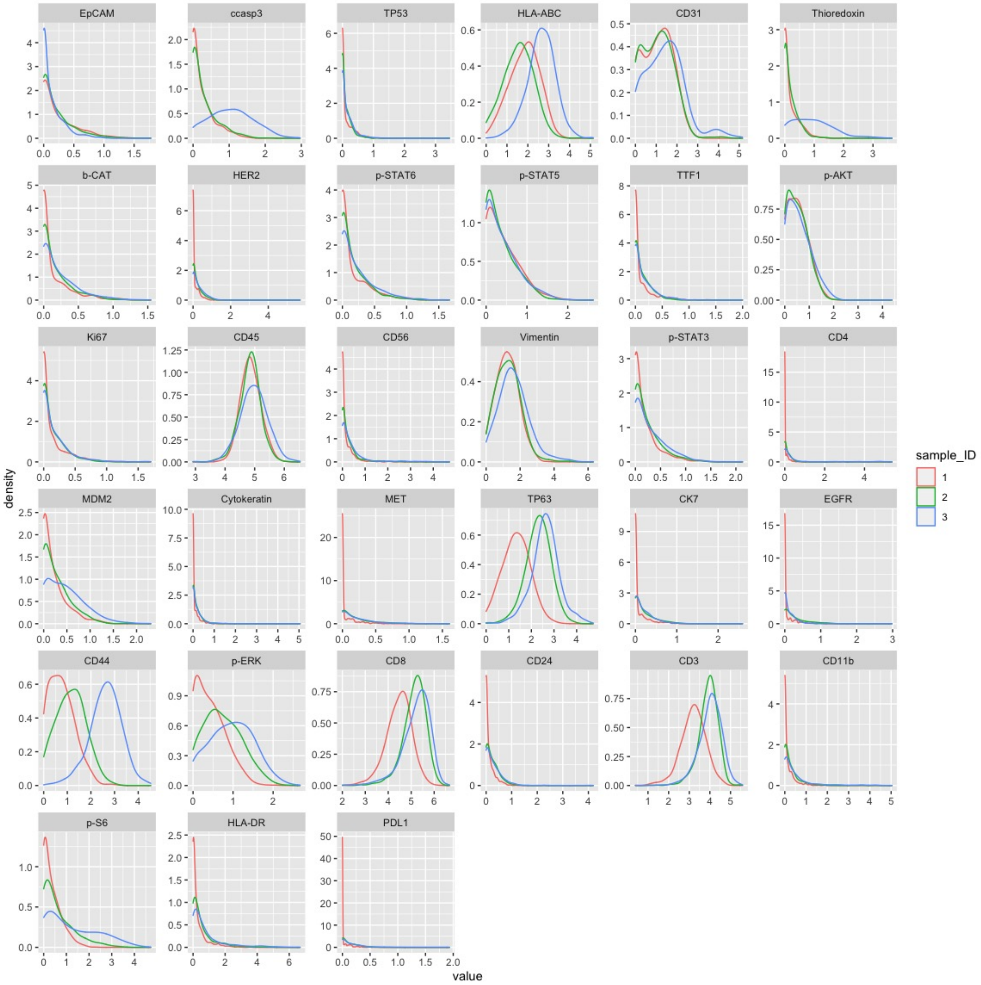


Figure S6. Cytotoxic T cells protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for CD8+ T cells from the PBMC sample.

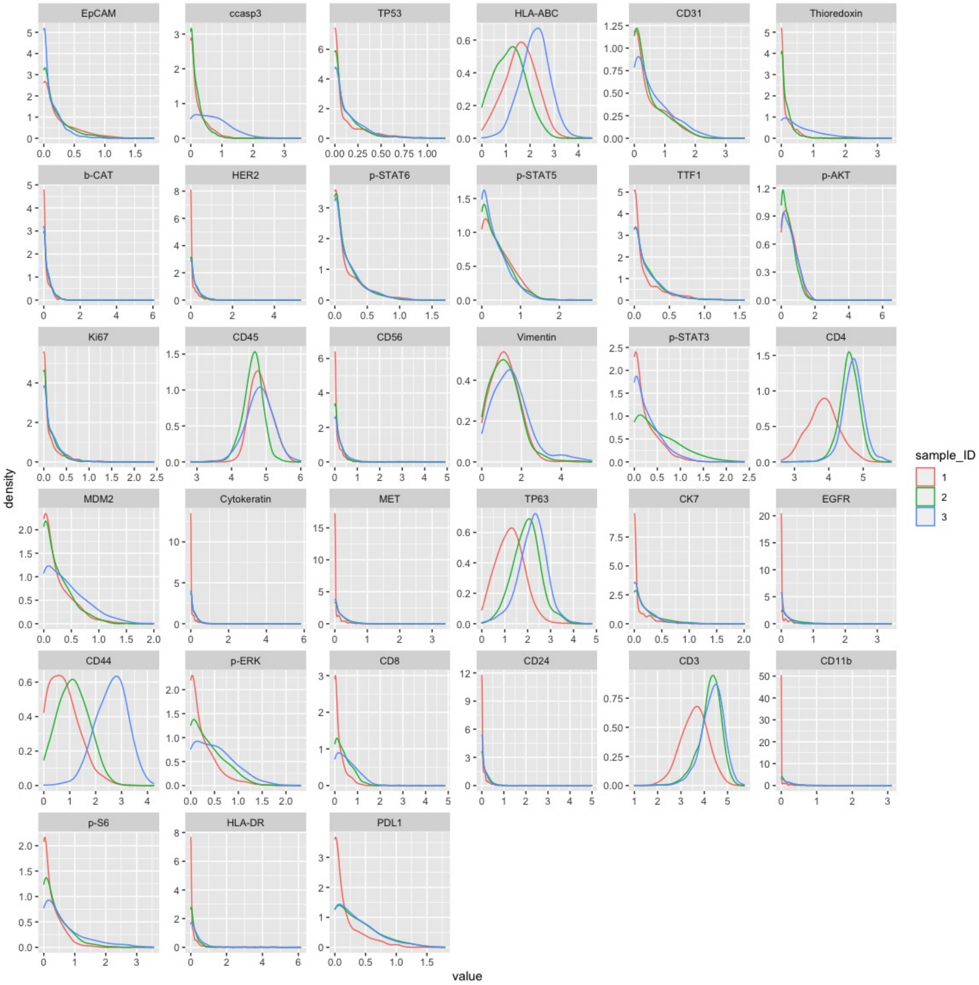
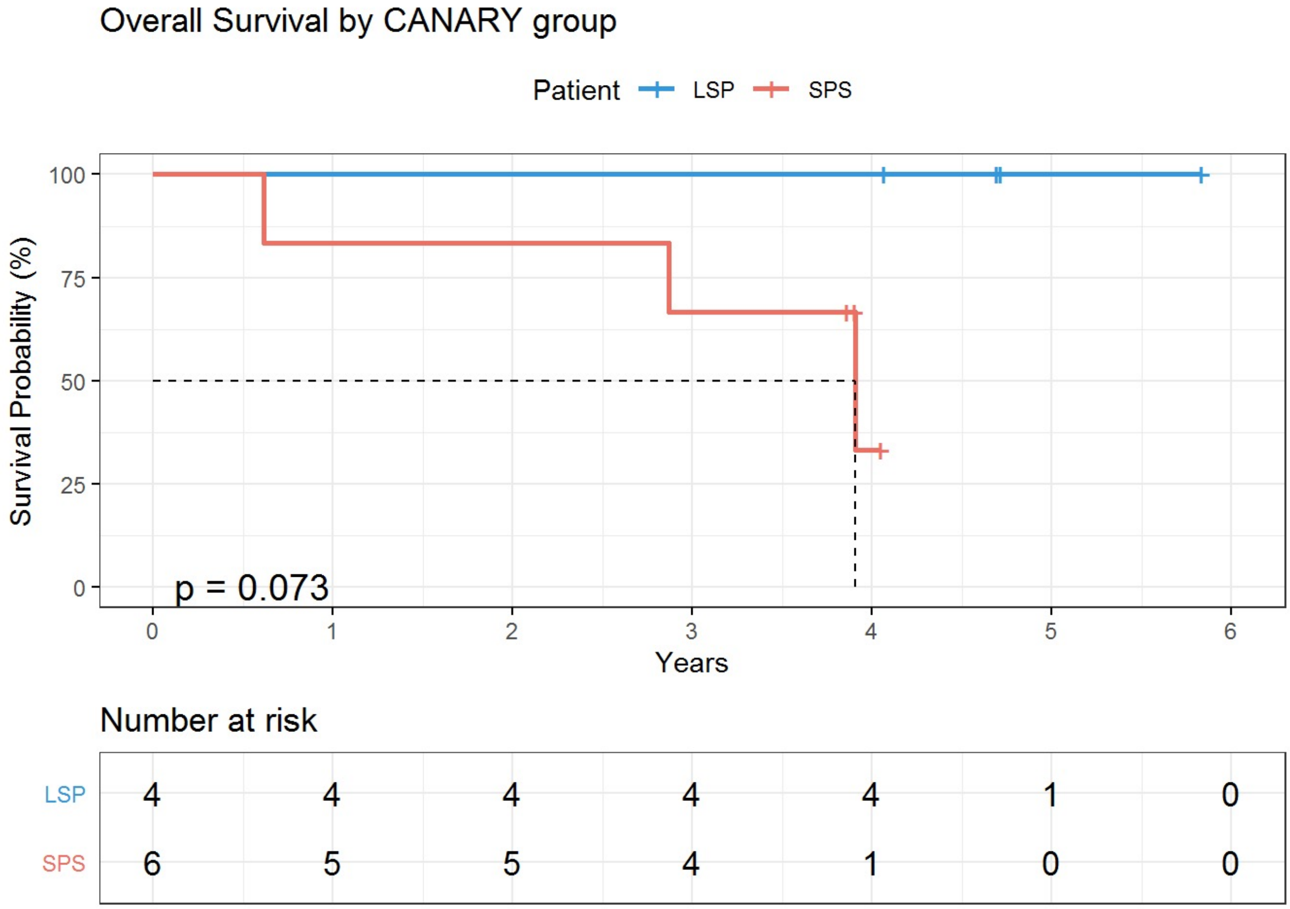


Figure S7. T helper cells protein expression across replicates. Samples 1 and 2 correspond to experimental mix of cell lines stained and run through the CyTOF machine separately. Sample 3 corresponds to a computational mix, for which each cell line and PBMCs were stained and run through the CyTOF machine independently and then files were concatenated to obtain a labeled mix. All samples were analyzed in Cytobank, where cell types were manually gated and annotated based on protein expression. This is the data for CD4+ T cells from the PBMC sample.



	0	1	2	3	4	5	6
LSP	4	4	4	4	4	1	0
SPS	6	5	5	4	1	0	0

Figure S8. Survival analysis of LPS vs SPS ADC samples. Survival curves were generated using the Kaplan-Meier method, and statistically significant differences were analyzed with the log rank test.

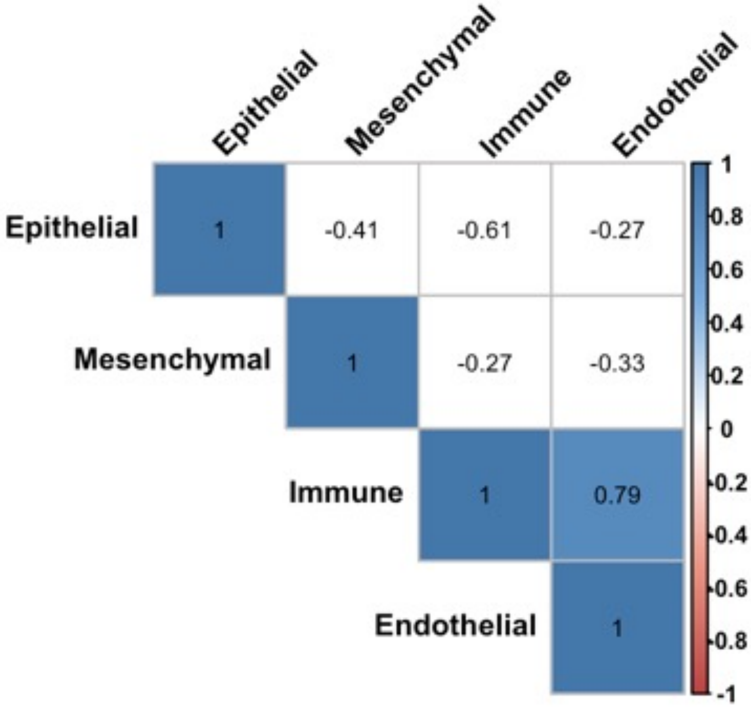


Figure S9. Spearman correlation of main cell types. Only significant correlations (p value >0.05) are colored. P values are adjusted for multiple hypothesis testing by Benjamini-Hochberg procedure.

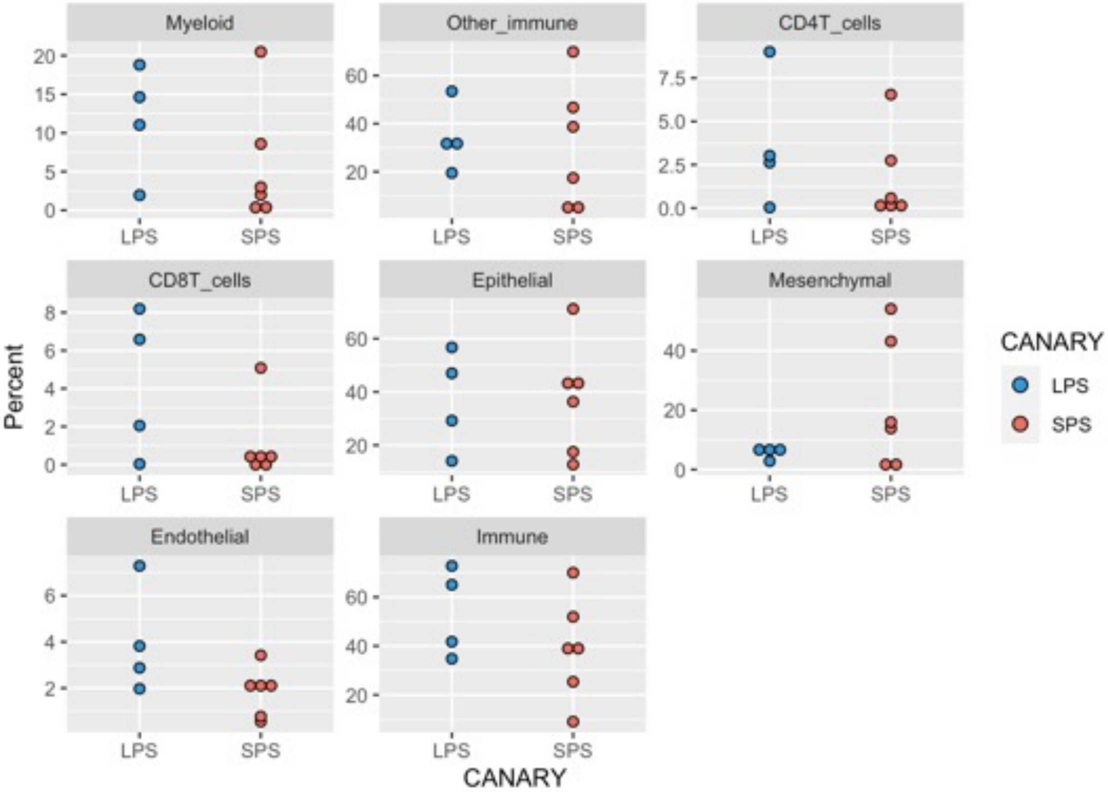


Figure S10. Differential abundance analysis. P value >0.05 for all comparisons. "Immune" correspond to the percentages of all immune subtypes added up per patient.

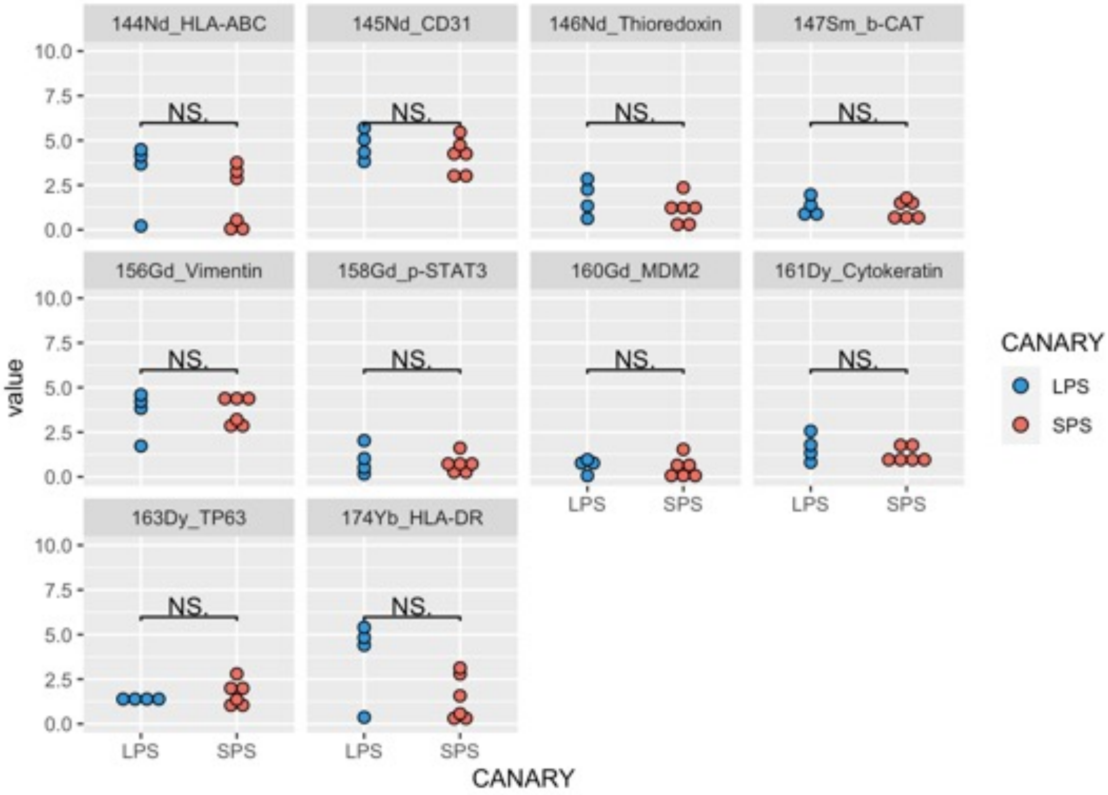


Figure S11. Protein expression comparison for endothelial cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.

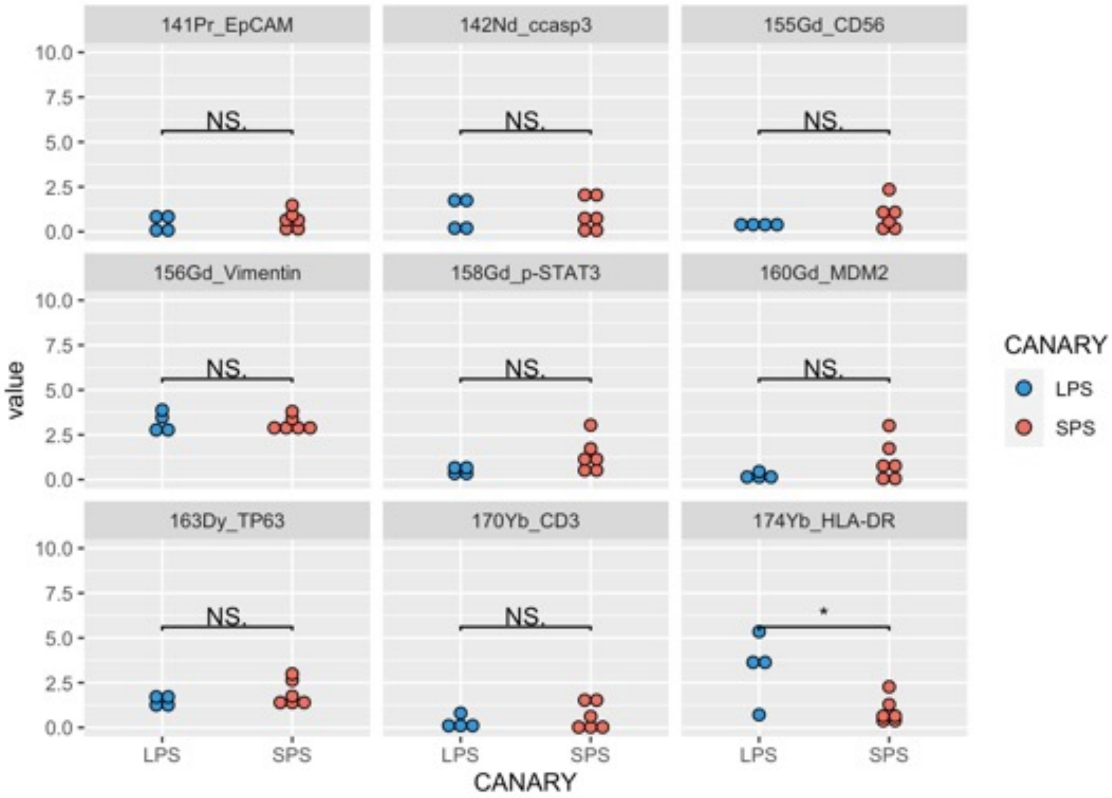


Figure S12. Protein expression comparison for fibroblasts/mesenchymal cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.

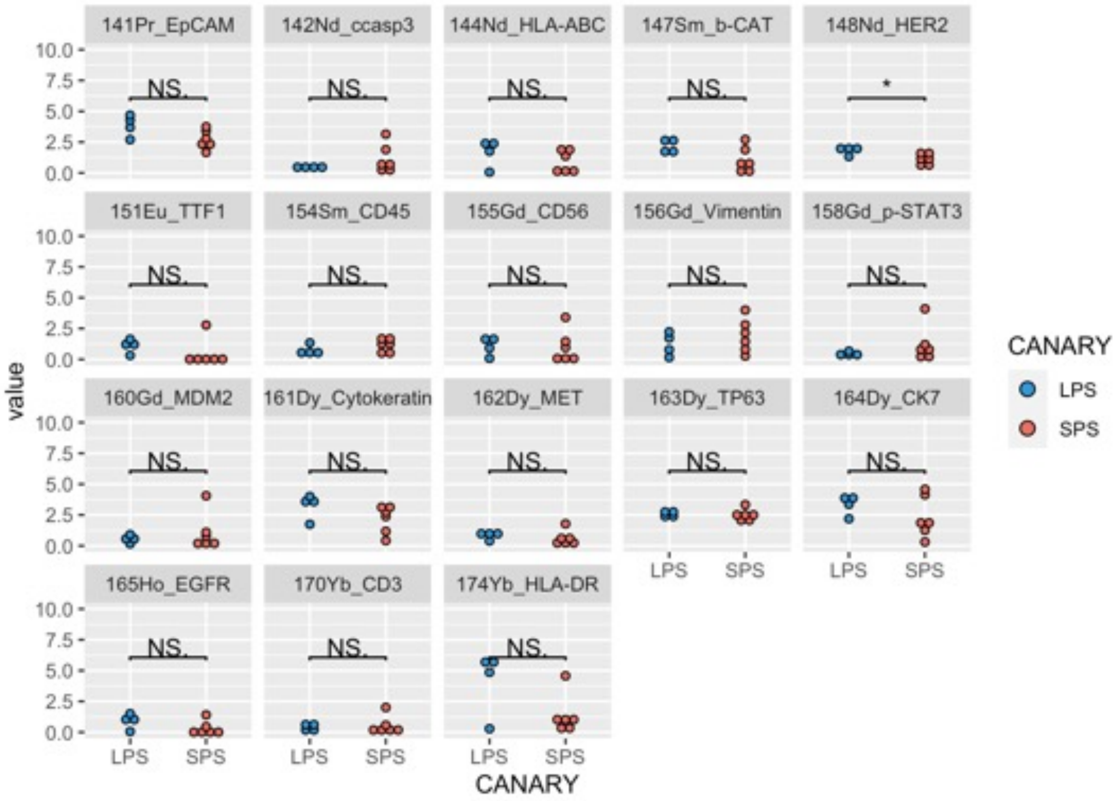


Figure S13. Protein expression comparison for epithelial cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.

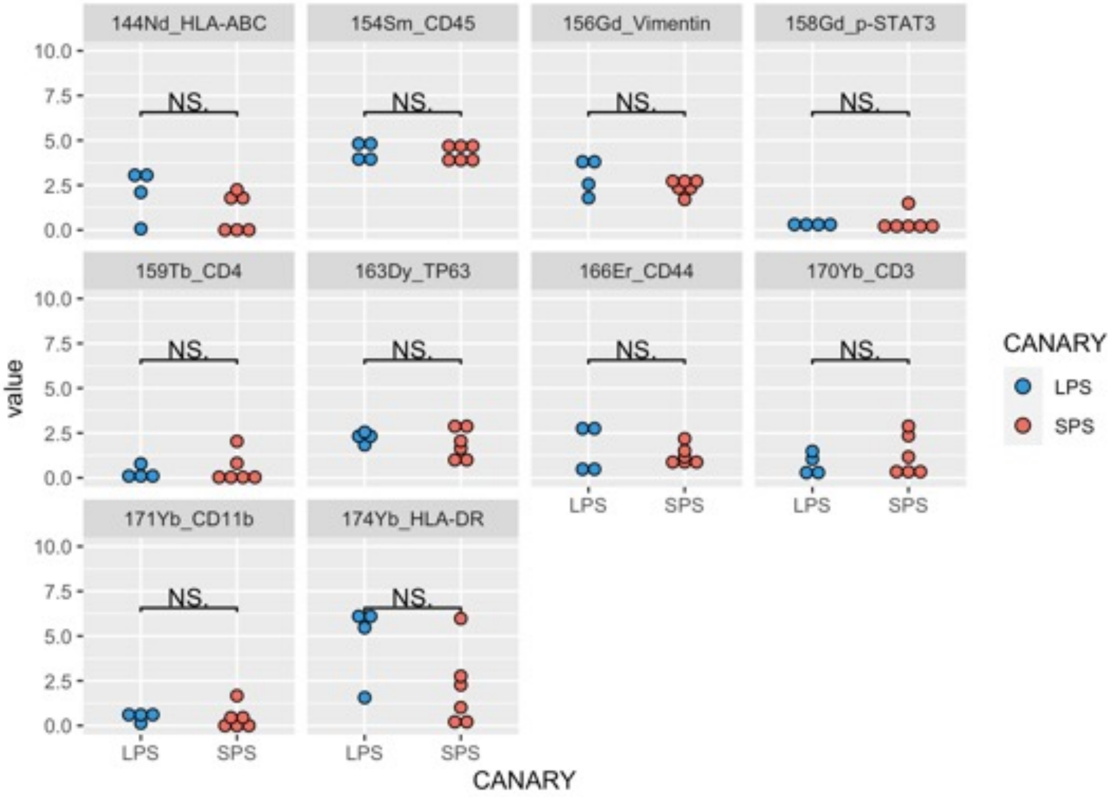


Figure S14. Protein expression comparison for immune cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.

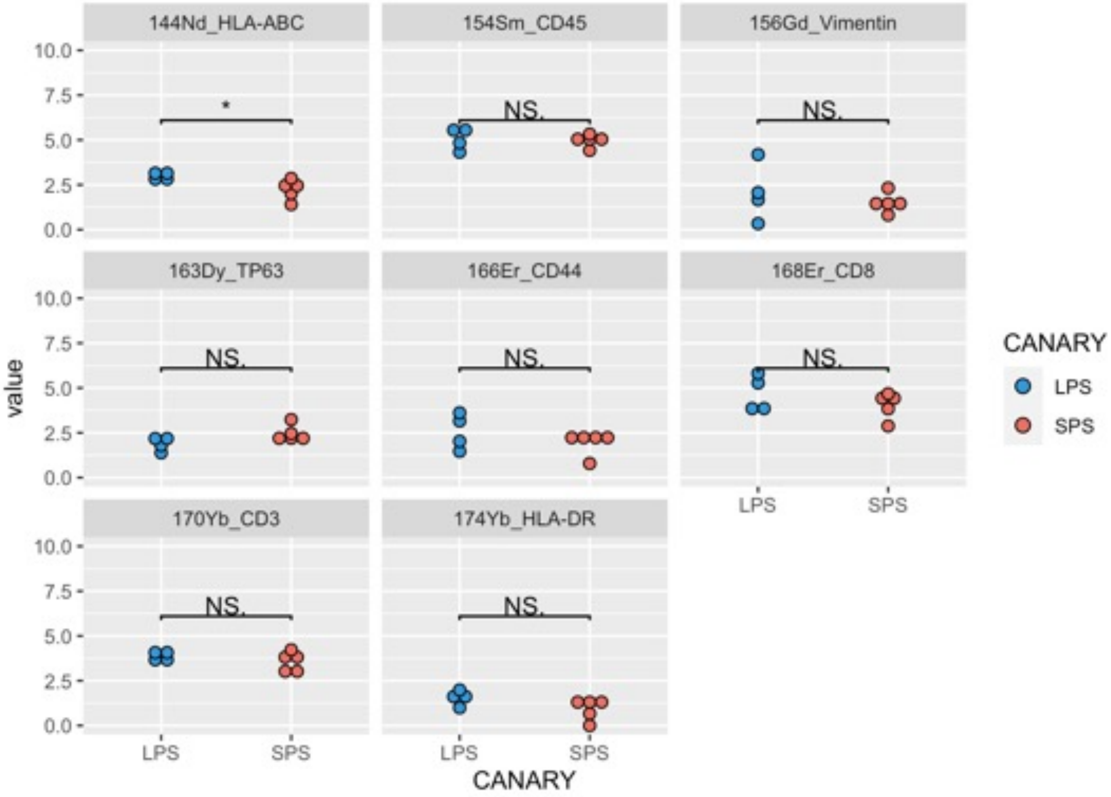


Figure S15. Protein expression comparison for CD8+ T cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.

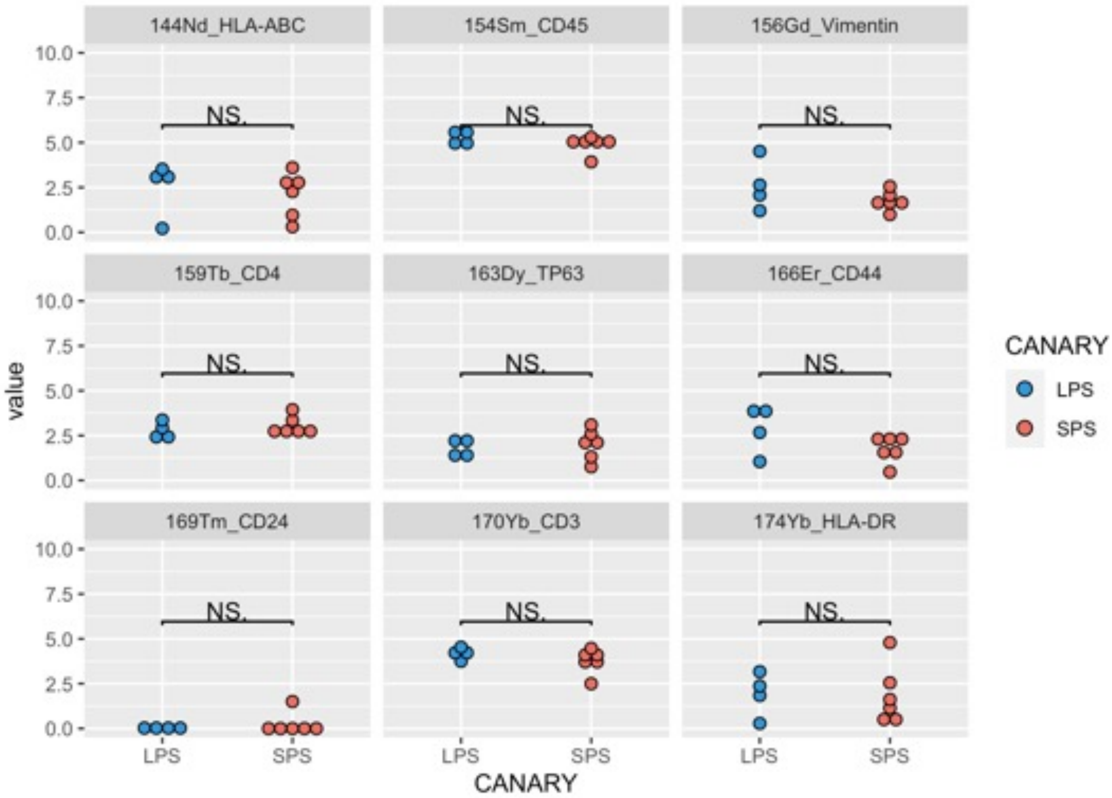


Figure S16. Protein expression comparison for CD4+ T cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.

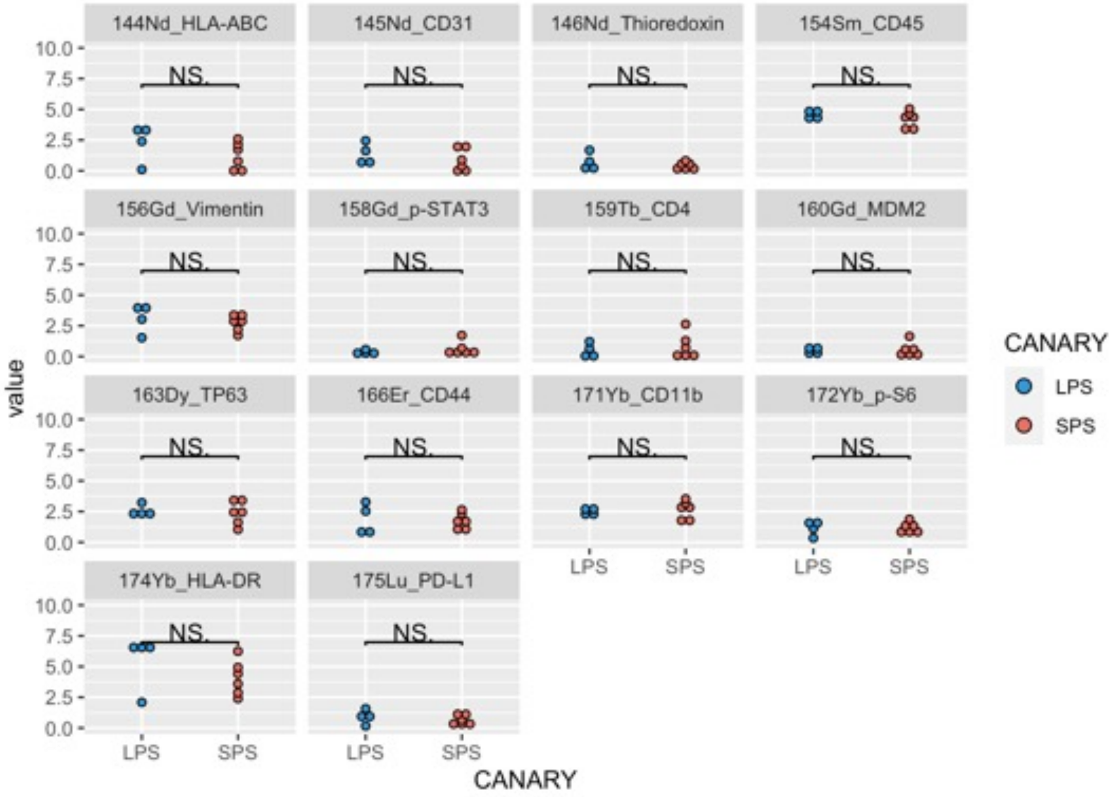
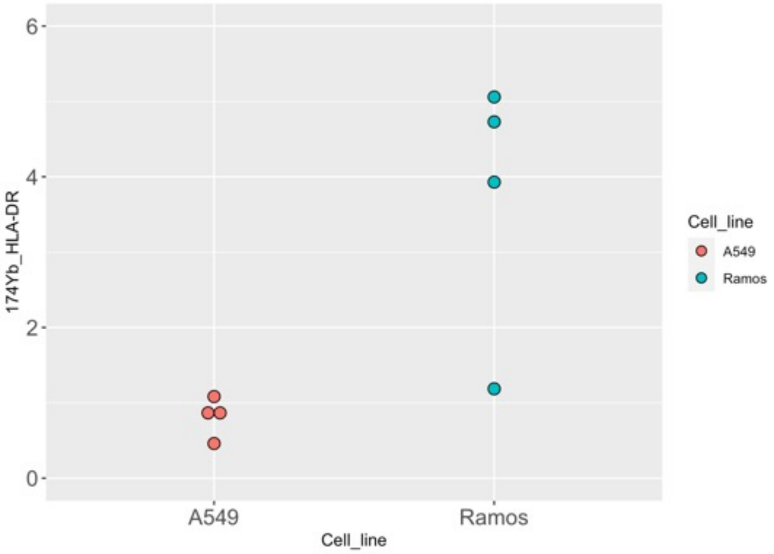


Figure S17. Protein expression comparison for myeloid cells. Only the protein markers which have an average protein expression > 1.4 for at least one patient are shown.



	Min	1st Qu.	Median	Mean	3rd Qu	Max
A549	0.4604	0.7593	0.8666	0.8193	0.9267	1.0836
Ramos	1.186	3.244	4.329	3.726	4.812	5.059

Figure S18. HLA-DR expression in batch control cell lines A549 and Ramos.

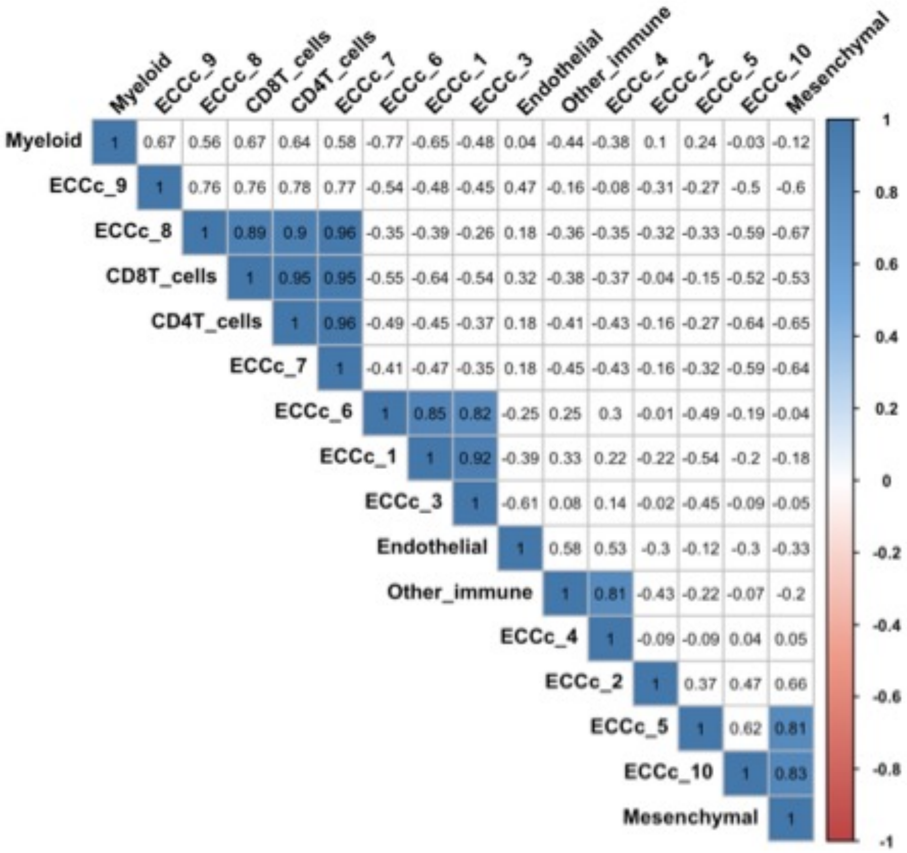


Figure S19. Spearman correlation of main all cell types and 10 epithelial clusters. Only significant correlations (p value >0.05) are colored. P values are adjusted for multiple hypothesis testing by Benjamini-Hochberg procedure.

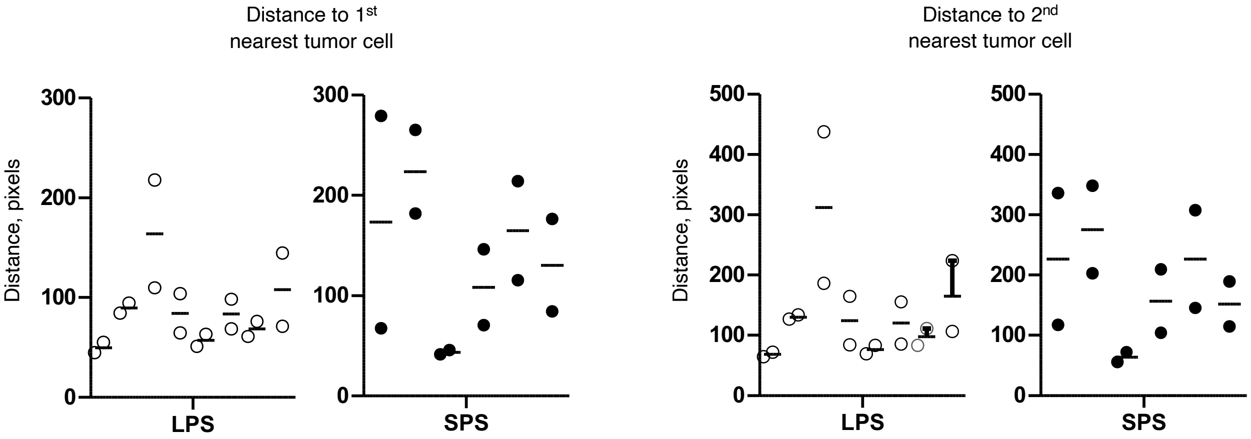


Figure S20. Extended Figure 5C showing results for each individual patient (2 cores/patient).

Appendix B

Supplementary material for Chapter 4.

Pt ID	SILA score	Group	Age at collection	Sex	Race	Smoking Status	Pack Years	Family History Cancer Type	Chest CT Location	Path_T	Path_N	Path_M	8th ed path stage	Path Nodule Size (cm)	Histology predominant	Histology other patterns
7984	0.049	Indolent	66	Female	Caucasian	Ex-smoker	50	Unknown	NA	T1a	N0	M0	Stage IA1	0.8	solid	acinar, lepidic
8356	0.115	Indolent	72	Female	Caucasian	Ex-smoker	37	Pancreatic	LLL	Tis	N0	M0	Stage 0	2.1	lepidic	NA
11424	0.75	Aggressive	72	Male	Caucasian	Current smoker	61	Unknown	RUL	T1c	N0	M0	Stage IA3	2.1	solid	NA
11522	0.23	Indolent	57	Female	Caucasian	Ex-smoker	20.5	Unknown	RUL	T1c	N0	M0	Stage IA3	2.2	acinar	lepidic
11538	0.758	Aggressive	69	Male	African American	Ex-smoker	60	Unknown	LLL	T2	N0	NA	Stage IB	2	micropapillary	NA
11561	0.555	Intermediate	58	Male	Caucasian	Ex-smoker	28	Breast	LUL	T1c	N0	M0	Stage IA3	2.2	papillary	acinar, micropapillary
11601	0.61	Aggressive	53	Male	Caucasian	Current smoker	38	Hematological Cancer	RUL	T1a	N0	M0	Stage IA1	1.7	solid	acinar
11646	0.695	Aggressive	71	Male	Caucasian	Ex-smoker	100	Pancreatic	RUL	T1b	N0	Mx	Stage IA2	2.5	micropapillary	solid, acinar
11652	0.479	Intermediate	66	Male	Caucasian	Ex-smoker	50	Colon	RLL	T1a	N0	M0	Stage IA1	1.8	papillary	micropapillary
11728	0.408	Intermediate	65	Male	Caucasian	Ex-smoker	50	Melanoma Skin Cancer	RUL	T1b	N0	NA	Stage IA2	1.7	micropapillary	acinar, solid
11759	0.634	Aggressive	80	Male	Caucasian	Ex-smoker	48	Unknown	RUL	T1b	N0	M0	Stage IA2	2	acinar	micropapillary
11813	0.547	Intermediate	60	Female	Caucasian	Ex-smoker	10	Lung - Small Cell	RLL	T1b	N0	M0	Stage IA2	1.3	acinar	lepidic
11817	0.797	Aggressive	68	Female	Caucasian	Ex-smoker	48	Unknown	LUL	T2b	N1	M0	Stage IIB	4.4	acinar	solid
11820	0.665	Aggressive	80	Male	Caucasian	Ex-smoker	NA	Unknown	RUL	T1a	N0	M0	Stage IA1	1.3	acinar	solid
11851	0.428	Intermediate	81	Male	Caucasian	Ex-smoker	90	Unknown	LUL	T2a	N0	M0	Stage IB	3.8	acinar	NA
11855	0.761	Aggressive	59	Male	Caucasian	Ex-smoker	15	Esophagus	RLL	T1c	N0	M0	Stage IA3	2.5	acinar	NA
11886	0.315	Indolent	59	Female	Caucasian	Ex-smoker	10	Esophagus	RUL	T2a	N0	M0	Stage IB	0.9	acinar	solid
11901	0.647	Aggressive	83	Male	Caucasian	Ex-smoker	NA	Other	RUL	T1c	N0	M0	Stage IA3	2.7	acinar	solid
11906	0.505	Intermediate	62	Female	Caucasian	Ex-smoker	27	Pancreatic	RLL	T1b	N0	M0	Stage IA2	1.4	acinar	lepidic
11918	0.36	Indolent	68	Male	African American	Ex-smoker	25	Gastrointesti l Cancer	RUL	T1a	N0	M0	Stage IA1	1.3	solid	acinar, lepidic
11938	0.417	Intermediate	63	Male	Caucasian	Current smoker	72	Hematological Cancer	RUL	T2a	N0	M0	Stage IB	1.4	acinar	lepidic
11952	0.822	Aggressive	61	Female	Caucasian	Ex-smoker	63	Melanoma Skin Cancer	RUL	T1c	N0	M0	Stage IA3	2.1	acinar	NA
11957	0.712	Aggressive	68	Male	Caucasian	Ex-smoker	24	Unknown	RLL	T3	N1	M0	Stage IIIA	4.3	acinar	solid
12177	0.649	Aggressive	74	Male	Caucasian	Ex-smoker	15	Bone	LUL	T1b	N2	M0	Stage IIIA	2	acinar	solid
12281	0.34	Indolent	60	Female	Caucasian	Ex-smoker	25	Unknown	RUL	T1b	N0	M0	Stage IA2	1.4	acinar	lepidic, micropapillary
12323	0.794	Aggressive	59	Female	Caucasian	Ex-smoker	20	Unknown	LLL	T4	N0	M0	Stage IIIA	1.8	acinar	micropapillary
12546	0.476	Intermediate	69	Male	Caucasian	Ex-smoker	60	Breast	RUL	T1a	N0	Mx	Stage IA1	1.8	papillary	acinar
12889	0.768	Aggressive	74	Female	Caucasian	Ex-smoker	50	Breast	RUL	T3	N0	M0	Stage IIB	6.1	solid	NA
12890	0.791	Aggressive	64	Male	Caucasian	Current smoker	48	Unknown	NA	T2b	N0	M0	Stage IIA	4.5	acinar	micropapillary
12911	0.322	Indolent	72	Male	African American	Ex-smoker	15	Unknown	LUL	T1b	N0	M0	Stage IA2	1.5	acinar	micropapillary, lepidic
12915	0.435	Intermediate	61	Female	Caucasian	Never smoked	NA	Unknown	NA	T2a	N0	M0	Stage IB	3.5	acinar	lepidic
12924	0.669	Aggressive	70	Male	Caucasian	Ex-smoker	53	Melanoma Skin Cancer	LLL	T2a	N1	M0	Stage IIB	4	papillary	acinar
12929	0.622	Aggressive	86	Female	Caucasian	Ex-smoker	37.5	Unknown	LLL	T3	N0	M0	Stage IIB	3.3	acinar	lepidic
12931	0.79	Aggressive	65	Male	African American	Ex-smoker	50	Lung Cancer	RLL	T2b	N0	M0	Stage IIA	4.9	solid	acinar
12935	0.441	Intermediate	58	Male	Caucasian	Ex-smoker	96	Liver	NA	T1c	Nx	M1a	Stage IV	2.4	solid	NA
12994	0.75	Aggressive	76	Male	Caucasian	Ex-smoker	24	Brain	RUL	T3	N2	M0	Stage IIIB	3.5	micropapillary	acinar
13014	0.735	Aggressive	82	Female	Caucasian	Ex-smoker	27	Lung Cancer	LUL	T2b	N0	M0	Stage IIA	4.1	acinar	micropapillary
13034	0.551	Intermediate	70	Male	Caucasian	Ex-smoker	39	Lung Cancer	RUL	T2a	N0	M0	Stage IB	3.7	micropapillary	acinar
13055	0.418	Intermediate	79	Female	Caucasian	Never smoked	NA	Unknown	RUL	T1c	N0	M0	Stage IA3	2.4	acinar	micropapillary
13074	0.742	Aggressive	76	Female	Caucasian	Never smoked	NA	Unknown	RUL	T1b	N0	M0	Stage IA2	1.7	solid	acinar
13155	0.731	Aggressive	64	Male	Caucasian	Ex-smoker	144	Unknown	RUL	T1c	N0	M0	Stage IA3	3	acinar	micropappilary
13197	0.697	Aggressive	78	Male	Caucasian	Ex-smoker	35	Lung Cancer	RLL	T1b	N1	M0	Stage IIB	1.3	acinar	micropapillary
13207	0.675	Aggressive	60	Female	Caucasian	Ex-smoker	43	Prostate	RLL	T1c	N0	M0	Stage IA3	2.3	micropapillary	NA
13276	0.459	Intermediate	63	Female	Caucasian	Ex-smoker	47	Breast	RUL	T1b	N0	M0	Stage IA2	1.8	acinar	micropapillary
13317	0.697	Aggressive	62	Male	Caucasian	Current smoker	78	Prostate	RUL	T1c	N0	M0	Stage IA3	2.5	micropapillary	solid
13356	0.621	Aggressive	50	Female	Caucasian	Ex-smoker	78	Unknown	RUL	T3	N0	M0	Stage IIB	1.7	acinar	solid
13376	0.274	Indolent	64	Female	Caucasian	Current smoker	20	Bladder	RUL	T2a	Nx	M0	Stage IB	3.2	lepidic	acinar
13436	0.739	Aggressive	56	Male	Caucasian	Ex-smoker	45	Gynecological Cancer	RLL	T3	N0	M0	Stage IIB	6.9	acinar	NA
13536	0.502	Intermediate	76	Female	Caucasian	Ex-smoker	46	Prostate	RUL	T1b	N0	M0	Stage IA2	1.7	acinar	NA
13538	0.368	Indolent	62	Male	Caucasian	Ex-smoker	30	Unknown	LLL	T1a	N0	M0	Stage IA1	0.9	papillary	lepidic
13579	0.664	Aggressive	52	Male	Caucasian	Ex-smoker	45	Lung Cancer	RLL	T3	N0	M0	Stage IIB	5.5	acinar	NA
13622	0.799	Aggressive	67	Female	Caucasian	Ex-smoker	110	Lung Cancer	RUL	T2b	N0	M0	Stage IIA	4.1	solid	NA
13634	0.25	Indolent	67	Female	Caucasian	Current smoker	NA	Other	RUL	T3	N2	M0	Stage IIIB	2	solid	NA
13636	0.853	Aggressive	75	Female	Caucasian	Ex-smoker	120	Breast	RUL	T4	N0	M0	Stage IIIA	7.3	papillary	NA
13651	0.548	Intermediate	59	Female	Caucasian	Ex-smoker	35	Bladder	RUL	T1b	N2	M0	Stage IIIA	1.9	acinar	lepidic
13724	0.789	Aggressive	75	Female	Caucasian	Ex-smoker	15	Melanoma Skin Cancer	RLL	T2b	N0	M0	Stage IIA	5	solid	acinar
13746	0.821	Aggressive	64	Male	Caucasian	Ex-smoker	30	Prostate	LUL	T4	N0	M1a	Stage IV	NA	acinar	NA
13769	0.699	Aggressive	66	Female	Caucasian	Ex-smoker	10.5	Lung Cancer	RUL	T2b	N0	M0	Stage IIA	4.6	acinar	lepidic
13771	0.554	Intermediate	74	Female	Caucasian	Ex-smoker	NA	Prostate	LLL	T1c	N0	M0	Stage IA3	2.2	papillary	acinar
13774	0.735	Aggressive	54	Female	Caucasian	Ex-smoker	35.25	Unknown	LUL	T1c	N0	M0	Stage IA3	2.7	acinar	papillary
13801	0.705	Aggressive	65	Female	Caucasian	Ex-smoker	20	Unknown	RUL	T1b	N0	M0	Stage IA2	1.7	acinar	NA
13922	0.359	Indolent	82	Female	Caucasian	Current smoker	60	UNknown	RUL	T1a	N1	M0	Stage IIB	1	acinar	solid
13988	0.575	Intermediate	56	Male	Caucasian	Ex-smoker	28.5	Melanoma Skin Cancer	NA	T2b	N0	M0	Stage IIA	4.1	acinar	NA
14048	0.774	Aggressive	62	Female	Caucasian	Ex-smoker	35	Lung Cancer	LLL	T2a	N0	M0	Stage IB	3.5	acinar	NA
14201	0.416	Intermediate	82	Male	Caucasian	Ex-smoker	1.25	Breast	LLL	T1b	N0	M0	Stage IA2	1.8	acinar	lepidic
14301	0.826	Aggressive	82	Female	Caucasian	Ex-smoker	40	Brain	RUL	T2b	N1	M0	Stage IIB	4.8	micropapillary	solid, acinar
14330	0.732	Aggressive	50	Female	Caucasian	Ex-smoker	39	Colon	LUL	T2b	N0	M0	Stage IIA	4.8	acinar	lepidc
14428	0.493	Intermediate	73	Male	Caucasian	Current smoker	45	Gynecological Cancer	RUL	T1b	N0	NA	Stage IA2	3.9	acinar	micropapillary, lepidic
14610	0.758	Aggressive	64	Female	Caucasian	Never smoked	31	Unknown	RUL	T2b	N0	M0	Stage IIA	4.5	mucinous acinar	lepidic
14813	0.627	Aggressive	69	Male	Caucasian	Ex-smoker	61.5	Prostate	RUL	T1b	N0	NA	Stage IA2	1.6	solid	acinar, lepidic
14825	0.418	Intermediate	59	Female	Asian	Never smoked	NA	Unknown	LUL	T1c	N1	M0	Stage IIB	2.7	acinar	micropapillary
14836	0.773	Aggressive	66	Female	Caucasian	Never smoked	NA	Lung Cancer	RUL	T1c	N0	M0	Stage IA3	2.2	solid	acinar
14855	0.513	Intermediate	79	Male	Caucasian	Ex-smoker	50	Other	LLL	T2a	N0	M0	Stage IB	3.5	acinar	micropapillary
14933	0.599	Intermediate	75	Male	Caucasian	Ex-smoker	80	Other	RLL	T1c	N0	NA	Stage IA3	2.6	acinar	micropapillary
14953	0.821	Aggressive	73	Male	Caucasian	Ex-smoker	57	Unknown	LUL	T2b	N0	M0	Stage IIA	4.4	solid	NA
14955	0.56	Intermediate	79	Female	Caucasian	Ex-smoker	27	Breast	RLL	T3	N0	M0	Stage IIB	1.3	solid	acinar
14958	0.148	Indolent	68	Female	Caucasian	Current smoker	23.5	Other	LUL	T2a	N0	M0	Stage IB	3.2	solid	acinar
14962	0.721	Aggressive	68	Female	Caucasian	Ex-smoker	32	Breast	LLL	T2b	N1	M0	Stage IIB	5	papillary	acinar
14965	0.35	Indolent	62	Female	Caucasian	Never smoked	NA	Other	LUL	T1c	N0	NA	Stage IA3	2.5	acinar	lepidic
15001	0.532	Intermediate	66	Female	Caucasian	Never smoked	NA	Unknown	RUL	T2a	N0	M0	Stage IB	2.3	micropapillary	acinar
15002	0.724	Aggressive	90	Male	Asian	Ex-smoker	28.5	Prostate	RUL	T2b	N0	M0	Stage IIA	4.2	solid	NA
15083	0.421	Intermediate	75	Male	Caucasian	Current smoker	75	Unknown	RUL	T1b	N0	M0	Stage IA2	2.1	acinar	micropapillary, lepidic
15187	0.616	Aggressive	70	Male	Caucasian	Never smoked	NA	Unknown	RLL	T1b	N0	M0	Stage IA2	1.9	acinar	lepidic
15224	0.716	Aggressive	48	Female	Caucasian	Never smoked	NA	Colon	LLL	T1b	N0	Mx	Stage IA2	1.9	acinar	papillary, micropapillary
15325	0.622	Aggressive	82	Female	Caucasian	Ex-smoker	30	Colon	RML	T1a	N0	M0	Stage IA1	0.9	papillary	NA
15326	0.612	Aggressive	55	Female	Caucasian	Ex-smoker	3	Pancreatic	RLL	T1c	N0	M0	Stage IA3	2.1	acinar	micropapillary
15467	0.699	Aggressive	61	Female	African American	Never smoked	NA	Colon	LUL	T1c	N0	M0	Stage IA3	2.4	mucinous acinar	NA
15506	0.777	Aggressive	81	Female	Caucasian	Never smoked	NA	Gastrointesti l Cancer	RUL	T2b	N0	M0	Stage IIA	4.6	mucinous acinar	NA
15569	0.687	Aggressive	62	Male	Caucasian	Never smoked	NA	Prostate	RUL	T2b	N1	M0	Stage IIA	4.9	acinar	solid
15626	0.699	Aggressive	71	Female	Caucasian	Never smoked	NA	Lung Cancer	RLL	T1c	N1	M0	Stage IIB	2.2	acinar	micropapilary, papillary, lepidic
15641	0.38	Indolent	60	Male	Caucasian	Ex-smoker	52	Lung Cancer	LUL	T1b	N1	M0	Stage IIB	1.8	acinar	solid, micropapillary
15741	0.405	Intermediate	56	Female	African American	Never smoked	NA	Head and Neck Cancer	RUL	T1b	N0	NA	Stage IA2	1.8	acinar	lepidic

Table S1. Detailed patient clinical characteristics

Pt ID	CyTOF	RNA-Seq	WES	scRNA-Seq	MxIF
7984	1	0	0	0	1
8356	1	1	0	1	1
11424	0	1	1	0	1
11522	1	0	0	1	1
11538	1	1	1	0	1
11561	1	1	1	0	1
11601	0	1	1	0	1
11646	1	1	1	0	1
11652	1	1	1	0	1
11728	0	0	0	0	1
11759	1	1	0	0	1
11813	1	1	0	0	1
11817	1	1	1	1	1
11820	0	1	1	0	1
11851	1	1	0	0	1
11855	1	1	1	0	1
11886	1	1	0	0	1
11901	1	0	0	0	1
11906	1	1	0	0	1
11918	1	0	0	1	1
11938	1	1	1	0	1
11952	1	1	0	0	1
11957	0	1	1	0	1
12177	0	1	0	0	1
12281	1	1	1	0	1
12323	1	1	0	0	1
12546	1	1	1	0	1
12889	1	1	1	1	1
12890	1	1	1	0	1
12911	0	0	0	0	1
12915	1	1	1	0	1
12924	1	1	1	0	1
12929	1	1	0	1	1
12931	1	1	1	0	1
12935	1	0	0	1	1
12994	1	1	1	0	1
13014	1	1	1	0	1
13034	1	0	0	0	1
13055	1	0	0	0	1
13074	1	1	1	0	1
13155	1	1	0	0	1
13197	1	0	0	0	1
13207	1	1	1	0	1
13276	1	1	1	0	1
13317	1	1	1	0	1
13356	0	0	1	0	1
13376	1	1	1	0	1
13436	1	1	1	0	1
13536	1	1	0	0	1
13538	1	1	0	0	1
13579	0	1	1	0	1
13622	1	1	1	0	1
13634	0	1	0	1	1
13636	1	0	0	1	1
13651	1	1	0	0	1
13724	1	1	1	0	1
13746	0	0	0	0	1
13769	1	1	1	0	1
13771	1	1	1	0	1
13774	1	0	0	1	1
13801	0	1	0	0	1
13922	0	1	0	0	1
13988	1	1	1	0	1
14048	1	1	1	0	1
14201	0	1	1	0	1
14301	1	1	1	0	1
14330	0	1	0	0	1
14428	0	0	0	1	1
14610	1	1	1	0	1
14813	0	1	1	0	1
14825	0	1	1	0	1
14836	1	1	1	0	1
14855	1	1	1	0	1
14933	0	1	1	0	1
14953	0	1	1	0	1
14955	1	1	1	0	1
14958	1	1	1	1	1
14962	1	1	0	0	1
14965	1	1	1	1	1
15001	1	1	1	0	1
15002	1	1	1	1	1
15083	1	1	1	0	1
15187	1	1	1	0	1
15224	1	1	0	0	1
15325	1	1	1	0	1
15326	1	1	1	0	1
15467	1	1	1	1	1
15506	1	1	1	0	1
15569	1	1	1	0	1
15626	0	1	0	0	1
15641	0	1	1	0	1
15741	1	1	0	0	1

Table S2. Data collection by patient. (0=No, 1=Yes)

Reference group	Test group	ENSEMBL ID	Symbol	log2FoldChange	p value	p value (adj)
Indolent	Aggressive	ENSG00000108576.5	SLC6A4	-3.631980607	7.02E-08	0.001515953
Indolent	Aggressive	ENSG00000130294.10	KIF1A	4.33966982	4.68E-07	0.003527126
Indolent	Aggressive	ENSG00000149948.9	HMGA2	3.844206007	6.30E-07	0.003527126
Indolent	Aggressive	ENSG00000118322.8	ATP10B	2.905940066	9.28E-07	0.003527126
Indolent	Aggressive	ENSG00000197301.3	RP11-366L20.2	3.200667758	9.55E-07	0.003527126
Indolent	Aggressive	ENSG00000100413.12	POLR3H	-1.820271699	9.80E-07	0.003527126
Indolent	Aggressive	ENSG00000155974.7	GRIP1	1.630370299	8.97E-06	0.026885401
Indolent	Aggressive	ENSG00000164669.8	INTS4L1	2.072296832	9.96E-06	0.026885401
Indolent	Aggressive	ENSG00000065618.12	COL17A1	2.72584176	1.13E-05	0.027183728
Indolent	Aggressive	ENSG00000152669.8	CCNO	2.330335431	1.30E-05	0.028040319
Indolent	Aggressive	ENSG00000178343.4	SHISA3	3.387560179	2.33E-05	0.040539858
Indolent	Aggressive	ENSG00000270358.1	IGHV4-61	2.960817134	2.49E-05	0.040539858
Indolent	Aggressive	ENSG00000167588.8	GPD1	-1.898779478	2.57E-05	0.040539858
Indolent	Aggressive	ENSG00000173432.6	SAA1	2.36048576	2.73E-05	0.040539858
Indolent	Aggressive	ENSG00000021826.10	CPS1	3.322625723	2.94E-05	0.040539858
Indolent	Aggressive	ENSG00000223532.5	HLA-B	3.228191816	3.09E-05	0.040539858
Indolent	Aggressive	ENSG00000187950.4	OVCH1	-1.899695931	3.19E-05	0.040539858
Indolent	Aggressive	ENSG00000211936.2	IGHV4-4	2.843440501	3.45E-05	0.041405362
Indolent	Aggressive	ENSG00000263001.1	GTF2I	-2.175708756	4.30E-05	0.048867056
Indolent	Aggressive	ENSG00000159263.11	SIM2	2.137299378	5.03E-05	0.054335045
Indolent	Intermediate	ENSG00000114455.9	HHLA2	3.297350463	1.20E-06	0.015688676
Indolent	Intermediate	ENSG00000155974.7	GRIP1	1.807818465	1.37E-06	0.015688676
Indolent	Intermediate	ENSG00000233008.1	RP11-475O6.1	1.815365747	2.14E-05	0.119137532
Indolent	Intermediate	ENSG00000261520.1	DLGAP1-AS5	2.271273463	2.57E-05	0.119137532
Indolent	Intermediate	ENSG00000164669.8	INTS4L1	1.920458303	2.81E-05	0.119137532
Indolent	Intermediate	ENSG00000170927.10	PKHD1	2.549510284	3.12E-05	0.119137532
Indolent	Intermediate	ENSG00000270358.1	IGHV4-61	2.541562662	6.77E-05	0.189433624
Indolent	Intermediate	ENSG00000106278.7	PTPRZ1	2.240801218	7.43E-05	0.189433624
Indolent	Intermediate	ENSG00000118322.8	ATP10B	2.369693016	9.77E-05	0.205018187
Indolent	Intermediate	ENSG00000117983.13	MUC5B	2.690841405	9.83E-05	0.205018187
Indolent	Intermediate	ENSG00000065618.12	COL17A1	1.649864878	0.000149181	0.220366412
Indolent	Intermediate	ENSG00000168143.8	FAM83B	2.531825427	0.000173943	0.233516248
Indolent	Intermediate	ENSG00000211670.2	IGLV3-9	2.5512349	0.000190422	0.233516248
Indolent	Intermediate	ENSG00000171724.2	VAT1L	-1.844174433	0.000203624	0.233516248
Indolent	Intermediate	ENSG00000170579.10	DLGAP1	1.799082807	0.00021949	0.239724509
Indolent	Intermediate	ENSG00000133063.11	CHIT1	-1.993177776	0.000301518	0.286879103
Indolent	Intermediate	ENSG00000149948.9	HMGA2	2.669493727	0.000351529	0.298617436
Indolent	Intermediate	ENSG00000223532.5	HLA-B	3.241208651	0.000392787	0.321717861
Indolent	Intermediate	ENSG00000136883.8	KIF12	1.982682171	0.000406776	0.321717861
Indolent	Intermediate	ENSG00000197301.3	RP11-366L20.2	2.283482651	0.000782513	0.492954247
Intermediate	Aggressive	ENSG00000023839.6	ABCC2	2.48154518	4.53E-08	0.000837745
Intermediate	Aggressive	ENSG00000171560.10	FGA	3.761217094	9.69E-08	0.000837745
Intermediate	Aggressive	ENSG00000135454.9	B4GALNT1	2.424251387	1.19E-07	0.000837745
Intermediate	Aggressive	ENSG00000145794.12	MEGF10	1.911906096	4.73E-07	0.002500355
Intermediate	Aggressive	ENSG00000021826.10	CPS1	2.837783429	9.10E-07	0.003732077
Intermediate	Aggressive	ENSG00000113739.6	STC2	1.531248536	1.06E-06	0.003732077
Intermediate	Aggressive	ENSG00000240216.3	CPHL1P	2.087821087	6.91E-06	0.013871314
Intermediate	Aggressive	ENSG00000173432.6	SAA1	1.855845949	7.02E-06	0.013871314
Intermediate	Aggressive	ENSG00000025423.7	HSD17B6	-1.518457392	8.20E-06	0.013871314
Intermediate	Aggressive	ENSG00000164283.8	ESM1	1.731011117	8.52E-06	0.013871314
Intermediate	Aggressive	ENSG00000179603.13	GRM8	1.557691984	1.09E-05	0.014425084
Intermediate	Aggressive	ENSG00000160862.8	AZGP1	2.192218506	1.94E-05	0.020323777
Intermediate	Aggressive	ENSG00000136231.9	IGF2BP3	1.91423555	2.01E-05	0.020323777
Intermediate	Aggressive	ENSG00000167779.3	IGFBP6	1.559799682	2.02E-05	0.020323777
Intermediate	Aggressive	ENSG00000101057.11	MYBL2	1.610033045	2.34E-05	0.022491299
Intermediate	Aggressive	ENSG00000145920.10	CPLX2	1.817142293	2.54E-05	0.022985436
Intermediate	Aggressive	ENSG00000144452.10	ABCA12	1.763964205	3.25E-05	0.025447451
Intermediate	Aggressive	ENSG00000106236.3	NPTX2	2.004396803	3.47E-05	0.026222061
Intermediate	Aggressive	ENSG00000152578.8	GRIA4	1.68515163	3.62E-05	0.026433707
Intermediate	Aggressive	ENSG00000206557.5	TRIM71	-1.605200172	4.82E-05	0.030873329

Table S3. Top 20 differentially expressed per group comparison

Reference group	Test group	Regulon	Size	NES	p.value	FDR
Indolent	Aggressive	FOXO1	34	-2.78	0.00541	0.233
Indolent	Aggressive	SPI1	81	-2.5	0.0123	0.264
Indolent	Intermediate	HIF1A	128	-2.1	0.036	0.797
Indolent	Intermediate	SPI1	81	-2.08	0.0377	0.797
Intermediate	Aggressive	FOXM1	32	2.37	0.0178	0.766
Intermediate	Aggressive	HIF1A	128	1.89	0.0588	0.892

Table S4. Transcription factor activity inferred with VIPER

database	pathway	pval	padj	log2err	NES	size	state	pvlabel
HALLMARK	HALLMARK_TNFA_SIGNALING_VIA_NFKB	1.27E-25	6.35E-24	1.31101476	-3.0500513	189	down	***
HALLMARK	HALLMARK_E2F_TARGETS	2.31E-15	5.77E-14	1.00731796	2.40633076	182	up	***
HALLMARK	HALLMARK_GLYCOLYSIS	1.30E-12	1.62E-11	0.91011973	2.32426556	172	up	***
HALLMARK	HALLMARK_TGF_BETA_SIGNALING	1.33E-06	1.11E-05	0.64355184	-2.3235829	52	down	***
HALLMARK	HALLMARK_G2M_CHECKPOINT	3.24E-13	5.39E-12	0.93259521	2.27985762	181	up	***
HALLMARK	HALLMARK_INFLAMMATORY_RESPONSE	2.29E-09	2.29E-08	0.77493903	-2.105593	179	down	***
HALLMARK	HALLMARK_APOPTOSIS	1.12E-05	8.03E-05	0.59332548	-1.8363303	144	down	***
HALLMARK	HALLMARK_MYC_TARGETS_V2	0.00075123	0.00313011	0.47727082	1.82136349	56	up	**
HALLMARK	HALLMARK_KRAS_SIGNALING_UP	6.59E-05	0.00041206	0.5384341	-1.6794626	170	down	***
HALLMARK	HALLMARK_COMPLEMENT	8.11E-05	0.00045029	0.5384341	-1.6701543	168	down	***
HALLMARK	HALLMARK_ALLOGRAFT_REJECTION	0.0001696	0.00084798	0.51884808	-1.6607444	156	down	***
HALLMARK	HALLMARK_ESTROGEN_RESPONSE_LATE	0.00036704	0.00166838	0.49849311	1.65497782	174	up	**
HALLMARK	HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.00253117	0.00844963	0.4317077	-1.6526158	67	down	**
HALLMARK	HALLMARK_IL6_JAK_STAT3_SIGNALING	0.00424795	0.01132538	0.40701792	-1.6374488	73	down	*
HALLMARK	HALLMARK_SPERMATOGENESIS	0.00277594	0.00867482	0.31827968	1.63514195	87	up	**
HALLMARK	HALLMARK_KRAS_SIGNALING_DN	0.00345843	0.01017187	0.27986565	1.59126171	114	up	*
HALLMARK	HALLMARK_MITOTIC_SPINDLE	0.00430365	0.01132538	0.24169839	1.49227903	187	up	*
HALLMARK	HALLMARK_P53_PATHWAY	0.00115331	0.00443579	0.45505987	-1.4903422	182	down	**
HALLMARK	HALLMARK_UV_RESPONSE_DN	0.0063868	0.015967	0.40701792	-1.4698557	135	down	*
HALLMARK	HALLMARK_IL2_STAT5_SIGNALING	0.00253489	0.00844963	0.4317077	-1.4338598	182	down	**
REACTOME	Nuclear Events (kinase and transcription factor activation)	1.35E-07	2.85E-05	0.69013246	-2.4255931	54	down	***
REACTOME	NGF-stimulated transcription	2.07E-06	0.00025325	0.62725674	-2.4108058	34	down	***
REACTOME	Activation of the pre-replicative complex	1.21E-06	0.00016557	0.64355184	2.23448667	31	up	***
REACTOME	DNA strand elongation	9.04E-07	0.0001311	0.6594444	2.22474877	30	up	***
REACTOME	Interleukin-3, Interleukin-5 and GM-CSF signaling	2.01E-05	0.00146279	0.57561026	-2.1745332	40	down	**
REACTOME	DAP12 interactions	5.61E-05	0.0031775	0.55733224	-2.1729129	34	down	**
REACTOME	Defective C1GALT1C1 causes Tn polyagglutination syndrome (TNPS)	2.37E-05	0.00157126	0.57561026	2.12940602	12	up	**
REACTOME	Cell-extracellular matrix interactions	0.00066629	0.01982641	0.47727082	-2.1096974	16	down	*
REACTOME	Leishmania infection	1.33E-09	7.71E-07	0.78818681	-2.1082288	187	down	***
REACTOME	Activation of ATR in response to replication stress	2.51E-05	0.0016169	0.57561026	2.10696801	36	up	**
REACTOME	Unwinding of DNA	4.94E-05	0.00286659	0.55733224	2.09279556	12	up	**
REACTOME	Signaling by BMP	0.00027553	0.01031466	0.49849311	-2.0891351	23	down	*
REACTOME	FOXO-mediated transcription	2.25E-05	0.00153646	0.57561026	-2.0867589	53	down	**
REACTOME	RUNX3 regulates NOTCH signaling	0.00122689	0.0289566	0.45505987	-2.083939	12	down	*
REACTOME	DNA Double-Strand Break Repair	6.49E-08	1.67E-05	0.70497572	2.07315574	120	up	***
REACTOME	Incretin synthesis, secretion, and inactivation	0.00069843	0.02026323	0.47727082	-2.0682918	11	down	*
REACTOME	Synthesis, secretion, and inactivation of Glucagon-like Peptide-1 (GLP-1)	0.00069843	0.02026323	0.47727082	-2.0682918	11	down	*
REACTOME	Kinesins	7.58E-06	0.00073264	0.61052688	2.06489404	50	up	***
REACTOME	Defective GALNT12 causes colorectal cancer 1 (CRCS1)	9.39E-05	0.0048445	0.5384341	2.06032814	12	up	**
REACTOME	FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes	0.00123512	0.0289566	0.45505987	-2.0559994	19	down	*
REACTOME	DNA Replication	1.72E-07	3.12E-05	0.69013246	2.04900338	118	up	***
REACTOME	Regulation of signaling by CBL	0.00098723	0.02633757	0.45505987	-2.0441653	20	down	*
REACTOME	IRAK4 deficiency (TLR2/4)	0.00097442	0.02633757	0.47727082	-2.0298925	10	down	*
REACTOME	Homology Directed Repair	1.79E-06	0.0002311	0.64355184	2.02929939	94	up	***
REACTOME	Scavenging of heme from plasma	6.79E-05	0.00375496	0.5384341	2.02296598	10	up	**
REACTOME	O-linked glycosylation of mucins	0.000103	0.00517635	0.5384341	1.993646	51	up	**
REACTOME	GPVI-mediated activation cascade	0.00053526	0.01701839	0.47727082	-1.975532	31	down	*
REACTOME	COPI-dependent Golgi-to-ER retrograde traffic	6.52E-06	0.00068756	0.61052688	1.97196418	88	up	***
REACTOME	Cell Cycle Checkpoints	2.77E-09	1.07E-06	0.77493903	1.96189352	233	up	***
REACTOME	DNA Replication Pre-Initiation	2.16E-05	0.00151982	0.57561026	1.96158714	77	up	**
REACTOME	Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	0.00056732	0.01779406	0.47727082	1.95890832	12	up	*
REACTOME	Platelet activation, signaling and aggregation	5.66E-08	1.64E-05	0.71951283	-1.9574309	220	down	***
REACTOME	Anti-inflammatory response favouring Leishmania parasite infection	1.49E-05	0.00128326	0.59332548	-1.9315853	111	down	**
REACTOME	Leishmania parasite growth and survival	1.49E-05	0.00128326	0.59332548	-1.9315853	111	down	**
REACTOME	HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA)	1.84E-05	0.00142717	0.57561026	1.92443859	88	up	**
REACTOME	Cell recruitment (pro-inflammatory response)	0.00119399	0.02886708	0.45505987	-1.9131797	23	down	*
REACTOME	Purinergic signaling in leishmaniasis infection	0.00119399	0.02886708	0.45505987	-1.9131797	23	down	*
REACTOME	G2/M Checkpoints	6.97E-06	0.00070355	0.61052688	1.90920141	121	up	***
REACTOME	Cell surface interactions at the vascular wall	1.79E-05	0.00142717	0.57561026	-1.9080789	105	down	**
REACTOME	Synthesis of DNA	2.02E-05	0.00146279	0.57561026	1.90445803	111	up	**
REACTOME	Deposition of new CENPA-containing nucleosomes at the centromere	0.00098485	0.02633757	0.45505987	1.89385046	22	up	*
REACTOME	Nucleosome assembly	0.00098485	0.02633757	0.45505987	1.89385046	22	up	*
REACTOME	ADORA2B mediated anti-inflammatory cytokines production	7.43E-05	0.00391948	0.5384341	-1.893381	79	down	**
REACTOME	Resolution of Abasic Sites (AP sites)	0.00060691	0.01878175	0.47727082	1.89207068	37	up	*
REACTOME	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	3.32E-05	0.00203034	0.55733224	-1.8844226	93	down	**
REACTOME	Cell Cycle	3.33E-13	7.74E-10	0.93259521	1.87569927	560	up	***
REACTOME	Homologous DNA Pairing and Strand Exchange	0.00041154	0.01384326	0.49849311	1.87256783	40	up	*
REACTOME	Chromosome Maintenance	0.00011223	0.00542683	0.5384341	1.86602503	77	up	**
REACTOME	Resolution of Sister Chromatid Cohesion	3.30E-05	0.00203034	0.55733224	1.86547463	105	up	**
REACTOME	HDR through Homologous Recombination (HRR)	0.00033003	0.011606	0.49849311	1.85740471	63	up	*
REACTOME	Resolution of D-Loop Structures	0.00106369	0.02683507	0.45505987	1.85019033	31	up	*
REACTOME	CDC6 association with the ORC:origin complex	0.0017821	0.03977167	0.45505987	1.84890144	11	up	*
REACTOME	Mitotic Prometaphase	5.17E-06	0.00057116	0.61052688	1.84206804	174	up	***
REACTOME	DNA Repair	2.63E-07	4.36E-05	0.67496286	1.83108338	260	up	***
REACTOME	Cell Cycle, Mitotic	2.57E-10	1.99E-07	0.81403584	1.82777151	452	up	***
REACTOME	G1/S-Specific Transcription	0.00140525	0.03229303	0.45505987	1.82660188	26	up	*
REACTOME	Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	0.00013728	0.00637259	0.51884808	1.82618397	84	up	**
REACTOME	Amplification of signal from the kinetochores	0.00013728	0.00637259	0.51884808	1.82618397	84	up	**
REACTOME	Resolution of D-loop Structures through Holliday Junction Intermediates	0.00113307	0.02827805	0.45505987	1.82533528	30	up	*
REACTOME	Mitochondrial translation elongation	0.00015651	0.00698583	0.51884808	1.81290592	82	up	**
REACTOME	Presynaptic phase of homologous DNA pairing and strand exchange	0.00165654	0.03732841	0.45505987	1.81172358	37	up	*
REACTOME	Processing of DNA double-strand break ends	0.00049027	0.01602712	0.47727082	1.8110077	56	up	*
REACTOME	Cilium Assembly	8.67E-06	0.00080469	0.59332548	1.80801633	171	up	***
REACTOME	Mitochondrial translation	0.00019339	0.00846918	0.51884808	1.80158903	88	up	**
REACTOME	Signaling by NTRK1 (TRKA)	0.00014127	0.00642926	0.51884808	-1.799561	104	down	**
REACTOME	FCGR3A-mediated phagocytosis	0.00106247	0.02683507	0.45505987	-1.7995403	55	down	*
REACTOME	Leishmania phagocytosis	0.00106247	0.02683507	0.45505987	-1.7995403	55	down	*
REACTOME	Parasite infection	0.00106247	0.02683507	0.45505987	-1.7995403	55	down	*
REACTOME	Golgi-to-ER retrograde transport	7.38E-05	0.00391948	0.5384341	1.79642641	121	up	**
REACTOME	Ca2+ pathway	0.00066278	0.01982641	0.47727082	-1.7951431	57	down	*
REACTOME	Neutrophil degranulation	1.66E-09	7.71E-07	0.78818681	-1.7951194	394	down	***
REACTOME	Extra-nuclear estrogen signaling	0.00081884	0.02317726	0.47727082	-1.786353	67	down	*
REACTOME	Mitochondrial translation termination	0.00024427	0.00989219	0.49849311	1.78346669	82	up	**
REACTOME	Mitochondrial translation initiation	0.0002472	0.00989219	0.49849311	1.7819652	82	up	**
REACTOME	G alpha (s) signalling events	0.00020404	0.00876984	0.51884808	-1.7707473	104	down	**
REACTOME	EML4 and NUDC in mitotic spindle formation	0.00032996	0.011606	0.49849311	1.76329106	100	up	*
REACTOME	Mitotic Spindle Checkpoint	0.00028045	0.01033202	0.49849311	1.76124987	101	up	*
REACTOME	Anchoring of the basal body to the plasma membrane	0.00046732	0.01549496	0.49849311	1.76124411	86	up	*
REACTOME	Peptide hormone metabolism	0.00195738	0.04308285	0.4317077	-1.751031	54	down	*
REACTOME	G1/S Transition	0.00025243	0.0099305	0.49849311	1.7346799	123	up	**
REACTOME	Signaling by TGFB family members	0.00034791	0.01205218	0.49849311	-1.7308303	94	down	*
REACTOME	Assembly of the pre-replicative complex	0.0016547	0.03732841	0.45505987	1.71971142	62	up	*
REACTOME	Signaling by Interleukins	1.32E-07	2.85E-05	0.69013246	-1.7021115	365	down	***
REACTOME	Mitotic G1 phase and G1/S transition	0.00028651	0.01039048	0.49849311	1.69987083	141	up	*
REACTOME	S Phase	0.00021969	0.00910525	0.51884808	1.69408643	150	up	**
REACTOME	GPCR ligand binding	2.65E-06	0.00030729	0.62725674	-1.6940602	257	down	***
REACTOME	Innate Immune System	2.82E-12	3.28E-09	0.89867123	-1.6910333	814	down	***
REACTOME	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	0.00120911	0.02893137	0.45505987	-1.6696285	88	down	*
REACTOME	Class A/1 (Rhodopsin-like receptors)	0.00010482	0.00517635	0.5384341	-1.6613907	183	down	**
REACTOME	Signaling by NTRKs	0.00071853	0.02058906	0.47727082	-1.645622	123	down	*
REACTOME	Separation of Sister Chromatids	0.0002086	0.00880293	0.51884808	1.64117651	164	up	**
REACTOME	G alpha (q) signalling events	0.00027369	0.01031466	0.49849311	-1.6336741	141	down	*
REACTOME	Signaling by GPCR	3.62E-08	1.20E-05	0.71951283	-1.6311137	536	down	***
REACTOME	M Phase	1.61E-05	0.00133695	0.57561026	1.62827778	315	up	**
REACTOME	Toll Like Receptor 9 (TLR9) Cascade	0.00241048	0.04995293	0.4317077	-1.6111875	81	down	*
REACTOME	Hemostasis	1.75E-07	3.12E-05	0.69013246	-1.5989495	488	down	***
REACTOME	RHO GTPases Activate Formins	0.00201998	0.04381652	0.4317077	1.59770504	120	up	*
REACTOME	Intra-Golgi and retrograde Golgi-to-ER traffic	0.00037493	0.01279732	0.49849311	1.59128784	184	up	*
REACTOME	Organelle biogenesis and maintenance	0.00026677	0.01031466	0.49849311	1.58863657	236	up	*
REACTOME	GPCR downstream signalling	2.98E-07	4.61E-05	0.67496286	-1.5815009	489	down	***
REACTOME	Mitotic Metaphase and Anaphase	0.00050328	0.01622373	0.47727082	1.57064193	207	up	*
REACTOME	ESR-mediated signaling	0.00101266	0.0267088	0.45505987	-1.5677657	148	down	*
REACTOME	Mitotic Anaphase	0.00090891	0.02541673	0.47727082	1.56260303	206	up	*
REACTOME	Mitotic G2-G2/M phases	0.00196759	0.04308285	0.35481951	1.53409666	176	up	*
REACTOME	Intracellular signaling by second messengers	0.00117335	0.02886708	0.45505987	-1.3869739	267	down	*
REACTOME	Cytokine Signaling in Immune system	3.51E-05	0.0020876	0.55733224	-1.3703931	710	down	**

Table S5. Pathway analysis Indolent vs Aggressive

database	pathway	pval	padj	log2err	NES	size	state	pvlabel
HALLMARK	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	9.18E-30	4.59E-28	1.4172759	-3.1252756	184	down	***
HALLMARK	HALLMARK_TNFA_SIGNALING_VIA_NFKB	3.43E-16	8.57E-15	1.03769616	-2.5234635	189	down	***
HALLMARK	HALLMARK_MYOGENESIS	1.85E-12	2.31E-11	0.89867123	-2.4106909	162	down	***
HALLMARK	HALLMARK_APICAL_JUNCTION	8.54E-13	1.42E-11	0.921426	-2.4033342	163	down	***
HALLMARK	HALLMARK_TGF_BETA_SIGNALING	1.07E-07	6.68E-07	0.70497572	-2.3688587	52	down	***
HALLMARK	HALLMARK_COAGULATION	4.56E-08	3.80E-07	0.71951283	-2.2260988	97	down	***
HALLMARK	HALLMARK_ANGIOGENESIS	1.51E-05	5.02E-05	0.59332548	-2.1844579	29	down	***
HALLMARK	HALLMARK_CHOLESTEROL_HOMEOSTASIS	8.66E-06	3.09E-05	0.59332548	-2.0948751	67	down	***
HALLMARK	HALLMARK_UV_RESPONSE_UP	1.32E-07	7.35E-07	0.69013246	-2.0916139	133	down	***
HALLMARK	HALLMARK_COMPLEMENT	1.76E-08	1.76E-07	0.73376199	-2.0873569	168	down	***
HALLMARK	HALLMARK_INFLAMMATORY_RESPONSE	7.80E-08	5.57E-07	0.70497572	-2.0063356	179	down	***
HALLMARK	HALLMARK_APOPTOSIS	7.88E-07	3.94E-06	0.6594444	-1.988896	144	down	***
HALLMARK	HALLMARK_HYPOXIA	1.65E-06	6.87E-06	0.64355184	-1.9040023	168	down	***
HALLMARK	HALLMARK_KRAS_SIGNALING_UP	2.02E-06	7.75E-06	0.62725674	-1.877361	170	down	***
HALLMARK	HALLMARK_MITOTIC_SPINDLE	9.89E-07	4.50E-06	0.64355184	-1.8755572	187	down	***
HALLMARK	HALLMARK_UV_RESPONSE_DN	3.95E-05	0.00012333	0.55733224	-1.8272038	135	down	***
HALLMARK	HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.00191833	0.00417027	0.45505987	-1.8122958	38	down	**
HALLMARK	HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.00173325	0.0039392	0.45505987	-1.769168	44	down	**
HALLMARK	HALLMARK_KRAS_SIGNALING_DN	0.00020797	0.00061168	0.51884808	1.75752017	114	up	***
HALLMARK	HALLMARK_IL6_JAK_STAT3_SIGNALING	0.00229528	0.00478183	0.4317077	-1.6949363	73	down	**
HALLMARK	HALLMARK_HEDGEHOG_SIGNALING	0.00728863	0.01349746	0.24518806	-1.6835071	28	down	*
HALLMARK	HALLMARK_XENOBIOTIC_METABOLISM	0.00029223	0.00081175	0.49849311	-1.6669577	159	down	***
HALLMARK	HALLMARK_P53_PATHWAY	0.0005545	0.0014592	0.47727082	-1.6014024	182	down	**
HALLMARK	HALLMARK_INTERFERON_ALPHA_RESPONSE	0.00374352	0.00719907	0.4317077	-1.5952132	87	down	**
HALLMARK	HALLMARK_ANDROGEN_RESPONSE	0.0034353	0.00687061	0.4317077	-1.5762568	90	down	**
HALLMARK	HALLMARK_INTERFERON_GAMMA_RESPONSE	0.00089535	0.00223837	0.47727082	-1.5186357	185	down	**
HALLMARK	HALLMARK_MTORC1_SIGNALING	0.0016116	0.00383714	0.45505987	-1.5093321	187	down	**
HALLMARK	HALLMARK_ADIPOGENESIS	0.00924001	0.01650002	0.22908938	-1.4211755	178	down	*
HALLMARK	HALLMARK_IL2_STAT5_SIGNALING	0.00973349	0.01678188	0.22347912	-1.405853	182	down	*
HALLMARK	HALLMARK_HEME_METABOLISM	0.01654032	0.0275672	0.17000428	-1.3675008	162	down	*
REACTOME	Extracellular matrix organization	9.28E-21	1.08E-17	1.17789326	-2.5838576	252	down	***
REACTOME	Smooth Muscle Contraction	3.33E-09	4.07E-07	0.77493903	-2.518818	33	down	***
REACTOME	Degradation of the extracellular matrix	2.48E-11	5.75E-09	0.86341539	-2.4368108	116	down	***
REACTOME	Post-translational protein phosphorylation	6.64E-10	1.03E-07	0.80121557	-2.4210396	78	down	***
REACTOME	Cell-extracellular matrix interactions	3.59E-07	2.88E-05	0.67496286	-2.388933	16	down	***
REACTOME	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	2.00E-09	2.73E-07	0.77493903	-2.3634061	88	down	***
REACTOME	Platelet degranulation	1.09E-09	1.58E-07	0.78818681	-2.3374584	103	down	***
REACTOME	Elastic fibre formation	3.35E-07	2.78E-05	0.67496286	-2.3251077	41	down	***
REACTOME	Collagen formation	1.50E-08	1.66E-06	0.73376199	-2.3009142	78	down	***
REACTOME	Collagen degradation	4.33E-07	3.36E-05	0.67496286	-2.2897096	54	down	***
REACTOME	Response to elevated platelet cytosolic Ca2+	5.78E-09	6.72E-07	0.7614608	-2.2857702	108	down	***
REACTOME	Neutrophil degranulation	3.26E-17	1.51E-14	1.06720999	-2.2586806	394	down	***
REACTOME	Semaphorin interactions	7.51E-07	5.29E-05	0.6594444	-2.2420796	60	down	***
REACTOME	ECM proteoglycans	4.83E-07	3.62E-05	0.6594444	-2.2354289	62	down	***
REACTOME	Integrin cell surface interactions	2.08E-07	1.85E-05	0.69013246	-2.2257978	77	down	***
REACTOME	Collagen biosynthesis and modifying enzymes	7.84E-07	5.35E-05	0.6594444	-2.2161643	58	down	***
REACTOME	Molecules associated with elastic fibres	7.75E-06	0.00047392	0.59332548	-2.1968759	34	down	***
REACTOME	RHO GTPases Activate ROCKs	3.51E-05	0.00159267	0.55733224	-2.1620137	19	down	**
REACTOME	Crosslinking of collagen fibrils	5.24E-05	0.00221382	0.55733224	-2.152575	16	down	**
REACTOME	Platelet activation, signaling and aggregation	2.88E-11	6.08E-09	0.86341539	-2.1455071	220	down	***
REACTOME	Assembly of collagen fibrils and other multimeric structures	1.08E-05	0.00064368	0.59332548	-2.133815	53	down	***
REACTOME	RHO GTPases activate CIT	6.86E-05	0.00279676	0.5384341	-2.1236945	20	down	**
REACTOME	Non-integrin membrane-ECM interactions	1.66E-05	0.00083742	0.57561026	-2.1062156	49	down	***
REACTOME	Laminin interactions	0.0001422	0.00532786	0.51884808	-2.061494	28	down	**
REACTOME	Leishmania infection	2.09E-08	2.21E-06	0.73376199	-2.0540501	187	down	***
REACTOME	RHO GTPases activate PAKs	0.00028273	0.00925031	0.49849311	-2.0169528	22	down	**
REACTOME	Innate Immune System	1.29E-19	9.95E-17	1.1421912	-2.0155004	814	down	***
REACTOME	Hemostasis	1.82E-13	6.05E-11	0.94363223	-2.0046589	488	down	***
REACTOME	Other semaphorin interactions	0.00066233	0.0158619	0.47727082	-1.9959244	18	down	*
REACTOME	Sema4D induced cell migration and growth-cone collapse	0.00043874	0.01273998	0.49849311	-1.984	20	down	*
REACTOME	Antigen processing-Cross presentation	1.36E-05	0.0007525	0.59332548	-1.9818782	81	down	***
REACTOME	Signal regulatory protein family interactions	0.00106653	0.02337313	0.45505987	-1.972806	14	down	*
REACTOME	RHO GTPases activate PKNs	0.00043176	0.01273998	0.49849311	-1.9715525	32	down	*
REACTOME	L1CAM interactions	1.87E-05	0.00092533	0.57561026	-1.9615634	96	down	***
REACTOME	Sema4D in semaphorin signaling	0.00055557	0.01452075	0.47727082	-1.9550738	24	down	*
REACTOME	Nervous system development	3.18E-12	9.22E-10	0.89867123	-1.9513299	491	down	***
REACTOME	Cell surface interactions at the vascular wall	1.62E-05	0.00083742	0.57561026	-1.948758	105	down	***
REACTOME	NR1H2 and NR1H3-mediated signaling	0.00021118	0.00743305	0.51884808	-1.9445925	39	down	**
REACTOME	Chondroitin sulfate/dermatan sulfate metabolism	0.00023921	0.00805331	0.51884808	-1.9425917	41	down	**
REACTOME	Signaling by VEGF	1.64E-05	0.00083742	0.57561026	-1.9414436	99	down	***
REACTOME	ROS and RNS production in phagocytes	0.00033667	0.01071355	0.49849311	-1.9385759	31	down	*
REACTOME	Signaling by Receptor Tyrosine Kinases	4.13E-11	7.38E-09	0.85133906	-1.934458	451	down	***
REACTOME	RHO GTPase Effectors	3.45E-08	3.49E-06	0.71951283	-1.931323	234	down	***
REACTOME	Defective C1GALT1C1 causes Tn polyagglutination syndrome (TNPS)	0.00182703	0.03478846	0.45505987	1.92937477	12	up	*
REACTOME	Signaling by Rho GTPases	2.53E-09	3.27E-07	0.77493903	-1.9293146	347	down	***
REACTOME	Axon guidance	1.58E-11	4.08E-09	0.86341539	-1.9249773	472	down	***
REACTOME	Pre-NOTCH Processing in Golgi	0.0009298	0.02117575	0.47727082	-1.9178488	17	down	*
REACTOME	Plasma lipoprotein assembly	0.00115645	0.02464624	0.45505987	-1.9146705	10	down	*
REACTOME	Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	0.00203585	0.03582789	0.4317077	1.91300704	12	up	*
REACTOME	Cell-Cell communication	5.20E-05	0.00221382	0.55733224	-1.906027	95	down	**
REACTOME	EPH-Ephrin signaling	5.39E-05	0.00223642	0.55733224	-1.9042689	81	down	**
REACTOME	Collagen chain trimerization	0.00049943	0.01352655	0.47727082	-1.8993863	38	down	*
REACTOME	Signaling by NTRK1 (TRKA)	2.17E-05	0.00105113	0.57561026	-1.8939485	104	down	**
REACTOME	Signaling by BMP	0.00132968	0.02733483	0.45505987	-1.8928437	23	down	*
REACTOME	IRAK4 deficiency (TLR2/4)	0.00166846	0.03229864	0.45505987	-1.8859702	10	down	*
REACTOME	Caspase activation via extrinsic apoptotic signalling pathway	0.00152106	0.03063498	0.45505987	-1.8803012	24	down	*
REACTOME	MET activates PTK2 signaling	0.00190173	0.03485247	0.45505987	-1.8784132	28	down	*
REACTOME	NOTCH4 Intracellular Domain Regulates Transcription	0.00210857	0.03655381	0.4317077	-1.8778881	16	down	*
REACTOME	Trafficking and processing of endosomal TLR	0.00186302	0.03485247	0.45505987	-1.8762937	10	down	*
REACTOME	Extra-nuclear estrogen signaling	0.00017462	0.0063383	0.51884808	-1.8679957	67	down	**
REACTOME	Regulation of actin dynamics for phagocytic cup formation	0.00042803	0.01273998	0.49849311	-1.8668407	55	down	*
REACTOME	Integrin signaling	0.0010598	0.02337313	0.45505987	-1.8661049	26	down	*
REACTOME	Signal Transduction	4.55E-26	1.06E-22	1.3267161	-1.8580881	1901	down	***
REACTOME	FCGR3A-mediated phagocytosis	0.00050077	0.01352655	0.47727082	-1.8527082	55	down	*
REACTOME	Leishmania phagocytosis	0.00050077	0.01352655	0.47727082	-1.8527082	55	down	*
REACTOME	Parasite infection	0.00050077	0.01352655	0.47727082	-1.8527082	55	down	*
REACTOME	MET promotes cell motility	0.00075783	0.0179637	0.47727082	-1.8496116	39	down	*
REACTOME	Signaling by NTRKs	3.66E-05	0.00160465	0.55733224	-1.8456228	123	down	**
REACTOME	RHO GTPases Activate WASPs and WAVEs	0.0009711	0.02190163	0.47727082	-1.8453884	33	down	*
REACTOME	Muscle contraction	1.16E-05	0.00067594	0.59332548	-1.8364811	149	down	***
REACTOME	Metabolism of fat-soluble vitamins	0.00130805	0.02713029	0.45505987	-1.834388	31	down	*
REACTOME	Defective B3GALTL causes Peters-plus syndrome (PpS)	0.00112857	0.02427477	0.45505987	-1.8322655	33	down	*
REACTOME	Signaling by NOTCH1	0.00042851	0.01273998	0.49849311	-1.8231585	65	down	*
REACTOME	NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux	0.00155059	0.03063498	0.45505987	-1.821391	31	down	*
REACTOME	Potential therapeutics for SARS	0.0005589	0.01452075	0.47727082	-1.8155625	71	down	*
REACTOME	Diseases associated with glycosaminoglycan metabolism	0.00195778	0.03525517	0.4317077	-1.8092494	36	down	*
REACTOME	Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants	0.00058758	0.01452075	0.47727082	-1.8076456	50	down	*
REACTOME	Constitutive Signaling by NOTCH1 PEST Domain Mutants	0.00058758	0.01452075	0.47727082	-1.8076456	50	down	*
REACTOME	Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer	0.00058758	0.01452075	0.47727082	-1.8076456	50	down	*
REACTOME	Signaling by NOTCH1 PEST Domain Mutants in Cancer	0.00058758	0.01452075	0.47727082	-1.8076456	50	down	*
REACTOME	Signaling by NOTCH1 in Cancer	0.00058758	0.01452075	0.47727082	-1.8076456	50	down	*
REACTOME	Peptide hormone metabolism	0.00112628	0.02427477	0.45505987	-1.796514	54	down	*
REACTOME	Signaling by TGFB family members	0.00018166	0.00649221	0.51884808	-1.7815595	94	down	**
REACTOME	EPHB-mediated forward signaling	0.00176685	0.03392057	0.45505987	-1.7807072	39	down	*
REACTOME	Fcgamma receptor (FCGR) dependent phagocytosis	0.00045814	0.01308856	0.49849311	-1.7780155	76	down	*
REACTOME	N-glycan trimming in the ER and Calnexin/Calreticulin cycle	0.00191663	0.03485247	0.45505987	-1.7769467	33	down	*
REACTOME	Infectious disease	7.47E-11	1.24E-08	0.83908894	-1.7730144	670	down	***
REACTOME	Anti-inflammatory response favouring Leishmania parasite infection	0.00022021	0.0075227	0.51884808	-1.7704998	111	down	**
REACTOME	Leishmania parasite growth and survival	0.00022021	0.0075227	0.51884808	-1.7704998	111	down	**
REACTOME	Signaling by Interleukins	3.31E-07	2.78E-05	0.67496286	-1.7688599	365	down	***
REACTOME	Developmental Biology	3.93E-11	7.38E-09	0.85133906	-1.764564	733	down	***
REACTOME	Immune System	2.15E-19	1.25E-16	1.1421912	-1.7604151	1697	down	***
REACTOME	Nuclear Events (kinase and transcription factor activation)	0.00155615	0.03063498	0.45505987	-1.7596944	54	down	*
REACTOME	Synthesis of substrates in N-glycan biosythesis	0.00160773	0.03138453	0.45505987	-1.7565131	54	down	*
REACTOME	Glycosaminoglycan metabolism	0.00046202	0.01308856	0.49849311	-1.7549442	105	down	*
REACTOME	Signaling by Nuclear Receptors	1.46E-05	0.00079083	0.59332548	-1.7521654	206	down	***
REACTOME	Binding and Uptake of Ligands by Scavenger Receptors	0.00283042	0.04597537	0.4317077	-1.7479678	37	down	*
REACTOME	Interleukin-4 and Interleukin-13 signaling	0.00056322	0.01452075	0.47727082	-1.7391344	89	down	*
REACTOME	Oncogenic MAPK signaling	0.0012457	0.02607002	0.45505987	-1.7328605	77	down	*
REACTOME	Clathrin-mediated endocytosis	0.00010199	0.00401567	0.5384341	-1.7314737	126	down	**
REACTOME	SARS-CoV Infections	0.00011508	0.00439857	0.5384341	-1.7285216	133	down	**
REACTOME	Regulation of PTEN gene transcription	0.00308442	0.04941449	0.4317077	-1.7284962	53	down	*
REACTOME	VEGFA-VEGFR2 Pathway	0.00039517	0.01223984	0.49849311	-1.7266565	91	down	*
REACTOME	ADORA2B mediated anti-inflammatory cytokines production	0.00081836	0.01886382	0.47727082	-1.7191098	79	down	*
REACTOME	ER-Phagosome pathway	0.0020897	0.03649912	0.4317077	-1.694822	69	down	*
REACTOME	Unfolded Protein Response (UPR)	0.00228537	0.03847033	0.4317077	-1.6889865	83	down	*
REACTOME	Signaling by NOTCH	0.00025852	0.00857908	0.49849311	-1.6825778	161	down	**
REACTOME	Signaling by NOTCH4	0.00201987	0.03581794	0.4317077	-1.6820226	73	down	*
REACTOME	Disease	1.54E-13	5.97E-11	0.94363223	-1.678673	1273	down	***
REACTOME	G alpha (12/13) signalling events	0.00215959	0.03688777	0.4317077	-1.6576341	70	down	*
REACTOME	Vesicle-mediated transport	8.68E-08	8.40E-06	0.70497572	-1.6511391	597	down	***
REACTOME	Beta-catenin independent WNT signaling	0.00043789	0.01273998	0.49849311	-1.6509538	134	down	*
REACTOME	ESR-mediated signaling	0.00080858	0.01886382	0.47727082	-1.6342273	148	down	*
REACTOME	Programmed Cell Death	0.00039045	0.01223984	0.49849311	-1.6331326	164	down	*
REACTOME	Diseases of signal transduction by growth factor receptors and second messengers	3.43E-05	0.00159267	0.55733224	-1.6203755	347	down	**
REACTOME	Toll-like Receptor Cascades	0.00066105	0.0158619	0.47727082	-1.6181034	133	down	*
REACTOME	Rho GTPase cycle	0.0018999	0.03485247	0.45505987	-1.596725	123	down	*
REACTOME	Signaling by GPCR	1.70E-06	0.00011255	0.64355184	-1.5958335	536	down	***
REACTOME	G alpha (s) signalling events	0.00199222	0.0355994	0.4317077	-1.5954721	104	down	*
REACTOME	Transcriptional regulation by RUNX2	0.00274598	0.04492197	0.4317077	-1.5951699	106	down	*
REACTOME	Cytokine Signaling in Immune system	1.49E-07	1.38E-05	0.69013246	-1.594206	710	down	***
REACTOME	Apoptosis	0.00106432	0.02337313	0.45505987	-1.5905708	154	down	*
REACTOME	GPCR downstream signalling	7.16E-06	0.00044928	0.61052688	-1.5774446	489	down	***
REACTOME	PTEN Regulation	0.00268722	0.04427245	0.4317077	-1.5756801	123	down	*
REACTOME	Membrane Trafficking	4.72E-06	0.00030485	0.61052688	-1.5487379	562	down	***
REACTOME	Cellular responses to external stimuli	8.05E-05	0.00322436	0.5384341	-1.5115655	463	down	**
REACTOME	MAPK family signaling cascades	0.00082017	0.01886382	0.47727082	-1.4991865	282	down	*
REACTOME	RAF/MAP kinase cascade	0.00154048	0.03063498	0.45505987	-1.4964892	243	down	*
REACTOME	MAPK1/MAPK3 signaling	0.00188394	0.03485247	0.45505987	-1.4851827	249	down	*
REACTOME	GPCR ligand binding	0.00265749	0.0440953	0.4317077	-1.4597466	257	down	*
REACTOME	Cellular responses to stress	0.00031655	0.01021308	0.49849311	-1.4489898	455	down	*
REACTOME	Asparagine N-linked glycosylation	0.00239451	0.04001764	0.4317077	-1.4407313	268	down	*
REACTOME	Neuronal System	0.00192041	0.03485247	0.45505987	-1.4402002	293	down	*
REACTOME	Metabolism of proteins	5.88E-07	4.27E-05	0.6594444	-1.4167175	1616	down	***
REACTOME	Post-translational protein modification	1.25E-05	0.00071044	0.59332548	-1.4057764	1147	down	***
REACTOME	Transport of small molecules	0.00212986	0.03664937	0.4317077	-1.3522642	561	down	*
REACTOME	Adaptive Immune System	0.00284996	0.04597537	0.35481951	-1.3252782	637	down	*

Table S6. Pathway analysis Indolent vs Intermediate

database	pathway	pval	padj	log2err	NES	size	state	pvlabel
HALLMARK	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	7.85E-32	3.92E-30	1.46752398	2.96464727	184	up	***
HALLMARK	HALLMARK_G2M_CHECKPOINT	8.14E-24	2.04E-22	1.26273989	2.75793334	181	up	***
HALLMARK	HALLMARK_E2F_TARGETS	2.63E-21	4.38E-20	1.19534448	2.6413625	182	up	***
HALLMARK	HALLMARK_GLYCOLYSIS	1.84E-14	2.30E-13	0.97599468	2.37906185	172	up	***
HALLMARK	HALLMARK_APICAL_JUNCTION	2.65E-14	2.65E-13	0.97599468	2.36438684	163	up	***
HALLMARK	HALLMARK_MITOTIC_SPINDLE	6.05E-13	5.04E-12	0.921426	2.26840735	187	up	***
HALLMARK	HALLMARK_HYPOXIA	3.19E-10	2.28E-09	0.81403584	2.17106762	168	up	***
HALLMARK	HALLMARK_MYOGENESIS	1.02E-09	6.40E-09	0.78818681	2.12772364	162	up	***
HALLMARK	HALLMARK_ANGIOGENESIS	3.26E-05	0.00013568	0.55733224	2.09364182	29	up	***
HALLMARK	HALLMARK_MTORC1_SIGNALING	2.01E-08	1.12E-07	0.73376199	2.00440556	187	up	***
HALLMARK	HALLMARK_DNA_REPAIR	7.00E-07	3.50E-06	0.6594444	1.93899792	134	up	***
HALLMARK	HALLMARK_COAGULATION	1.18E-05	5.38E-05	0.59332548	1.91598583	97	up	***
HALLMARK	HALLMARK_SPERMATOGENESIS	0.00029827	0.00099425	0.49849311	1.79579906	87	up	***
HALLMARK	HALLMARK_ALLOGRAFT_REJECTION	9.06E-05	0.00034842	0.5384341	-1.6917798	156	down	***
HALLMARK	HALLMARK_MYC_TARGETS_V1	0.00018124	0.00064727	0.51884808	1.66622261	183	up	***
HALLMARK	HALLMARK_MYC_TARGETS_V2	0.01216924	0.03053655	0.15535473	1.57911431	56	up	*
HALLMARK	HALLMARK_HEDGEHOG_SIGNALING	0.01898148	0.04126409	0.12954747	1.57727787	28	up	*
HALLMARK	HALLMARK_XENOBIOTIC_METABOLISM	0.00307844	0.00962012	0.29109092	1.54628242	159	up	**
HALLMARK	HALLMARK_INTERFERON_ALPHA_RESPONSE	0.0090812	0.02522555	0.17500402	1.53404824	87	up	*
HALLMARK	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.01228437	0.03053655	0.14850014	1.49434909	102	up	*
HALLMARK	HALLMARK_UV_RESPONSE_UP	0.01282535	0.03053655	0.14247037	1.45846464	133	up	*
HALLMARK	HALLMARK_ESTROGEN_RESPONSE_LATE	0.00896752	0.02522555	0.16765853	1.45834726	174	up	*
HALLMARK	HALLMARK_KRAS_SIGNALING_UP	0.01387368	0.03153108	0.13464697	1.42511621	170	up	*
REACTOME	Collagen formation	2.96E-17	2.30E-14	1.06720999	2.72546534	78	up	***
REACTOME	Collagen biosynthesis and modifying enzymes	8.39E-14	3.25E-11	0.95454163	2.64806459	58	up	***
REACTOME	Assembly of collagen fibrils and other multimeric structures	2.63E-12	7.65E-10	0.89867123	2.54397651	53	up	***
REACTOME	Collagen chain trimerization	1.52E-10	3.21E-08	0.8266573	2.48494283	38	up	***
REACTOME	Collagen degradation	8.12E-10	1.45E-07	0.80121557	2.41935313	54	up	***
REACTOME	Extracellular matrix organization	7.89E-21	9.17E-18	1.17789326	2.40670989	252	up	***
REACTOME	Degradation of the extracellular matrix	1.90E-12	6.31E-10	0.89867123	2.34826544	116	up	***
REACTOME	ECM proteoglycans	3.77E-09	4.61E-07	0.7614608	2.33366953	62	up	***
REACTOME	Post-translational protein phosphorylation	1.01E-09	1.68E-07	0.78818681	2.31452917	78	up	***
REACTOME	MET activates PTK2 signaling	2.66E-07	1.77E-05	0.67496286	2.29515363	28	up	***
REACTOME	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	2.04E-09	2.94E-07	0.77493903	2.27703604	88	up	***
REACTOME	Integrin cell surface interactions	6.57E-09	7.27E-07	0.7614608	2.26816878	77	up	***
REACTOME	Non-integrin membrane-ECM interactions	1.53E-07	1.11E-05	0.69013246	2.25301281	49	up	***
REACTOME	G2/M Checkpoints	6.52E-11	1.68E-08	0.83908894	2.2494859	121	up	***
REACTOME	Activation of ATR in response to replication stress	8.00E-07	4.65E-05	0.6594444	2.2374103	36	up	***
REACTOME	DNA Replication	1.45E-10	3.21E-08	0.8266573	2.22565686	118	up	***
REACTOME	Activation of the pre-replicative complex	1.18E-06	6.36E-05	0.64355184	2.21255291	31	up	***
REACTOME	DNA strand elongation	1.15E-06	6.35E-05	0.64355184	2.21186884	30	up	***
REACTOME	Cell Cycle Checkpoints	2.07E-15	9.63E-13	1.00731796	2.21184814	233	up	***
REACTOME	Translocation of ZAP-70 to Immunological synapse	0.00029363	0.00688993	0.49849311	-2.1817147	10	down	**
REACTOME	Homology Directed Repair	2.69E-08	2.41E-06	0.73376199	2.17752243	94	up	***
REACTOME	Synthesis of DNA	8.02E-09	8.46E-07	0.74773966	2.16788417	111	up	***
REACTOME	Unwinding of DNA	9.92E-06	0.00041808	0.59332548	2.15105225	12	up	***
REACTOME	Crosslinking of collagen fibrils	7.55E-06	0.00032798	0.61052688	2.14820394	16	up	***
REACTOME	Generation of second messenger molecules	0.00019511	0.00492665	0.51884808	-2.1385729	22	down	**
REACTOME	MET promotes cell motility	2.94E-06	0.00014532	0.62725674	2.12885125	39	up	***
REACTOME	Phosphorylation of CD3 and TCR zeta chains	0.00072505	0.01439555	0.47727082	-2.1271819	13	down	*
REACTOME	Syndecan interactions	1.09E-05	0.00042809	0.59332548	2.11716413	24	up	***
REACTOME	DNA Replication Pre-Initiation	5.41E-07	3.31E-05	0.6594444	2.11189343	77	up	***
REACTOME	DNA Double-Strand Break Repair	1.95E-08	1.97E-06	0.73376199	2.10507761	120	up	***
REACTOME	Laminin interactions	1.76E-05	0.0006375	0.57561026	2.09531526	28	up	***
REACTOME	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	2.15E-06	0.00010868	0.62725674	-2.0902903	93	down	***
REACTOME	HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA)	9.18E-07	5.20E-05	0.6594444	2.07567528	88	up	***
REACTOME	NCAM1 interactions	1.23E-05	0.00046741	0.59332548	2.07492209	33	up	***
REACTOME	Binding and Uptake of Ligands by Scavenger Receptors	1.53E-05	0.0005624	0.59332548	2.06400311	37	up	***
REACTOME	Resolution of D-Loop Structures	2.26E-05	0.00075856	0.57561026	2.0636425	31	up	***
REACTOME	Cell Cycle	1.68E-21	3.89E-18	1.20397524	2.05312697	560	up	***
REACTOME	Resolution of D-loop Structures through Synthesis-Dependent Strand Annealing (SDSA)	3.51E-05	0.00111792	0.55733224	2.0352638	25	up	**
REACTOME	Resolution of D-loop Structures through Holliday Junction Intermediates	3.50E-05	0.00111792	0.55733224	2.03304206	30	up	**
REACTOME	HDR through Homologous Recombination (HRR)	4.25E-06	0.00019768	0.61052688	2.03291587	63	up	***
REACTOME	Presynaptic phase of homologous DNA pairing and strand exchange	3.07E-05	0.00100299	0.55733224	2.03053447	37	up	**
REACTOME	Cell Cycle, Mitotic	4.24E-17	2.46E-14	1.06720999	2.02519999	452	up	***
REACTOME	Homologous DNA Pairing and Strand Exchange	5.15E-05	0.00153379	0.55733224	2.01566225	40	up	**
REACTOME	Diseases of glycosylation	3.13E-07	2.02E-05	0.67496286	2.00869184	122	up	***
REACTOME	Mitotic Prometaphase	3.17E-08	2.73E-06	0.71951283	1.99229043	174	up	***
REACTOME	Chondroitin sulfate biosynthesis	0.00025775	0.00618287	0.49849311	1.97834361	16	up	**
REACTOME	S Phase	1.75E-07	1.23E-05	0.69013246	1.97016902	150	up	***
REACTOME	Kinesins	5.38E-05	0.00158055	0.55733224	1.96133665	50	up	**
REACTOME	Assembly of the pre-replicative complex	4.66E-05	0.00142316	0.55733224	1.96011854	62	up	**
REACTOME	Separation of Sister Chromatids	1.41E-07	1.05E-05	0.69013246	1.95845125	164	up	***
REACTOME	Costimulation by the CD28 family	7.33E-05	0.00202759	0.5384341	-1.9572152	57	down	**
REACTOME	Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins	2.29E-05	0.00075856	0.57561026	1.95690616	70	up	***
REACTOME	Processing of DNA double-strand break ends	3.95E-05	0.00123949	0.55733224	1.95547655	56	up	**
REACTOME	Resolution of Sister Chromatid Cohesion	3.20E-06	0.00015488	0.62725674	1.95361547	105	up	***
REACTOME	Switching of origins to a post-replicative state	1.06E-05	0.00042809	0.59332548	1.94955295	84	up	***
REACTOME	Diseases associated with O-glycosylation of proteins	4.96E-05	0.00149756	0.55733224	1.94943526	55	up	**
REACTOME	G1/S-Specific Transcription	0.00025817	0.00618287	0.49849311	1.94914861	26	up	**
REACTOME	APC/C-mediated degradation of cell cycle proteins	1.97E-05	0.00067196	0.57561026	1.94731723	81	up	***
REACTOME	Regulation of mitotic cell cycle	1.97E-05	0.00067196	0.57561026	1.94731723	81	up	***
REACTOME	Polo-like kinase mediated events	0.00055146	0.01133662	0.47727082	1.93758979	16	up	*
REACTOME	Mitotic Spindle Checkpoint	1.25E-05	0.00046741	0.59332548	1.93522836	101	up	***
REACTOME	Smooth Muscle Contraction	0.00015343	0.00396009	0.51884808	1.93440599	33	up	**
REACTOME	Signaling by MET	1.82E-05	0.0006422	0.57561026	1.93398525	73	up	***
REACTOME	G1/S Transition	1.36E-06	7.19E-05	0.64355184	1.92462139	123	up	***
REACTOME	Mitotic Metaphase and Anaphase	2.70E-08	2.41E-06	0.73376199	1.92306695	207	up	***
REACTOME	L1CAM interactions	1.01E-05	0.00041808	0.59332548	1.92139157	96	up	***
REACTOME	APC/C:Cdc20 mediated degradation of mitotic proteins	6.82E-05	0.00190901	0.5384341	1.91251566	69	up	**
REACTOME	Mitotic Anaphase	3.91E-08	3.24E-06	0.71951283	1.91057245	206	up	***
REACTOME	Protein-protein interactions at synapses	6.60E-05	0.0018702	0.5384341	1.89855806	64	up	**
REACTOME	RHO GTPase Effectors	4.19E-08	3.36E-06	0.71951283	1.88229733	234	up	***
REACTOME	Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	5.83E-05	0.00167144	0.55733224	1.87849122	84	up	**
REACTOME	Amplification of signal from the kinetochores	5.83E-05	0.00167144	0.55733224	1.87849122	84	up	**
REACTOME	Regulation of APC/C activators between G1/S and early anaphase	8.72E-05	0.00238415	0.5384341	1.87697881	74	up	**
REACTOME	RHO GTPases Activate Formins	1.17E-05	0.00045294	0.59332548	1.87666738	120	up	***
REACTOME	Retrograde neurotrophin signalling	0.00117469	0.0214866	0.45505987	1.87440863	12	up	*
REACTOME	Anchoring fibril formation	0.00104554	0.01958709	0.45505987	1.87374501	13	up	*
REACTOME	M Phase	4.64E-10	8.98E-08	0.81403584	1.87340485	315	up	***
REACTOME	Phosphorylation of the APC/C	0.00075522	0.01474267	0.47727082	1.87174401	20	up	*
REACTOME	COPI-dependent Golgi-to-ER retrograde traffic	9.21E-05	0.00248656	0.5384341	1.86292783	88	up	**
REACTOME	Condensation of Prometaphase Chromosomes	0.00131402	0.02384737	0.45505987	1.86151047	10	up	*
REACTOME	Elastic fibre formation	0.00040568	0.00897511	0.49849311	1.85948316	41	up	**
REACTOME	Mitotic G1 phase and G1/S transition	6.96E-06	0.0003109	0.61052688	1.85723691	141	up	***
REACTOME	APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	0.00010103	0.00269759	0.5384341	1.8518421	68	up	**
REACTOME	Diseases associated with glycosaminoglycan metabolism	0.00108071	0.01999687	0.45505987	1.84922082	36	up	*
REACTOME	O-glycosylation of TSR domain-containing proteins	0.00041358	0.00906375	0.49849311	1.8464064	34	up	**
REACTOME	Orc1 removal from chromatin	0.00016509	0.00421431	0.51884808	1.84554391	64	up	**
REACTOME	Chondroitin sulfate/dermatan sulfate metabolism	0.00053789	0.01115651	0.47727082	1.84391226	41	up	*
REACTOME	Defective B3GALTL causes Peters-plus syndrome (PpS)	0.00062595	0.01275501	0.47727082	1.84016555	33	up	*
REACTOME	APC:Cdc20 mediated degradation of cell cycle proteins prior to satisfation of the cell cycle checkpoint	0.00030687	0.00705798	0.49849311	1.83895395	67	up	**
REACTOME	DNA Repair	1.14E-07	8.85E-06	0.70497572	1.83638718	260	up	***
REACTOME	Mitotic G2-G2/M phases	1.68E-06	8.66E-05	0.64355184	1.83630309	176	up	***
REACTOME	Receptor-type tyrosine-protein phosphatases	0.00184155	0.03216488	0.45505987	1.83524473	14	up	*
REACTOME	Cell junction organization	0.00046619	0.01006489	0.49849311	1.83086384	66	up	*
REACTOME	G2/M Transition	4.13E-06	0.00019578	0.61052688	1.82673007	174	up	***
REACTOME	Dissolution of Fibrin Clot	0.0022413	0.03745706	0.4317077	1.82406189	10	up	*
REACTOME	EML4 and NUDC in mitotic spindle formation	4.51E-05	0.00139814	0.55733224	1.82359413	100	up	**
REACTOME	G2/M DNA damage checkpoint	0.00089332	0.01700978	0.47727082	1.82341382	54	up	*
REACTOME	Signaling by PDGF	0.00049695	0.0105909	0.47727082	1.81705658	53	up	*
REACTOME	FCERI mediated Ca+2 mobilization	0.00323428	0.04702775	0.4317077	-1.8157686	32	down	*
REACTOME	Deposition of new CENPA-containing nucleosomes at the centromere	0.00254165	0.03858996	0.34452129	1.80793158	22	up	*
REACTOME	Nucleosome assembly	0.00254165	0.03858996	0.34452129	1.80793158	22	up	*
REACTOME	Signaling by Rho GTPases	4.20E-09	4.88E-07	0.7614608	1.80627341	347	up	***
REACTOME	EPH-Ephrin signaling	0.00024433	0.006103	0.49849311	1.80398443	81	up	**
REACTOME	Synaptic adhesion-like molecules	0.00231228	0.03768318	0.4317077	1.80162303	17	up	*
REACTOME	Cdc20:Phospho-APC/C mediated degradation of Cyclin A	0.00068713	0.01376035	0.47727082	1.79955163	66	up	*
REACTOME	TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest	0.00253656	0.03858996	0.35481951	1.79713333	13	up	*
REACTOME	Inactivation of APC/C via direct inhibition of the APC/C complex	0.0024189	0.03771205	0.35481951	1.79398478	21	up	*
REACTOME	Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components	0.0024189	0.03771205	0.35481951	1.79398478	21	up	*
REACTOME	Plasma lipoprotein remodeling	0.0030356	0.04520325	0.4317077	1.78615245	17	up	*
REACTOME	Antigen processing-Cross presentation	0.00033744	0.0076105	0.49849311	1.78531045	81	up	**
REACTOME	Centrosome maturation	0.00025613	0.00618287	0.49849311	1.78525776	73	up	**
REACTOME	Recruitment of mitotic centrosome proteins and complexes	0.00025613	0.00618287	0.49849311	1.78525776	73	up	**
REACTOME	Defective GALNT12 causes colorectal cancer 1 (CRCS1)	0.0033389	0.04817553	0.31077692	1.78336331	12	up	*
REACTOME	AURKA Activation by TPX2	0.00046793	0.01006489	0.49849311	1.77249201	64	up	*
REACTOME	Resolution of Abasic Sites (AP sites)	0.00143293	0.02580391	0.45505987	1.77048599	37	up	*
REACTOME	CDK-mediated phosphorylation and removal of Cdc6	0.00108463	0.01999687	0.45505987	1.77000107	66	up	*
REACTOME	Platelet degranulation	0.00014748	0.00384946	0.51884808	1.76555422	103	up	**
REACTOME	Diseases of metabolism	4.89E-06	0.00022253	0.61052688	1.76071603	204	up	***
REACTOME	A tetrasaccharide linker sequence is required for GAG synthesis	0.00353008	0.0493998	0.29109092	1.75973012	22	up	*
REACTOME	Recruitment of NuMA to mitotic centrosomes	0.00051851	0.01094993	0.47727082	1.75235933	81	up	*
REACTOME	WNT5A-dependent internalization of FZD4	0.0032391	0.04702775	0.31077692	1.75093469	15	up	*
REACTOME	MHC class II antigen presentation	0.00053229	0.01113975	0.47727082	1.74812809	98	up	*
REACTOME	Axon guidance	1.17E-09	1.81E-07	0.78818681	1.7470376	472	up	***
REACTOME	Plasma lipoprotein assembly, remodeling, and clearance	0.00212812	0.03582336	0.4317077	1.74463263	52	up	*
REACTOME	Chromosome Maintenance	0.00081447	0.01563657	0.47727082	1.74208258	77	up	*
REACTOME	Response to elevated platelet cytosolic Ca2+	0.00034735	0.00775852	0.49849311	1.74060717	108	up	**
REACTOME	Neurexins and neuroligins	0.00249082	0.03857455	0.33506856	1.73848857	42	up	*
REACTOME	Resolution of AP sites via the multiple-nucleotide patch replacement pathway	0.00338362	0.04851941	0.29723292	1.73391539	24	up	*
REACTOME	Nervous system development	2.15E-09	2.94E-07	0.77493903	1.72158647	491	up	***
REACTOME	Golgi-to-ER retrograde transport	0.00032047	0.00729848	0.49849311	1.71545853	121	up	**
REACTOME	Autodegradation of Cdh1 by Cdh1:APC/C	0.00231971	0.03768318	0.4317077	1.7148864	60	up	*
REACTOME	NCAM signaling for neurite out-growth	0.00229563	0.03768318	0.34452129	1.71221178	53	up	*
REACTOME	Regulation of TP53 Activity through Phosphorylation	0.00078764	0.01524747	0.47727082	1.71032456	82	up	*
REACTOME	APC/C:Cdc20 mediated degradation of Securin	0.00168784	0.02970337	0.45505987	1.70723816	62	up	*
REACTOME	The role of GTSE1 in G2/M progression after G2 checkpoint	0.00188768	0.03272448	0.45505987	1.70632025	65	up	*
REACTOME	ABC transporter disorders	0.00211649	0.03582336	0.4317077	1.69930069	65	up	*
REACTOME	O-linked glycosylation	0.00065785	0.01328849	0.47727082	1.69786686	92	up	*
REACTOME	Loss of Nlp from mitotic centrosomes	0.00238142	0.03771205	0.4317077	1.68297167	61	up	*
REACTOME	Loss of proteins required for interphase microtubule organization from the centrosome	0.00238142	0.03771205	0.4317077	1.68297167	61	up	*
REACTOME	Regulation of PLK1 Activity at G2/M Transition	0.00240964	0.03771205	0.32635161	1.65035813	79	up	*
REACTOME	Cell-Cell communication	0.00200779	0.03454895	0.35481951	1.64430863	95	up	*
REACTOME	PCP/CE pathway	0.00227736	0.03768318	0.33506856	1.62856264	85	up	*
REACTOME	Mitochondrial translation initiation	0.00348767	0.04910221	0.26984231	1.60656287	82	up	*
REACTOME	Intra-Golgi and retrograde Golgi-to-ER traffic	0.00029131	0.00688993	0.49849311	1.60654029	184	up	**
REACTOME	Metabolism of carbohydrates	0.00030196	0.00701451	0.49849311	1.54789275	241	up	**
REACTOME	Developmental Biology	2.48E-07	1.70E-05	0.67496286	1.52419154	733	up	***
REACTOME	Signaling by Receptor Tyrosine Kinases	1.80E-05	0.0006422	0.57561026	1.51408645	451	up	***
REACTOME	Vesicle-mediated transport	7.62E-06	0.00032798	0.61052688	1.502957	597	up	***
REACTOME	Post-translational protein modification	3.96E-07	2.49E-05	0.67496286	1.42526866	1147	up	***
REACTOME	Transcriptional Regulation by TP53	0.00300487	0.04515546	0.26984231	1.41410483	323	up	*
REACTOME	Signal Transduction	3.57E-09	4.60E-07	0.77493903	1.40068291	1901	up	***
REACTOME	Metabolism of proteins	2.48E-08	2.40E-06	0.73376199	1.39415494	1616	up	***
REACTOME	Disease	7.54E-07	4.49E-05	0.6594444	1.38725974	1273	up	***
REACTOME	Membrane Trafficking	0.00074406	0.01464787	0.47727082	1.38349814	562	up	*
REACTOME	Infectious disease	0.00301296	0.04515546	0.26521689	1.30890624	670	up	*
REACTOME	Metabolism	0.00309969	0.04586357	0.26082057	1.20509032	1675	up	*

Table S7. Pathway analysis Intermediate vs Aggressive

X	Y	r	p-adjust	Feature description
SILA	ANTPOST_LENGTH_MM	0.425	0.007	A measure of the anterior-posterior distance.
SILA	AUTO_LARGEST_PLANAR_DIAMETER_MM	0.447	0.002	A measure of the longest straight line that can fit entirely inside an XY-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters, computed by the program.
SILA	AUTO_LARGEST_PLANAR_ORTHO_DIAMETER_MM	0.401	0.020	A measure of the longest orthogonal line to the longest planar line, that can fit entirely inside an XY-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters, computed by the program.
SILA	AUTO_CORONAL_LONG_AXIS_MM	0.379	0.049	A measure of the longest straight line that can fit entirely inside an XZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters, computed by the program.
SILA	AUTO_CORONAL_SHORT_AXIS_MM	0.426	0.007	A measure of the longest orthogonal line to the longest planar line, that can fit entirely inside an XZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters, computed by the program.
SILA	AUTO_SAGITTAL_LONG_AXIS_MM	0.409	0.014	A measure of the longest straight line that can fit entirely inside an YZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters, computed by the program.
SILA	AUTO_SAGITTAL_SHORT_AXIS_MM	0.453	0.002	A measure of the longest orthogonal line to the longest planar line, that can fit entirely inside an YZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters, computed by the program.
SILA	AVG_AXIAL_DIAMETER_MM	0.434	0.004	The average of largest axial planar and orthogonal diameters, in millimeters
SILA	AVG_CORONAL_DIAMETER_MM	0.401	0.020	The average of largest coronal planar and orthogonal diameters, in millimeters
SILA	AVG_DENSITY_OF_SOLID_REGION	0.888	0.000	The average density of voxels identified as Solid (-450HU <= voxel < 1050).
SILA	AVG_SAGITTAL_DIAMETER_MM	0.442	0.003	The average of largest sagittal planar and orthogonal diameters, in millimeters
SILA	LARGEST_PLANAR_DIAMETER_MM	0.447	0.002	A measure of the longest straight line that can fit entirely inside an XY-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters.
SILA	LARGEST_PLANAR_ORTHO_DIAMETER_MM	0.401	0.020	A measure of the longest orthogonal line to the longest planar line, that can fit entirely inside an XY-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters
SILA	COMPACTNESS1_MM	0.444	0.003	Dimensionfull measure of compactness of ROI, independent of scale and orientation (first of three implementations), using standard unit shape-derived information.
SILA	CORONAL_LONG_AXIS_MM	0.379	0.049	A measure of the longest straight line that can fit entirely inside an XZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters.
SILA	CORONAL_SHORT_AXIS_MM	0.426	0.007	A measure of the longest orthogonal line to the longest planar line, that can fit entirely inside an XZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters
SILA	CRANIALCAUDAL_LENGTH_MM	0.427	0.006	A measure of the cranial-caudal distance.
SILA	ENERGY_VOXELS	0.619	0.000	A measure of the magnitude of raw voxel values in an image. A greater amount of larger values implies a greater sum of the squares of these values.
SILA	FOOTPRINT_Y_MM	0.401	0.020	The Y dimensions of the bounding box of the ROI, in millimeters.
SILA	GLCM_COL_MEAN	0.838	0.000	Average column means of GLCM in all 26 directions.
SILA	GLCM_ENTROPY	0.380	0.048	Average entropies of GLCM in all 26 directions.
SILA	GLCM_HOMOGENEITY	0.504	0.000	Average homogeneities of GLCM in all 26 directions.
SILA	GLCM_ROW_MEAN	0.807	0.000	Average row means of GLCM in all 26 directions.
SILA	KURTOSIS_HU	0.493	0.000	A measure of the 'peakedness' of the distribution of HU values in the image ROI. A higher kurtosis implies that the mass of the distribution is concentrated towards the tail(s) rather than towards the mean. A lower kurtosis implies the reverse, that the mass of the distribution is concentrated towards a spike the mean.
SILA	KURTOSIS_VOXELS	0.493	0.000	A measure of the 'peakedness' of the distribution of raw voxel values in the image ROI. A higher kurtosis implies that the mass of the distribution is concentrated towards the tail(s) rather than towards the mean. A lower kurtosis implies the reverse, that the mass of the distribution is concentrated towards a spike the mean.
SILA	L1_DISTANCE_MM	0.388	0.035	The length of the long (L1) full principal axis, in millimeters, from edge to edge of the ROI.
SILA	L2_DISTANCE_MM	0.410	0.013	The length of the short (L2) full principal axis, in millimeters, from edge to edge of the ROI.
SILA	L3_DISTANCE_MM	0.467	0.001	The length of the normal (L3) full principal axis, in millimeters, from edge to edge of the ROI.
SILA	PART_SOLID_DIAMETER_MM	0.571	0.000	The average diameter of the solid portions of a part-solid lesion.
SILA	LESION_TYPE	0.490	0.000	The density classification of the lesion. A value of: 3 == SOLID, 2 == PART_SOLID, 1 == GGO.
SILA	LUNG_RADS_DIAMETER_MM	0.434	0.004	The average of largest planar and largest planar orthogonal diameters, in millimeters
SILA	LUNG_RADS_ISOLATION	0.423	0.007	The Lung-RADS estimate for this structure isolating the study from its priors (treating the current study as a baseline scan). NOTE: This metric ranges from 0 to 5, corresponding respectively to a Lung-RADS score of 0, 1, 2, 3, 4A, and 4B.)
SILA	LUNG_RADS	0.423	0.007	The Lung-RADS estimate taking priors into account. NOTE: This metric ranges from 0 to 5, corresponding respectively to a Lung-RADS score of 0, 1, 2, 3, 4A, and 4B.)
SILA	MAX_HU	0.494	0.000	The maximum of the HU values within the image ROI.
SILA	MAX_VOXELS	0.517	0.000	The maximum raw voxel values within the image ROI.
SILA	MEAN_HU	0.801	0.000	The mean of the HU values within the image ROI.
SILA	MEAN_VOXELS	0.797	0.000	The mean of the raw voxel values within the image ROI.
SILA	MEDIAN_HU	0.861	0.000	The median of the HU values within the image ROI.
SILA	MEDIAN_VOXELS	0.858	0.000	The median of the raw voxel values within the image ROI.
SILA	NORMALIZED_ABOVE_MEAN_DEVIATION_VOXELS	0.463	0.001	Another uniformity measurement.
SILA	PERCENT_GGO	-0.609	0.000	The estimated percent ground glass density of this ROI.
SILA	PERCENT_SOLID	0.609	0.000	The estimated percent solid density of this ROI.
SILA	PERCENT_SOLID_INCL_AIR	0.592	0.000	The estimated percent solid density of this ROI including AIR in structure as part of volume.
SILA	ROOT_MEAN_SQUARE	-0.651	0.000	The square-root of the mean of the squares of the HU values in the image ROI. It is another measure of the magnitude of the image values.
SILA	ROOT_MEAN_SQUARE_VOXELS	0.828	0.000	The square-root of the mean of the squares of the raw voxel values in the image ROI. It is another measure of the magnitude of the image values.
SILA	SAGITTAL_LONG_AXIS_MM	0.409	0.014	A measure of the longest straight line that can fit entirely inside an YZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters.
SILA	SAGITTAL_SHORT_AXIS_MM	0.453	0.002	A measure of the longest orthogonal line to the longest planar line, that can fit entirely inside an YZ-planar slice of the 3D structure (from edge to edge, without ever leaving structure), in millimeters
SILA	SKEWNESS_HU	-0.815	0.000	Measures the asymmetry of the distribution of HU values in the image ROI about the mean of the values. Depending on where the tail is elongated and the mass of the distribution is concentrated, this value can be positive or negative.
SILA	SKEWNESS_VOXELS	-0.815	0.000	Measures the asymmetry of the distribution of raw voxel values in the image ROI about the mean of the values. Depending on where the tail is elongated and the mass of the distribution is concentrated, this value can be positive or negative.
SILA	SOLID_VOLUME_ML	0.602	0.000	Volume of the solid density of the specified ROI in milliliters.
SILA	SOLID_VOLUME_MM3	0.602	0.000	Volume of the solid density of the specified ROI in cubic millimeters.
SILA	SOLID_VOLUME_VOXELS	0.599	0.000	Volume of the solid density of the specified ROI in voxels.
SILA	SURFACE_AREA_MM2	0.431	0.005	Surface area of the specified ROI of the image in square millimeters.
SILA	SURFACE_AREA_TO_VOLUME_RATIO_MM	-0.451	0.002	Ratio of surface area to volume, in standard units.
SILA	TRANSVERSE_LENGTH_MM	0.403	0.019	A measure of the transverse distance.
SILA	UNIFORMITY_HU	0.678	0.000	A measure of the sum of the squares of each discrete HU value in the image ROI. This is a measure of the heterogeneity of an image, where a greater uniformity implies a greater heterogeneity or a greater range of discrete image values.
SILA	UNIFORMITY_ACR	0.462	0.001	A uniformity measurement as defined by the American College of Radiology.
SILA	VOLUME_ML	0.433	0.005	Volume of the specified ROI of the image in milliliters.
SILA	VOLUME_MM3	0.433	0.005	Volume of the specified ROI of the image in cubic millimeters.
SILA	VOLUME_VOXELS	0.435	0.004	Volume derived from voxel count inside ROI
SILA	VOLUMETRIC_LENGTH_MM	0.385	0.039	A measure of the longest straight line that can fit entirely inside the 3D structure (from edge to edge, without ever leaving structure).

Table S8. Pairwise Spearman correlation between SILA score and HealthMyne Radiomics Features

Pt ID	Cluster
11938	1
13376	2
13436	3
8356	2
12994	2
12929	4
12924	1
13622	3
13771	1
13651	2
13074	2
11817	4
13536	2
11906	2
13276	2
13207	4
13317	1
12915	1
13769	1
11855	1
11851	2
11538	2
12889	4
12931	4
11813	1
11646	2
11759	2
13014	1
14855	3
11952	2
11561	1
11886	2
13724	4
14958	2
12281	2
12323	1
14955	1
15001	1
14048	3
15224	2
14965	2
15325	1
14962	2
15187	1
15506	3
14301	3
13538	2
15326	1
15569	4
14610	3
13988	3
13155	4
15083	1
11652	1
15002	4
12546	1
12890	1
15467	1
15741	2

Table S9. Data integration patient clusters

Feature name	Dataset of origin	Cluster
ECC_3	CyTOF	I
ECC_5	CyTOF	II
fmes_3	CyTOF	I
OtherI_4	CyTOF	I
HLA DR	CyTOF	I
ANTPOST_LENGTH_MM	HealthMyne (radiomics)	II
AUTO_LARGEST_PLANAR_DIAMETER_MM	HealthMyne (radiomics)	II
AUTO_LARGEST_PLANAR_ORTHO_DIAMETER_MM	HealthMyne (radiomics)	II
AUTO_CORONAL_LONG_AXIS_MM	HealthMyne (radiomics)	II
AUTO_CORONAL_SHORT_AXIS_MM	HealthMyne (radiomics)	II
AUTO_SAGITTAL_LONG_AXIS_MM	HealthMyne (radiomics)	II
AUTO_SAGITTAL_SHORT_AXIS_MM	HealthMyne (radiomics)	II
AVG_AXIAL_DIAMETER_MM	HealthMyne (radiomics)	II
AVG_CORONAL_DIAMETER_MM	HealthMyne (radiomics)	II
AVG_DENSITY_OF_SOLID_REGION	HealthMyne (radiomics)	II
AVG_SAGITTAL_DIAMETER_MM	HealthMyne (radiomics)	II
LARGEST_PLANAR_DIAMETER_MM	HealthMyne (radiomics)	II
LARGEST_PLANAR_ORTHO_DIAMETER_MM	HealthMyne (radiomics)	II
COMPACTNESS1_MM	HealthMyne (radiomics)	II
CORONAL_LONG_AXIS_MM	HealthMyne (radiomics)	II
CORONAL_SHORT_AXIS_MM	HealthMyne (radiomics)	II
CRANIALCAUDAL_LENGTH_MM	HealthMyne (radiomics)	II
ENERGY_VOXELS	HealthMyne (radiomics)	II
FOOTPRINT_Y_MM	HealthMyne (radiomics)	II
GLCM_COL_MEAN	HealthMyne (radiomics)	II
GLCM_ENTROPY	HealthMyne (radiomics)	II
GLCM_HOMOGENEITY	HealthMyne (radiomics)	III
GLCM_ROW_MEAN	HealthMyne (radiomics)	II
KURTOSIS_HU	HealthMyne (radiomics)	II
KURTOSIS_VOXELS	HealthMyne (radiomics)	II
L1_DISTANCE_MM	HealthMyne (radiomics)	II
L2_DISTANCE_MM	HealthMyne (radiomics)	II
L3_DISTANCE_MM	HealthMyne (radiomics)	II
PART_SOLID_DIAMETER_MM	HealthMyne (radiomics)	II
LUNG_RADS_DIAMETER_MM	HealthMyne (radiomics)	II
MAX_HU	HealthMyne (radiomics)	II
MAX_VOXELS	HealthMyne (radiomics)	II
MEAN_HU	HealthMyne (radiomics)	II
MEAN_VOXELS	HealthMyne (radiomics)	II
MEDIAN_HU	HealthMyne (radiomics)	II
MEDIAN_VOXELS	HealthMyne (radiomics)	II
NORMALIZED_ABOVE_MEAN_DEVIATION_VOXELS	HealthMyne (radiomics)	II
PERCENT_GGO	HealthMyne (radiomics)	I
PERCENT_SOLID	HealthMyne (radiomics)	II
PERCENT_SOLID_INCL_AIR	HealthMyne (radiomics)	II
ROOT_MEAN_SQUARE	HealthMyne (radiomics)	I
ROOT_MEAN_SQUARE_VOXELS	HealthMyne (radiomics)	II
SAGITTAL_LONG_AXIS_MM	HealthMyne (radiomics)	II
SAGITTAL_SHORT_AXIS_MM	HealthMyne (radiomics)	II
SKEWNESS_HU	HealthMyne (radiomics)	I
SKEWNESS_VOXELS	HealthMyne (radiomics)	I
SOLID_VOLUME_ML	HealthMyne (radiomics)	II
SOLID_VOLUME_MM3	HealthMyne (radiomics)	II
SOLID_VOLUME_VOXELS	HealthMyne (radiomics)	II
SURFACE_AREA_MM2	HealthMyne (radiomics)	II
SURFACE_AREA_TO_VOLUME_RATIO_MM	HealthMyne (radiomics)	I
TRANSVERSE_LENGTH_MM	HealthMyne (radiomics)	II
UNIFORMITY_HU	HealthMyne (radiomics)	II
VOLUME_ML	HealthMyne (radiomics)	II
VOLUME_MM3	HealthMyne (radiomics)	II
VOLUME_VOXELS	HealthMyne (radiomics)	II
VOLUMETRIC_LENGTH_MM	HealthMyne (radiomics)	II
HALLMARK_ALLOGRAFT_REJECTION	RNA-Seq	I
HALLMARK_ANDROGEN_RESPONSE	RNA-Seq	I
HALLMARK_ANGIOGENESIS	RNA-Seq	III
HALLMARK_APICAL_JUNCTION	RNA-Seq	III
HALLMARK_APOPTOSIS	RNA-Seq	I
HALLMARK_CHOLESTEROL_HOMEOSTASIS	RNA-Seq	I
HALLMARK_COAGULATION	RNA-Seq	I
HALLMARK_COMPLEMENT	RNA-Seq	I
HALLMARK_DNA_REPAIR	RNA-Seq	IV
HALLMARK_E2F_TARGETS	RNA-Seq	IV
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	RNA-Seq	III
HALLMARK_ESTROGEN_RESPONSE_LATE	RNA-Seq	I
HALLMARK_G2M_CHECKPOINT	RNA-Seq	IV
HALLMARK_GLYCOLYSIS	RNA-Seq	IV
HALLMARK_HEDGEHOG_SIGNALING	RNA-Seq	III
HALLMARK_HYPOXIA	RNA-Seq	I
HALLMARK_IL6_JAK_STAT3_SIGNALING	RNA-Seq	I
HALLMARK_INFLAMMATORY_RESPONSE	RNA-Seq	I
HALLMARK_INTERFERON_ALPHA_RESPONSE	RNA-Seq	I
HALLMARK_INTERFERON_GAMMA_RESPONSE	RNA-Seq	I
HALLMARK_KRAS_SIGNALING_DN	RNA-Seq	II
HALLMARK_KRAS_SIGNALING_UP	RNA-Seq	I
HALLMARK_MITOTIC_SPINDLE	RNA-Seq	IV
HALLMARK_MTORC1_SIGNALING	RNA-Seq	I
HALLMARK_MYC_TARGETS_V1	RNA-Seq	I
HALLMARK_MYC_TARGETS_V2	RNA-Seq	IV
HALLMARK_MYOGENESIS	RNA-Seq	III
HALLMARK_P53_PATHWAY	RNA-Seq	I
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	RNA-Seq	I
HALLMARK_SPERMATOGENESIS	RNA-Seq	IV
HALLMARK_TGF_BETA_SIGNALING	RNA-Seq	I
HALLMARK_TNFA_SIGNALING_VIA_NFKB	RNA-Seq	I
HALLMARK_UV_RESPONSE_DN	RNA-Seq	III
HALLMARK_UV_RESPONSE_UP	RNA-Seq	I
HALLMARK_WNT_BETA_CATENIN_SIGNALING	RNA-Seq	I
HALLMARK_XENOBIOTIC_METABOLISM	RNA-Seq	I
Cell Cycle	RNA-Seq	IV
Innate Immune System	RNA-Seq	I
Cell Cycle Mitotic	RNA-Seq	IV
Leishmania infection	RNA-Seq	I
Neutrophil degranulation	RNA-Seq	I
Cell Cycle Checkpoints	RNA-Seq	IV
Signaling by GPCR	RNA-Seq	I
Platelet activation signaling and aggregation	RNA-Seq	I
DNA Double Strand Break Repair	RNA-Seq	IV
Nuclear Events kinase and transcription factor activation	RNA-Seq	I
Signaling by Interleukins	RNA-Seq	I
DNA Replication	RNA-Seq	IV
Hemostasis	RNA-Seq	I
DNA Repair	RNA-Seq	IV
GPCR downstream signalling	RNA-Seq	I
DNA strand elongation	RNA-Seq	IV
Activation of the pre replicative complex	RNA-Seq	IV
Homology Directed Repair	RNA-Seq	IV
NGF stimulated transcription	RNA-Seq	I
GPCR ligand binding	RNA-Seq	I
Mitotic Prometaphase	RNA-Seq	IV
COPI dependent Golgi to ER retrograde traffic	RNA-Seq	IV
G2 M Checkpoints	RNA-Seq	IV
Kinesins	RNA-Seq	IV
Cilium Assembly	RNA-Seq	IV
Anti inflammatory response favouring Leishmania parasite infection	RNA-Seq	I
Leishmania parasite growth and survival	RNA-Seq	I
M Phase	RNA-Seq	IV
HDR through Homologous Recombination HRR or Single Strand Annealing SSA	RNA-Seq	IV
Cell surface interactions at the vascular wall	RNA-Seq	I
Interleukin 3 Interleukin 5 and GM CSF signaling	RNA-Seq	I
Synthesis of DNA	RNA-Seq	IV
DNA Replication Pre Initiation	RNA-Seq	IV
FOXO mediated transcription	RNA-Seq	I
Activation of ATR in response to replication stress	RNA-Seq	IV
Immunoregulatory interactions between a Lymphoid and a non Lymphoid cell	RNA-Seq	I
Resolution of Sister Chromatid Cohesion	RNA-Seq	IV
DAP12 interactions	RNA-Seq	I
ADORA2B mediated anti inflammatory cytokines production	RNA-Seq	I
Golgi to ER retrograde transport	RNA-Seq	IV
O linked glycosylation of mucins	RNA-Seq	II
Class A 1 Rhodopsin like receptors	RNA-Seq	I
Chromosome Maintenance	RNA-Seq	IV
Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	RNA-Seq	IV
Amplification of signal from the kinetochores	RNA-Seq	IV
Signaling by NTRK1 TRKA	RNA-Seq	I
Mitochondrial translation elongation	RNA-Seq	IV
Mitochondrial translation	RNA-Seq	IV
G alpha s signalling events	RNA-Seq	I
Separation of Sister Chromatids	RNA-Seq	IV
S Phase	RNA-Seq	IV
Mitochondrial translation termination	RNA-Seq	IV
Mitochondrial translation initiation	RNA-Seq	IV

Feature name	Dataset of origin	Cluster
G1 S Transition	RNA-Seq	IV
G alpha q signalling events	RNA-Seq	I
Organelle biogenesis and maintenance	RNA-Seq	IV
Mitotic Spindle Checkpoint	RNA-Seq	IV
Mitotic G1 phase and G1 S transition	RNA-Seq	IV
HDR through Homologous Recombination HRR	RNA-Seq	IV
EML4 and NUDC in mitotic spindle formation	RNA-Seq	IV
Signaling by TGFB family members	RNA-Seq	I
Intra Golgi and retrograde Golgi to ER traffic	RNA-Seq	IV
Homologous DNA Pairing and Strand Exchange	RNA-Seq	IV
Anchoring of the basal body to the plasma membrane	RNA-Seq	IV
Processing of DNA double strand break ends	RNA-Seq	IV
Mitotic Metaphase and Anaphase	RNA-Seq	IV
GPVI mediated activation cascade	RNA-Seq	I
Resolution of Abasic Sites AP sites	RNA-Seq	IV
Ca2 pathway	RNA-Seq	I
Signaling by NTRKs	RNA-Seq	I
Extra nuclear estrogen signaling	RNA-Seq	I
Mitotic Anaphase	RNA-Seq	IV
ESR mediated signaling	RNA-Seq	I
Resolution of D Loop Structures	RNA-Seq	IV
FCGR3A mediated phagocytosis	RNA-Seq	I
Leishmania phagocytosis	RNA-Seq	I
Parasite infection	RNA-Seq	I
Resolution of D loop Structures through Holliday Junction Intermediates	RNA-Seq	IV
PI5P PP2A and IER3 Regulate PI3K AKT Signaling	RNA-Seq	I
Presynaptic phase of homologous DNA pairing and strand exchange	RNA-Seq	IV
Assembly of the pre replicative complex	RNA-Seq	IV
Peptide hormone metabolism	RNA-Seq	I
Mitotic G2 G2 M phases	RNA-Seq	IV
RHO GTPases Activate Formins	RNA-Seq	IV
Toll Like Receptor 9 TLR9 Cascade	RNA-Seq	I
Extracellular matrix organization	RNA-Seq	III
Disease	RNA-Seq	I
Nervous system development	RNA-Seq	I
Axon guidance	RNA-Seq	I
Degradation of the extracellular matrix	RNA-Seq	III
Signaling by Receptor Tyrosine Kinases	RNA-Seq	I
Developmental Biology	RNA-Seq	I
Infectious disease	RNA-Seq	I
Post translational protein phosphorylation	RNA-Seq	III
Platelet degranulation	RNA-Seq	I
Regulation of Insulin like Growth Factor IGF transport and uptake by Insulin like Growth Factor Binding Proteins IGFBPs	RNA-Seq	III
Signaling by Rho GTPases	RNA-Seq	I
Smooth Muscle Contraction	RNA-Seq	III
Response to elevated platelet cytosolic Ca2	RNA-Seq	I
Collagen formation	RNA-Seq	III
RHO GTPase Effectors	RNA-Seq	I
Vesicle mediated transport	RNA-Seq	I
Cytokine Signaling in Immune system	RNA-Seq	I
Integrin cell surface interactions	RNA-Seq	III
Elastic fibre formation	RNA-Seq	III
Collagen degradation	RNA-Seq	III
ECM proteoglycans	RNA-Seq	III
Semaphorin interactions	RNA-Seq	I
Collagen biosynthesis and modifying enzymes	RNA-Seq	III
Membrane Trafficking	RNA-Seq	I
Molecules associated with elastic fibres	RNA-Seq	III
Assembly of collagen fibrils and other multimeric structures	RNA-Seq	III
Muscle contraction	RNA-Seq	III
Antigen processing Cross presentation	RNA-Seq	I
Signaling by Nuclear Receptors	RNA-Seq	I
Non integrin membrane ECM interactions	RNA-Seq	III
Signaling by VEGF	RNA-Seq	I
L1CAM interactions	RNA-Seq	III
Diseases of signal transduction by growth factor receptors and second messengers	RNA-Seq	I
Cell Cell communication	RNA-Seq	I
EPH Ephrin signaling	RNA-Seq	I
Cellular responses to external stimuli	RNA-Seq	I
Clathrin mediated endocytosis	RNA-Seq	I
SARS CoV Infections	RNA-Seq	I
Chondroitin sulfate dermatan sulfate metabolism	RNA-Seq	III
Signaling by NOTCH	RNA-Seq	I
ROS and RNS production in phagocytes	RNA-Seq	I
VEGFA VEGFR2 Pathway	RNA-Seq	I
Programmed Cell Death	RNA-Seq	I
Regulation of actin dynamics for phagocytic cup formation	RNA-Seq	I
Signaling by NOTCH1	RNA-Seq	I
Beta catenin independent WNT signaling	RNA-Seq	I
Fcgamma receptor FCGR dependent phagocytosis	RNA-Seq	I
Glycosaminoglycan metabolism	RNA-Seq	III
Collagen chain trimerization	RNA-Seq	III
Potential therapeutics for SARS	RNA-Seq	I
Constitutive Signaling by NOTCH1 HD PEST Domain Mutants	RNA-Seq	I
Constitutive Signaling by NOTCH1 PEST Domain Mutants	RNA-Seq	I
Signaling by NOTCH1 HD PEST Domain Mutants in Cancer	RNA-Seq	I
Signaling by NOTCH1 PEST Domain Mutants in Cancer	RNA-Seq	I
Signaling by NOTCH1 in Cancer	RNA-Seq	I
Interleukin 4 and Interleukin 13 signaling	RNA-Seq	I
Toll like Receptor Cascades	RNA-Seq	I
MET promotes cell motility	RNA-Seq	III
RHO GTPases Activate WASPs and WAVEs	RNA-Seq	I
Apoptosis	RNA-Seq	I
Defective B3GALTL causes Peters plus syndrome PpS	RNA-Seq	III
Oncogenic MAPK signaling	RNA-Seq	I
Synthesis of substrates in N glycan biosythesis	RNA-Seq	I
EPHB mediated forward signaling	RNA-Seq	I
MET activates PTK2 signaling	RNA-Seq	III
Rho GTPase cycle	RNA-Seq	III
Diseases associated with glycosaminoglycan metabolism	RNA-Seq	III
Signaling by NOTCH4	RNA-Seq	I
ER Phagosome pathway	RNA-Seq	I
G alpha 12 13 signalling events	RNA-Seq	III
Unfolded Protein Response UPR	RNA-Seq	I
PTEN Regulation	RNA-Seq	I
Transcriptional regulation by RUNX2	RNA-Seq	I
Binding and Uptake of Ligands by Scavenger Receptors	RNA-Seq	I
Regulation of PTEN gene transcription	RNA-Seq	I
Diseases of glycosylation	RNA-Seq	III
G2 M Transition	RNA-Seq	IV
Diseases of metabolism	RNA-Seq	III
Switching of origins to a post replicative state	RNA-Seq	IV
Signaling by MET	RNA-Seq	III
APC C mediated degradation of cell cycle proteins	RNA-Seq	IV
Regulation of mitotic cell cycle	RNA-Seq	IV
Activation of APC C and APC C Cdc20 mediated degradation of mitotic proteins	RNA-Seq	IV
Diseases associated with O glycosylation of proteins	RNA-Seq	III
Protein protein interactions at synapses	RNA-Seq	III
APC C Cdc20 mediated degradation of mitotic proteins	RNA-Seq	IV
Costimulation by the CD28 family	RNA-Seq	I
Regulation of APC C activators between G1 S and early anaphase	RNA-Seq	IV
APC C Cdh1 mediated degradation of Cdc20 and other APC C Cdh1 targeted proteins in late mitosis early G1	RNA-Seq	IV
Orc1 removal from chromatin	RNA-Seq	IV
Centrosome maturation	RNA-Seq	IV
Recruitment of mitotic centrosome proteins and complexes	RNA-Seq	IV
Metabolism of carbohydrates	RNA-Seq	IV
APC Cdc20 mediated degradation of cell cycle proteins prior to satisfation of the cell cycle checkpoint	RNA-Seq	IV
O glycosylation of TSR domain containing proteins	RNA-Seq	III
Cell junction organization	RNA-Seq	III
AURKA Activation by TPX2	RNA-Seq	IV
Signaling by PDGF	RNA-Seq	III
Recruitment of NuMA to mitotic centrosomes	RNA-Seq	IV
MHC class II antigen presentation	RNA-Seq	IV
O linked glycosylation	RNA-Seq	III
Cdc20 Phospho APC C mediated degradation of Cyclin A	RNA-Seq	IV
Regulation of TP53 Activity through Phosphorylation	RNA-Seq	IV
G2 M DNA damage checkpoint	RNA-Seq	IV
CDK mediated phosphorylation and removal of Cdc6	RNA-Seq	IV
APC C Cdc20 mediated degradation of Securin	RNA-Seq	IV
The role of GTSE1 in G2 M progression after G2 checkpoint	RNA-Seq	IV
Plasma lipoprotein assembly remodeling and clearance	RNA-Seq	IV
ABC transporter disorders	RNA-Seq	IV
Autodegradation of Cdh1 by Cdh1 APC C	RNA-Seq	IV
NCAM signaling for neurite out growth	RNA-Seq	III
PCP CE pathway	RNA-Seq	IV
Loss of Nlp from mitotic centrosomes	RNA-Seq	IV
Loss of proteins required for interphase microtubule organization from the centrosome	RNA-Seq	IV
Regulation of PLK1 Activity at G2 M Transition	RNA-Seq	IV
Neurexins and neuroligins	RNA-Seq	III
A tetrasaccharide linker sequence is required for GAG synthesis	RNA-Seq	III

Table S10. Data integration features clusters

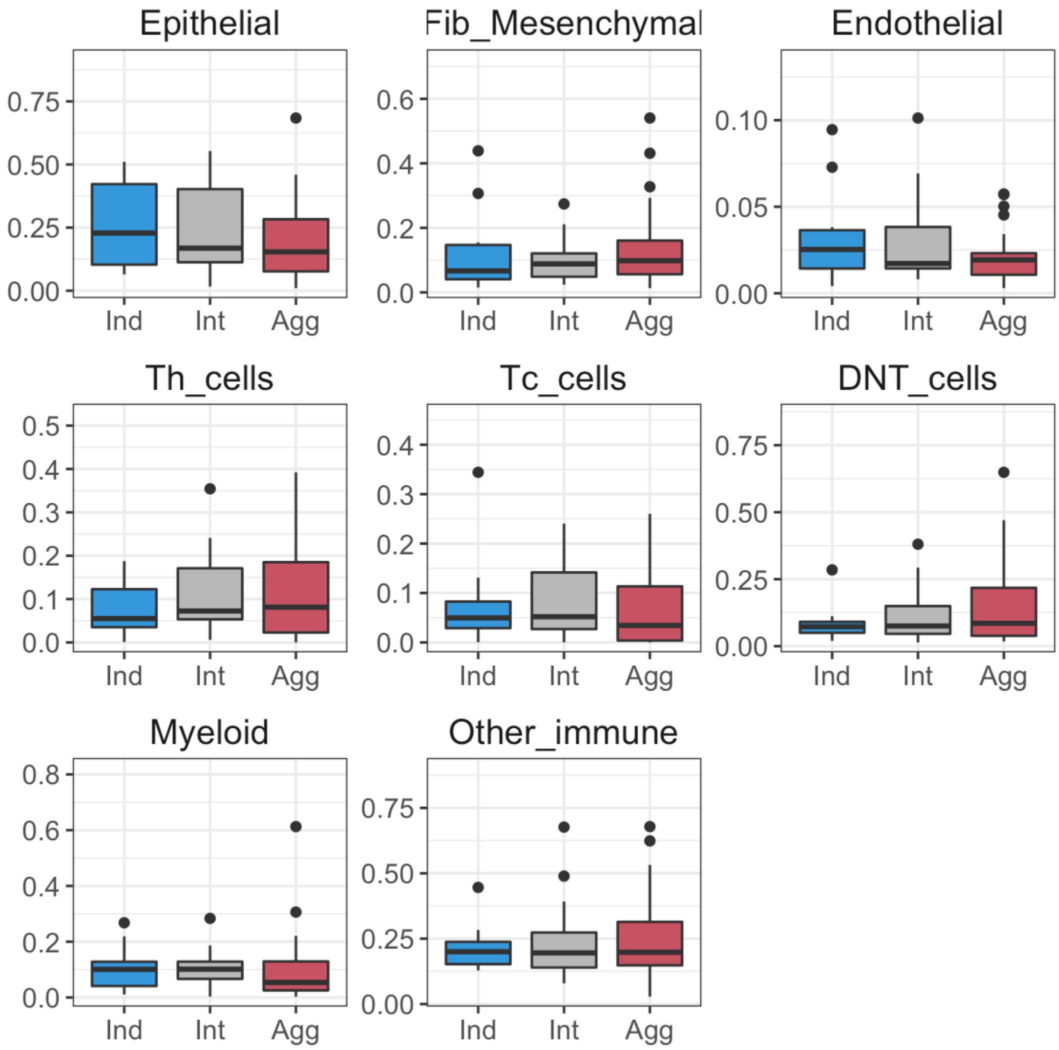
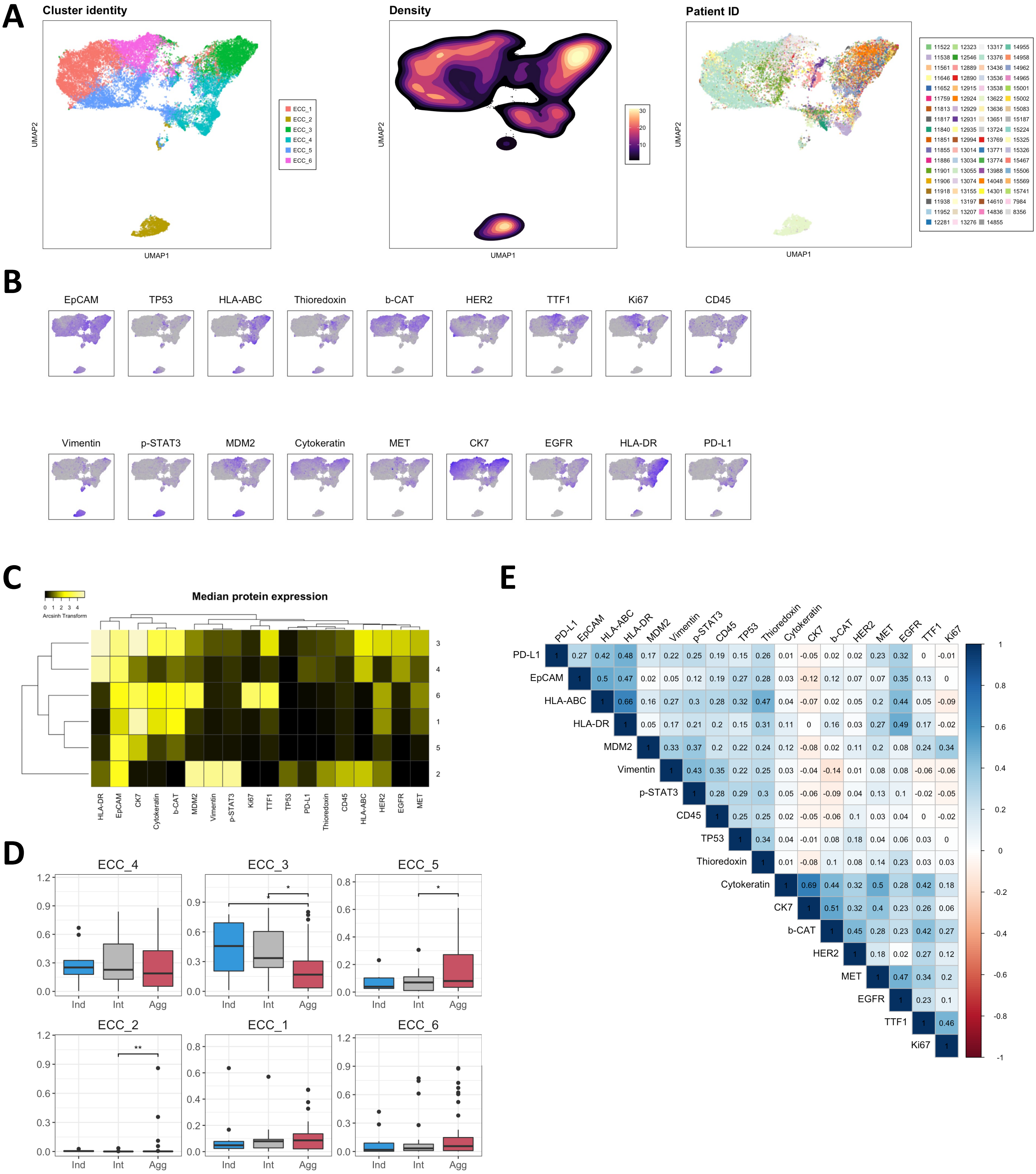


Figure S1. Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. P value >0.05 for all comparisons.



	PD-L1	EpCAM	HLA-ABC	HLA-DR	MDM2	Vimentin	p-STAT3	CD45	TP53	Thioredoxin	Cytokeratin	CK7	b-CAT	HER2	MET	EGFR	TTF1	Ki67
PD-L1	1	0.27	0.42	0.48	0.17	0.22	0.25	0.19	0.15	0.26	0.01	-0.05	0.02	0.02	0.23	0.32	0	-0.01
EpCAM		1	0.5	0.47	0.02	0.05	0.12	0.19	0.27	0.28	0.03	-0.12	0.12	0.07	0.07	0.35	0.13	0
HLA-ABC			1	0.66	0.16	0.27	0.3	0.28	0.32	0.47	0.04	-0.07	0.02	0.05	0.2	0.44	0.05	-0.09
HLA-DR				1	0.05	0.17	0.21	0.2	0.15	0.31	0.11	0	0.16	0.03	0.27	0.49	0.17	-0.02
MDM2					1	0.33	0.37	0.2	0.22	0.24	0.12	-0.08	0.02	0.11	0.2	0.08	0.24	0.34
Vimentin						1	0.43	0.35	0.22	0.25	0.03	-0.04	-0.14	0.01	0.08	0.08	-0.06	-0.06
p-STAT3							1	0.28	0.29	0.3	0.05	-0.06	-0.09	0.04	0.13	0.1	-0.02	-0.05
CD45								1	0.25	0.25	0.02	-0.05	-0.06	0.1	0.03	0.04	0	-0.02
TP53									1	0.34	0.04	-0.01	0.08	0.18	0.04	0.06	0.03	0
Thioredoxin										1	0.01	-0.08	0.1	0.08	0.14	0.23	0.03	0.03
Cytokeratin											1	0.69	0.44	0.32	0.5	0.28	0.42	0.18
CK7												1	0.51	0.32	0.4	0.23	0.26	0.06
b-CAT													1	0.45	0.28	0.23	0.42	0.27
HER2														1	0.18	0.02	0.27	0.12
MET															1	0.47	0.34	0.2
EGFR																1	0.23	0.1
TTF1																	1	0.46
Ki67																		1

Figure S2. Epithelial cancer cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

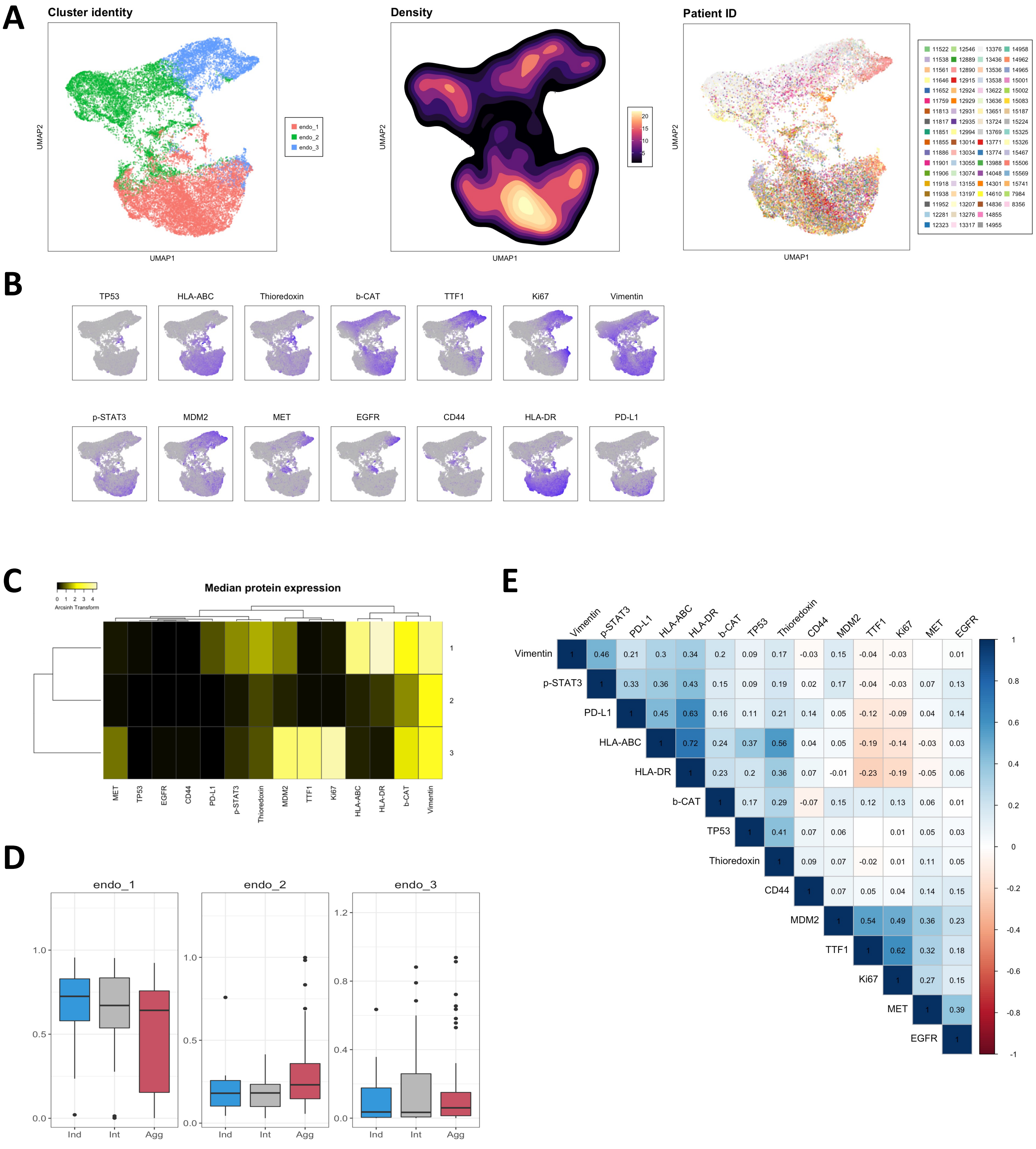


Figure S3. Endothelial cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

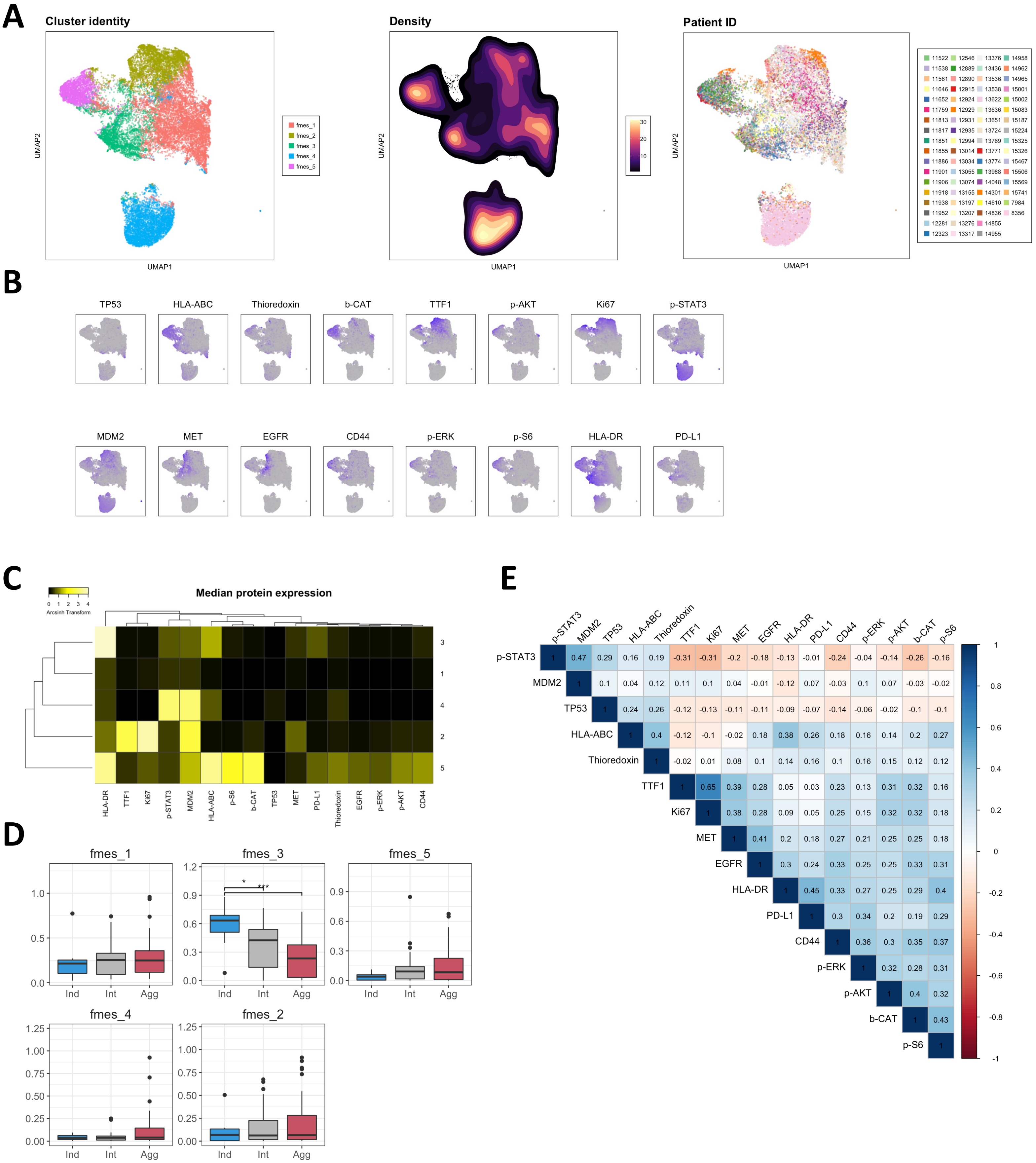


Figure S4. Fibroblasts/Mesenchymal cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

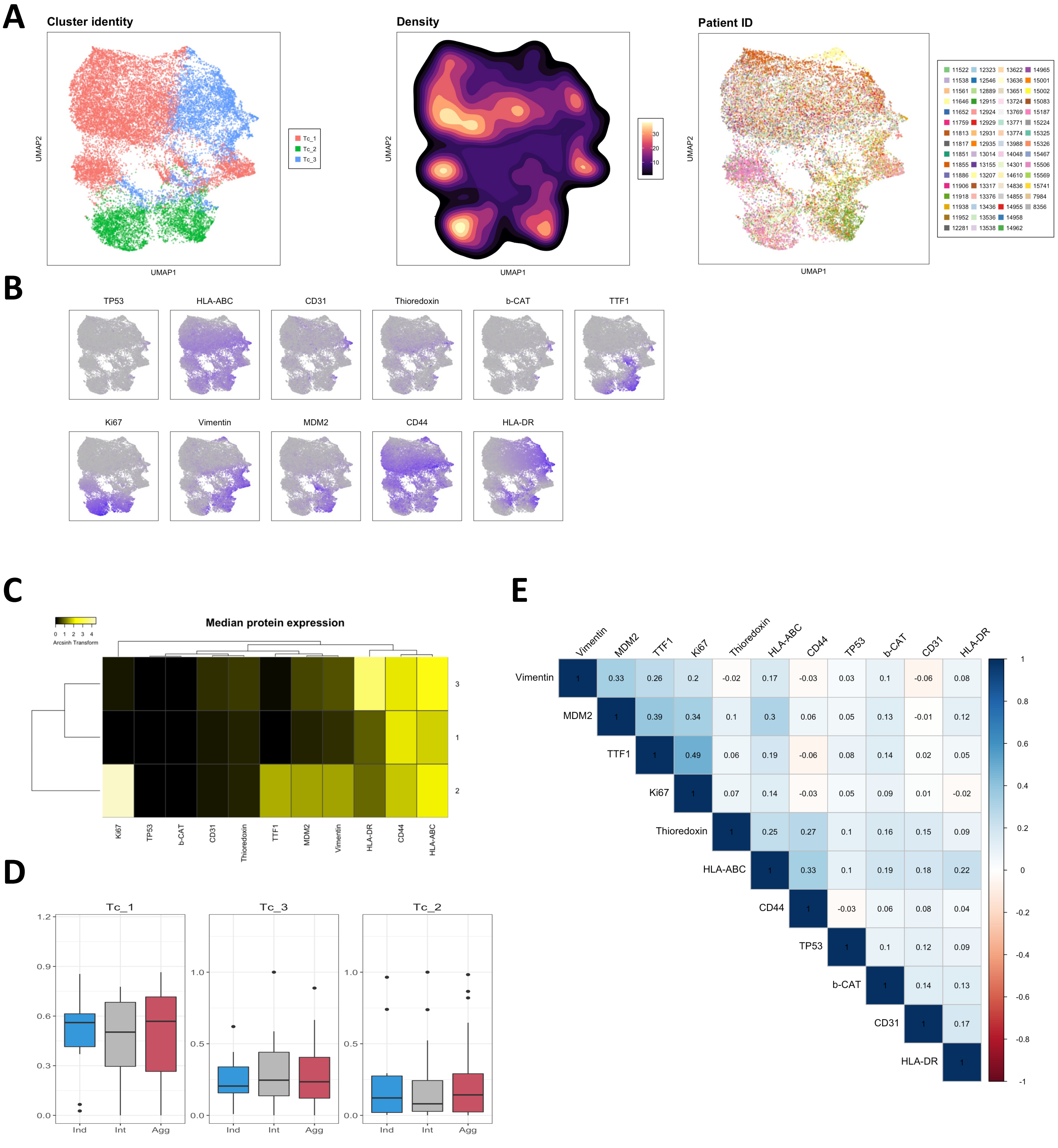


Figure S5. CD8+ T cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

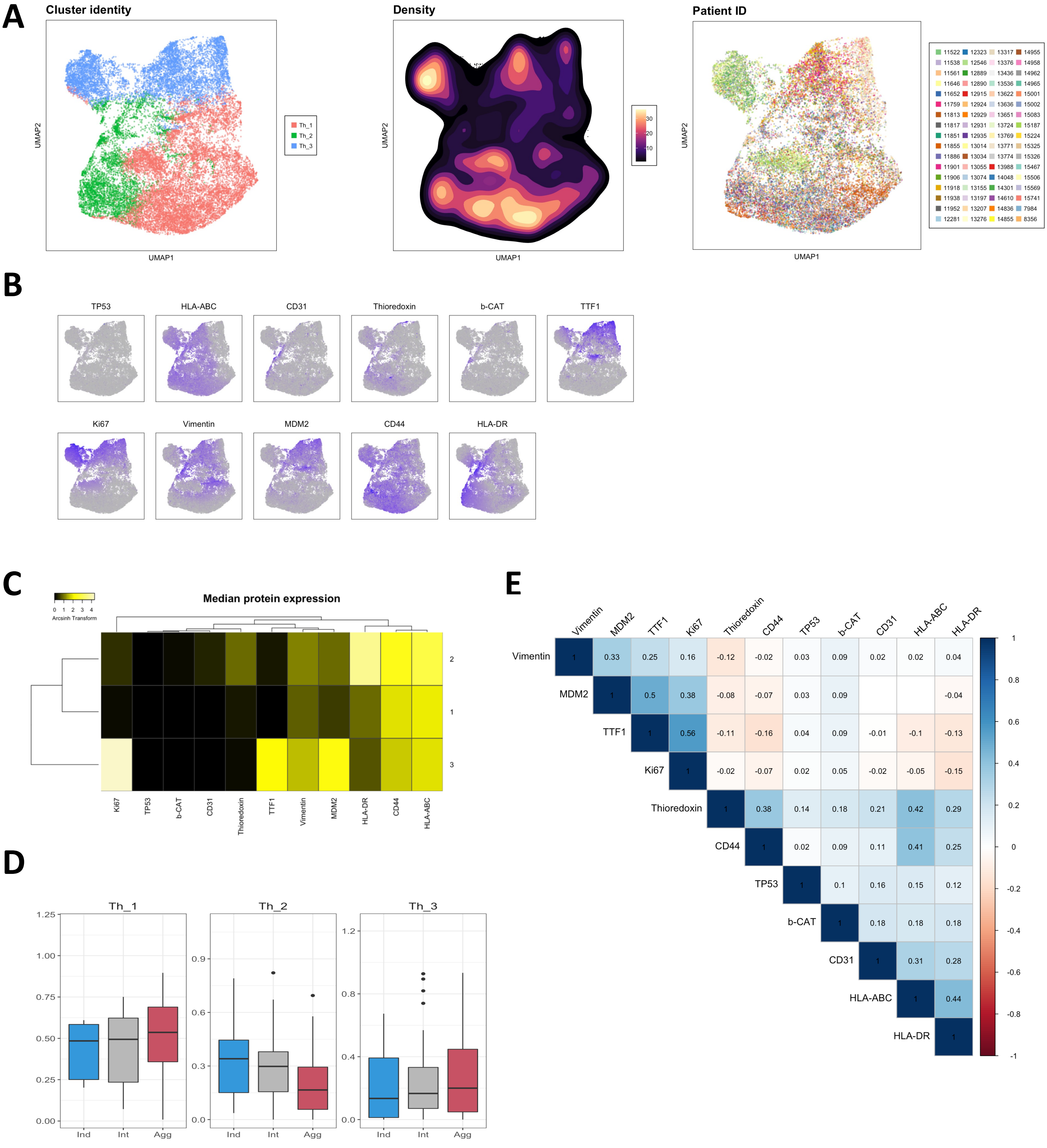


Figure S6. CD4+ T cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

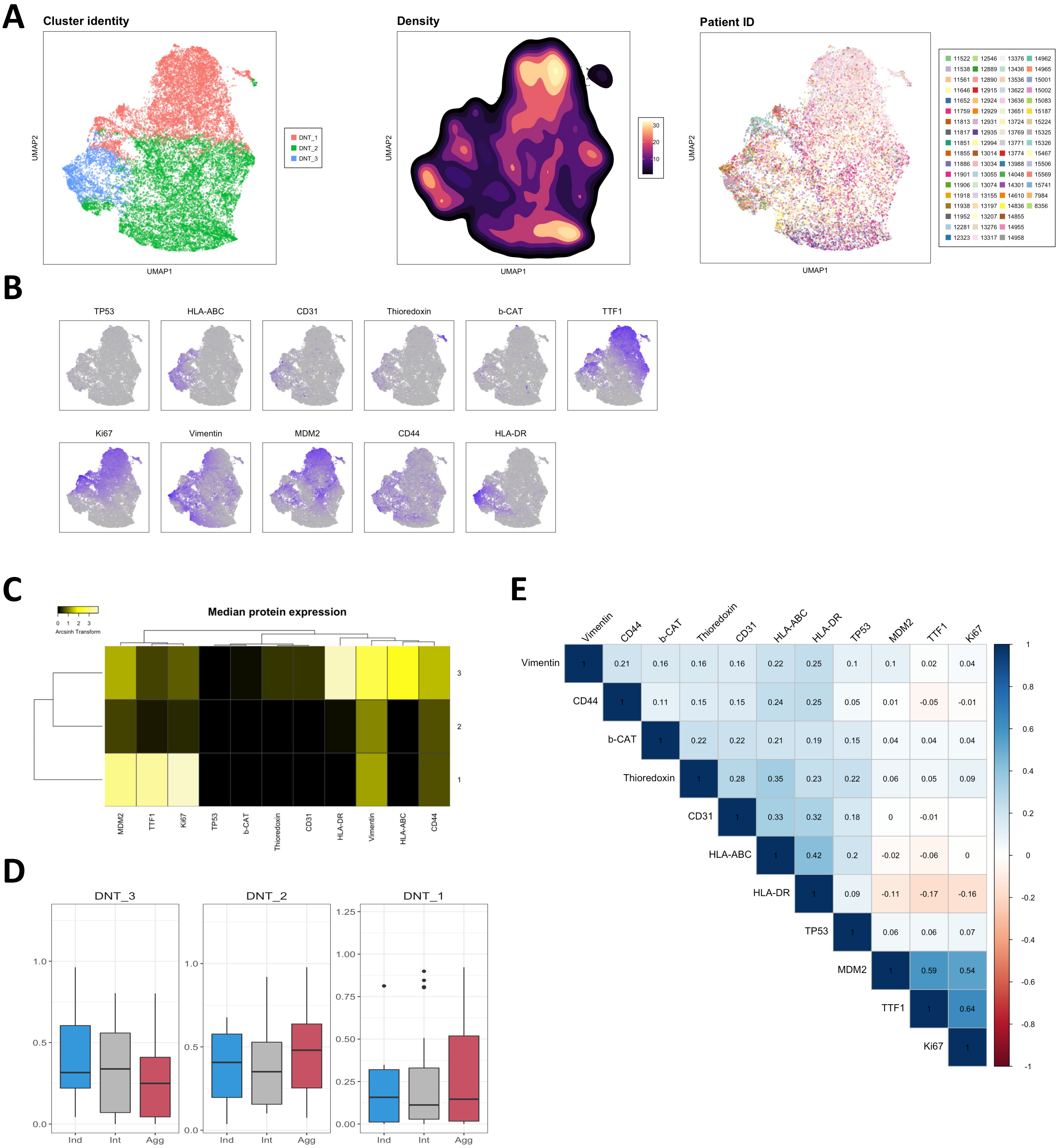


Figure S7. CD8-/CD4- T cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

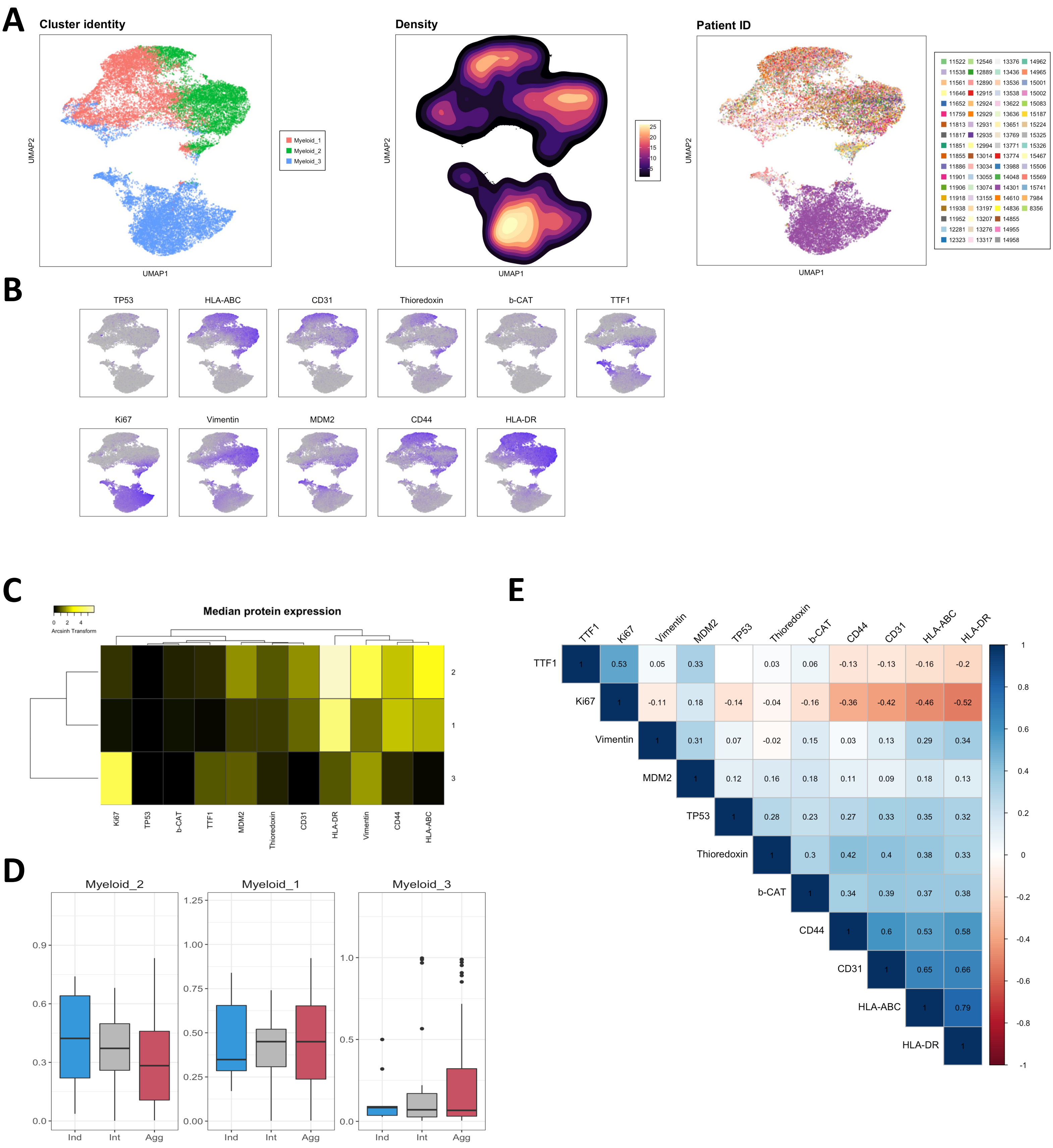


Figure S8. Myeloid cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

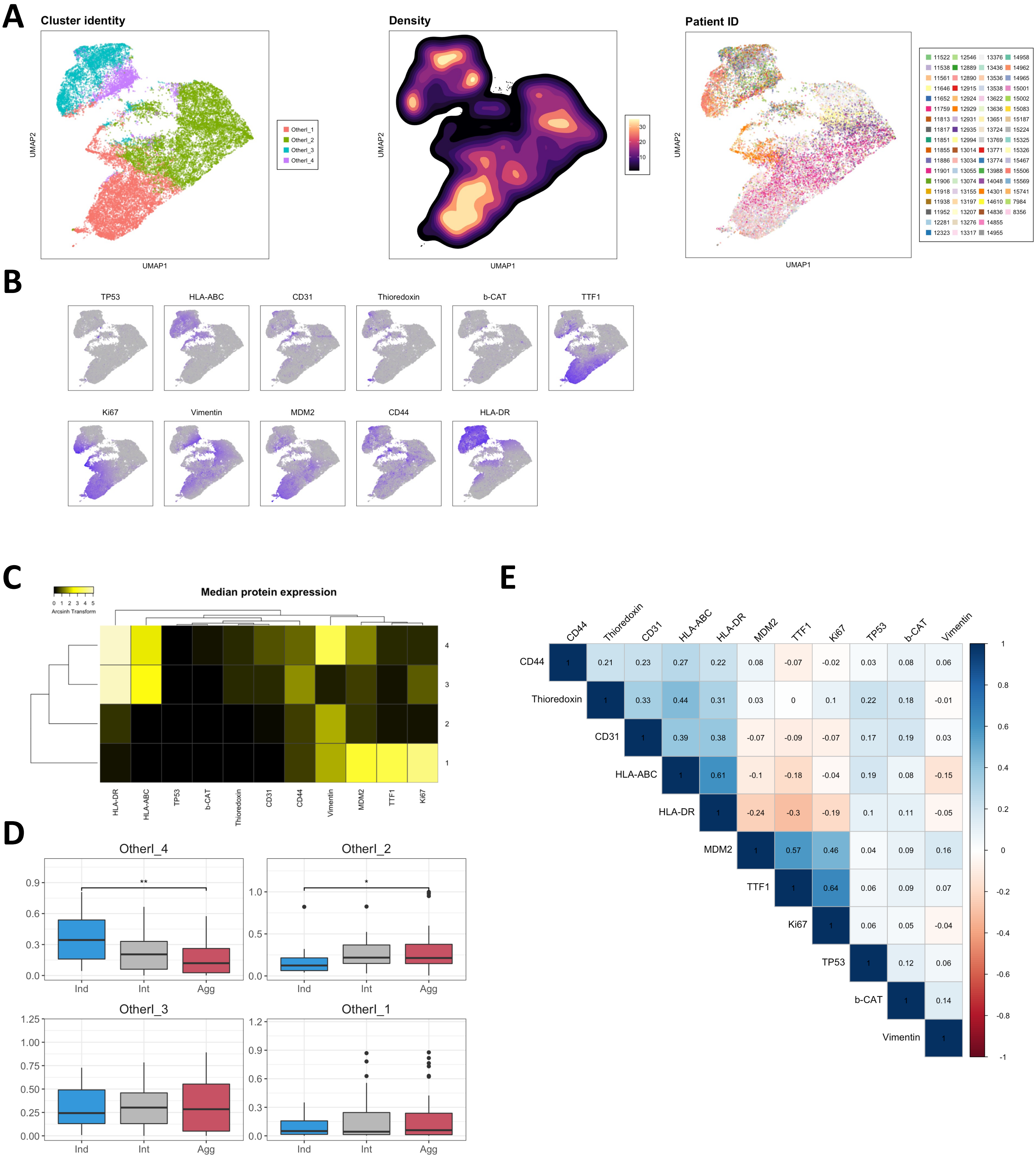


Figure S9. Other immune cells cluster analysis. (A) UMAP representation of the clusters colored by cluster identity, density, and patient ID. (B) UMAP representation of the clusters colored by protein expression intensity. These are the features used for both clustering and UMAP visualization. (C) Heatmap of median protein expression per protein marker per cluster. (D) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.01, ***=pvalue<0.001. (E) Protein-protein Spearman correlation analysis. Only significant correlations (p value >0.05) are colored.

Figure E19.

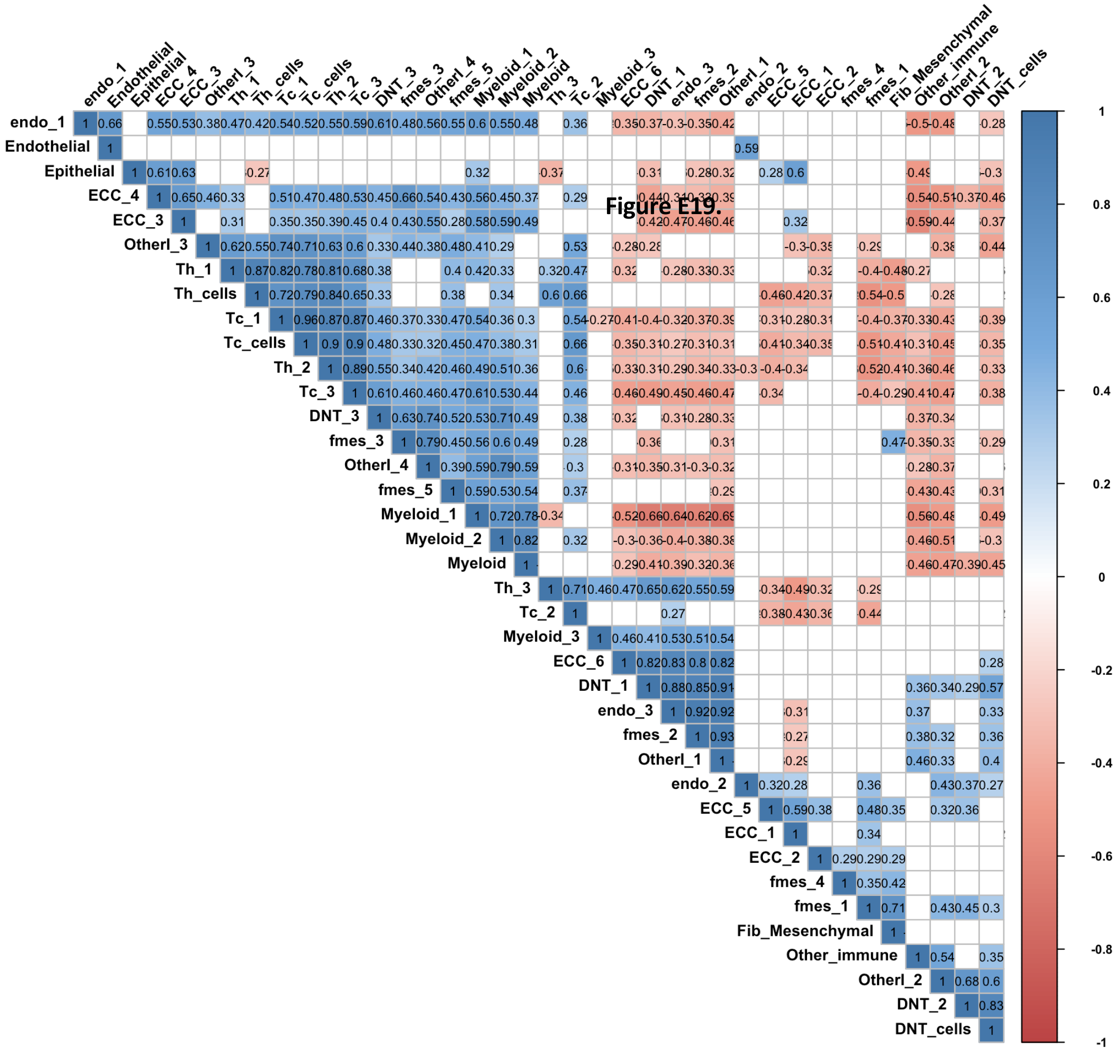


Figure S10. Spearman correlation of fraction per patient sample of main cell types and cell types clusters. Only significant correlations (p value >0.05) are colored.

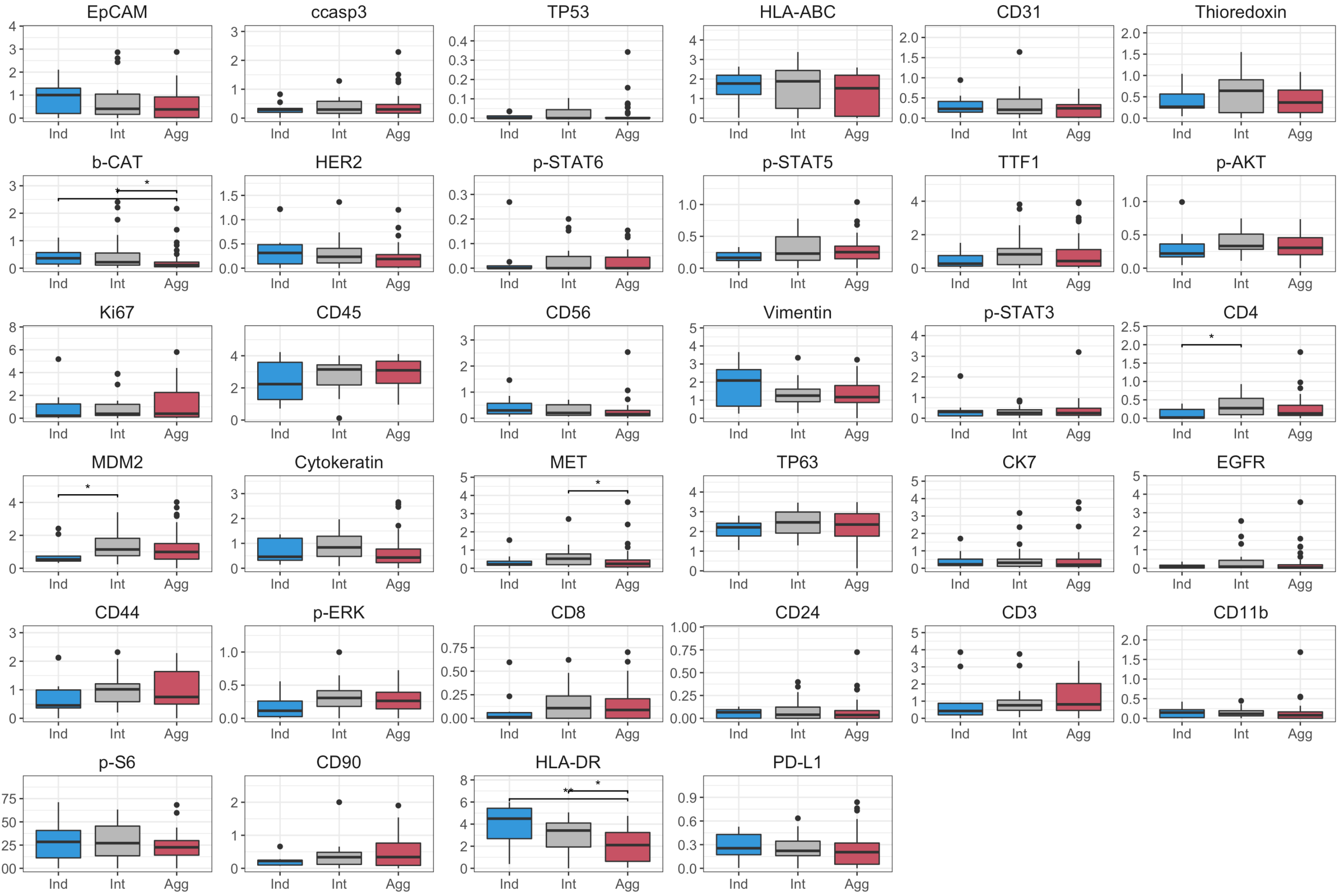
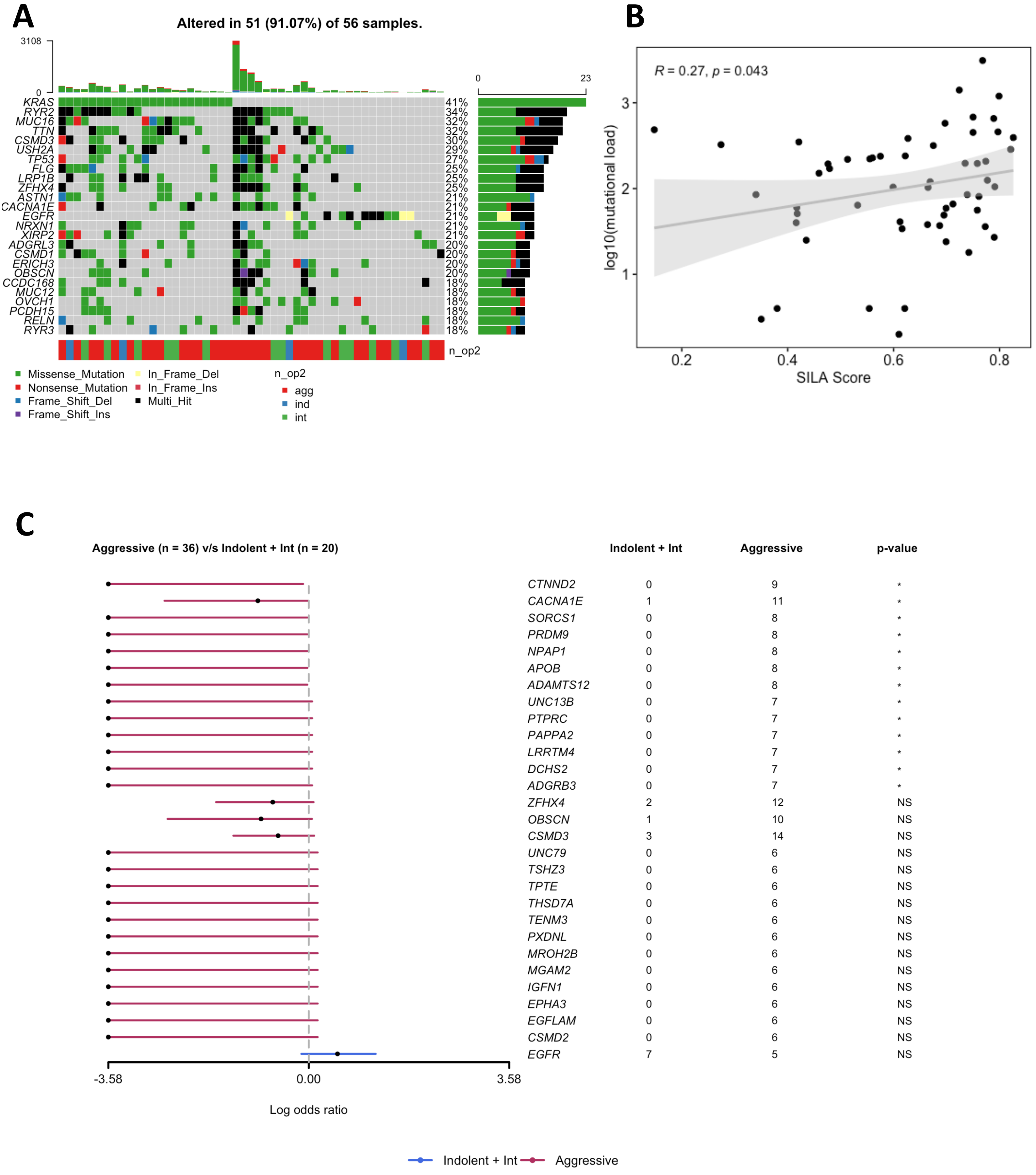


Figure S11. Differential bulk protein expression analysis per patient sample. No star=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001.



	Indolent + Int	Aggressive	p-value
CTNND2	0	9	*
CACNA1E	1	11	*
SORCS1	0	8	*
PRDM9	0	8	*
NPAP1	0	8	*
APOB	0	8	*
ADAMTS12	0	8	*
UNC13B	0	7	*
PTPRC	0	7	*
PAPPA2	0	7	*
LRRTM4	0	7	*
DCHS2	0	7	*
ADGRB3	0	7	*
ZFHX4	2	12	NS
OBSCN	1	10	NS
CSMD3	3	14	NS
UNC79	0	6	NS
TSHZ3	0	6	NS
TPTE	0	6	NS
THSD7A	0	6	NS
TENM3	0	6	NS
PXDNL	0	6	NS
MROH2B	0	6	NS
MGAM2	0	6	NS
IGFN1	0	6	NS
EPHA3	0	6	NS
EGFLAM	0	6	NS
CSMD2	0	6	NS
EGFR	7	5	NS

Figure S12. Whole Exome Sequencing data analysis. (A) Oncoplot showing top 25 mutated genes. (B) Spearman correlation of SILA score and Log10 of mutational load per patient. (C) Clinical enrichment analysis of mutations comparing Indolent+Intermediate versus Aggressive tumor samples.

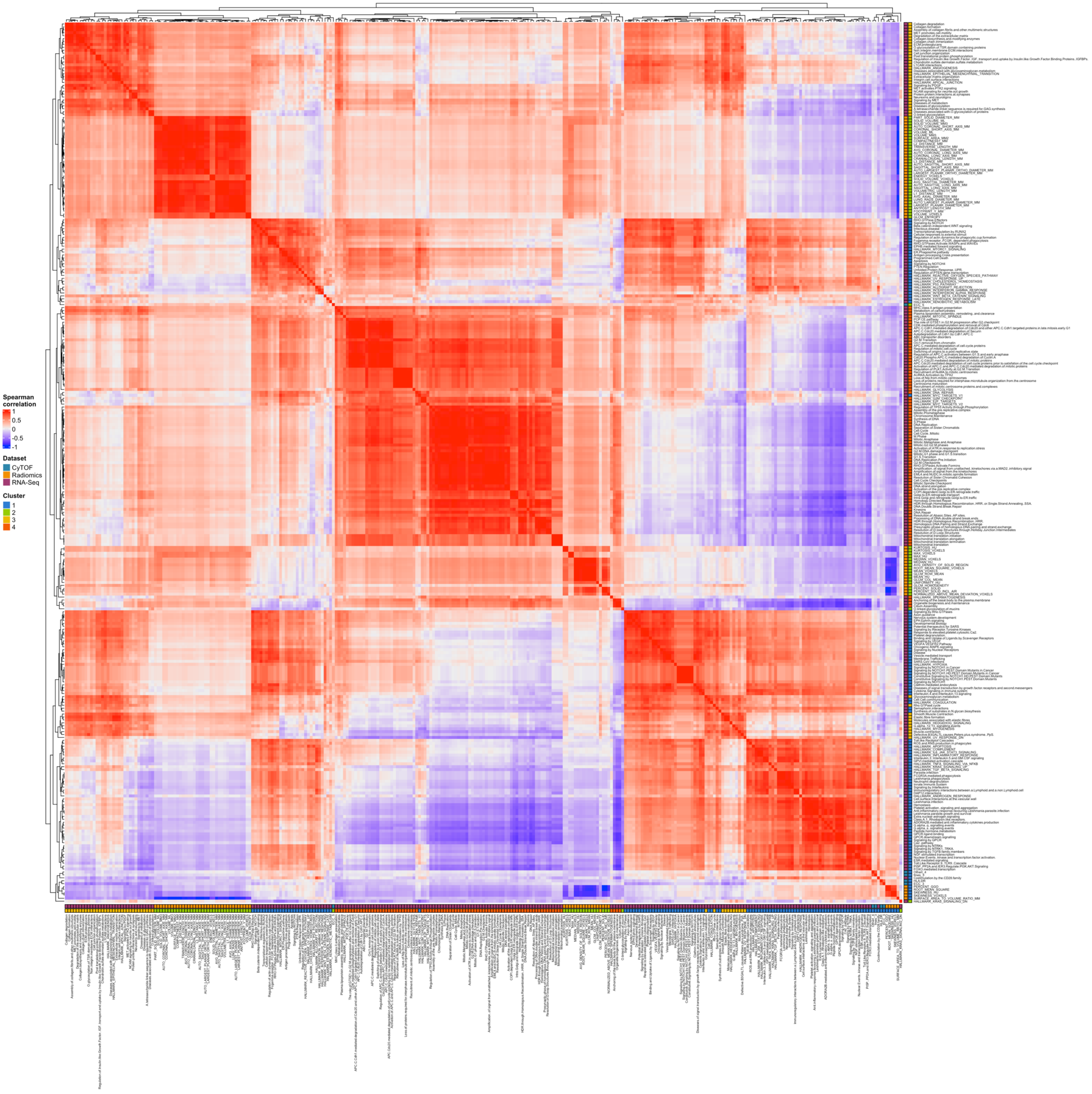


Figure S13. Similarity matrix of features used for data integration.

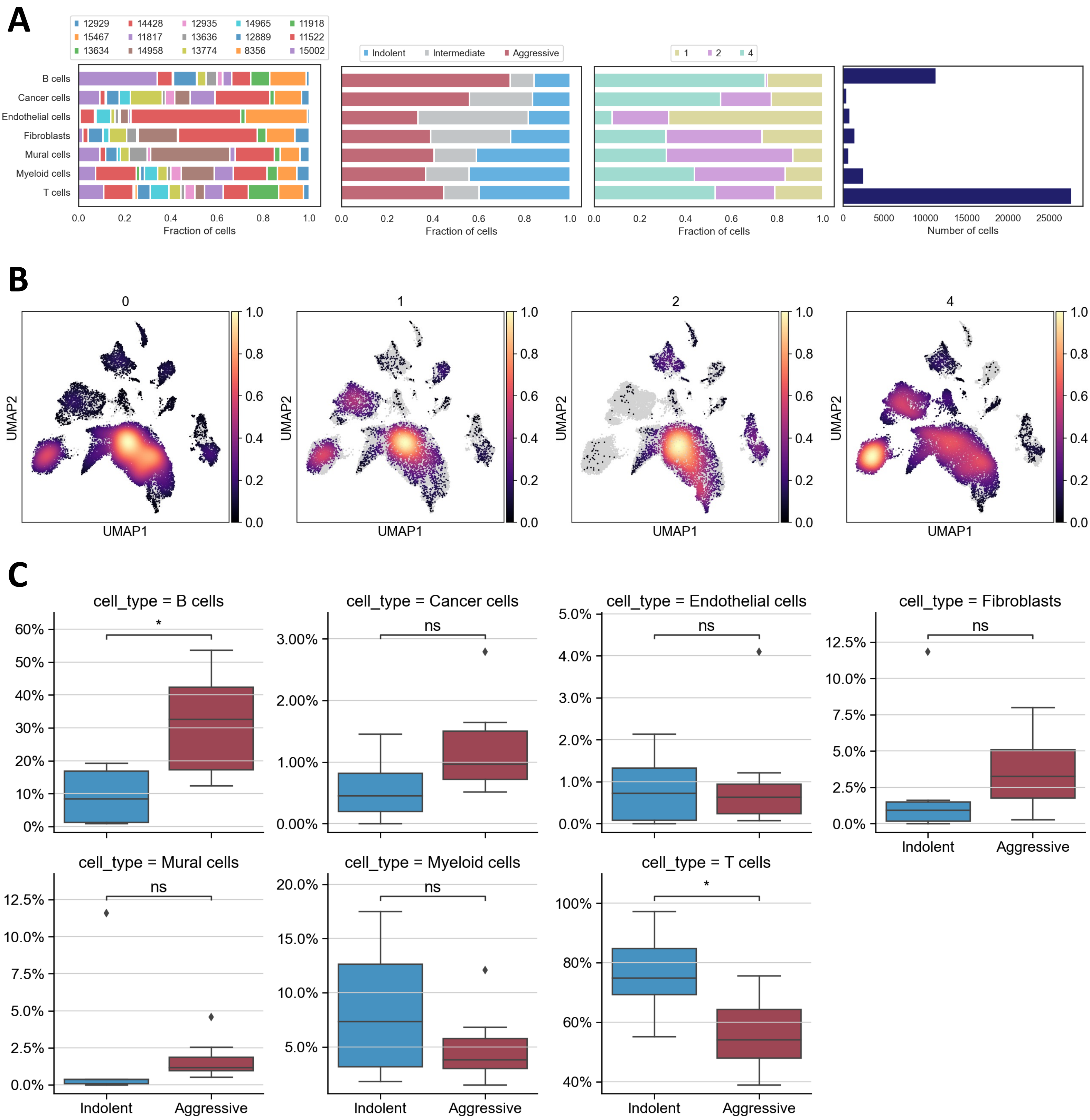


Figure S14. Single cell RNA-Seq analysis of 15 tumor samples. (A) Fraction of cells per cell type colored by patient ID, risk group, data integration patient cluster, and number of cells per cell type. (B) UMAP representation of 44867 cells from 15 patients colored by cell density. Labels correspond to data integration patient cluster 1=P1, 2=P2, 4=P4, 0=patients not included in data integration. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001.

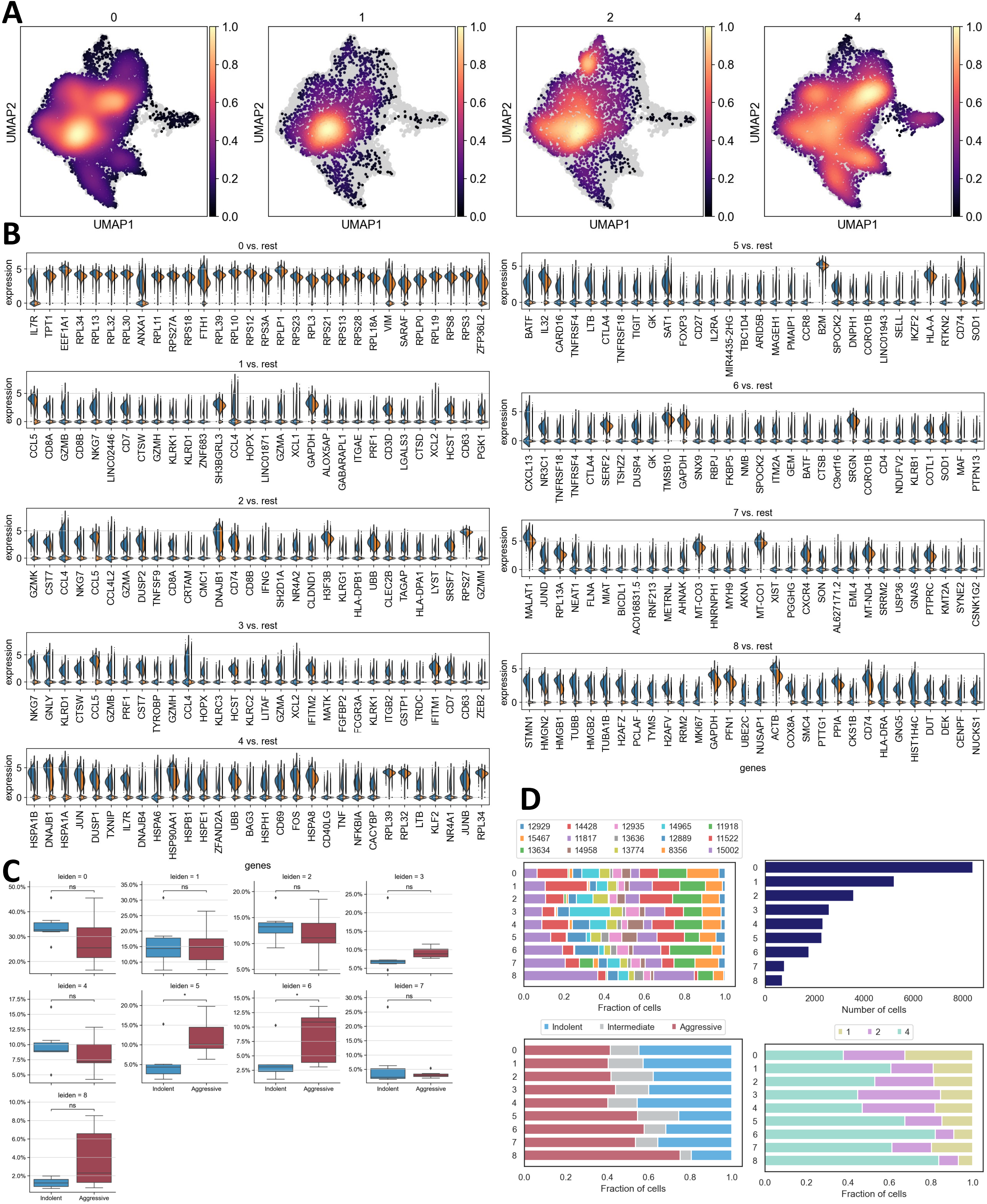


Figure S15. T cells cluster analysis. (A) UMAP representation of 27708 cells from 15 patients colored by cell density. Labels correspond to data integration patient cluster 1=P1, 2=P2, 4=P4, 0=patients not included in data integration. (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.

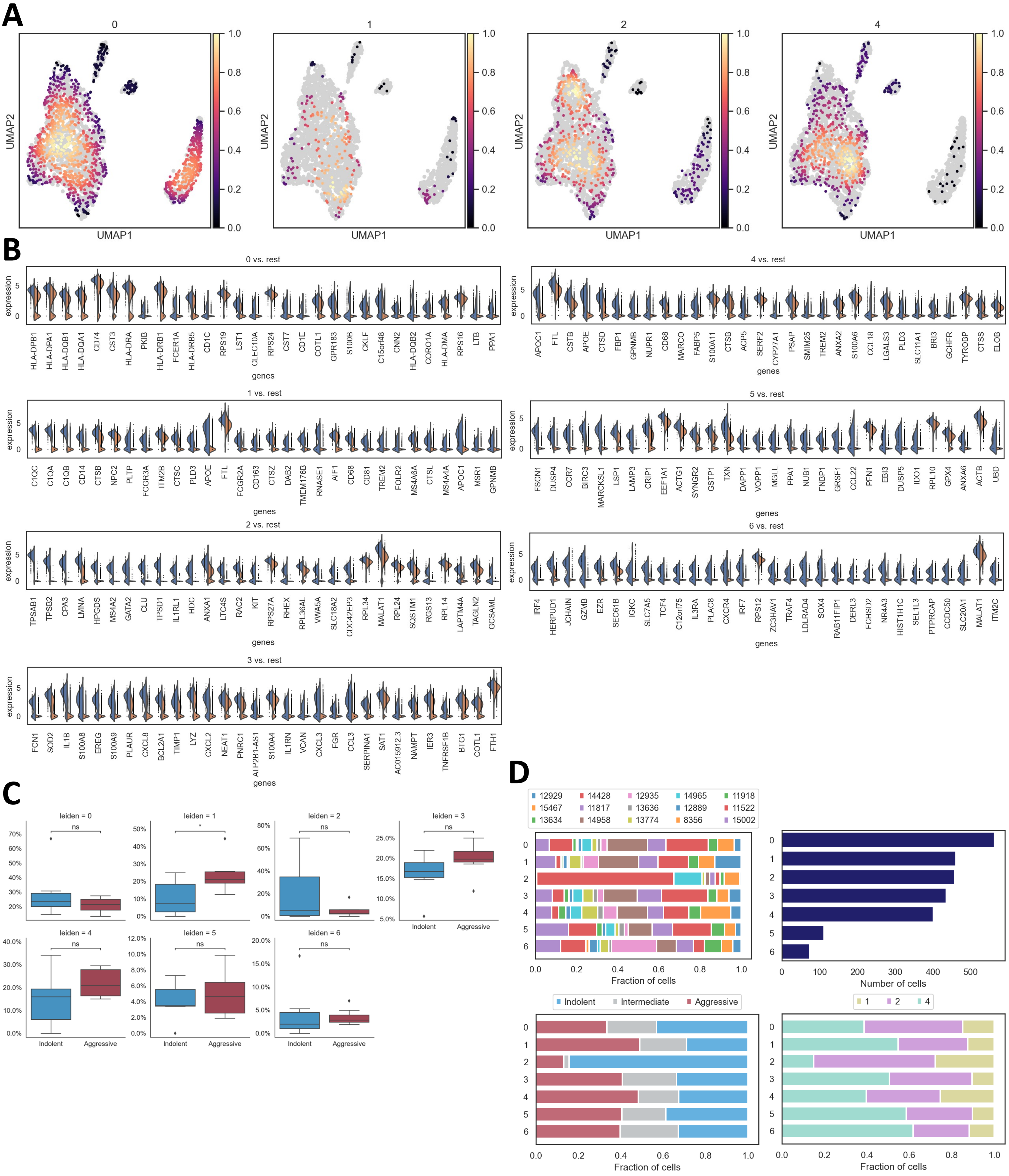


Figure S16. Myeloid cells cluster analysis. (A) UMAP representation of 2497 cells from 15 patients colored by cell density. Labels correspond to data integration patient cluster 1=P1, 2=P2, 4=P4, 0=patients not included in data integration. (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.

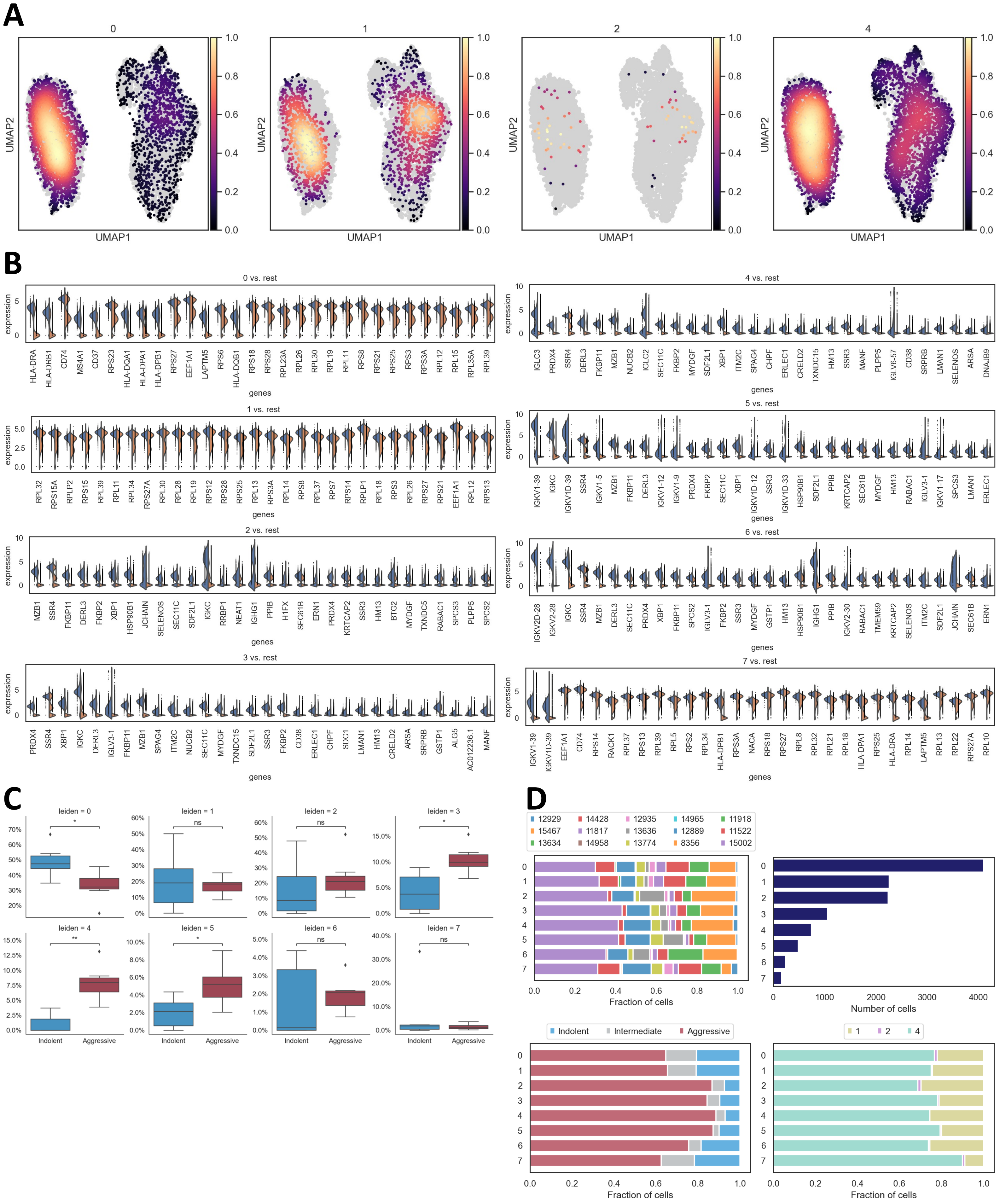


Figure S17. B cells cluster analysis. (A) UMAP representation of 11246 cells from 15 patients colored by cell density. Labels correspond to data integration patient cluster 1=P1, 2=P2, 4=P4, 0=patients not included in data integration. (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.

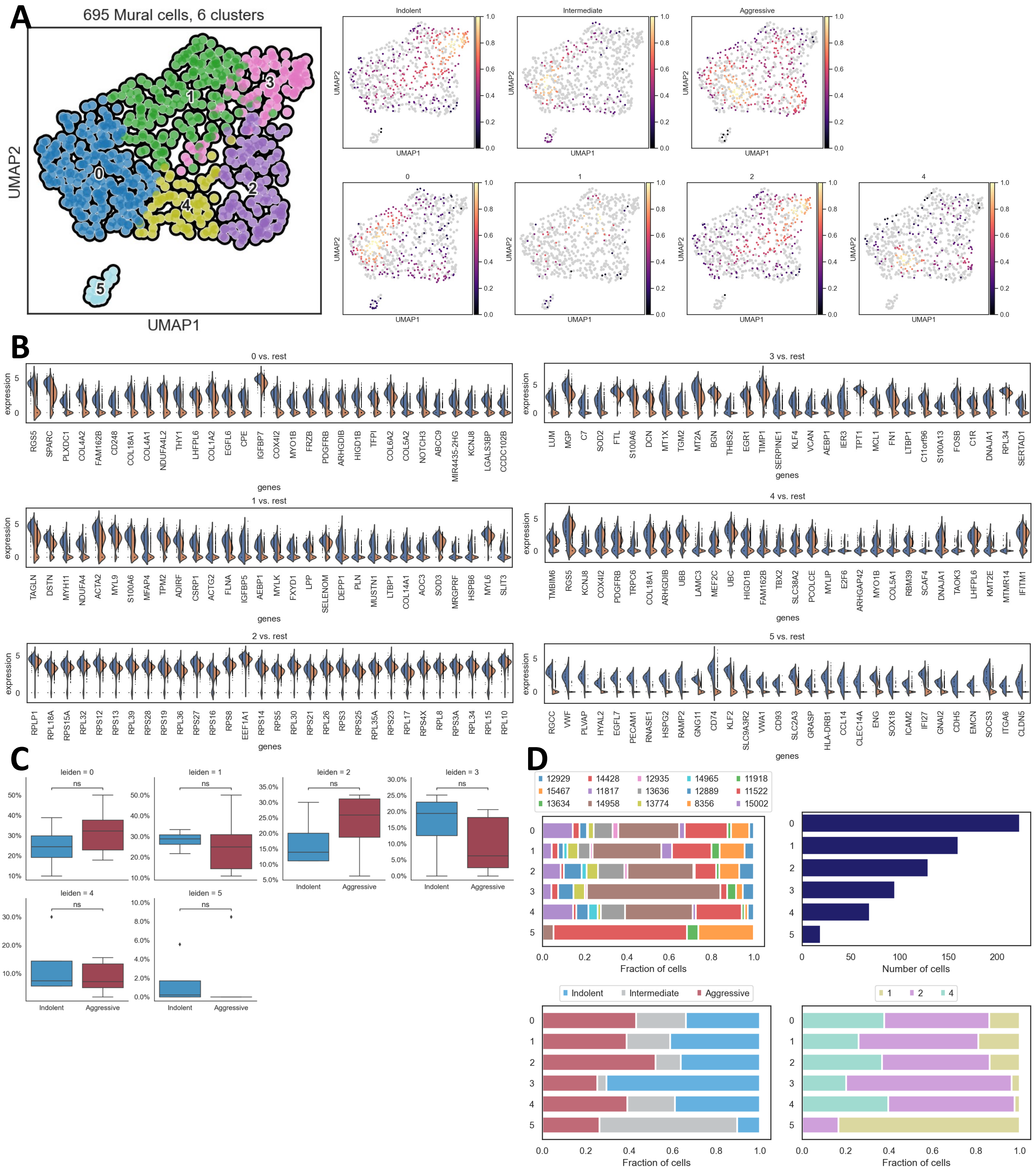


Figure S18. Mural cells cluster analysis. (A) UMAP representation of 695 cells from 15 patients colored by cluster identity, and cell density devided by risk group (up) or data integration patient clusters (down) (1=P1, 2=P2, 4=P4, 0=patients not included in data integration). (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.

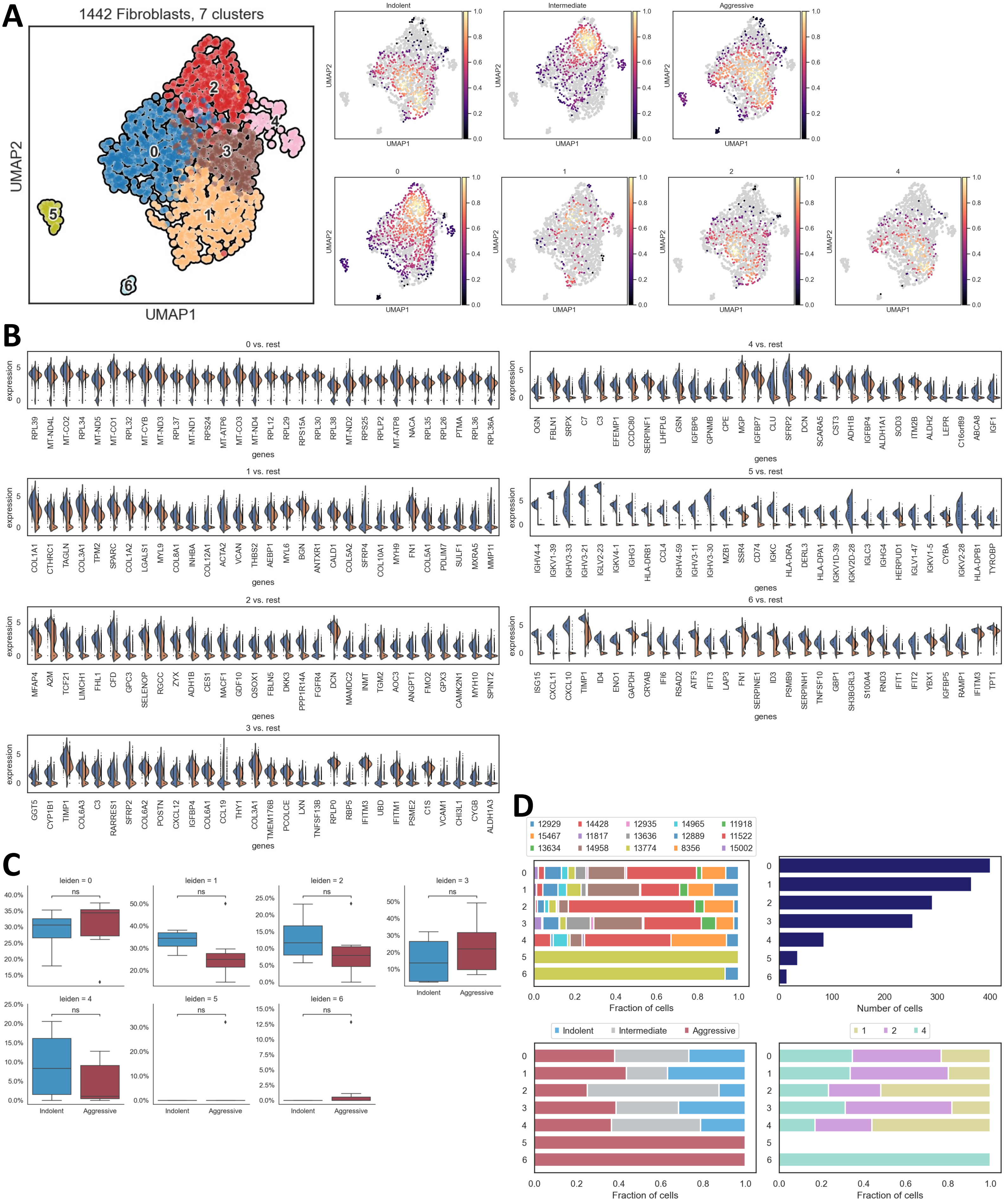


Figure S19. Fibroblasts cells cluster analysis. (A) UMAP representation of 1442 cells from 15 patients colored by cluster identity, and cell density devided by risk group (up) or data integration patient clusters (down) (1=P1, 2=P2, 4=P4, 0=patients not included in data integration). (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.

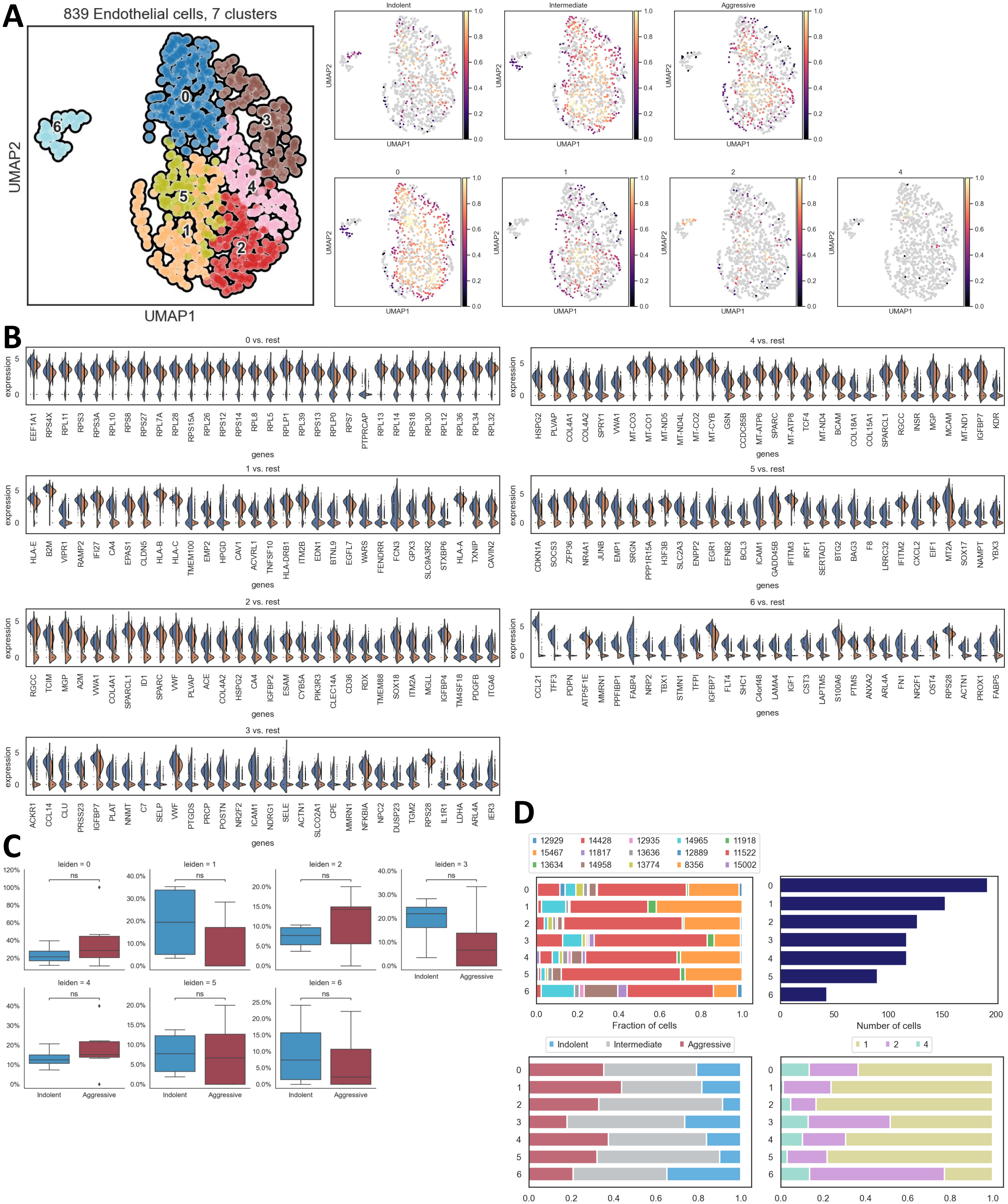


Figure S20. Endothelial cells cluster analysis. (A) UMAP representation of 839 cells from 15 patients colored by cluster identity, and cell density devided by risk group (up) or data integration patient clusters (down) (1=P1, 2=P2, 4=P4, 0=patients not included in data integration). (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.

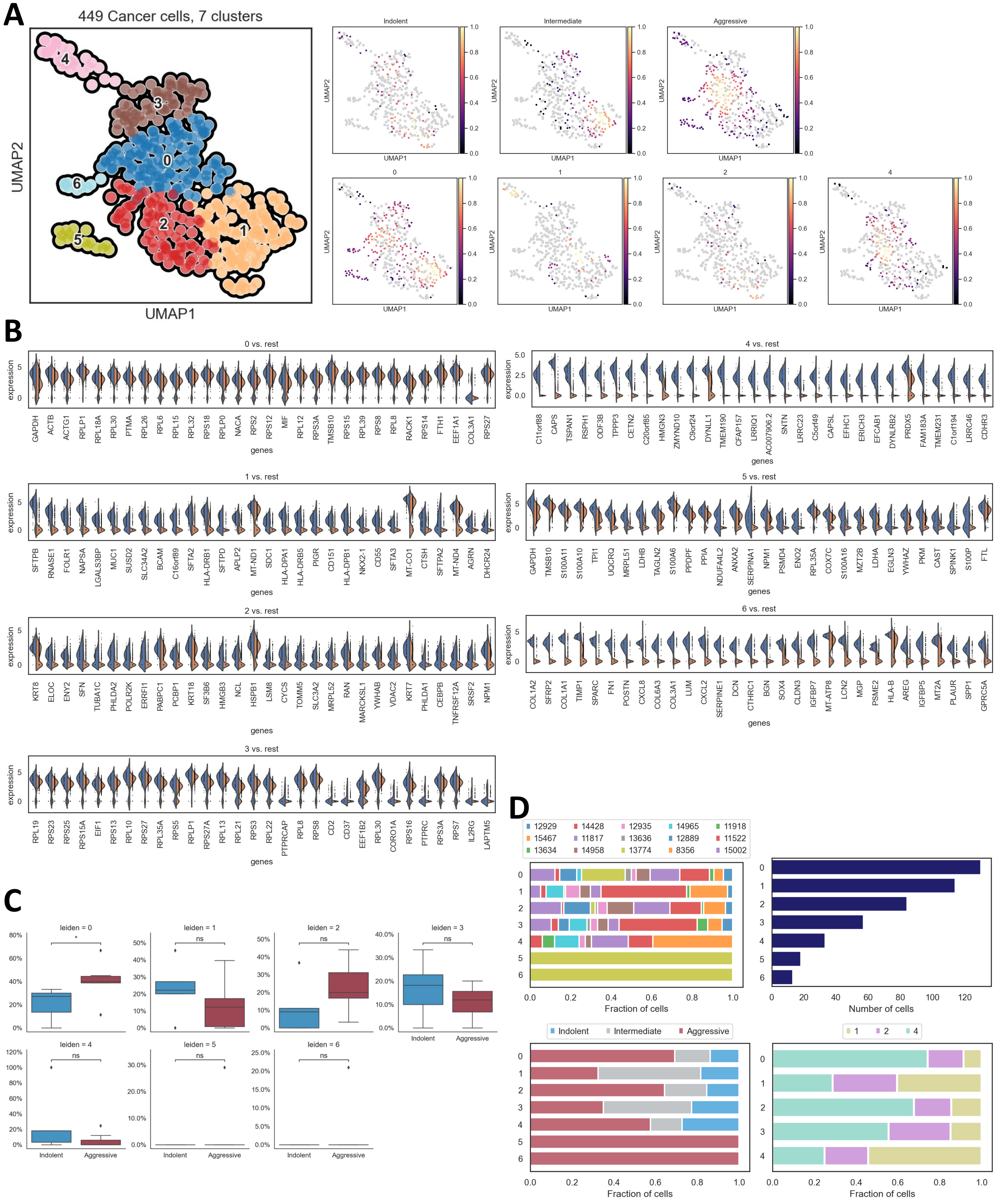


Figure S21. Cancer cells cluster analysis. (A) UMAP representation of 449 cells from 15 patients colored by cluster identity, and cell density devided by risk group (up) or data integration patient clusters (down) (1=P1, 2=P2, 4=P4, 0=patients not included in data integration). (B) Split violin visualization showing the top 30 marker genes for each cluster when compared to the rest.. (C) Differential abundance analysis. Y axis corresponds to the fraction of cells per patient sample. ns=pvalue>0.05, *=pvalue<0.05, **=pvalue<0.001. (D) Fraction of cells per cluster colored by patient ID, risk group, data integration patient cluster, and number of cells per cluster.