Computational dissection and prediction of cancer immunotherapy response

by

Alvin Shi

B.S., Chemistry
Duke University (2013)

Submitted to the Department of Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author .....
Department of Computational and Systems Biology
August 30, 2021

Certified by .....
Manolis Kellis
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Christopher B. Burge
Professor of Biology
Director, Computational and Systems Biology Graduate Program

Computational dissection and prediction of cancer immunotherapy response

by

Alvin Shi

Submitted to the Department of Computational and Systems Biology
on August 30, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

Abstract

Checkpoint blockade immunotherapies have transformed the standard of care and outcomes for many cancer types; however. more than 60% of patients still do not experience a durable clinical response from these treatments. To address this problem, the development of novel biomarkers and more effective combinatorial therapies are needed. In this thesis, we first explore and validate the use of extracellular vesicular (EV) RNA as a potential biomarker for immunotherapy response. We discover differentially expressed genes and pathways within the plasma-derived EV RNA that is concordant with known biology. We also show that mutational information contained within EV RNA can stratify responders and non-responders. We leverage a Bayesian probabilistic model to deconvolve the tissue-of-origin of EV RNA transcripts, allowing greater interpretability for differentially expressed genes and pathways. Next, we performed large-scale epigenomics profiling in two cohorts of immunotherapy patients, and we discovered a non-responder enhancer signature that is lost in responders. Many genes contained within this epigenetic signature are associated with immunotherapy resistance, and we reasoned targeting this signature with acetylation-reader bromodomain inhibitors would allow suppression of multiple resistance mechanisms at once. We show that bromodomain inhibitors exhibit considerable synergism with anti-PD1 in reducing tumor volume in murine melanoma transplantation models, and this synergism also improves anti-tumor killing by tumor infiltrating lymphocytes. Using the same cohort, we also identify 189 peaks with differential activity in both the responders and non-responders, and we show these peaks are potentially predictive biomarkers of immunotherapy response. Finally, we leverage three transgenic mice lines to investigate the effect of T-cell receptor repertoire on cell fate commitment by CD4+ SP T-cells into either the thymic conventional (Tconv) or thymic T regulator (tTreg) lineages. We show based on overlap and machine learning analysis that T-cell receptors are not the sole determining factor in Tconv vs. tTreg cell fate decisions. Together, these projects offer new biomarkers and novel combinatorial treatment options for checkpoint blockade immunotherapies.

Thesis Supervisor: Manolis Kellis
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I have been blessed with the opportunity to work with and learn from some of the best and brightest during my Ph.D. experience. I am grateful to all the individuals who have contributed towards my growth as an sctienst and as a human being.

First, I would like to thank my Ph.D. advisor Manolis Kellis for his support, mentorship, and guidance throughout my Ph.D. His mentorship has assisted me tremendously during the worst times and the best of times, and I could not have asked for a better mentor for my Ph.D. I will always strive to share the same level of passion for science that Manolis possesses. I am also grateful for the friendly and constructive lab environment that Manolis has created. I would also like to thank my committee members Stefani Spranger and Stefanie Jegelka for their advice and feedback.

Second, I would like to thank past and present members of the Kellis lab. I especially would like to thank Yue Li, Yongjin Park, and Liang He who offered great mentorship to me early in my graduate career. I would like to thank Li-lun Ho and Kiki Galani for providing experimental support. I would also like to thank Zhizhuo Zhang, Yaping Liu, Vinod Yadav, Shahin Mohammadi, Irwin Jungreis, Lei Hou, Jackie Yang, Remy Tuyeras, Peter Nguyen, Maria Kousi, Kai Kang, Brad Ruzicka, Nicola Rinaldi, and Patty Purcell for providing sage advice and guidance. I would like to thank the past Ph.D. students from the Kellis lab - Khoi Nguyen, Xinchen Wang, Abhishek Sarkar, Max Wolf, Angela Yen, Kunal Bhutani, and Bob Altshuler - for providing camaraderie and support. I would also like to acknowledge my former UROPs and summer students - Isabel Chien, Roger Jin, Larry Zhang, and Wesley Lo - for teaching me much about mentorship. Finally I like to thank Carles Boix, who started this long journey with me six years ago, and whose friendship and companionship has made the Ph.D. journey more bearable.

I also want to thank members of the Boland, Rai, and Weng labs, without whom this thesis would not exist. In particular, I would like to thank Genevieve Boland, for being an excellent collaborator and a caring mentor. From the Boland lab, I would like to thank Gyulnara Kasumova, Dennie T. Frederick, and William Michaud

for driving the experimental part of our collaboration forward. I also want to thank Kunal Rai for being a patient collaborator, a fastidious scientist, and a great mentor. I would also like to thank Mayinuer Maitituoheti and Ming Tang for being awesome collaborators to work with. Lastly, I would like to thank Nan-ping Weng for giving me a chance to work in his lab in 2013, and for being a source of advice regarding both science and life over the years.

Many other people have made this thesis possible. I would like to thank Jacqueline Carota, Christopher Burge, and the Computational Systems Biology Department for the opportunity to study and work at MIT. I would like to thank the members of the CSB class of 2015 for providing camaraderie during a period of growth and change. I would also like to thank my Duke friends - Maggie Chu, Tim Visutipol, Jeffrey Ding, and Tony Cao - for providing entertainment and companionship over the past six years.

Finally, I like to thank my parents, Wenyan and Zhongkai, for their constant support and confidence in me over the years. I am forever indebted to them for the sacrifices they have made on my behalf, and I shall never forget the unconditional love they have showered upon me since the beginning. I cannot imagine doing any of this without their love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of several key topics necessary to contextualize this thesis. It starts with an overview of cancer immunotherapy, extravesicular profiling and analysis, epigenetic profiling and analysis, TCR repertoire profiling and analysis, and computational methods. It concludes with an outline of the thesis in section 1.2.

1.1 Background

1.1.1 Melanoma

Melanoma is a cancer of melanocytes, UV-absorbing, pigment-producing cells found throughout the body. Cutaneous melanoma is the most common in the Western world, with a global incidence of 15-25 per 100,000 individuals. In 2019, 96,480 new cases of melanoma was diagnosed, and 7,230 people died in the U.S alone [1]. There are two types of melanin produced by melanocytes: black pigment eumelanin and a red/yellow pigment pheomelanin. The ratio of eumelanin to pheomelanin determines skin color, and, since eumelanin is a better UV absorber, the level of melanoma cancer risk. Also, increased melanoma risk is associated with physical characteristics, such as blond or red hair and light eye color.

Melanomas carry the highest mutation rates of any cancer, and they contain an overwhelming number of UV-induced mutations, such as C>T or G>T transitions.

A familial background occurs in 8% of patients with melanoma, a significant fraction of which carry high-risk, high-penetrance mutations in the *CKDN2A* locus; however, the vast majority of melanomas are sporadic, arising from low penetrance, low risk alleles. Population-level Genome Wide Association Studies (GWAS) have revealed *MC1R*, a master regulator of pigmentation transcription factor *MITF*, as a risk gene. *MITF* alone is amplified in 4-21% of melanomas [2]. Besides *MITF*, there are a number of other driving mutations that tend to converge on recurrently mutated genes, including genes related to proliferation (*BRAF*, *NRAS*, and *NF1*), growth and metabolism (*PTEN* and *KIT*), resistance to apoptosis (*TP53*), replicative lifespan (*TERT*), and cell cycle (*CDKN2A*). Another key pathway with recurrent mutation is the MAPK pathway. which is involved in controlling cell proliferation and survival; *MAPK* mutational events are associated with nearly 70% of melanomas [1].

1.1.2 Checkpoint blockade immunotherapy

Checkpoint blockade immunotherapies are a revolutionary class of immunothearpies that have transformed treatment options for a number of cancer types, including melanoma, colorectal cancers, and non-small cell lung cancers. They work by reversing negative immune regulation induced by expression of ligands against inhibitory receptors on T-cells. These inhibitory receptors include CTLA4, PD1, TIM3, BTLA, VISTA, and LAG-3 [3]. Since the first 2011 trial of ipilimumab (anti-CTLA4) [4], checkpoint blockade immunotherapies have provided immense relief for 40% of the patient population experiencing durable clinical responses to these treatments [5]. The success of checkpoint blockade immunotherapies earned its inventors, Jim Allison and Tasuku Honjo, the 2018 Nobel Prize in Physiology or Medicine.

Checkpoint blockade immunotherapies have also transformed how cancers are managed. Clinical trials are now far more cognizant of the role that immune systems play in mediating successful treatments. This is in contrast with previous thinking, in which preclinical cancer drugs were routeinly tested on cultured cancer cell lines or immune-compromised mouse lines; current preclinical models are widely using immune-competent animals [5]. Checkpoint blockade immunotherapies can have a

delayed effect after an initial increase in the size of tumor metastases. These pseudo-progressions may reflect the increased time it takes to activate the immune system and effect an anti-tumor response. This understanding has resulted in the development of immune RECIST (iRECIST) evaluation system, which has replaced the traditional radiological evaluation criteria RECIST-1.1 for immunotherapy patients [5].

The poor overall efficacy of checkpoint blockade immunotherapies has led to a search for actionable biomarkers to stratify the patient population, with the hopes of engendering overall better response for a subset of patients with positive biomarker status. One such biomarker is Tumor Mutational Burden (TMB), the sum of synonymous and non-synonymous mutations present within a patient's tumor. A meta-analysis of 27 cancer types showed that overall response rate was correlated with TMB [6]. Several studies, including that of KEYNOTE-061 [7], have confirmed the predictive value of TMB in the context of checkpoint blockade immunotherapy. There are also epigenetic changes associated with TMB. Cai et al. showed that association between TMB and DNA methylation have the potential to serve as complimentary biomarkers in the context of non-small cell lung cancer (NSCLC) immunotherapies. Specifically, they showed that high TMB NSCLCs had methylation aberrations and copy number changes, with the latter offering predictive potential [8].

Another biomarker that showed stratification between responders and non-responders to checkpoint blockade immunotherapy is neoantigen load. Neoantigen predictions are done computationally, with the focus being on major histocompatibility complex binding of peptides based on anchor residue identities. In general, predictions made by neoantigen load are not as accurate as those made by TMB [9]. Today, neoantigens can be measured by the difference in predicted MHC I binding affinities between the mutant and wild-type peptide, generating an index known as the differential agretopicity index (DAI). A high DAI suggests that the mutant peptide significantly increased in binding affinity relative to wild-type and can thus generate a stronger immune reaction. Studies have shown that DAI is superior to TMB for stratifying patient response to checkpoint blockade immunotherapies [10, 11].

In addition to TMB and neoantigen load, other predictive biomakers include tu-

mor PD-L1 expression, immunophenotyping of the tumor microenvironment, specific mutations exhibited by the tumor, diversity of immune repertoire within the tumor microenvironment, and a number of blood-based molecular biomarkers [12].

1.1.3 Extravesicular profiling and analysis

Extracellular vesicles (EVs) are lipid enclosed membranes released by all mammalian cells and contain cargo representative of their intracellular origins. EVs are found in all types of bodily fluids, including plasma, serum, blood, saliva, urine, and amniotic fluid. EVs refer to a generalized umbrella term that encompass microvesicles (150-1000nm), exosomes (40-150nm), apoptotic bodies (100-5000nm). EVs can contain a variety of biomolecular macromolecules, including DNA, RNA, miRNA, proteins, and lipids. Different types of EVs may enrich for specific macromolecules, such as the observed enrichment for miRNAs in exosomes [13].

EVs have recently emerged as attractive biomarkers for immunotherapy resistance due to the variety of intercellular communication roles EVs play in the tumor microenvironment. In a recent study, Chen et al. showed the presence of PD-L1 on melanoma derived exosomes and demonstrated that higher levels of circulating exosomal PD-L1 correlated negatively with response to checkpoint blockade immunotherapies [14]. Several additional have reported the presence of cancer-associated miRNAs responsible for generating immune tolerance and suppression in the microenvironment inside exosomes [13]. These findings collectively show that EVs - exosomes in particular - play a causal role in generating an immunosuppressive microenvironment that makes response to checkpoint blockade immunotherapies more difficult.

EVs have many natural characteristics that make them attractive biomarkers. Since they are secreted by all cell types, profiling EVs offers a snapshot of many different cell types at once, though this raises additional problems in deconvoluting the mixed profiles. Given that EVs are found in all bodily fluids, this makes EVs a minimally invasive or non-invasive biomarker. Moreover, EVs protect the biomacromolecules inside with a lipid bilayer, offering protection for RNAs and proteins against enzymatic activity. Finally, EV levels are increased during tumor progression, and

this increased EV shedding may offer potential for early diagnosis and monitoring of cancers [15].

1.1.4 Epigenetic profiling and chromatin state analysis

Epigenetic modifications control gene expression patterns in a cell, and these modifications are stable and somatically heritable. An epigenotype is defined as the ensemble of DNA methylation states, histone modifications, histone variant composition, and lncRNAs in a particular locus [16]. DNA methylation is the most well-studied of the epigenetic modifications. A repressive epigenetic signal at promoters, it occurs when carbon 5 of CpG dinucleotides are methylated. After cell division, the state of DNA methylation is maintained by DNA methyltransferase 1 (DNMT1), which also plays a key role in imprinting gene expression. DNA methylation is highly dysregulated in human cancers, with the loss of DNA methylation as one of the first epigenetic changes described in human cancers [17].

Chromatin modifications involve covalent post-translational modifications (PTM) of the amino-terminal histone tails by the addition or deletion of acetyl, methyl, phosphate, or other groups. These chromatin modifications impact gene expression via their effects on chromatin structure or the attraction/repulsion of binding proteins such as PTM writers and readers. Since chromatin is made up of DNA packed around histone cores, the folding patterns of DNA - as determined by chromatin PTMs - cause downstream changes in gene expression. Different PTMs are associated with different outcomes. For example, H3 lysine trimethylation (H3K3me3) is associated with promoter regions; H3 lysine 4 monomethylation (H3K4me1) is associated with active enhancer regions; H3 lysine 36 trimethylation (H3K36me3) is associated with transcribed regions; H3 lysine 27 trimethylation (H3K27me3) is associated with Polycomb repression; H3 lysine 9 trimethylation (H3K9me3) is associated with heterochromatin regions; H3K27ac and H3k9ac is associated with enhancer activation and promoter regions [18]. Chromatin modifications have a significant role in cancer. For example, For example, EZH2, a polycomb repressive complex 2 member with a methyltransferase and reader proteins that recognize H3K27me3, is overexpressed at both the

transcriptional and translational levels in many human cancers [17].

Major studies in the field now profile several chromatin marks in conjunction with one an other, such as done by the ROADMAP Epigenomics project [18]. A major issue in the combinatorial analysis of chromatin mark is how to reduce the complexity of the dataset by clustering it. The ChromHMM algorithm [19] is a multivariate Hidden Markov Model trained on combinatorial chromatin state data. ChromHMM learns a set of *de novo* chromatin state definitions, and then assigns each location in the genome to an instance of each state. One can then use the emission matrix (a matrix relating chromatin states with the combinatorial marks that reside in them) and genomic annotations to assign functional categories, such as enhancers or polycomb repressed, to individual chromatin states. One can then compare chromatin states between two different conditions using the program Epilogos. ChromHMM and Epilogos are extremely powerful tools for interpreting combinatorial chromatin state data and associating it with downstream functional effects.

1.1.5 TCR repertoire profiling and analysis

T-cell receptor activation requires the interaction of the T-cell receptor (TCR) with the antigen-major histocompatibility complex (MHC) molecules. TCRs are diverse heterodimers consisting of an α and β chain (expressed by the vast majority of T-cells) or γ and δ chains (expressed by mucosoal T-cells). The TCR consists of a variable region, crucial for antigen recognition, and a constant region, important for structural reasons. The variable region of TCRα and TCRδ chains consist of a number of variable (V) and joining (J) gene, while the TCRβ and TCRγ chains are additionally encoded by a set of diversity (D) genes. During V(D)J recombination, one random allele of a gene segment is recombined with others to form a functional variable region. Further recombination of the variable region with a constant gene segment creates a functional TCR chain transcript. Random nucleotides are added/deleted at the junction between gene segments, leading to additional diversity. Each TCR chain contains three hypervariable loop regions known as the complementarity determining regions (CDR1-3). CDR1-2 are coded by the V gene and are necessary for the interaction

between TCR and the MHC complex, whereas CDR3, a highly diverse region at the junction between the V/D/J genes, is necessary for the interaction between the TCR and the peptide-MHC complex [20].

The collective TCRs of an individual is known as the TCR repertoire or the TCR profile. The TCR profile can change with disease status, age, and a number of other physiological factors. This is driving interest in profiling the immune repertoire under different disease conditions, such as cancer, autoimmune, or infectious disease. In cancer, T-cells can kill tumor cells upon recognition tumor-specific antigens. Several studies have tried to identify the specific T-cell clonotypes responsible by analyzing the tumor infilitrating lymphocyte repertoire [21]. The major challenge to analyzing and interpreting immune repertoires in the context of disease is the large overall diversity of the T-cell receptor repertoire. VDJ recombination can yield a repertoire as large as 10^15 to 10^20 unique TCR chains. The actual diversity present in adult humans is around 10^13 different clonotypes. Moreover, individual TCRs are quite rare at the population level [20]. Today, next-generation sequencing methods are able to capture millions of TCR sequences from a given individual, allowing us to probe the overwhelming diversity of T-cells.

1.1.6 Bayesian inference and modeling

Bayesian statistics is a statistical methodology for data analysis centered on Bayes' theorem, in which the data set y and the data parameters θ are related to each other in the equation described in Eqn. 1.1. The typical Bayesian inference problem undergoes three steps: 1) capturing information regarding a given parameter in the prior distribution $p(\theta)$, 2) determining the likelihood function which relates information about the observed parameters from the data, and 3) updating both the likelihood and prior distribution with Bayes' rule to create the posterior distribution. The major subjective element of Bayesian inference is the choice of priors. Statisticians have the choice between informative priors, weakly informative priors, or diffuse priors in terms of model specification. Each choice has its own advantages and disadvantages, for example, diffuse priors will often produce results more aligned to that of the like-

lihood, and regardless of the choice, prior specification will have an impact on the posterior estimates [22].

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.1}$$

The Bayesian emphasis on estimating the entire posterior distribution of model parameters makes inference computationally challenging for sophisticated models, often due to the high dimensionality of the calculations. A key algorithm in tackling this problem is Markov chain Monte Carlo (MCMC) - a technique for sampling from a probability distribution. MCMC involves two separate steps: 1) obtaining a set of parameter values from the posterior distribution using Markov chains, and 2) obtaining a distributional estimate of the posterior with sampled parameters using Monte Carlo integration. There are many variants of MCMC, including Metropolis Hastings, Gibbs sampling, Hamiltonian Monte Carlo, in which a transition kernel for the Markov chain is defined such that the resulting stationary distribution is the distribution of interest.

1.1.7 Identification of differentially expressed genes

Gene expression and other high-dimensional datasets (e.g., from ChIP-seq studies) are often compared between multiple conditions to yield the subset of genes or markers that show differential expression relative to certain conditions. These powerful type of studies have yielded many fundamental insights into basic biology as well disease pathogenesis [23]. Study designs for these studies can be complex, with the potential for several experimental factors varying over multiple levels. Modeling such complex designs with statistical rigor is now the domain of specialized packages (often in R) designed to handle large-scale bioinformatics data with complex designs.

One such package is *limma*, which operates on a matrix of expression values, where each row represents a gene and each column corresponds to a sample. limma fits a linear model to each row of the data, and it allows sharing of information between samples to model correlations due to repeated measures and other causes. *limma* also has a built in parametric empircal Bayes module, allowing moderation of residual

variances by allowing linear models to borrow strength across genes [23]. The gene-level estimated variance becomes a compromise between gene-wise estimators and global variability (estimated by pooling the ensemble of all genes). This approach is beneficial in experiments with small sample sizes, ensuring inference is stable even in the case of low number of replicates.

1.2 Thesis outline

Chapters 2 and 3 of the thesis concerns the use of extravesicular (EV) RNA as a predictive biomarker in checkpoint blockade immunotherapy, with chapter 2 providing a set of predictive genes and pathways and chapter 3 providing a novel deconvolution algorithm for inferring tissue-of-origin for plasma-derived EVs. In chapters 4 and 5, we provide a detailed investigation into the epigenetic changes that stratify responders and non-responders to immunotherapy, with chapter 4 focused on *in vitro* and *in vivo* experimentation and chapter 5 focused on generating predictive biomarkers. Finally, chapter 6 investigates the impact of T-cell receptor repertoires on theymic T regulatory vs. T conventional T-cell differentiation.

1.3 Previously published work

The work in chapters 2 and 3 derived from work published by Shi et al. [24]. The work in chapter 6 appeared in work published by Ko et al. [25]. Work in chapters 4 and 5 derive from manuscripts currently under review for publication. Some passages in this thesis have been quoted verbatim from the above sources.

1.4 Collaborators and contributions

This thesis would not have been possible without the close collaboration of many people. In particular, chapters 2 and 3 would not have been possible without the help of Gyulnara Kasumova, William A. Michaud, Jessica Cintolo-Gonzalez, Marta

Diaz-Martinez, Jacqueline Ohmura, Arnav Mehta, Isabel Chien, Dennie T. Frederick, Sonia Cohen, Deborah Plana, Douglas Johnson, Keith T. Flaherty, Ryan J. Sullivan, Manolis Kellis, and Genevieve Boland. The work in chapters 4 and 5 would not have been possible without the help of Mayinuer Maitituoheti, Ming Tang, Lilun Ho, Christopher Terranova, Kyriaki Galani, Emily Z. Keung, Caitlin A. Creasy, Anand K. Singh, Apoorvi Chaudhri, Nazanin E. Anvar, Jiekun Yang, Ayush T. Raman, Sharmistha Sarka, Shan Jiang, Jared Malke, Lauren Haydu, Elizabeth Burton, Michael A. Davies, Jeffrey E. Gershenwald, Patrick Hwu, Alexander Lazar, David Liu, Jamie H Cheah, Christian K. Soule, Chantale Bernanthez, Jennifer Wargo, and Kunal Rai. Finally, the work in 6 would not have been possible without Annette Ko, Masashi Watanabe, Thomas Nguyen, Achouak Achour, Baojun Zhang, Xiaoping Sun, Qun Wang, Yuan Zhuang, Nan-ping Weng, and Richard J. Hodges.

Chapter 2

Extracellular vesicles are predictive biomarkers of immunotherapy response

2.1 Introduction

Historically, blood-based biomarkers for immunotherapy focused on cell-free DNA (cfDNA) or circulating tumor cells [26, 27], which solely reflect tumor-based properties and not changes in the immune system during treatment. To improve prediction and tracking of ICI resistance, simultaneous capture of transcriptomic features from both the tumor and immune system [28] is critical.

Extracellular Vesicles (EV) are produced by many cell types including tumor and immune cells, which contain a sub-transcriptome of their cell-of-origin. EVs are involved in oncogenesis, immune modulation, and serve as communicators of genetic and epigenetic signals. In cancers, tumor-secreted EVs modulate the tumor microenvironment, elicit anti-tumoral immune responses [29], and plasma-derived EV transcripts are markers of anti-tumor immune activity [30]. EVs are also secreted by many immune cell-types implicated in ICI response including CD4+/CD8+ T-cells, dendritic cells, regulatory T-cells, and macrophages [31, 32, 33, 34]. During

tumor progression, both overall and tumor-specific EVs are elevated in plasma10. We hypothesize that bulk, non-enriched plasma-derived EVs capture both tumor-derived EVs and non-tumor-derived EVs reflecting tumor-intrinsic and non-tumor signals. We analyzed pre-treatment and on-treatment peripheral blood-derived bulk EV RNA from 50 patients with metastatic melanoma (discovery cohort; N=33 responders, N=17 non-responders) treated with ICI via transcriptome microarray (Fig. 2-1a and Table S1). A subset of patients had post-treatment plasma samples (N=15) and tumors (N=26). Additionally, we profiled four melanoma cell lines and paired EV. We validated results from the discovery cohort using an independent validation cohort of 30 patients (N=14 responders, N=16 non-responders) using extracellular vesicle RNA-seq (evRNA-seq). To integrate transcriptomic data from two sequencing platforms, we utilized a multi-pronged statistical strategy to minimize cross-platform variance (Fig. 2-1b).

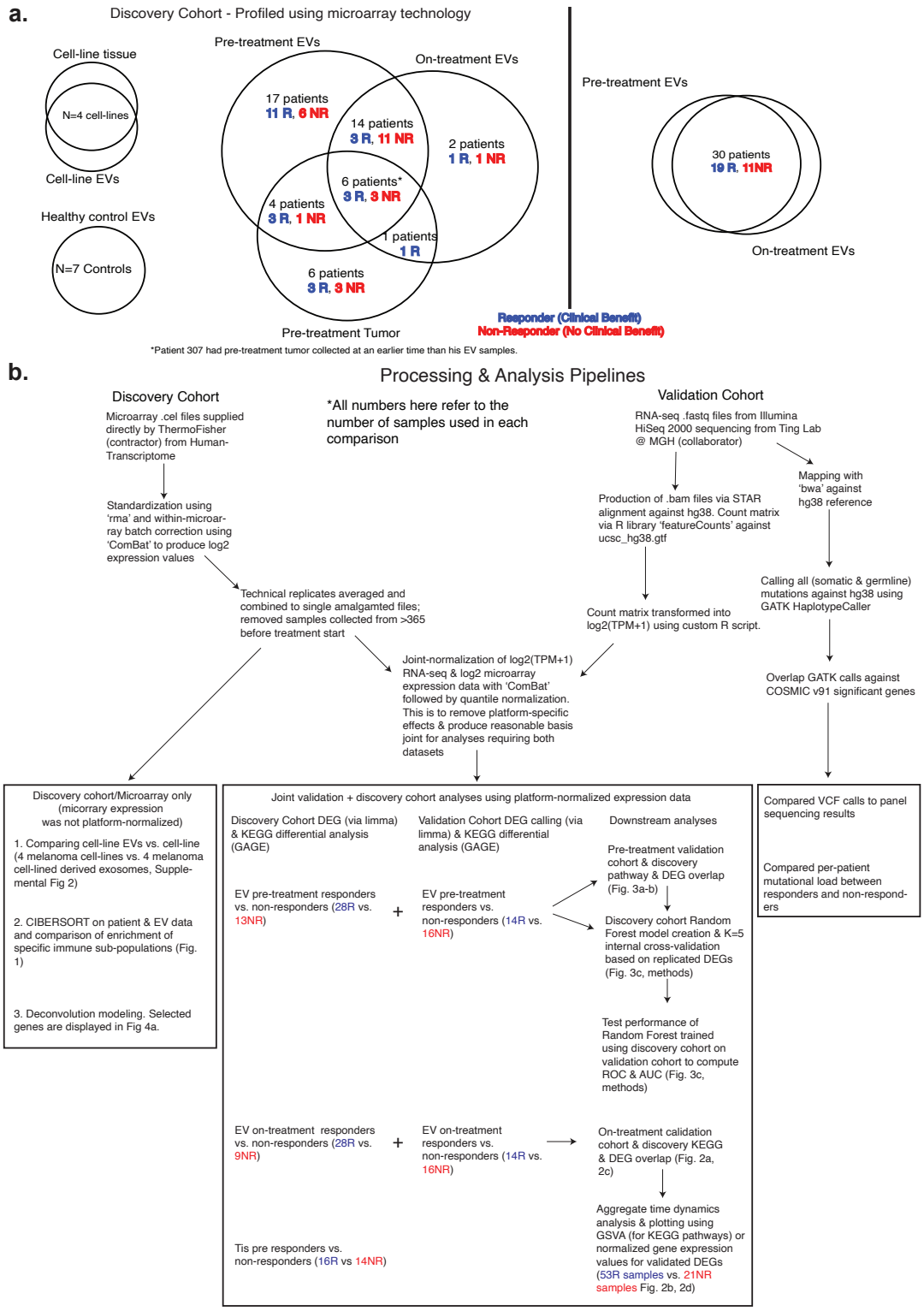


Figure 2-1: **EV cohort description and processing methodology**. (a) Outline of both the discovery and validation study cohort (b) Outline of the processing and analysis steps undertaken to generate the major results in the paper

2.2 Methods

2.2.1 Tumor cell lines

Melanoma cell lines (A375, RPMI 7951, SK-MEL-30, SK-MEL-2, MeWo) were purchased directly from the American Type Culture Collection (ATCC) and maintained in culture per ATCC recommendations. The A375 cell line was maintained in Dulbecco's minimal essential medium (DMEM), whereas RPMI 7951, SK-MEL-30, SK-MEL-2, and MeWo cell lines were cultured in RPMI-1640 media. All growth media consisted of media supplemented with 10% FBS, and 100 I.U./mL penicillin, 100 µg/ml streptomycin, and 0.292 mg/mL L-glutamine. Cells were grown on plates and incubated at 37°C with a humidified atmosphere of 5% CO2 in air.

2.2.2 Patient samples and plasma isolation

Serial tumor and blood samples were collected from patients with melanoma under protocols approved by the Institutional Review Board at the Massachusetts General Hospital. Patient samples were linked to clinical data in a retrospective electronic health records database. Blood was collected in sodium citrate cell preparation tubes, with plasma isolated after centrifugation at room temperature for 25-30 minutes at a relative centrifugal force of 1800. Plasma was then frozen and stored at -80°C until use. Peripheral blood mononuclear cells (PBMCs) are collected from the same sodium citrate tubes, washed in phosphate buffered saline, resuspended in DMSO with 90% FBS, slow frozen at 1°C per minute, and stored at -80°C until use.

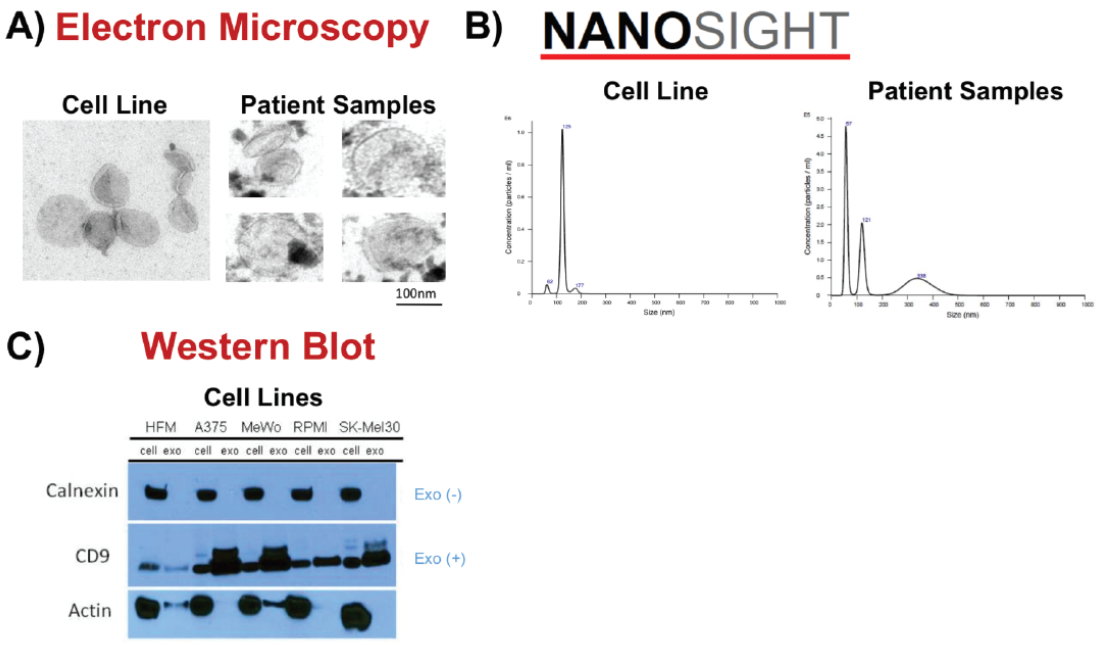


Figure 2-2: ***In vitro* EV characterization**.(a) Electron microscope images of representative cellline and patient-derived EVs. (b) Nanosight analysis of a representative cell line sample (RPMI) and patient sample. (c) Western blot of 5 cell line/paired EV proteins with no calnexin within EV, but high levels of CD9 within EV as compared to paired cells.

2.2.3 Isolation of EVs

Cell lines: When 150mm plates reached between 50 – 70% confluence, depending on the doubling time of the cell line, the media was replaced with the appropriate media containing EV-depleted FBS and the media was harvested after 48 hours. EVs were isolated from cell-conditioned media using serial centrifugation to remove cellular debris and filtration with a 0.8µM or smaller filter (Millipore) followed by ultracentrifugation as previously described [35]. Briefly, cell-conditioned media was collected and centrifuged at 3,000 revolutions per minute (rpm) for 10 minutes at 4°C after which the supernatant was decanted and filtered using a 0.45µM or 0.8µM filter. The filtered supernatant then underwent ultracentrifugation at 150,000xg for 120 minutes. The pellet was then washed in PBS and underwent a second round of ultracentrifugation for 90 minutes. The EVs were then resuspended in cold RPMI media on ice, and then frozen and stored at -80°C.

Plasma: All EV for RNA transcriptomic analysis from plasma were isolated from 1ml of plasma using column isolation (Qiagen exoRNAeasy midi kit). Column isolation using the Qiagen exoRNAeasy serum/plasma midi kit resulted in direct isolation of EV RNA (10–30ng RNA/ml). Approximately 1 in 10 patients had an additional 1ml of plasma isolated in parallel using ultracentrifugation (for quality control studies only, Fig. 2-2). For ultracentrifugation, plasma was thawed and filtered through a 0.2µM filter prior to ultracentrifugation as described for cell lines, above.

2.2.4 Nanoparticle tracking analysis, electron microscopy, Western Blot analysis

Nanoparticle tracking analysis: Nanoparticle tracking analysis (NTA) using the Nanosight LM10 (Malvern) was employed in order to assess size distribution and concentration of particles in cell cultures and selected patient samples. Samples were diluted in PBS either 1/500 or 1/1000 according to Nanosight instruction manual.

Transmission electron microscopy: Electron microscopy was used to confirm the presence of EVs in cell culture and selected patient samples by size and morphology. Isolated EV suspensions were diluted 2:1 in 1xPBS and 8-10µl aliquots of each diluted sample were placed on formvar-carbon coated Ni mesh grids; samples were allowed to adsorb for 15 minutes. Following adsorption, grid preparations were placed on drops of primary antibody CD9, rabbit monoclonal (D801A), Cell Signaling #13174, diluted 1:25 (dilutions made in DAKO antibody diluent). Samples were allowed to incubate in primary antibody for at least 1 hour at room temperature, then rinsed on drops of PBS and incubated in drops of a secondary gold conjugate at least 1 hour at room temperature: Goat anti-rabbit 10nm IgG (Ted Pella #15726). Grids were then rinsed on drops of 1xPBS, then distilled water, contrast-stained for 10 minutes in droplets of chilled tylose/uranyl acetate, and air-dried prior to examining in a JEOL JEM 1011 transmission electron microscope at 80 kV. Images were collected using an AMT digital camera and imaging system with proprietary image capture software (Advanced Microscopy Techniques, Danvers, MA).

Protein isolation from cell lines or EVs: protein was isolated using RRIPA buffer supplemented with protease inhibitors as described in Wright 198938. EVs were lysed directly in a sample buffer containing SDS and DTT.

Protein Quantification: Protein concentration of EV samples was determined using the DC™ Protein Assay (Bio-Rad) according to the manufacturer's protocol.

Western Blot: Western blot was performed on samples of EV isolated from cell lines (Fig. 2-2) to confirm their presence in the samples. Antibodies for EV markers CD939 were used as a positive marker given their consistency in expression across EV samples, whereas calnexin, an endoplasmic reticulum protein, is not found in EV and was used as a negative control. Actin was used as a loading control for cell lines, using standard protocols.

2.2.5 RNA extraction and sequencing

Cell lines: Cells were trypsinized, washed in media for trypsin deactivation, then washed in PBS, counted, and pelleted. The cell pellet was resuspended in Trizol at a concentration of 5 x 106 cells per 1mL and RNA isolated according to manufacturer's instructions (Invitrogen). RNA was then quantified using a NanoDrop spectrometer (Thermofisher).

Cell line derived EV: RNA was extracted from EV using Qiagen exoRNeasy kit.

Tumor: Formalin-fixed tissue was analyzed to confirm that viable tumor was present via standard hematoxylin and eosin (H & E) staining. DNA was extracted from snap frozen tissue using Qiagen's AllPrep DNA/RNA FFPE Kit.

Peripheral blood derived/plasma EV: Total EV RNA was extracted from thawed patient plasma using the Qiagen exoRNeasy Serum/Plasma Midi Kit as per the manufacturer's protocol. Briefly, 1mL of plasma was thawed and filtered using a 0.8µM or smaller filter (Millipore). The sample was then mixed with binding buffer and placed on a spin column. After a wash step, the EVs were lysed on the column using QIAzol and the eluate was then treated with chloroform to achieve phase separation. The aqueous phase was combined with 100% ethanol and then underwent column extraction with wash steps and final elution of EV RNA in RNase-free water.

Extracellular vesicle RNA-seq (evRNA-seq): Sequencing was performed using an Illumina HiSeq 2000 system in the Ting lab at Massachusetts General Hospital. RNA was prepared with the SMARTer Stranded Total RNA-Seq Kit v2 (Pico Input Mammalian) per supplier instructions. For EV sequencing, RNA with a RIN of 4 or less was not subjected to fragmentation. Libraries were purified using AMPure beads and ZapR v2 for the depletion of ribosomal cDNA.

2.2.6 Discovery cohort microarray processing

We performed microarray analysis (via a subcontract to Thermo Fisher Scientific) utilizing Applied Biosystems GeneChip Human Transcriptome Array 2.0. Raw .cel files provided by ThermoFisher were read using the R package 'oligo'. We performed background subtraction, quantile normalization, and summarization via the Robust Multichip Average (rma) algorithm using the R package 'oligo' [36]. We used the R package 'pd.hta.2.0' to provide functional annotations for the probes. We filtered all probes that did not map to an annotated gene, as well as duplicate probes. We corrected for batch effects using the 'ComBat' algorithm from the R 'sva' package [37]. For analyses that required validation by evRNA-seq, the resulting matrix was further corrected for platform-specific effects using 'ComBat' and was quantile normalized (fully described in the RNA-seq processing section). Analyses that only utilized microarray data (Fig. 2-1b) was performed using the non-platform-corrected data, since cross-platform normalization with evRNA-seq log2(TPM+1) may unduly bias microarray-only analyses.

2.2.7 Validation cohort RNA-seq processing

Raw Illumina .fastq files were first filtered using 'fastp' [38] with default settings and then aligned to human hg38 transcriptomic reference using STAR v2.4.1 with default settings. To summarize the count data from the aligned .bam files, we utilized the function 'featurecounts' function from the R package 'Rsubread' [39]. Next, we derived log2(TPM+1) values using a custom R function. To make the RNA-seq count

matrix co-analyzable with our microarray data, we essentially treated the resulting ComBat-corrected log2(TPM+1) matrix as an additional microarray batch. To minimize platform-based effects, we concatenated log2(TPM+1) matrix with the microarray expression matrix to make a joint data matrix. Next, we used the 'ComBat' [37] algorithm from the R 'sva' [37] package to perform a "batch"-correction step where the two platforms were treated as two distinct batches. Existing biological treatments were included in the 'ComBat' normalization step as an input argument to preserve true biological effects during the platform-correction step. Next, we quantile normalized the resulting data matrix using R 'preprocessCore' library. The resulting platform-normalized matrix was then split back into evRNA-seq and microarray matrices for relevant downstream analyses.

2.2.8 Differential expression analysis

To calculate differential expression in our microarray-based discovery and our evRNA-seq validation cohort, we used the R 'limma' package to compute the p-values that corresponded to the comparisons [40]. For the samples that had multiple replicates, we modeled biological replicates as a random effect. Confounding variables such as age and prior immunotherapy treatment were tested for association against ICI response and did not exhibit significant associations, and, as a result, they were not included as covariates. To find the top differentially expressed genes, we utilized limma's Empirical Bayes linear modeling framework [40]. Differentially expressed genes were defined by 1.5 log-fold change between responders and non-responders and a nominal p-value cutoff of p=0.1 from limma. In order to be considered validated, a gene has to fulfill both the nominal p-value cutoff and log-fold change cutoff in both the validation and discovery cohorts. We note that this nominal p-value cutoff is ordinarily insufficient to control for false positive discoveries in a single cohort study; however, we require explicit confirmation for putative discovery cohort DEGs in our validation cohort, thus the combined false positive rate for a gene to be both falsely discovered and falsely validated is substantially lower than what the nominal p-value cutoff would suggest.

2.2.9 Concordance and differential pathway analysis

Concordance analysis was performed by first binarizing the mean expression values of either the cell-line or patient data based on a 1.5 log expression cutoff using non-platform-corrected microarray data. If a gene is either present or absent in both groups, it is labeled as concordant. We averaged the expression signal from two pre-treatment patients present in the pre-treatment time point; the post-treatment replicates were considered separately since they were from separate time points. To find differential pathways that are different between patient tumors and patient EV, we used gene-set enrichment between responders and non-responders with default parameters and GO biological processes database11. To find the canonical (C2) MSigDB [41] pathways that are significantly different between responders and non-responders in both the discovery and validation cohorts, we utilized the Gene Set Variation Analysis (GSVA)16 program with default settings to generate per-patient GSVA scores (a normalized statistic summarizing enrichment relative to the entire cohort analogous to ssGSEA scores) across our platform-corrected discovery and validation cohorts datasets. We then used a Mann-Whitney U-Test to test for differential GSVA scores between responders and non-responders. Similar to the rationale used for DEG analysis, we utilized a nominal p-value cutoff of 0.1 to flag differential pathways. A pathway was considered validated if it achieved significance in both the discovery and validation cohorts.

2.2.10 Survival analysis and time-series analysis

To compute and plot the Kaplan-Meier curves, we utilized overall survival data and censoring information as inputs into the Kaplan-Meier computation and plotting functions in the R package 'survminer'. To generate the time-series plots, we utilized the R Gene Set Variation Analysis (GSVA)16 package to generate a normalized enrichment score for each sample for target KEGG pathways. To generate the per-patient time dynamic plots, we normalized discovery cohort expression data by subtracting the expression of a patient's first sample from all samples from the same patient.

2.2.11 Building a predictive classifier

To build a random forest predictive classifier from our selected pre-treatment DEGs, we first merged the post-platform-corrected evRNA-seq and microarray data into a single combined matrix with 71 samples (N=41 from discovery cohort, N=30 from validation cohort). In order to reduce potential bias from platform and batch effects not corrected for by our ComBat correction steps, we randomized the selection of samples in the training and testing cohort by including samples from both sequencing platforms in the training and testing groups. To accomplish this, we randomly selected N=30 samples to be (reflecting the size of our discovery cohort) the size of the held-out testing cohort and N=41 samples (reflecting the size of our validation cohort) to be the training set. In order to minimize variability due to random sampling and potential non-linear bias between platforms that remains uncorrected for by ComBat, we ran 100 trials for our machine learning pipeline, each with a N=30 random selection of pre-treatment samples as testing set and the remaining N=41 random samples. Note that the randomly partitioned training and testing sets will typically contain both evRNA-seq and microarray samples and may contain different proportions of responders and non-responders depending on the partitioning.

Within each trial, we first conducted K=5-fold internal cross-validation within the training set using the function 'StratifiedKFold' from the python library 'sklearn' to optimize the hyperparameters (number of trees) of a random forest model within the training set by optimizing cross-validation AUROC [42]. For each trial, We searched T=10,20,30 as a potential number of trees for our random forest classifier. The optimal hyperparameters within the internal cross-validation and the predicted probabilities for all K=5 cross-validation folds for each trial were saved. We next used the optimal hyperparameters to train a random forest model on the entire training set, and then evaluated the performance of this trained model on the testing set. Both the predicted probabilities for the internal and ground truths for each trial were saved. To generate the ROC plots, we concatenated the predicted probabilities and ground truths across 100 trials for both the internal training CV and testing sets. We then

used the python 'roc_curve' from the python library 'sklearn.metrics' to compute the receiver operating characteristic (ROC) and the area under the ROC (AUROC) for both the results from the internal training-set cross-validation and the performance of the best performing model on the held-out testing set across all 100 trials [42].

2.2.12 Mutational calling and analysis from evRNA-seq data

To derive the mutational information shown in Table S1 and Fig. 4g-h, we first mapped all reads using 'bwa' mem v0.7.17 (with default settings) against reference human hg38 reference; 'bwa' was chosen to include reads lying outside of the reference transcriptome with potentially useful mutational information. Next, we used GATK 'HaplotypeCaller' submodule (with default settings) to call mutations from our patient RNA-seq libraries against hg38 reference and compared the resulting mutational information with summaries of SNaPshot panel sequencing results from clinical records. SNaPshot is a multiplexed PCR assay aimed at identifying somatic variants in 70 different loci from 15 cancer genes; the SNaPshot assay was performed by the MGH's pathology department (Boston, USA). We utilized a custom python script to overlap the resulting .vcf files produced by GATK against the 'CosmicCodingMuts.vcf' file downloaded from the COSMIC mutation database v89 [43].

2.3 Results

2.3.1 EV correspondence with tissue-of-origin

We analyzed melanoma cell-lines and cell-line-derived EV to correlate EV transcriptomes with tumor transcriptomes. We observed high correlation between cell-lines and EV (average R^2=0.87, Fig. 2-3a). Cell lines shared similar concordance in gene expression (Fig. 2-3b), and the majority of genes had small differences in expression (Fig. 2-3c). Unsurprisingly, each cell line had the highest correlation with their EV (Fig. 2-3d), suggesting that EV are reasonable proxies for expression in cell-lines. To determine if a patient's plasma-derived EV profiles correlate with their tumor's

profile, we analyzed paired EV and tumor transcriptomes from N=9 patients and observed close correlation of expression between the two profiles (average R^2=0.82, Fig. 2-4a). Concordance analysis demonstrated that most genes in bulk tumors are detected in corresponding EV (Fig. 2-4b). By conducting gene-set enrichment analysis via GAGE11, we found enrichment of immune-related signatures exclusively in EV (e.g., T-cell activation, NK activation), while tumor-exclusive transcripts enriched for metabolic and tumor-related pathways (Fig. 2-4c). To identify cell populations represented in EV, we utilized CIBERSORT to infer immune cell-type enrichments in patient EV [44, 45]. We observed enrichment in 5 immune sub-populations exclusively in EV (neutrophils, NK cells, CD4+/CD8+ T-cells), and a relative depletion in macrophages/mast cells (Fig. 2-4d), suggesting EV are over-enriched for signals from immune populations important for anti-PD1 responses [28]. Therefore, we hypothesized that EV transcripts from patients prior to and during treatment would predict or reflect resistance to ICI

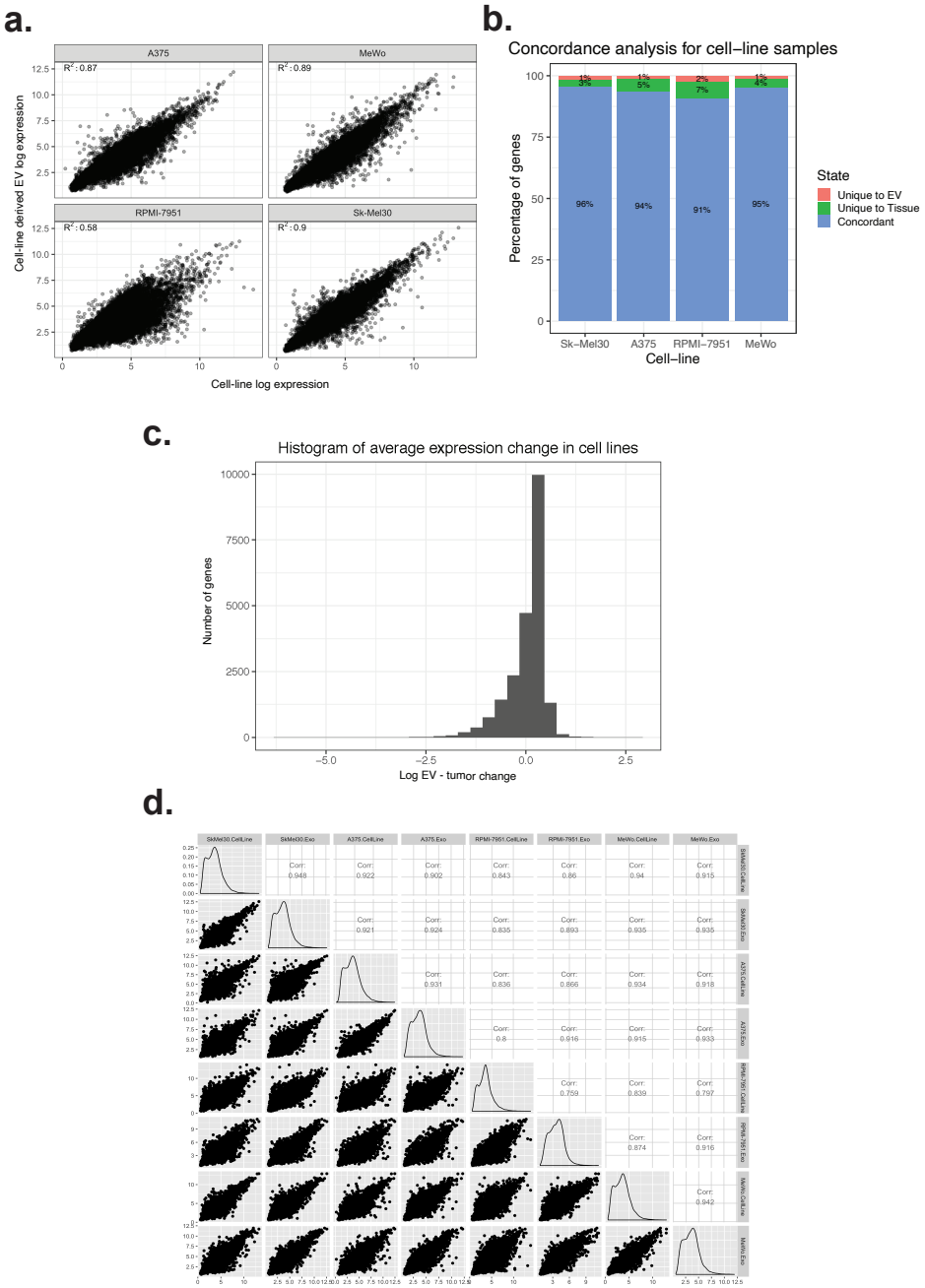


Figure 2-3: **Characterization of transcriptomic similarities between cell-line derived melanoma samples and their matched EV**.(a) Scatter plot visualizing differences between tissue and plasma-derived EV in patient samples. (b) Histogram visualizing the log fold changes between melanoma cell-lines and their EV counterparts. The profiles were compiled using the average of four expression profiles. (c) Concordance analysis across our 4 cell-line samples. Concordance was calculated by using a low expression cutoff as a cutoff for expressed vs. non-expressed status. Genes that were either expressed or not expressed in both tissue and EV compartments are considered concordant. (d) Correlation plot between all 4 cell-lines and their EV counterparts.

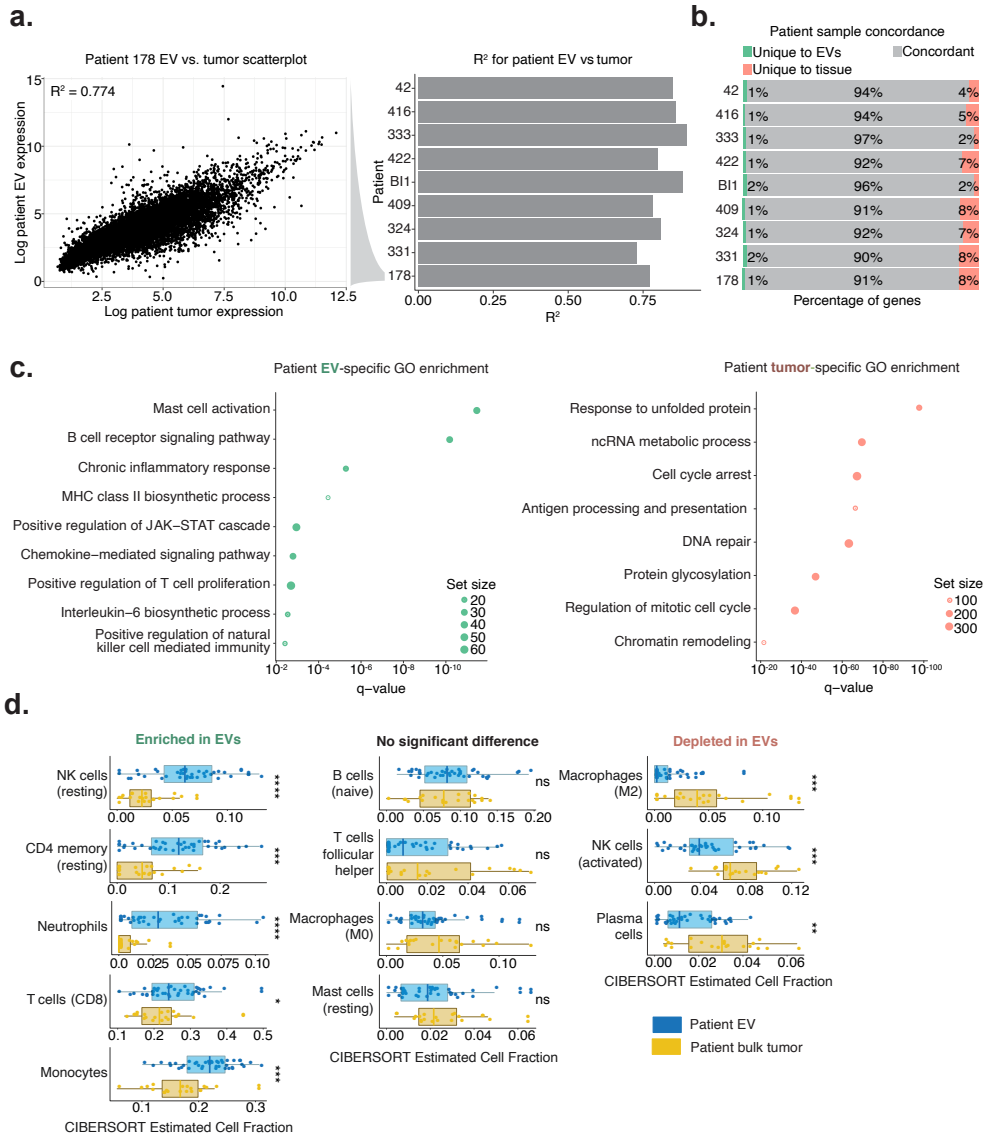


Figure 2-4: **Tumor and EV RNA concordance**. (a) Scatter plot displaying the relationship between expression values of tumors and plasma EV in a representative patient (Patient 178) and a histogram for R^2 between paired tumors and plasma EV across the cohort. (b) Concordance was calculated using a low-expression threshold cut-off for expressed versus non-expressed status (Section 2.2.9). Genes expressed or not expressed in tissue and EV were considered concordant (blue), while a subset of transcripts were unique EV or tumor. (c) Selected pathways from gene-set enrichment comparison of EV (left) vs. patient tumors (right). (d) CIBERSORT inferred deconvolution estimates for all pre-treatment patient tumor and pre-treatment patient plasma-derived EV samples using LM22 immune reference profiles [44].

2.3.2 On-treatment EV analysis

Since tumoral post-treatment signatures are more representative of ICI response than pretreatment values [27, 46], we correlated on-treatment EV transcripts with ICI response. Through differential gene set analysis utilizing the Molecular Signatures Database (MSigDB) [41] via Gene Set Variation Analysis (GSVA, section 2.2.9) [47], we observed 258 pathways significantly different between responders and non-responders in the discovery cohort, of which 25 pathways were significant in the validation cohort. Validated pathways (e.g. T-cell receptor, CTLA4, TGF-β, SMAD2/3, Notch, TNFR2, and VEGFR) (Fig. 2-5a, Fig. 2-6) are implicated in ICI resistance and melanoma progression [48, 49, 50]. With our longitudinal data, we visualized on-treatment pathway dynamics via single-sample GSVA16 and observed decreases in T-cell receptor (TCR) pathway activity during treatment in non-responders (Fig. 2-5c, SFig. 2-7a) and the CD28 costimulatory pathway (SFig. 2-8a). The CTLA4 pathway diverges over time (SFig. 2-8b), potentially resulting from peripheral tolerance during treatment [51], and similar changes are seen in tumor-related pathways (e.g. P53-Hypoxia (SFig. 2-8c) and Kinesin activity22 (SFig. 2-8d)). At the gene level, there were 1240 nominal and 43 FDR-corrected differentially expressed genes (DEGs) in the discovery cohort and 514 nominal and 3 FDR-corrected DEGs in the validation cohort. 80 nominal DEGs shared successful p-value validation and 47 nominal DEGs were successfully replicated at both p-value, minimum expression, and log-fold change levels (Section 2.2.6, Fig. 2-5c). The replication rate of the 47 DEGs is significantly above that expected by chance (p=0.00088, hypergeometric test). Many shared DEGs mirrored gene-set enrichment findings (e.g. KLF10: major actor in the TGF-β pathway [52]; WNT8B: impacts T-effector differentiation [53]). We detected cancer testis antigens (MAGEA1, MAGEA3) known to be expressed by melanoma cells [54] in validated DEGs. On-treatment DEGs (MAGEA1, MAGEA2, KLF10, and MIR4519) were plotted to illustrate time dynamics relative to normalized expression changes at first collection (Fig. 2d) and unnormalized expression (Fig. 2-7b-e).

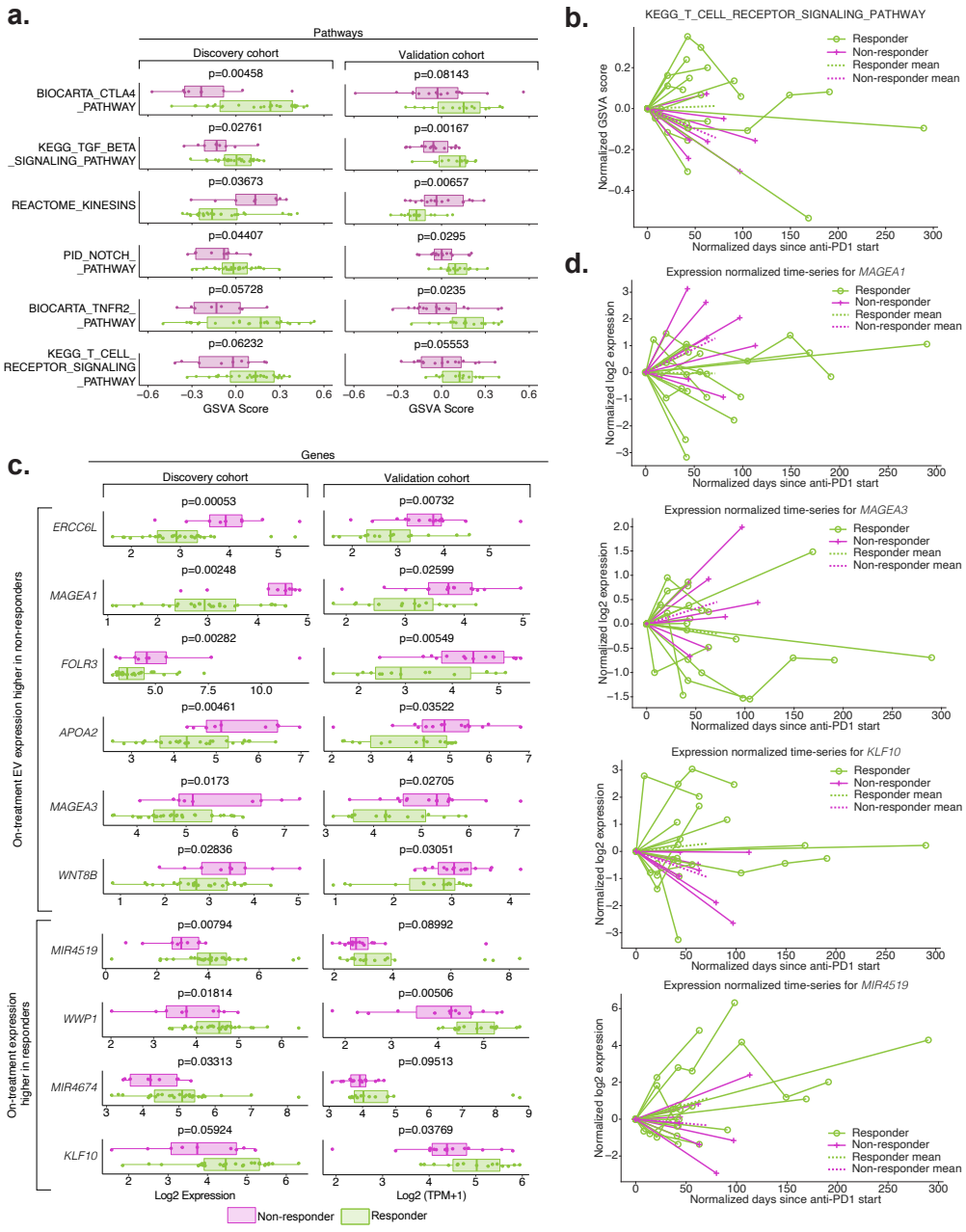


Figure 2-5: **Biological pathways and genes that stratify responders and non-responders in on-treatment EV profiles**. (a) MSigDB canonical (C2) pathway enrichments in on-treatment samples, (salmon color: discovery cohort; teal color: validation cohort). (b) Time dynamics of T-cell receptor KEGG signaling in responders versus non-responders visualized using single-sample gene set enrichment scores derived from GSVA analysis. Individual patient progressions are displayed as connected lines. (c) Expression comparison between responders (green) and non-responders (purple) for selected validated differentially expressed genes between responders (green) in on-treatment samples across both the discovery and validation cohort. The p-values displayed are generated from limma. (d) Time dynamics of several representative validated on-treatment DEGs, including MAGEA1, MAGEA3, KLF10, and MIR4519 in the discovery cohort. Individual patient progressions are displayed as connected lines.

Figure 2-6: **On-treatment validated pathways**. Box-plots and associated p-values for validated MSigDB canonical pathways that differ between responders (green) and non-responders (purple). The visualized points are individual GSVA scores inferred for each pathway. The p-values were generated by comparing responder vs. non-responder GSVA scores via a Mann-Whitney U-test.

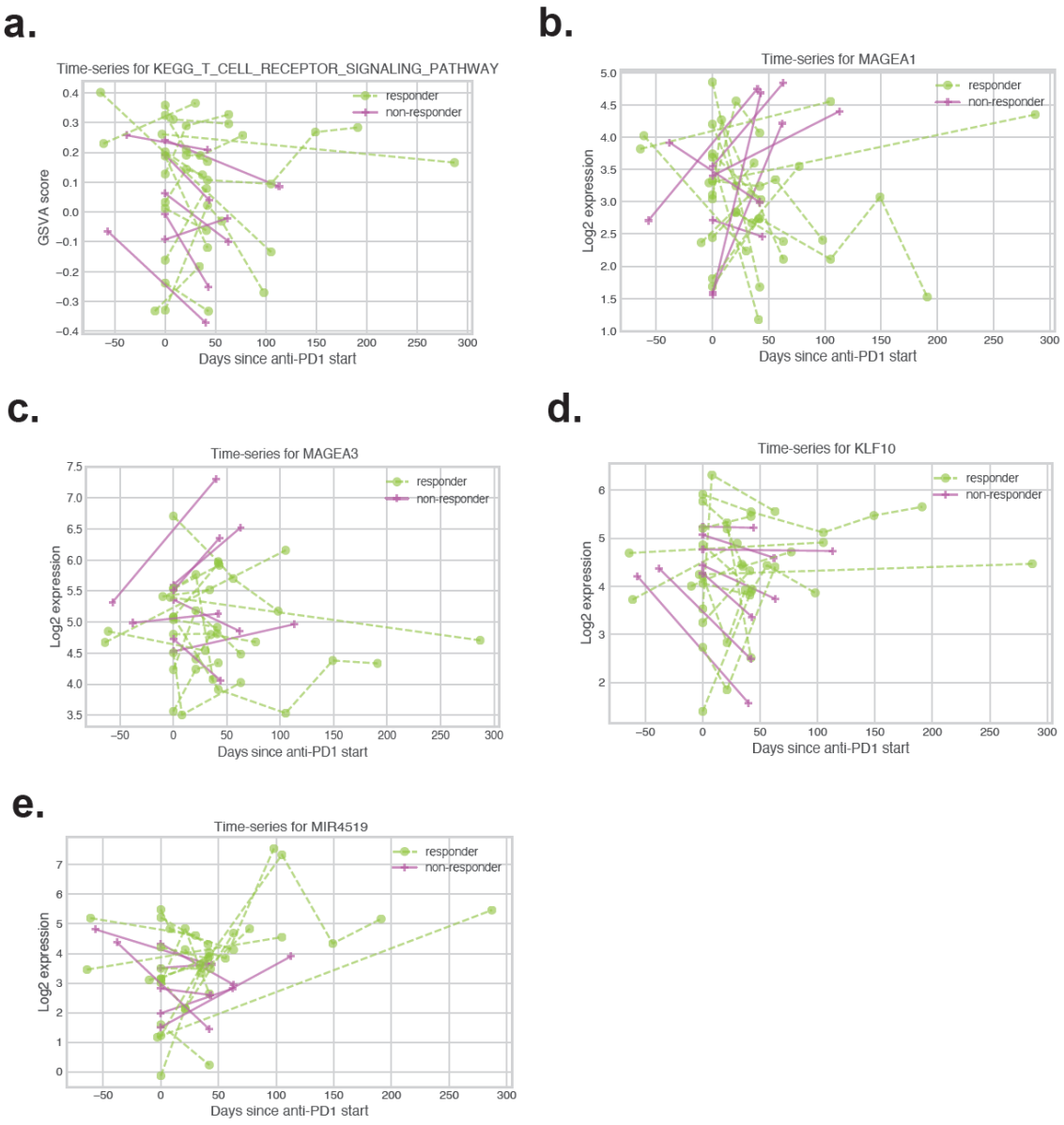


Figure 2-7: **Un-normalized time series plots for differential genes and pathways**. Un-normalized time-series plots showing time dynamics for pathways and genes discussed in Fig. 2-5b and Fig. 2-5d. Individual GSVA scores were used to plot the TCR KEGG pathway, while platform-normalized log2 expression values were used to plot the individual gene expression levels.

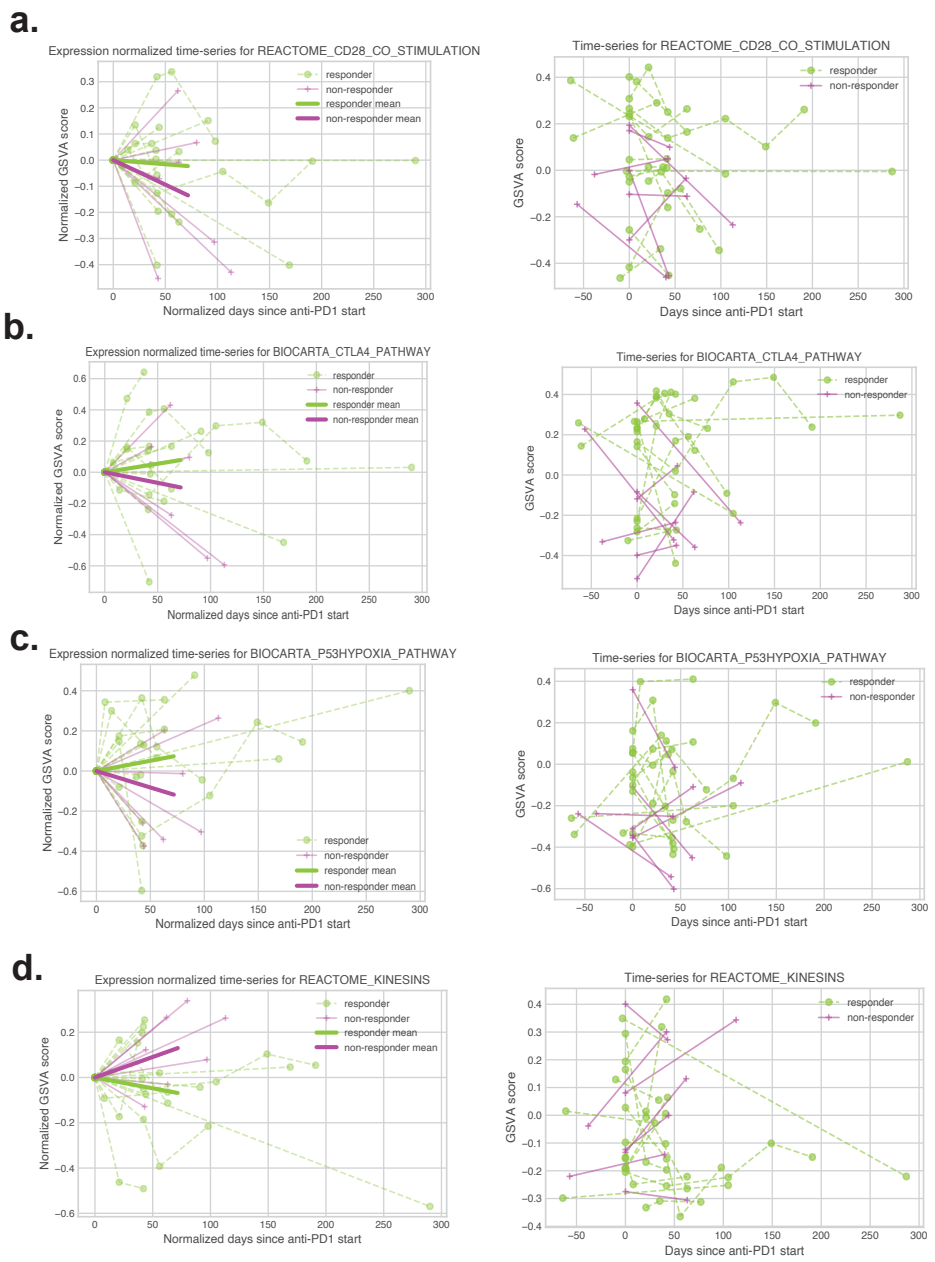


Figure 2-8: **Normalized and un-normalized time series plots for differential genes and pathways**.Normalized and unnormalized time-series plots showing time dynamics for selected pathways. The plotting was performed per methodology previously discussed in captions for Fig. 2-5b, Fig. 2-5d and Fig. 2-7.

2.3.3 Pre-treatment EV analysis

We assessed if pre-treatment EV transcriptomes are able to stratify responders from non-responders by performing gene-set enrichment via GSVA [55], showing 101 differentially expressed MSigDB pathways (discovery cohort), of which 26 replicated in the validation cohort (Fig. 2-9a, Fig. 2-10) [56]. Compared to on-treatment pathways, we see differential Notch and TGF-β signaling in the pre-treatment cohort and observe differences in MAPK-related signaling (ERRB4) and keratinization. Statistical testing revealed 366 nominal and 12 FDR-corrected DEGs in the pre-treatment discovery cohort, and 1406 nominal and 45 FDR-corrected differentially expressed genes (DEGs) in the pre-treatment validation cohort. 54 nominal DEGs had replicated p-values, while 38 nominal DEGs had replicated p-values and log fold changes, representing a replication rate significantly above that of random chance (p=0.0041, hypergeometric test). DEGs included members of both immune and tumor-related pathways implicated in ICI resistance or tumor growth (e.g. CD1A, MAP2K4, TRBV7-2 and IFGL1 [48, 57, 58]) (Fig. 2-9b), and cancer-associated miRNAs (e.g. miR551A) were enriched in non-responders. We constructed a pre-treatment random forest classifier to predict post-treatment response vs. non-response status from pre-treatment DEGs in order to quantify their predictive power. To minimize platform-specific differences and demonstrate the robustness of these markers, we pursued a machine learning framework that mixed samples across platforms in both the training and testing sets. To minimize variability from random sampling, we created 100 randomized partitionings ("trial") of the combined dataset into N=41 training and N=30 testing samples. Within each trial, we first conducted K=5 internal cross-validation to optimize the hyperparameters of a random forest model specific to that trial. We next used the best performing hyperparameter set to evaluate testing performance on the test set. By summarizing the results from all 100 trials, we are able to generate receiver operator characteristic (ROC) plots displayed for both the training cross-validation performance (top panel) and the testing set performance (bottom panel) in Fig. 2-9c. To evaluate binary classification accuracy, we utilized the area under

the receiver operator character (AUROC); this is the probability that a binary classifier will rank a randomly chosen responder patient higher than a randomly chosen non-responder one [59]. We observed moderate predictive power in both our internal cross-validation (AUROC=0.784) and our independent testing set (AUROC=0.737) for our pre-treatment DEGs (Methods). Multiple validated DEGs (IGFL1, TFF2, and MAP2K4) showed significant stratification in progression-free or overall survival (Fig. 2-9d, Fig. 2-11).

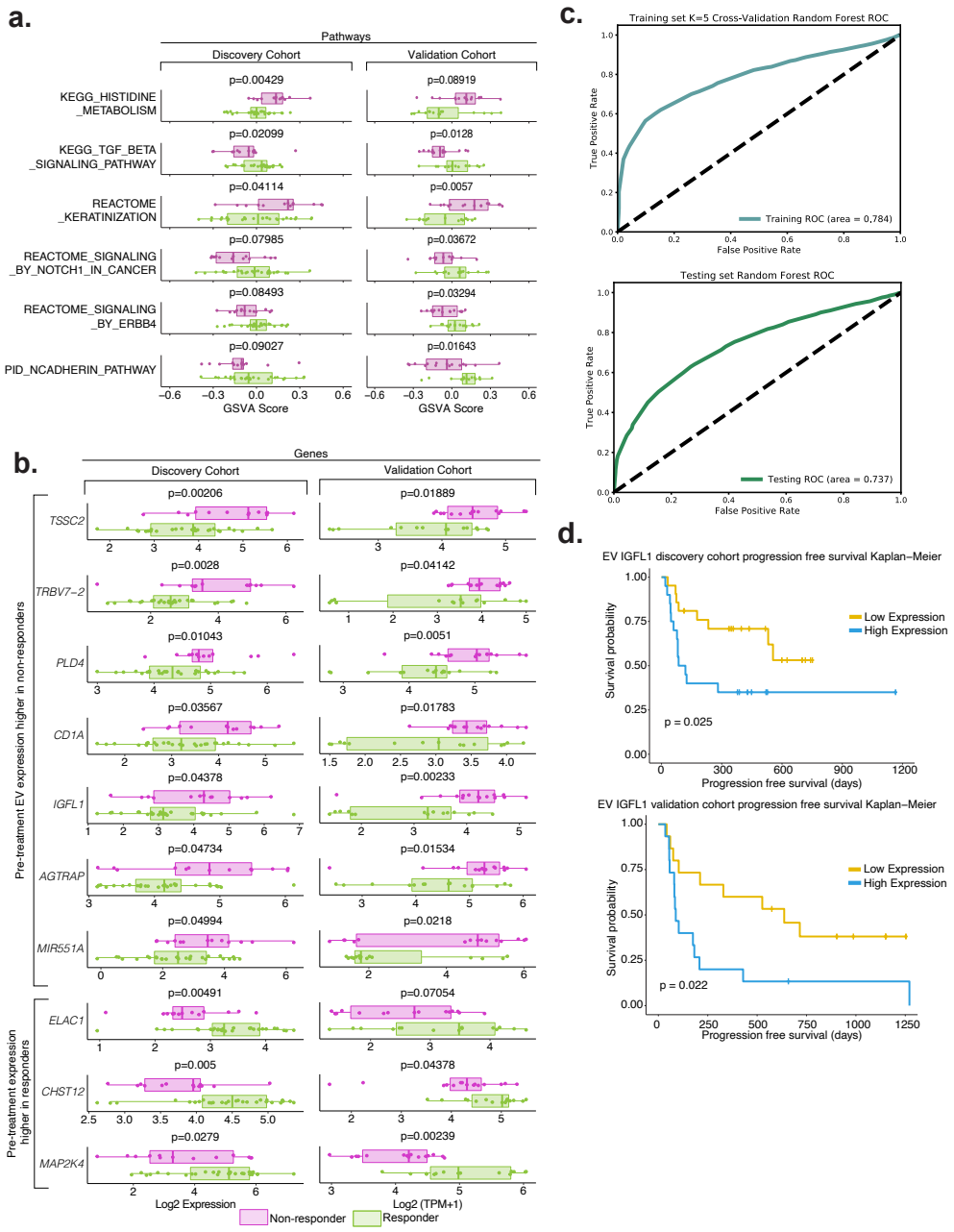


Figure 2-9: **Biological pathways and genes that stratify responders and non-responders in pre-treatment EV profiles**. (a) GSVA scores and associated p-values for selected MSigDB canonical pathways that differ significantly between responders and non-responders in pre-treatment EVs. The p-values were generated by performing a Mann-Whitney U-Test. (b) Boxplots of expression values of selected validated pre-treatment EV DEGs between responders and non-responders in both the discovery and validation cohort (purple color: non-responders, green color: responders). (c) Receiver operating characteristics (ROCs) generated by a random forest classifier [42] utilizing the validated pre-treatment DEGs genes as features to predict response vs. non-response status from pre-treatment plasma-derived EV transcriptomic profiles. (d) Kaplan-Meier overall survival plots and associated log-rank test p-values for IGFL1 in the discovery and validation cohorts.

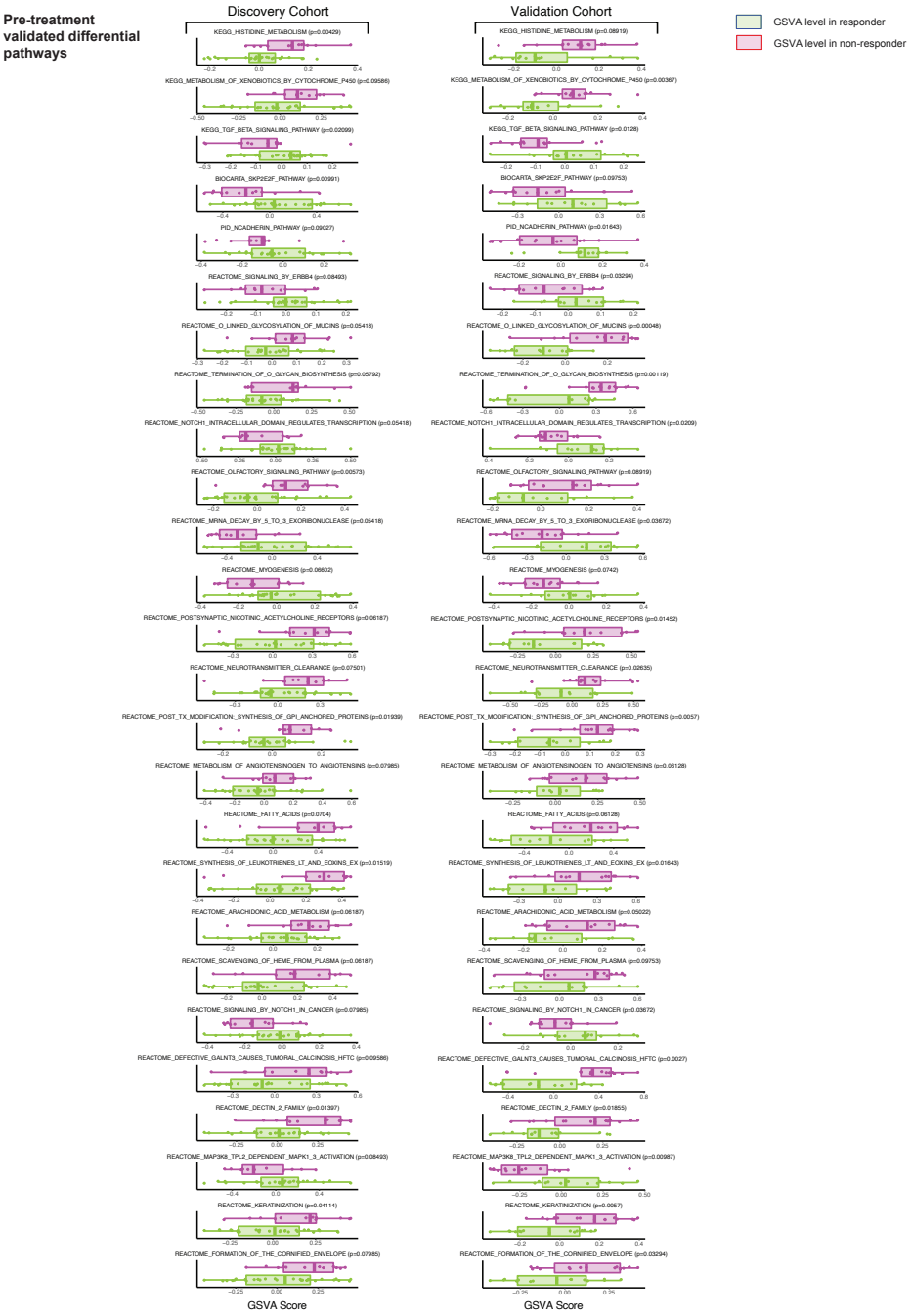


Figure 2-10: **Pre-treatment validated pathways**. Box-plots and associated p-values for validated MSigDB canonical pathways that differ between responders (green) and non-responders (purple). The visualized points are individual GSVA scores inferred for each pathway. The p-values were generated by comparing responder vs. non-responder GSVA scores via a Mann-Whitney U-test.

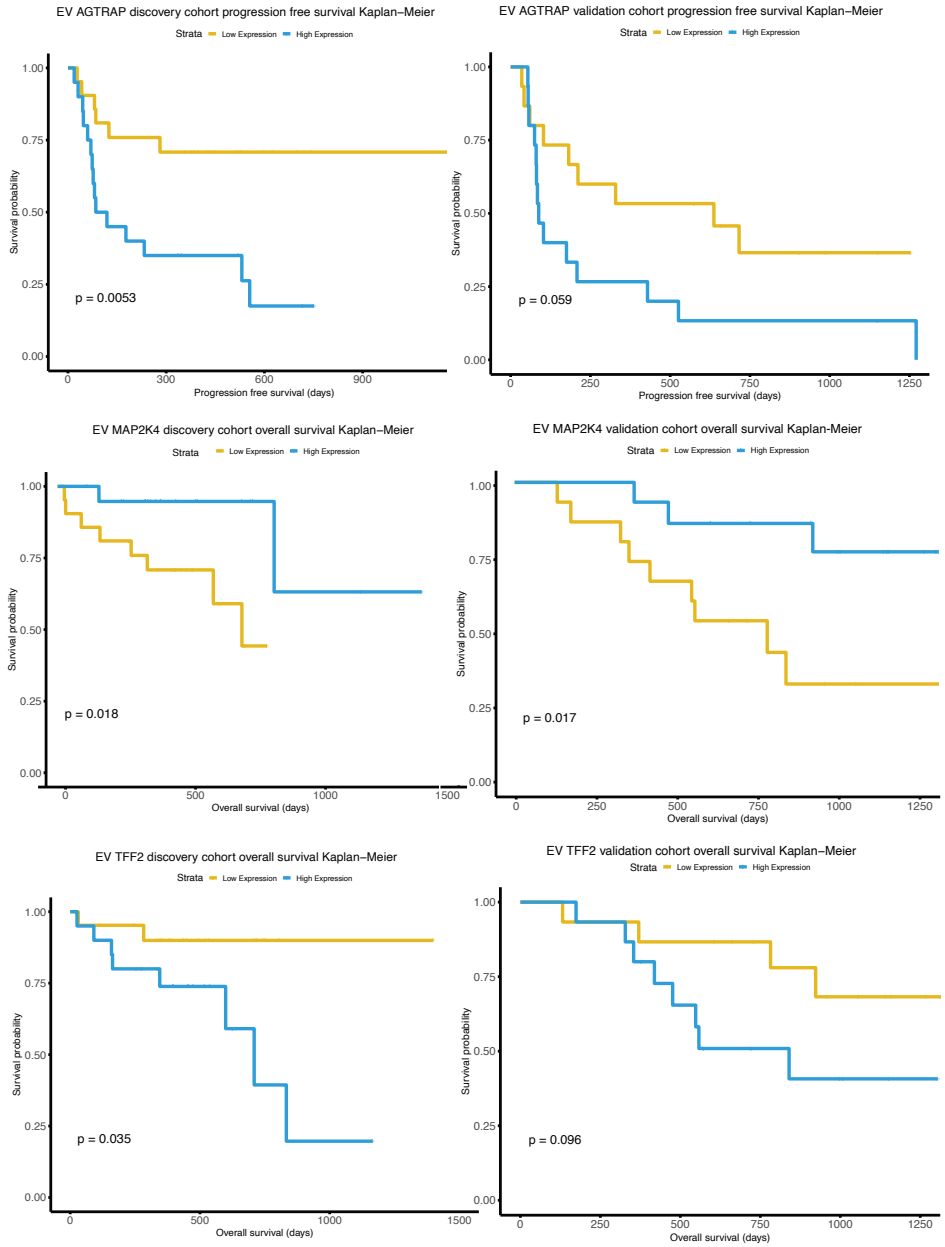


Figure 2-11: **Kaplan-Meier progression-free and overall survival curves for selected genes**.Kaplan-Meier progression-free and overall survival curves for selected genes that showed significant or near-significant differences between high-expressed and low-expressed patients.

2.3.4 EV RNA-seq mutational information

We investigated whether mutational information in the evRNA-seq can stratify responders and non-responders given data relating to tumor mutational burden (TMB) [60]. We surveyed the mutational landscape using hg38 genomic reference to call somatic and germline RNA-seq-associated mutations (section 2.2.12). By comparing evRNA-seq mutational calls against patient matched tumor panel sequencing results (MGH SNaPshot), we determined that three of our patients had specific driver mutations called by both panel sequencing and evRNA-seq mutational calling. Since panel data represents a small fraction of somatic tumor-associated mutations, we surveyed the entire somatic mutational landscape to evaluate differences in cancer driver mutational burden. Although patient samples varied in absolute number of mutations detected, we reasoned that differences in the fraction of somatic tumor-related mutations from COSMIC database relative to overall mutational pool can be attributed to changes in somatic mutation load and not differences in germline mutations [43]. We observed significantly higher COSMIC driver somatic mutational fraction in responders relative to non-responders (Fig. 2-12a-b). This was also reflected in the survival analysis. We observed significant stratification (p=0.014, log-rank test) between patients with high COSMIC mutational fraction (top 50% of cohort) vs. those patients with low COSMIC mutational fraction (Fig. 2-12c), with patients with higher mutational fractions experiencing overall longer-survival times. Interestingly, the progression-free survival log-rank test was not significant (Fig. 2-13, p=0.67), suggesting that evRNA-seq cancer driver information may be more effective for predicting long-term effects of ICI treatment rather than short-term effects.

2.4 Discussion

In this study, we explored the potential usage of plasma-derived EV transcriptomic profiles as a source of biomarkers for predicting and monitoring checkpoint blockade immunotherapy success. Our results show that EVs, in aggregate, correlate with certain aspects of bulk tumoral biology and reflect a number of previously identified

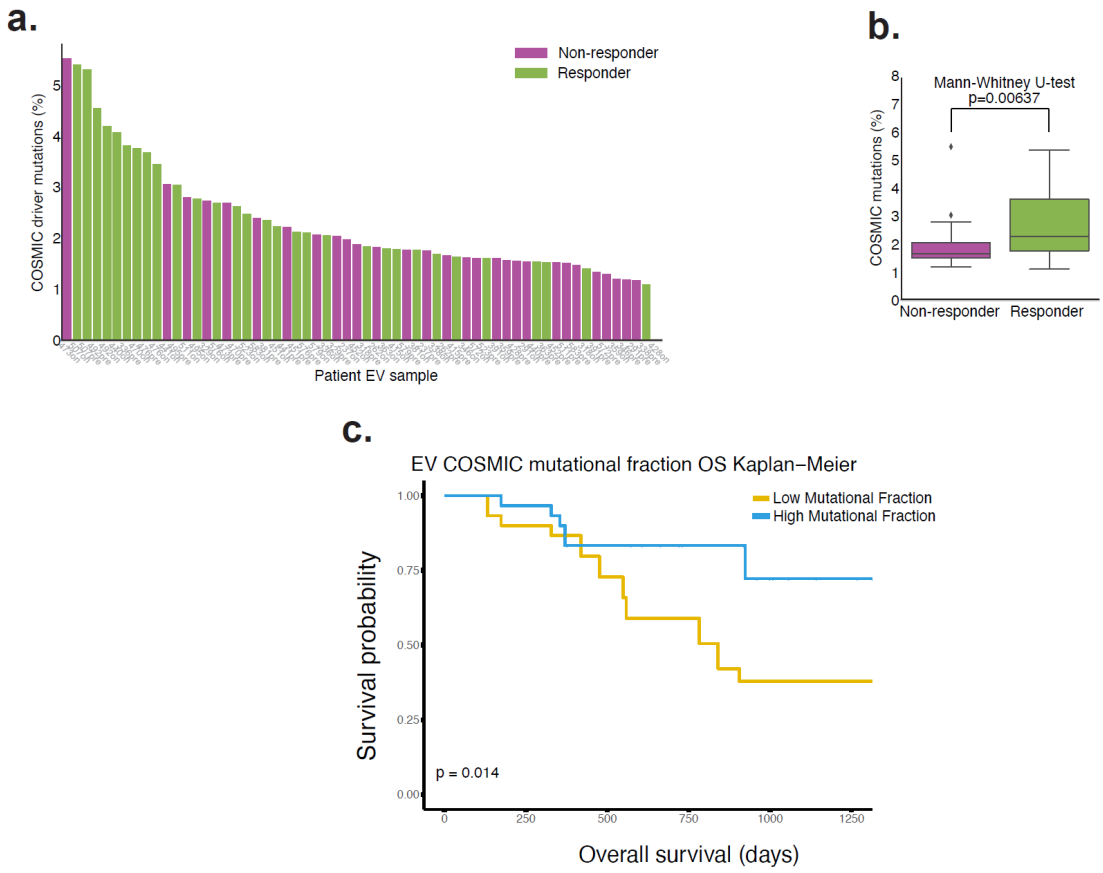


Figure 2-12: **Comparison of COSMIC mutational driver load between responders and non-responders**.Visualization of cohort-wide percentage of COSMIC driver mutations, as part of total mutations (a combination of somatic & germline) called for each patient's evRNA-seq profile in the validation cohort, a higher mutational load in responder patients (purple color: non-responders, green color: responders). (b) Boxplot summarizing the distribution of COSMIC driver mutation fraction between responder and non-responder profiles. A Mann-Whitney U-test was used to test for significant differences between the responder and non-responder distributions. (c) Kaplan-Meier overall survival plots for COSMIC survival fraction in the validation evRNA-seq cohort. High and low expression classifications were determined for each patient based on whether a particular patient's COSMIC mutational fraction was in the top half (teal) or bottom half (yellow) of the validation cohort's COSMIC mutational fraction distribution. A log-rank test was used to derive the p-value.

differentially regulated biological pathways implicated in ICI resistance or melanoma progression in responders vs. non-responders. The majority of differential pathways and genes at the pre-treatment time point primarily reflect differences in metabolic state as opposed to pre-existing immune-related differences, suggesting that plasma-

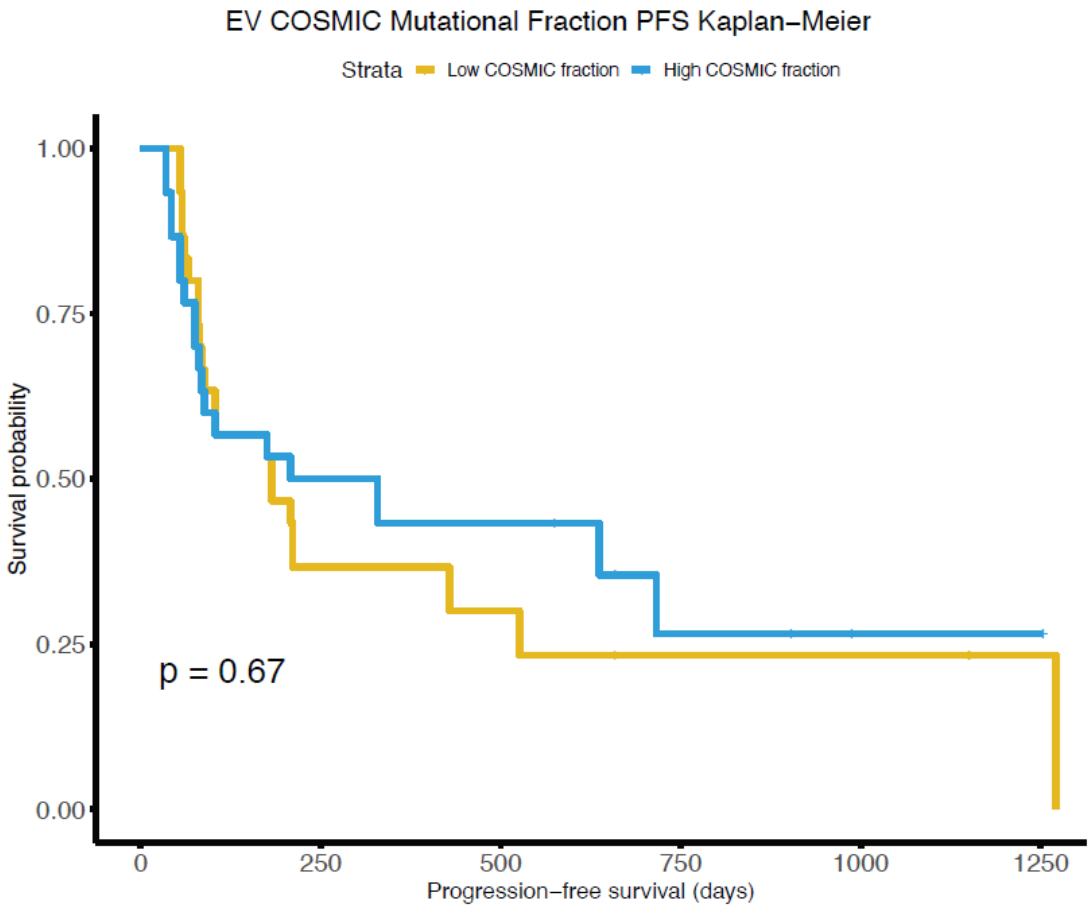


Figure 2-13: **Kaplan-Meier progression-free survival plots for COSMIC mutational fraction in the validation evRNA-seq cohort**.High and low expression classifications were determined for each patient based on whether a particular patient's COSMIC mutational fraction was in the top half (teal) or bottom half (yellow) of the validation cohort's COSMIC mutational fraction distribution.

derived EVs only start capturing immune-related differences between responders and non-responders after ICI treatment is administered. This is supported by the enrichment of immune-related pathways in our on-treatment differential pathway analysis, as well as the enrichment for non-tumor-derived DEGs as inferred by our deconvolution model. Though the validated pre-treatment DEGs we discovered are biologically informative and can be utilized to create predictive models of ICI response with moderate predictive ability, our predictive classifier still lags behind predictive classifiers created from transcriptomic profiles from bulk tumor biopsies in terms of performance; bulk-tumor profiles demonstrate performance in the 0.8-0.9 AUROC range

[61], compared to our testing AUROC of 0.737. Despite its lower predictive performance, the greater availability and ease of obtaining plasma-derived EV samples relative to bulk tumor biopsies may provide a viable clinical use case for EV profiling in the context of ICI treatment.

The higher levels of correlation between melanoma cell-lines and their EVs as compared to bulk patient tumors and corresponding plasma-derived EVs suggest that bulk plasma-derived EVs reflect a broader repertoire of EV sources. Indeed, the most robust enrichment in EVs is for immune-related pathways (Fig. 2-4c). This is reinforced by the relative enrichment of several key immune cell-types in our CIBERSORT deconvolution and are validated by our on-treatment DEGs and pathway enrichments. EV transcriptomic biomarkers may complement circulating tumor DNA (ctDNA) to gain transcriptomic information regarding tumor dynamics in addition to genomic information and may give a readout of both tumor-intrinsic and immunologic changes simultaneously.

Our utilization of evRNA-seq technology in the validation cohort brought additional challenges to our analysis when cross-comparing with discovery cohort microarray data. We reasoned that utilizing two separate sequencing technologies on two independent cohorts raises the bar for reproducibility and that findings replicated with distinct methodologies are likely to be robust. Additionally, we show that mutational information embedded in the evRNA-seq itself can potentially be exploited to stratify responder and non-responder populations and to serve as an orthogonal means to determine the tissue-of-origin of plasma-derived EV transcripts. This approach can be further enhanced through complementary WES of EV DNA. Although it is unlikely the utility of the mutational information from EV RNA-seq data will outstrip high-depth cfDNA or WES TMB data in the near future, this mutational information is embedded within a large amount of transcriptomic information provided, which reflects dynamic tumoral changes and can complement other DNA-based sequencing methods (ctDNA, evDNA-seq) in ICI monitoring and response prediction tasks.

Chapter 3

Deconvolution of extravesicular cargo enables tissue-of-origin identification

3.1 Introduction

Extravesicular vesicles (EVs) are key mediators of intercellular communication. Extracellular vesicles are phospholipid bilayered vesicles generated by almost all mammalian cell types. Consisting of exosomes (30-120nm in diameter), microvesicles (MVs 0.1-1.0µm), ectosomes (0.1-1 µm) and apoptotic bodies (0.8-5.0 µm) [62]. EVs is able to shuttle multiple types of cargo, including: membrane proteins, cytosolic proteins, lipids, DNA, mRNA, miRNA. These cargo can have a causal role and may serve as cancer biomarkers. Indeed, circulating EVs have been a source of liquid biopsies [63, 64] and EVs have been demonstrated to carry a variety of miRNAs with roles in tumor progression [62].

Given the importance of EVs, it is natural to wonder if it is possible to deconvolve the mixed EV profiles observed in bodily fluids such as blood into its tissues-of-origin components. To the best of our knowledge, there has been no explicit attempt to create deconvolution models with the expressed intent of deconvolving plasma-derived exosomal profiles. The closest approachesa re *in silico* deconvolution techniques such as CIBERSORT [44] and CIBERSORTx [45], which use cell-type specific expression profiles to deconvolve bulk mixtures. However, these approaches fail to account for

EV idiosyncrasies, such as the tumor-to-EV export process that may impact the RNA concentration. In this chapter, we provide the first such deconvolution model to explicitly account for such idiosyncrasies and provide an accurate estimate of tumor and non-tumor contributions to the plasma-derived EV mixture.

3.2 Methods

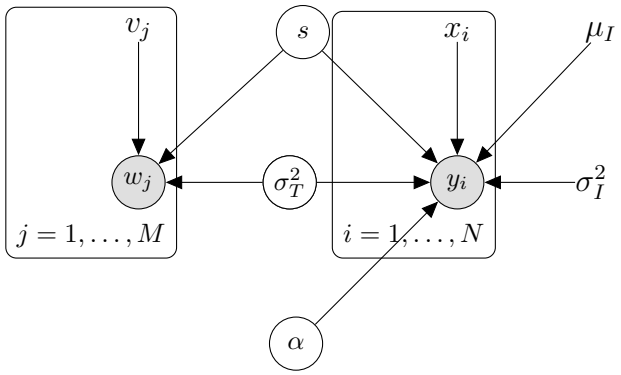
3.2.1 Deconvolution model justification

To summarize, we want to infer the contribution of the tumor-derived EV component and non-tumor derived (interchangeably referred to as "immune" and "non-tumor") component to the observed plasma-derived EV transcriptomic profile. In contrast to existing deconvolution models designed for bulk deconvolution (e.g., CIBERSORT[44]), our model explicitly models the changes in transcript abundance as a result of export/-packaging from the transcript abundance in the tumor to the transcript abundance. All of the data shown both here and in the main-figures related to deconvolution utilized only non-platform-corrected discovery cohort microarray data, since the discovery cohort exoRNA-seq included only plasma-derived EV samples and thus were not suitable inputs for our deconvolution model (see Fig. 2-1b for more detailed information regarding our analysis methodology). We created two versions of the deconvolution probabilistic model. In **single-gene mode**, the probabilistic model is fully fitted using the No-U-Turn Sampler (NUTS) Hamiltonian Monte Carlo (HMC) algorithm[65] and full posterior estimates for all the relevant parameters are returned. This model is fitted using the probabilistic programming language Stan[66]. This mode is designed for in-depth analysis of a single gene (or few genes), or situations where inferred parameters (e.g., scaling parameter, patient inferred tumor-EV expression) of interest requires a full posterior estimate. Due to the time and resource intensive nature of the fitting process, it is impractical to perform full MCMC inference when we want to analyze the deconvolution profiles for tens of thousands of genes. Thus, we included a second mode, a **multi-gene mode**, in which we fit a sim-

plified version of the single-gene model using Scipy's implementation sequential least squares programming (SLSQP) to return a point estimate of the mixing coefficient. The single-gene model not only returns posterior distribution mixing fraction, but also the full posterior distribution of the scaling coefficient, which allows per-patient imputation of the tumor-derived EV fractions (Supplemental Figure 9); however, the multi-gene model only returns a single maximum a posterori (MAP) estimate of the mixing fraction. We envision the usage of the single-gene model in cases when a specific gene needs to carefully dissected and more robust inference is required, whereas the multi-gene model can be used on large-scale transcriptomic datasets to infer population-wide mixing fractions.

3.2.2 Deconvolution model specification

- N: number of patient derived tumor profiles and tumor EV profiles
- M: number of cell-line tumor and tumor EV profiles
- x_i: the ith patient's observed tumor expression for the current gene
- y_i: the ith patient's observed peripheral-blood derived expression for the current gene
- w_j: the jth cell-line's
- α: mixing fraction between tumor-component and immune-component
- σ_T^2: variance component
- μ_I: Immune component mean (fixed parameter)
- σ_I^2: Immune variance component (fixed parameter)



Prior specification

$$s \sim \mathcal{N}(0, 2)$$

$$\sigma_T^2 \sim \mathcal{IG}(1, 1)$$

$$\alpha \sim \mathcal{B}(2, 2)$$

Where $\mathcal{B}$ denotes the Beta distribution and $\mathcal{IG}$ denotes the Inverse-Gamma distribution.

Data likelihood

$$p(y_i|\alpha, \sigma_T^2; \mu_I, \sigma_I^2) = \underbrace{\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2)}_{\text{Tumor EV component}} + \underbrace{(1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)}_{\text{Immune/background EV component}} \tag{3.1}$$

$$p(w_j|s, v_j, \sigma_T^2) = \mathcal{N}(w_j|s + v_j, \sigma_T^2) \tag{3.2}$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) = \prod_{i=1}^{N} \left[\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2) + (1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)\right] \prod_{j=1}^{M} \mathcal{N}(w_j|s + v_j, \sigma_T^2) \tag{3.3}$$

Full Posterior

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) \propto p(\mathbf{y}, \mathbf{w}, \mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) p(s, \sigma_T^2, \alpha) \tag{3.4}$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) \propto \prod_{i=1}^{N} \left[\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2) + (1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)\right] \prod_{j=1}^{M} \mathcal{N}(w_j|s + v_j, \sigma_T^2) \tag{3.5}$$

$$p(s,\alpha,\sigma_T^2|\mathbf{x},\mathbf{v},\mathbf{y},\mathbf{w};\mu_I,\sigma_I^2) \propto \prod_{i=1}^{N}\left[\alpha\mathcal{N}(y_i|x_i+s,\sigma_T^2)+(1-\alpha)\mathcal{N}(y_i|\mu_I,\sigma_I^2)\right]\prod_{j=1}^{M}\mathcal{N}(w_j|s+v_j,\sigma_T^2)p(s,\alpha,$$

(3.6)

3.3 Results

3.3.1 Validation of deconvolution model

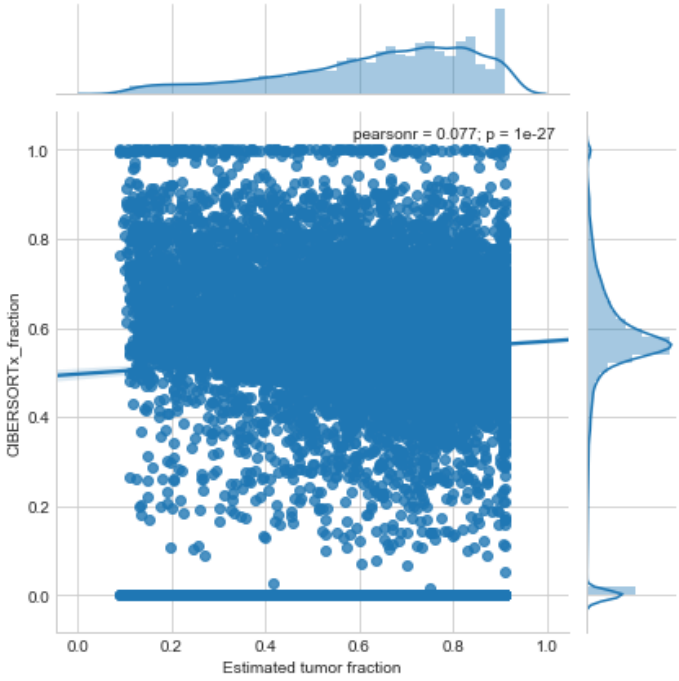


Figure 3-1: **Correlation between CIBERSORTx and EV deconvolution model**. Scatterplot with line-of-best-fit and correlation between inferred CIBERSORTx and our deconvolution model inferred tumor fraction estimates on discovery cohort data

In order to provide evidence that our model is correctly partitioning genes into tumor and non-tumor components, we utilized CIBERSORTx [45]- a recently published bulk deconvolution program from Newman et al. that attempts to separate bulk transcriptomic profiles into component cell-type-specific profiles. In order to run the program, we utilized first inputted our discovery cohort (non-cross-platform corrected) microarray data matrix into the online CIBERSORTx web portal and utilized the melanoma

reference profiles from Tirosh et al.'s *Science* 2016 single-cell dissection of metastatic melanoma that is provided by CIBERSORTx's default online profiles [44, 45]. Since the algorithm generates estimates for all component cell-types instead of tumor-only profiles, we averaged the non-tumor cell-types in order to make it comparable to our non-tumor component estimation. A direct comparison of our values can be found in Supplemental Note Figure 3-1. We see that there's a slight but significant correlation between CIBERSORTx inferred-tumor fraction and tumor fraction inferred from our deconvolution model; however, it is clear from the CIBERSORTx density plot (y-axis) that the inferred tumor fraction is roughly normally distributed, an assumption that our model does not make (see density plot on x-axis). This continuous coding of tumor fractions hinders direct comparison of model predictions between CIBERSORTx and our model; thus, in order to better compare our cell-type predictions, we binarized model predictions for each gene as either tumor-derived or non-tumor derived using a cut-off of 0.5 as the threshold between tumor and non-tumor (same threshold used in the main manuscript). Using this cutoff, we can generate the confusion matrix found in Supplementary Note Table 3.1.

	CIBERSORTx tumor	CIBERSORTx non-tumor
EV deconvolution tumor	12984	2319
EV deconvolution non-tumor	3719	1244

Table 3.1: **Confusion matrix between CIBERSORTx and deconvolution model**. Confusion matrix between CIBERSORTx and our deconvolution model using 0.5 tumor fraction as a cutoff between tumor and non-tumor binary classification of genes

We can assess the concordance between binary predictions generated by CIBERSORTx and our model using values from the confusion matrix. This corresponded to the following binary classification statistics shown in Supplementary Note Table 3.2, using CIBERSORTx tumor predictions as "ground" truth and our deconvolution model estimates as predictions. We see that overall accuracy (0.70), sensitivity (0.78), precision (0.85), and F1-score (0.81) all support the ability of our deconvolution model to properly classify CIBERSORTx predicted tumor-derived genes. However,

the two models' predictions diverge significantly in terms of specificity (0.35) and negative predictive value (0.25), suggesting that the two models differ significantly in the overall prediction of non-tumor derived genes, with our model predicting a higher fraction of tumor-derived genes relative to CIBERSORTx. This is likely a result of the different distributional assumptions regarding tumor vs. non-tumor distributions between the two models (see Supplementary Note Figure 3-1). We reason that our model is likely to approximate reality more closely, based on known literature regarding significant increases in both overall and tumor-derived EV load in plasma during progression[10]. Furthermore, as mentioned in the main text, our model explicitly accounts for EV-specific characteristics - such as the differential EV packaging process - that bulk deconvolution techniques like CIBERSORTx does not account for. Additionally, our the underlying reference profiles is trained directly or inferred utilizing EV data, which is likely a far better approximation of the underlying mixture profiles in the context of cell-type deconvolution than bulk references. Though *in silico* independent validation via CIBERSORTx provides evidence for the validity of our deconvolution model predictions, particularly the prediction of tumor-derived genes, *in vivo* experimental evidence gathered via tumor vs. non-tumor derived EV selection/enrichment remains the gold standard to validate our model.

Metrics	**Value**
Accuracy	0.70
Sensitivity	0.78
Specificity	0.35
Precision	0.85
Negative Predictive Value	0.25
False Positive Rate	0.65
F1 Score	0.81

Table 3.2: **Confusion matrix statistics between CIBERSORTx and deconvolution model**. Binary classification performance metrics generated from the confusion matrix in Table 3.1

.

3.3.2 Application of deconvolution model on experimental data

The bulk EV selection approach raises questions regarding how EVs from non-tumor sources impact tumor-derived EV contributions. Based upon CIBERSORT results (Fig. 1c), we hypothesized that both tumor-derived EVs and non-tumor-derived EVs are detected. Dissecting the contribution from tumor-derived EVs versus non-tumor-derived sources may reveal whether response-related changes reflect changes in the tumor microenvironment or non-tumoral changes (i.e. systemic changes in the immune system). Therefore, we developed a probabilistic deconvolution model to infer: (i) a "packaging" coefficient representing depletion/enrichment of transcripts during packaging/export into EVs, (ii) the unobserved non-tumor-derived EV profile (Fig. 3-3, and (iii) a mixing fraction between unobserved tumor-derived EV component and the non-tumor-derived components for each gene (Fig. 3-2a). To assess the accuracy of the predictions we analyzed known genes involved in EV function or immunotherapy response [67, 48] (Fig. 3-2b) followed by DEGs at pre- and on-treatment time points (Fig. 3-2c). Our model also probed enrichment across gene sets and calculated a gene-set level tumor fraction, demonstrating significant enrichment for tumor-derived transcripts (Fig. 3-2d). When assessing ICI and melanoma-relevant KEGG pathways, our results align with expected ranking (Fig. 3-2e) (e.g. melanoma-related pathways have higher tumor fraction). Validated pre-treatment DEGs enriched for tumor-derived genes, while on-treatment DEGs had greater non-tumor contribution (Fig. 3-2f). This suggests that on-treatment DEGs preferentially reflect changes induced by ICI, consistent with findings from on-treatment differential pathway analysis. To illustrate the utility of our deconvolution algorithm for interpreting DEGs, we use KLF10, a member of the TGF-β pathway that has both roles as a tumor suppressor [52] and as an inductor of Th1/Th17 polarity and CD8+ memory T-cell formation [68]. Our data suggests that KLF10 is derived from the non-tumor component, suggesting differential on-treatment changes may be related to the Th1/Th17 polarity-inducing function of KLF10 as opposed to a tumor suppressor role.

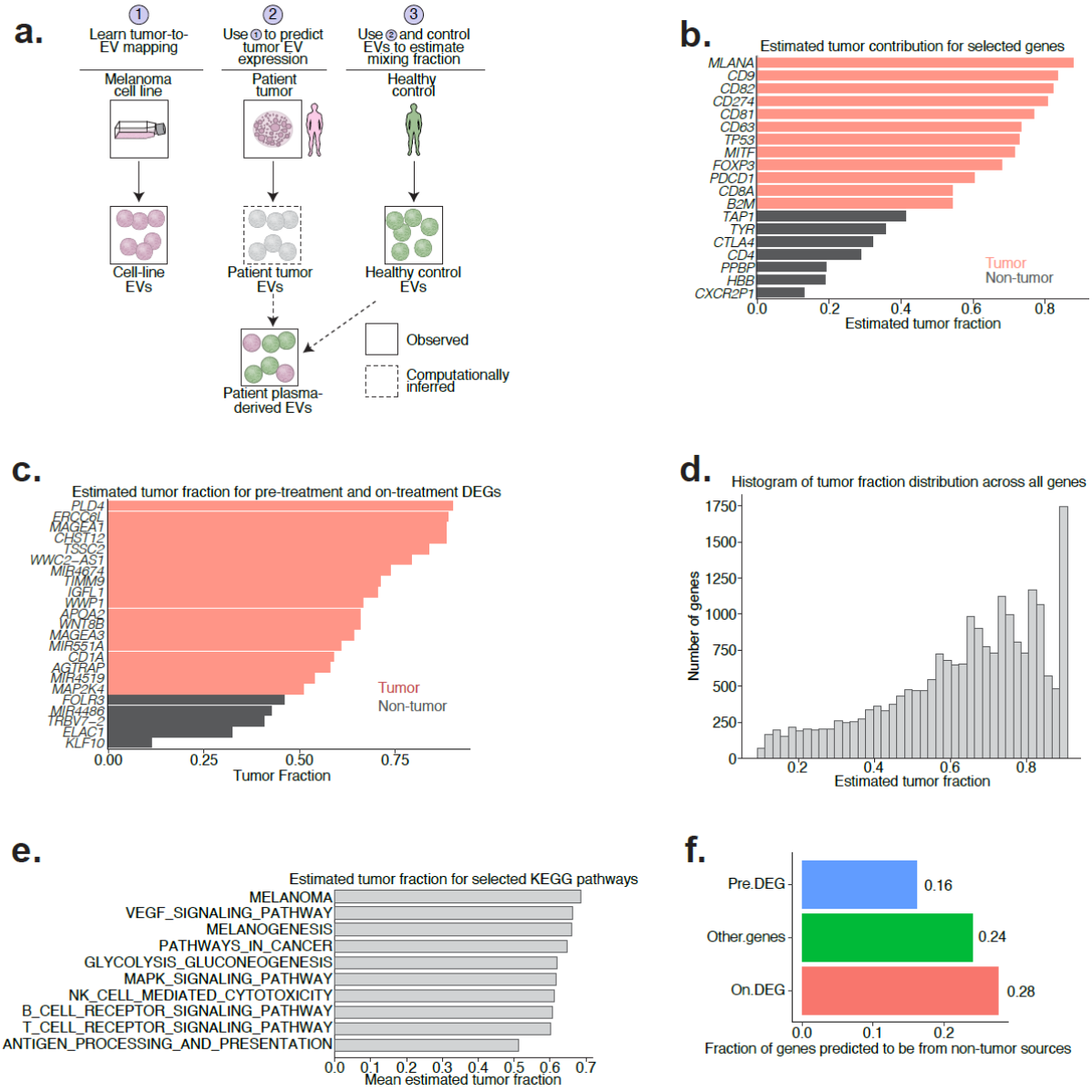


Figure 3-2: **Deconvolution of EV profiles and analysis of driver mutations in RNA-seq profiles**.(a) Schematic representation of our deconvolution model (see Supplementary Note). (b) Selected tumor contributions from known tumor and non-tumor genes. Red denotes predicted tumor and grey denotes predicted non-tumor tissue-of-origin for a particular gene. (c) Estimated tumor fraction for pre-treatment and on-treatment DEGs demonstrates that a majority of DEGs are predicted to come from tumor sources. (d) Histograms of maximum a posterori (MAP) estimates of tumor fraction from our model across all genes. (e) Average tumor fraction of all genes involved in several selected KEGG categories. (f) Predicted tumor fraction of validated pre-treatment DEGs, on-treatment DEGs, and all other genes predicted to be non-tumor derived (i.e., predicted tumor fractions of <0.5)

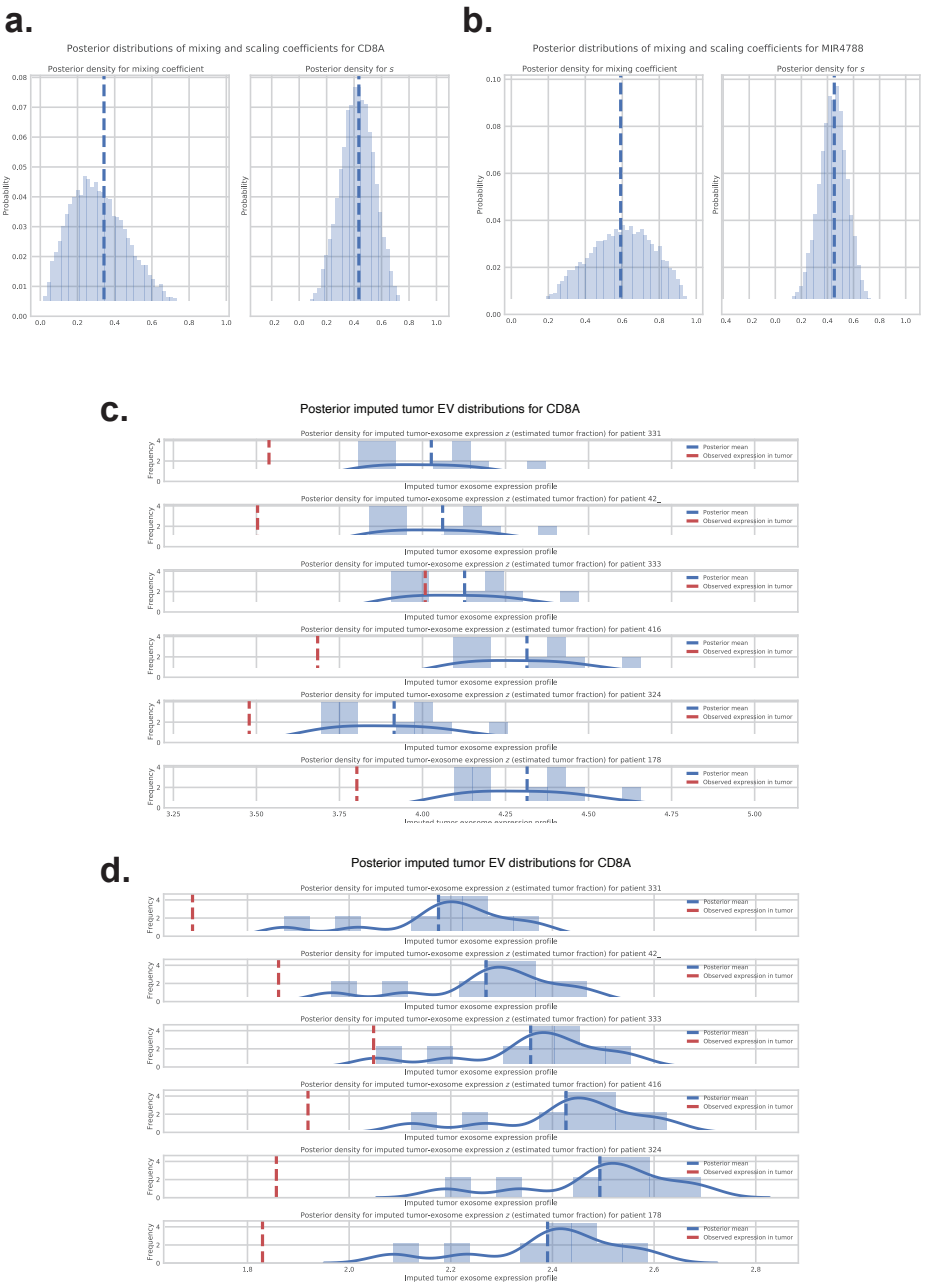


Figure 3-3: **Example of per-patient imputed tumor-EV expression from our Bayesian deconvolution model**. (a-b) Posterior estimates for two illustrative genes for high immune fraction (CD8A) and high tumor fraction (MIR47888). (c-d) Predicted tumor-derived EV expression from our deconvolution model for CD8A and MIR4788 for a subset of patients

3.4 Discussion

To address tissue of origin of EV transcripts, we developed a deconvolution model to characterize EV transcripts from tumor versus non-tumoral sources that explicitly accounts for differential EV transcript packaging. Currently, our model can only differentiate between tumor versus non-tumor contributions; however, ongoing experiments utilizing cell-specific EV selection and/or depletion may enable us to differentiate between specific sources. Our deconvolution model is limited by three major factors: (i) the simplifying assumptions regarding the linear nature of the packaging coefficient and how its shared between in vitro and in vivo samples, and (ii) lack of accounting for both tumor - especially potential immune infiltration in the tumor microenvironment - and patient heterogeneity, which is in part due to (iii) the limited number of samples. These limitations could potentially explain the classification of predominantly immune genes (e.g., CD8) into tumor compartments by our model. Thus, quantitative estimates of tumor purity should be interpreted in a qualitative fashion until in-depth in vitro experimental verification. As we continue to analyze data from more patients and perform in vitro EV selection experiments, we anticipate that our current model will serve as the foundation for more sophisticated models that can address these issues. Despite these limitations, our deconvolution model is the first to be able to pinpoint the potential source of EV expression and generate testable hypotheses regarding tissue-of-origin. Work is ongoing to experimentally validate the predicted source of circulating EVs via both tumor and immune cell-specific EV selection, which will be used to iteratively improve our deconvolution model and establish potential causal roles for EV transcripts in driving ICI resistance.

Chapter 4

Chromatin state changes during immunotherapy response reveals a non-responsive enhancer signature

4.1 Introduction

In recent years, there has been tremendous progress in melanoma immunotherapy including the FDA approval of anti-CTLA-4 antibodies (2011) and anti-PD-1 antibodies (2014). Though response rates for monotherapy with these agents are modest (15% for anti-CTLA-4 and 44% for anti-PD-1), a subset of responses are often durable [69, 70, 71, 72], with 2-year survival rates up to 43% among patients who receive anti-PD-1 monotherapy, and a 10-year survival rate of 20% for those who receive anti-CTLA-4 monotherapy [72, 73]. Although response rates are significantly increased [74] with combination anti-PD-1/anti-CTLA-4 therapy, a significant proportion of patients still do not achieve clinical response and the toxicity is high [74]. Therefore, there is a tremendous unmet need to identify biomarkers that predict response or resistance to immune checkpoint blockade (ICB) – either as monotherapy or in combination – and to identify actionable strategies that will enhance the effectiveness of these potent therapies in the patients most likely to benefit.

The epigenome consists of an array of chromatin modifications, including DNA methylation and histone marks, which collectively form a dynamic state that is referred to as a 'chromatin state'. The nature of chromatin states and their impact on associated genomic loci are determined by their constituent histone or DNA modification marks [75]. For example, the presence of the H3K27me3 (tri-methylation of lysine 27 on histone H3) mark in promoters is associated with transcriptional repression, whereas H3K4me3 (tri-methylation of lysine 4) is associated with transcriptionally active promoters. H3K4me1 and H3K27Ac modified nucleosomes are only present at enhancer elements, whereas the presence of H3K79me2 or H3K36me3 coincides with transcribed regions [76]. Thus, profiles of histone modification marks generate a comprehensive map of the epigenome.

Recent data indicate that responsiveness to ICB therapy may be associated with specific epigenetic processes. For example, regulation of histone modifications by HDAC, EZH2 or KMT2D has been proposed to modulate either response to these agents or antitumor activity of immune cells [77, 78, 79]. However, we do not have sufficient understanding of the epigenome content of sensitive and resistant patients to ICB. Furthermore, whether specific patterns of chromatin modification states are associated with response to immune checkpoint inhibitors has not been systematically investigated. As chromatin modification states are stable and heritable, specific pattern of chromatin modification states can potentially be used as biomarkers [80]. By generating epigenome profiles of 36 samples treated with ICB at MD Anderson Cancer Center (MDACC) followed by validation in an independent cohort of 30 samples treated with ICB at Massachusetts General Hospital (MGH), we demonstrate that the enhancer signature of 410 genomic loci in pre-treatment samples can predict non-response to ICB. Enhancer gains in non-responders were observed on a number of resistance-driving genes and enhancer-blocking bromodomain inhibitors synergized with anti-PD-1 antibody in pre-clinical models. Together, we identify enhancer gains as a key epigenetic mechanism driving resistance to anti-PD-1 therapy in melanoma which could also be leveraged for biomarker development or novel therapeutic combinations.

4.2 Methods

4.2.1 Patient samples

Tissue samples from metastatic melanoma patients were collected and viably frozen as part of an IRB-approved tissue banking protocol at the University of Texas MD Anderson Cancer Center and Massachusetts General Hospital. All patients signed written informed consent prior to having the sample collected. All patients received either pembrolizumab or nivolumab as the anti- PD-1 therapy for their metastatic melanoma. Thirty-six melanoma tumor samples (19 samples at baseline, and 17 samples at post treatment) from MDACC and 30 samples (10 samples at baseline, and 20 samples at post treatment) from MGH were analyzed by ChIP-seq. We also analyzed 12 MDACC pre-treatment samples by RNA-access for RNA expression. Response rate are accessed based on RECIST criteria.

4.2.2 Cell lines

Short term culture tumor cells and TILs were obtained from same anti PD-1 resistant melanoma patient. Short term culture tumor cells were cultured in RPMI + Glutamax supplemented with 10% FBS, Sodium Bicarbonate, HEPES, Human Transferred Insulin and b-Mercaptoethanol. TILs were cultured in RPMI + Glutamax supplemented with 10% Human serum, Sodium Bicarbonate, HEPES, Human Transferred Insulin and b-Mercaptoethanol. The B16-F10 and iBiP melanoma cell lines, 293T cells were cultured in complete DMEM high glucose, supplemented with 10% FBS. All cell lines were cultured at 37 C with 5% CO2.

4.2.3 Animal studies

All animal studies were performed according to University of Texas MD Anderson Cancer Center Institutional Animal Care and Use Committee (IACUC) approved protocols. Five million B16-F10 or iBiP melanoma cells were injected into 6-8 weeks old C57BL/6 mice (The Jackson Laboratory, 000664 |Black 6) via subcutaneously

and monitored every other day for tumor growth. Mice with established disease were randomly divided into 4 cohorts and treated every other day with IgG (100 µg/mouse), PD1 (100 µg/mouse), GSK-762 (7.5mg/kg) or vector PBS via subcutaneous injection. Tumor volume was measured every other day. Mice were treated for indicated number of days and euthanized once any arm of the treatment developed tumors approaching or beyond IACUC-approved limit of 1.5cm.

4.2.4 MDACC Chromatin immunoprecipitation

Chromatin immunoprecipitation was performed as described earlier33 with optimized shearing conditions and minor modifications. ChIP of 5-10 mg flash-frozen melanoma tumour was performed using 2mg antibody per ChIP experiment for H3K4me1 (Abcam ab8895), H3K27ac (Abcam ab4729), H3K4me3 (Abcam ab8580), H3K79me2 (Abcam ab3594) or 3mg antibody per ChIP experiment for H3K27me3 (Abcam ab6002). Enriched DNA was quantified using Qubit (Thermo Fisher Scientific) and ChIP libraries were amplified and barcoded using the NEBNext® Ultra™ II DNA library preparation kit (New England Biolabs) according to the manufacturer's recommendations. Following library amplification, DNA fragments were AMPure XP beads (Beckman Coulter) size-selected (200 - 600 bp), assessed using Bioanalyzer (Agilent Technologies) and sequenced at Sequencing & Microarray Facility (The University of Texas MD Anderson Cancer Center) using Illumina Hi-Seq 2000 36-bp single-end sequencing.

4.2.5 MGH Chromatin immunoprecipitation

20-50 mg of Snap-frozen melanoma tissues were pulverized by GenoGrinder for 2 min at 1500 rpm and then fixed with 1% methanol-free formaldehyde plus protease inhibitors cocktails (Roche, Inc) for 10 min at room temperature, and quenched by 125 uM glycine for 5 min at room temperature. Samples were incubated in cold RIPA buffer supplemented with protease inhibitor, and sonicated using Covaris E220. Supernatants were quantified using BioRad protein assay kits, and 1 mg of protein were

loaded on 96 well plates for chromatin immunoprecipitation.

Protein A/G coated silica columns embedded pipet tips were used for immunoprecipiating H3K27Ac antibody bounded proteins instead of Protein A/G beads. The DNA was eluted in 100 ul 50 mM Tris pH 8.0, 10 mM EDTA + 1% SDS after several washes and the elutes were treated with proteinase K for 16 hours at 65C before library synthesis using NEBNext Ultra II DNA library preparation kits (NEB, Inc). The samples were sequenced on Hiseq 2000 (Illumina, Inc) and 30-50 million paired-ended reads from each sample were recorded.

4.2.6 ChIP-seq analysis

ChIP-seq data were quality controlled and processed by pyflow-ChIPseq [81] a snakemake [82] based ChIPseq pipeline. Briefly, raw reads were mapped by bowtie1 [83] to hg19. Duplicated reads were removed and only uniquely mapped reads were retained. RPKM normalized bigwigs were generated by deep tools [84] and tracks were visualized with IGV [85]. Peaks were called using macs1.4 [86] with a p-value of 1e-9. Chromatin state was called using ChromHMM [19] and the emission profile was plotted by ComplexHeatmap [87]. Heatmaps were generated using R package EnrichedHeatmap. ChIP-seq peaks were annotated with the nearest genes using ChIPseeker [88]. Super-enhancers were identified using ROSE [89] based on H3K27ac ChIP-seq data.

4.2.7 Epilogos analysis

chromHMM profiles of 5 pre-treatment non-responders and 6 pre-treatment responders are consolidated using epilogos. With the output of Epilogos, the chromatin state for each bin is chosen for the state that contains the greatest weights.

4.2.8 Chromatin transition circos plot and transition heatmap

The consolidated chromHMM profiles by epilogos were compared. The number of bins that switch from one state to a different state in one group to the other group

was obtained. The number of bases that showed the transition change was obtained by multiply the number of bins with the bin size (1000bp). A circos transition plot was made by circlize R package.

The consolidated chromHMM profiles by epilogos were read into R package EnrichedHeatmap. The chromatin state (categorical variable) was plotted in a 25kb window centered on the active enhancer bins (chromatin state E7). Only bins that have E7 in one of the group were retained to plot. For two group comparisons, the bins were merged together if the same state change occurs in consecutive bins.

4.2.9 RNA-access sequencing and analysis of MDACC tumor

mRNA libraries of the melanoma tumor (n = 12) samples were prepared from 200 ng of total RNA with the use of the TruSeq Stranded mRNA HT sample preparation kit. Samples were dual-indexed before pooling. Libraries were quantified by qPCR with the use of the NGS Library Quantification Kit. Pooled libraries were sequenced by using the HiSeq2000 (Illumina) according to the manufacturer's instructions. An average of approximately 30 million paired-end reads per sample were obtained. The quality of raw reads was assessed by using FastQC. The raw reads were aligned to the Homo sapiens genome (hg19) using STAR v2.4.2a [90]. The mappability of unique reads on average was 89% RNA-seq dataset. The raw counts were computed with the use of quantMode function in STAR. The read counts that were obtained are analogous to the expression level of each gene across all the samples. The differential expression analysis was done by using DESeq2 [91]. Genes with raw mean reads of greater than 10 were used for normalization and differential gene expression analysis using DESeq2 package in R. Genes with absolute log2 fold-change greater than log2(1.5) and p value < 0.05 were called as differentially expressed genes. SKCM TCGA RNA-seq transcription comparison analysis was performed on the UALCAN website [92].

4.2.10 RNA-seq sequencing and analysis of MGH tumors

Total RNA from 5-20 mg of melanoma primary and metastatic tissues were extracted using AllPrep DNA/RNA mini isolation kit (Qiagen, Inc). 100 ng of total RNA was used as input for RNA-seq libraries using SMARTer Stranded Total RNA-seq - Pico input (Takara Bio USA, Inc) to remove rRNA transcripts. Each library was sequenced on Hiseq 2000 (Illumina, Inc) and approximately 20 million single-ended reads were recorded. Reads were aligned to Homo sapiens reference hg38 using STAR v2.5.3. Read counts were quantified using featureCounts. Differential expression was performed via limma-voom [23]. Multiple biological replicates stemming from the same patient was treated as a random effects, whereas batch effects were treated as a fixed effect.

4.2.11 HiChIP and Data Analysis

HiChIP experiments were performed as previously described by Mumbach et al. [93], with minor modifications. Briefly,1 x 10^7 STC2765 cells were crosslinked. In situ contacts were generated in isolated and pelleted nuclei by DNA digestion with MboI restriction enzyme, followed by biotinylation of digested DNA fragments with biotin–dATP, dCTP, dGTP, and dTTP. Thereafter, DNA was sheared with Covaris E220 with the following parameters: fill level =10, duty cycle = 5, PIP = 140, cycles/burst = 200, and time = 4 min; chromatin immunoprecipitation was done for H3K27Ac with use of anti-H3K27ac antibody. After reverse-crosslinking, 150 ng of eluted DNA was taken for biotin capture with Streptavidin C1 beads followed by transposition with Tn5. In addition, transposed DNA was used for library preparation with Nextera Ad1_noMX, Nextera Ad2.X primers, and Phusion HF 2XPCR master mix. The following PCR program was performed: 72C for 5 minutes, 98C for 1 minute, then cycle at 98C for 15 seconds, 63C for 30 seconds, 70C for 1 minute. Afterward, libraries were two-sided size selected with AMPure XP beads. Finally, libraries were paired-end sequenced with reading lengths of 76 nucleotides. HiChIP paired-end reads were aligned to the MboI digested hg19 genome using the HiC-Pro

pipeline with default conditions. The default setting of HiC-Pro removes duplicate reads, assigns reads to MboI fragments, identifies valid interactions and generates hi-resolution interaction matrices. HiChIP for H3K27ac generated high-resolution contact maps containing 65 million valid interactions in STC2765 cells. Files for Juicebox visualization were generated using the HiC-Pro hicpro2juicebox.sh command based on the total valid interactions. Identification of H3K27ac mediated loops was performed with the hichipper/diffloop programs using the HiC-Pro [94] output and ChIP-seq peaks from H3K27ac as anchor loci. Hichipper identifies intrachromosmal looping between anchor loci within 5kb-2MB and produces a per-loop FDR value from the loop proximity bias correction implemented by Mango. Using the mango output from hichipper [95], diffloop was used to filter significant loops (FDR < 0.01, width > 5000, loop-count > 2) and define enhancer-enhancer and enhancer-promoter interactions.

4.2.12 In vitro inhibitor assays

Melanoma short-term culture line STC2765 were treated with crizotinib (2µM, 24hrs) or iBET762 (1µM, 72hrs) prior to co-culture with TIL2765.

4.2.13 Enhancer modulation using CRISPR-dCas9-KRAB

In order to modulate gene expression without altering the target DNA sequences, an RNA-guided, catalytically inactive Cas9 (dCas9) fused to a transcriptional repressor domain (KRAB) was used to silence genomic regions identified as enhancers via KRAB repression at the promoter region. To generate a dCas9-KRAB effector stable cell line, we produced lentiviral particles from pHAGE EF1α dCas9-KRAB (Addgene plasmid 50919) using a standard protocol. Transduced cells were selected for 6 days with the use of antibiotic resistance and were expanded to generate a stable cell line. Next, gRNAs were designed by using the GPP Web Portal of the Broad Institute. Annealed gRNA oligos were ligated to pLKO.1-puro U6 sgRNA BfuAI stuffer (Addgene plasmid 50920), and lentiviral particles were generated. A transduction procedure

was performed in the stable dCas9-KRAB cell line, and transduced cells having both dCas9-KRAB and gRNA constructs were selected with the use of antibiotic resistance. To evaluate the effects of the recruitment of dCas9-KRAB to the target enhancer's genomic region, H3K27ac ChIP followed by quantitative PCR for enhancer regions was performed to assess the enrichment level of H3K27ac at the enhancer site in modulated cells compared with the non-modulated parental control cells. To investigate the impact of enhancers' modulation on the corresponding gene expression, qRT-PCR was performed for the target gene.

4.2.14 Tumor Infiltrating Lymphocytes (TILs) and matched Tumor cells co-culture

Harvested target tumor cells were labeled with DDAO-SE, then add effector cells suspension to achieve the desired E:T (Eeffector:Target) ratio. The mixtures were incubated at 37 °C, 5% CO2 in a humidified incubator for 3 hours. The cells were fixed and permeabilized with Fix/Perm solution (BD Biosciences, Cat. No. 554722) 20 min at RT immediately. The cells were stained for 30 min on ice with 5 µl biotin-labeled anti-cleaved caspase 3 monoclonal antibody (BD Biosciences, Cat. No. 550821). The cells were washed in Perm/WashTM buffer ((BD Biosciences, Cat. No. 554723) 2 times and re-suspended in D-PBS, 1% BSA for analysis on a flow cytometer.

4.2.15 Flow cytometry

TIL were stained with fluorochrome-conjugated monoclonal antibodies (CD3, CD4, and CD8 from BD Bioscience) in FACS Wash Buffer (Dulbecco's phosphate buffered saline 1x with 1% bovine serum albiumin) for 30 min on ice for surface staining. Dead cells were excluded using Ghost 450 cell viability dye from Tonbo Biosciences. For intracellular staining of active caspase-3, cells were fixed and permeabilized using Cytofix/Cytoperm (BD Bioscience) and stained with cleaved anti-caspase-3 (BD Bioscience) on ice as well. Acquisition of stained cells was done using BD FACSCanto II

and analyzed using FlowJo software (Tree star).

4.2.16 Pathway analysis

Differential enhancers associated genes in each group or in each cluster were imported into the ClusterProfiler [96] for pathway analysis, restricted to GO, KEGG, Hallmark and WiKi gene sets. The Enrichplot package was used to generate dotplot and networks for genesets enriched with a false discovery rate (FDR) cut-off of < 0.05.

4.2.17 Survival and statistical analysis

The survminer package was used for drawing the Kaplan-Meier plots and defining the optimal threshold (function surv). The outcome is overall survival censored at 10 years. P-values reported for the univariate model correspond to the logrank test.
The two-tailed Student's t-test was used to determine the statistical significance of two groups of data using GraphPad Prism. Data are presented as means ± standard error of the mean (SEM; error bars) of at least three independent experiments or three biological replicates. P-values less than 0.05 were considered statistically significant. *, P <0.05; **, P <0.01; and ***, P <0.001 indicate statistically significant differences.

4.3 Results

4.3.1 Defining chromatin states

To directly address whether epigenetic changes are associated with response to ICB therapy, we first performed epigenomic profiling of 36 metastatic melanoma samples from patients treated with nivolumab or pembrolizumab (anti-PD-1 antibody) at MDACC (Figure 4-1a). Response in these patients was documented using RECIST criteria, which identified 4 samples from patients achieved complete response, 4 partial response, 5 had stable disease and 23 had progressive disease in response to ICB therapy (Figures 4-2a-b). Overall, 13 samples from patients with the complete or

partial response or stable disease were annotated as "Responders (R)" and 23 samples from patients with progressive disease were labelled as "Non-responders (NR)" (Figures 4-2a-b). Samples were collected at three timepoints: 1. Pre-treatment, 2. On-treatment, and 3. Post-treatment.

To identify basic epigenome elements, we profiled 6 reference histone modifications that mark promoter (H3K4me3), enhancer (H3K4me1 and H3K27Ac), transcribed (H3K79me2) and repressed (H3K27me3 and H3K9me3) states using high-throughput ChIP-sequencing methodology [97, 98] in all 36 samples, generating 148 chromatin maps (Figure 4-2c). As histone modifications exert their function in a combinatorial fashion, we identified such chromatin states using the ChromHMM algorithm [19]. A 15 chromatin state model was chosen for more in-depth interrogation into the biology of chromatin in anti-PD-1 response as it presented sufficient resolution needed for biological interpretation (Figure 4-1b and Figure 4-2d). Annotation of these states based on the content of histone marks and their genomic locations revealed the presence of active promoter states (E1, E2, E3), active enhancer states (E6, E7), transcribed states (E4, E5), polycomb-enriched (E11), heterochromatin/bivalent (E9), poised (E8, E10) low (E12, E13, E14; merged as E12 afterwards) states (Figure 4-1b).

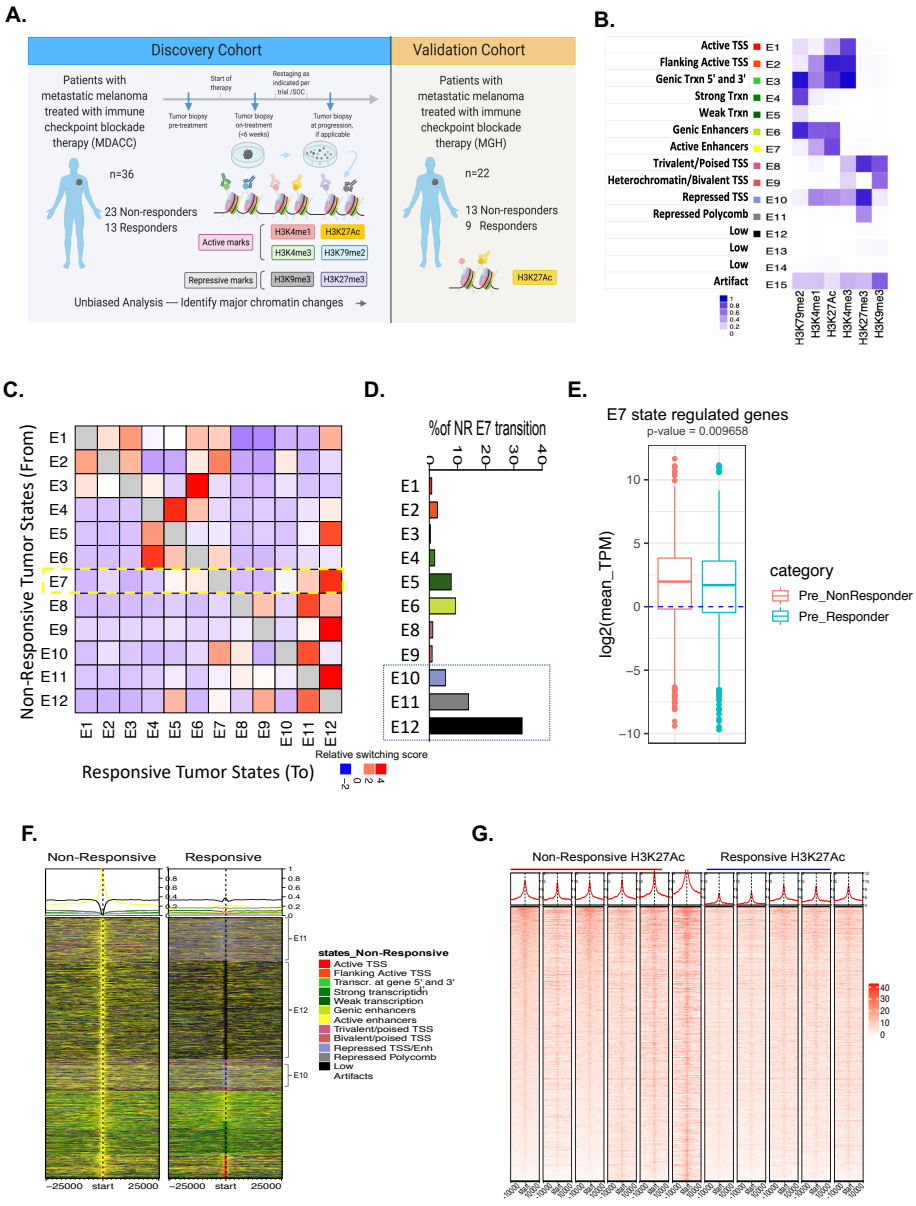


Figure 4-1: **Comprehensive epigenome profiling of anti-PD-1 treated melanoma patients**. (a) Outline of the study and description of patient sample (b) Emission parameters of the 15-state chromatin state model called from ChromHMM. (c) Heat map showing the fold enrichment of chromatin state transitions between responder and non-responder pre-treatment samples for the 15-state model defined by the ChromHMM. (d) Bar graph showing the percentage of non-responder active enhancer state E7 switch to any other states in responder. (e)Box plots showing the log2 mean expression levels (TPM) of genes associated with Enhancer State E7, genes were linked using H3K27ac HiChIP data. (f) Heatmap of chromatin state intensities for 31,155 loci that show switch from E7 in nonresponder pre-treatment samples to any other state in responder pre-treatment samples.

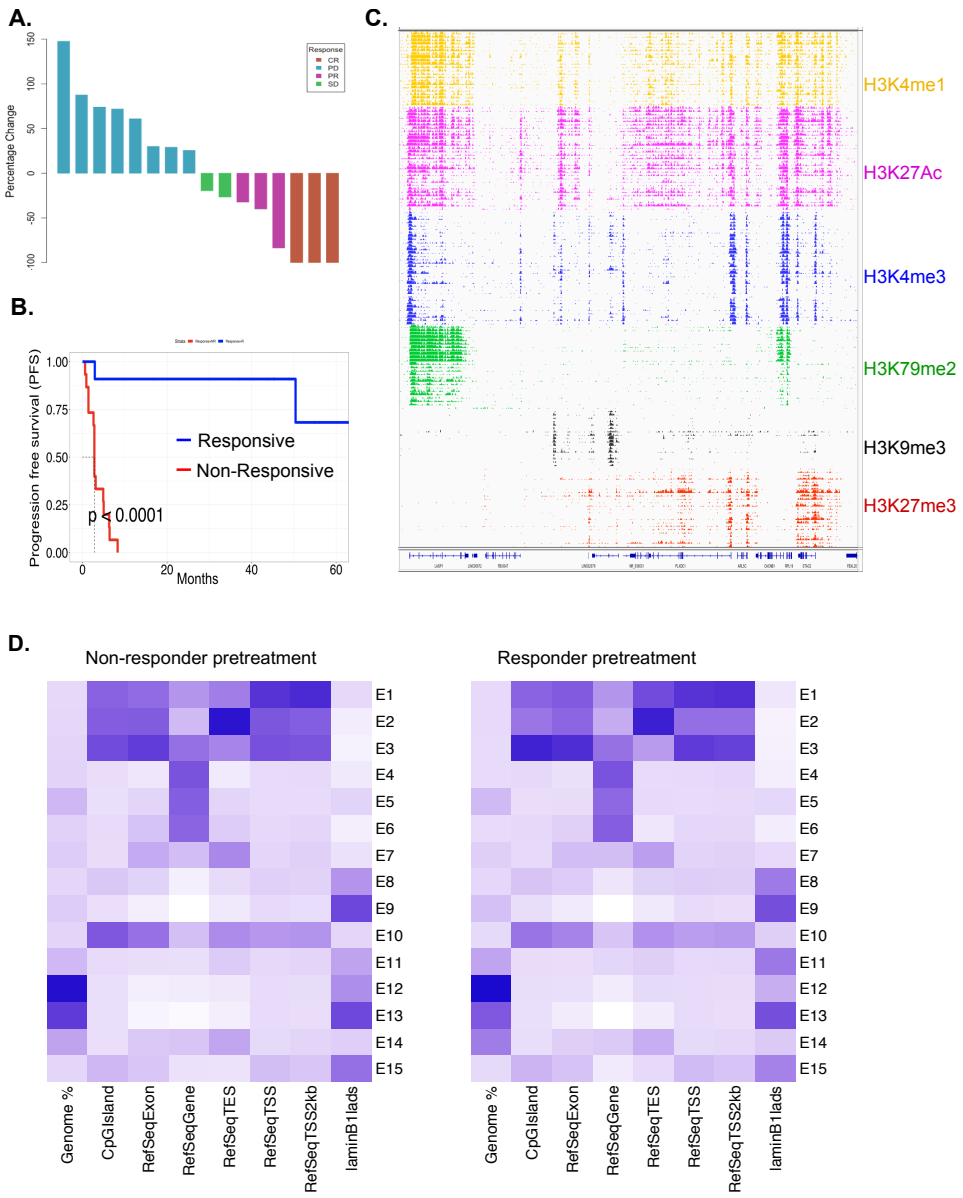
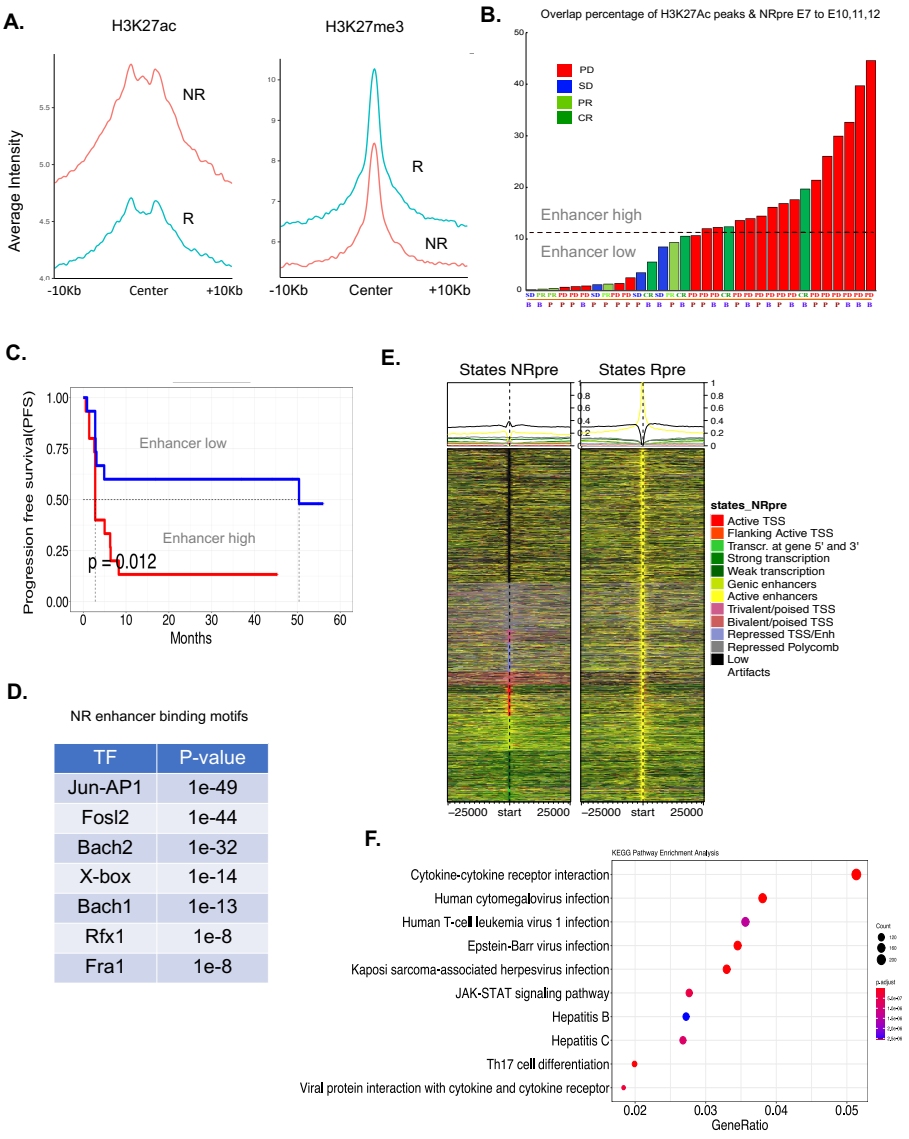


Figure 4-2: **Characteristics and chromatin maps of ICB-treated melanoma patients**. (a) Waterfall plot showing responsiveness of MDACC patients to anti-PD-1 therapy in a subset of patients where 6 months RECIST data was available. (b) Difference in progression free survival between responders and non-responders to anti-PD-1 therapy in melanoma. (c) IGV view of 6 different histone mark (as noted on right side) profiles on the shown chromosomal region in all the anti-PD-1 treated patients. (d) Genomic annotation enrichments for each chromatin state in anti-PD-1 responder and non-responder tumor samples.

4.3.2 Chromatin state transitions between sensitive and resistant lesions

In order to interrogate an epigenomic signature of response, we unbiasedly identified chromatin states that differentiate between pre-treatment samples belonging to the responsive and non-responsive groups. Epilogos (see section 4.2.7) based convergence of chromatin states followed by computation of transitions between them in the responder versus non-responder samples showed important differences in epigenomic features of anti-PD-1 response (Figure 4-1c). The major transition consisted of those in the active enhancer state E7 in non-responder samples to low (E12), polycomb (E11) or repressed states (E10) in responders based on number of switching bins as well as differences in expression of the associated genes in the responder and non-responder groups (Figure 4-1d-e). We identified 31,555 bins (1kb segments) that showed transition between active enhancer E7 state in non-responsive patients to Low, repressive states E10, E11 and E12 in responsive samples (Figure 4-1f). Analyzing signal of H3K27ac on these set of loci also showed similar loss of H3K27ac signal in pre-treatment samples from responder population compared to those from non-responder patients (Figure 4-1g and Figure 4-3a). Overall, average intensity profile of H3K27ac on these enhancers showed drastic increase in non-responder samples in comparison to responder samples, whereas average intensity profiles for H3K27me3 occupancy on these enhancers was significantly increased (Figure 4-3a).



TF	P-value
Jun-AP1	1e-49
Fosl2	1e-44
Bach2	1e-32
X-box	1e-14
Bach1	1e-13
Rfx1	1e-8
Fra1	1e-8

Figure 4-3: **Enhancer activation in non-responders to ICB in melanoma**. (a) Average intensity plots for H3K27ac (left) and H3K27me3 (right) in loci that lose H3K27ac marks (b) Bar plot showing percent overlap of the active enhancer loci which was calculated in each sample to define an enhancer-presence score. (c) Progression-free survival for enhancer-low or enhancer-high groups as described in panel B above. (d) List of enriched transcription factor (TF) motifs in non-responder specific active enhancers. (e) Heatmap of chromatin state intensities for 20,194 loci that show switch from E7 in responder pre-treatment samples to any other state in non-responder pre-treatment samples as shown by colors for each state. (f) Dot plot showing the significantly enriched pathways in genes targeted by responder E7 state enhancers that depleted in non-responder.

4.3.3 An enhancer signature predicts response to anti-PD-1 therapy in melanoma

Overlap percentage of active enhancer loci were calculated in each sample to define an enhancer-presence score. Using median score to guide cut-off points, the enhancer-high group was defined as samples above 11.3% (n=16), while enhancer-low group constituted the remaining samples. Objective response rate, defined as either partial response (PR) or complete response (CR) to therapy, was higher in the enhancer-low group than in the enhancer-high group (Figure 4-3b). Notably, PFS and OS were also statistically significantly different between the enhancer-low and enhancer-high groups [PFS: 2.8 months (enhancer-high) versus 34 months (enhancer-low), p = 0.0032; OS: 17.4 months (enhancer-high) versus 37 months (enhancer-low), p = 0.0063] (Figure 4-3c). These enhancers were enriched for binding motifs of Jun-AP1, Fosl2 and Bach1/2 among others (Figure 4-3d). We also noted that a set of enhancers were enriched in the E7 (active enhancer) state in responder samples in comparison to non-responders which mostly enriched surrounding genes involved in T-cells which suggests increased lymphocyte infiltration in responder samples (Figure 4-3e-f).

4.3.4 Enhancer activation upregulates genes contributing to anti-PD-1 resistance

Pathway analysis of active enhancers that transitioned from E7 in non-responders to E10/E11/E12 in responders revealed enrichment in various signaling pathways including those earlier shown to alter anti-tumor immune response or modulate immunotherapy response such as MET pathway, Notch pathway, NCAM/integrin signaling (Figure 4-4a) and included c-MET, (Figure 4-4b and Figure 4-5a) with associated change in expression (Figure 4-4c and Figure 4-5b). To understand the source of H3K27ac peaks observed to be enriched in NR tumors, we also generated the H3K27ac ChIP-Seq data on 10 short-term melanoma cultures (STCs) and 8 cognate tumor-infiltrating T cells (TILs) derived from patients. To identify gene targets of H3K27ac marked enhancers, we generated HiChIP data that identifies 3D interaction

of H3K27ac peaks with distal genomic loci. C-MET locus showed multiple distal enhancers that were present in non-responsive tumors, but not in responsive tumors and HiChIP data showed 4 distal enhancers E1, E2, E3 and E4 looping to the gene body/TSS of c-MET gene (Figure 4-4b, 4-4d). These enhancers were also present in a melanoma culture derived from ICB non-responder patient (STC2765) as suggested from overlapping H3K27ac peaks (Figure 4-4b) and c-MET expression was localized to melanomas cells (Figure 4-5C). Silencing of these enhancers using specific gRNAs and dCas9-KRAB22 repressed the gene expression in STC2765 (Figure 4-4d). The cell lines with dCas9-KRAB mediated enhancer suppression also showed increased tumor killing by cognate T cells (TIL2765) that were derived from the same tumor as STC2765 in a co-culture assay, thus demonstrating enhancer functionality (Figure 4-4e). Consistently, suppression of c-MET activity via Crizotinib also showed enhanced T-cell mediated killing of STC2765 cells by TIL2765 (Figure 4-4f).

To gain a better understanding of enhancer gains in gene expression, we generated gene expression (RNA-seq) data on 47 ICB treated samples consisting of 25 pre-treatment (14 NR and 11 R) and 22 on-treatment (16 NR and 6 R) samples (Figure 4-5d). We identified 922 gene targets of reproducibly enriched 752 enhancers (FDR <0.1) in anti-PD-1 non-responsive samples by overlapping H3K27ac HiChIP data from STC2765 and by leveraging the enhancer-promoter annotation from 935 samples [99]. To gain a better understanding of the heterogeneity of the enhancer patterns, we overlapped the 752 NR-specific and 747 R-specific peaks with those H3K27ac ChIP-Seq peaks in patient-derived melanoma cells or T cells. We noted gains of enhancers either in tumor cells or in T cells surrounding numerous genes that are linked to regulation of immune microenvironment or anti-tumor immune responses (For example, NOTCH1, AKT1, USP22, MYC in melanoma cells and CISH, LEF1 in T cells) or other potentially novel regulators (TGF-β2, MITF, FAM20C, RFPL2, MAMDC2, SPATA2 in melanoma cells and FKBP3, LGALSL, LARP1 in T cells) (Figure 4-4g, 4-5e, 4-6a-d). We noted concomitant upregulation of gene expression of a subset of these genes either at the pre-treatment stage or during on- or post-treatment stage (Figure 4-4h and Figure 4-6e). Importantly, we also identified enhancers on

other important inhibitory checkpoint receptors such as LAG3 [100] and BTLA [101], or their key partners such as CEACAM-1 (required for function of TIM-3 [102]) which were present in non-responder tumors and isolated infiltrating T-cells (TILs) (Figure 4-4i-k and Figure 4-7a). In addition, we also noted enhancer enrichment surrounding other receptors from T cells, such as CD244, CD48, CADM3, HVEM [103] that mediate key interactions with antigen presenting cells or tumor cells (Figures 4-7b-d). Finally, we also observed enhancer gains on important transcriptions such as NR4A1 which known to drive T cell exhaustion 28 and others such as CEBPβ and KLF6 (Figures 4-7e). Overall, these data suggest that activation of enhancers could be a key epigenetic mechanism for activation of regulators and processes that promote resistance to ICB.

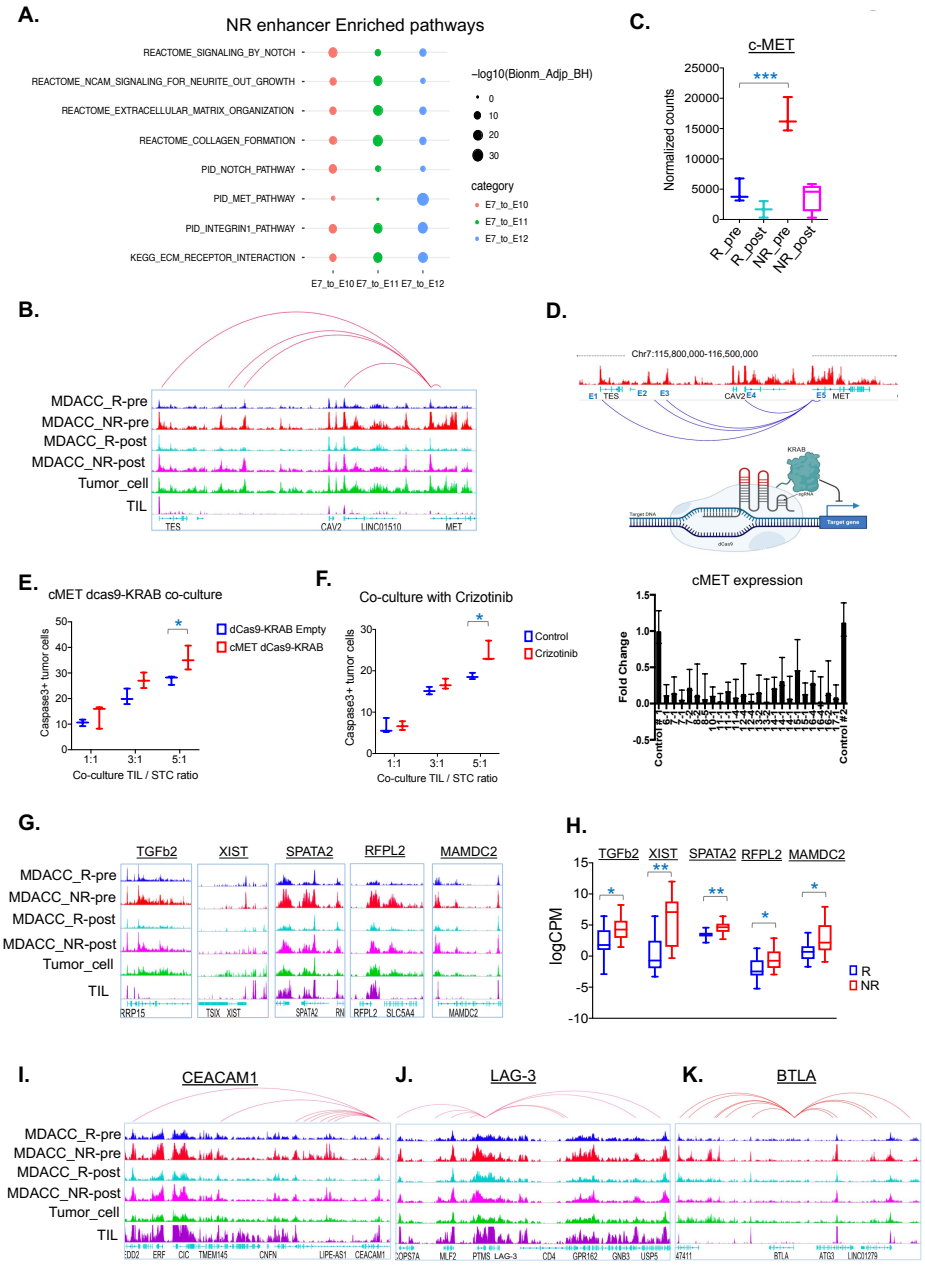


Figure 4-4: **Enhancer activation marks a number of resistance-associated genes in anti-PD-1 non-responders**. (a) Pathway analysis (GREAT) based on differentially enriched loci in E7 state in non-responder to E10, E11 or E12 in responders (b) IGV snapshot of aggregate H3K27ac profiles around c-MET. (c) Normalized RNA counts as a representation of gene expression for c-MET between non-responder and responder samples at pre- and post-treatment stages. (d) Enhancer locations and HiChIP-derived loops between enhancers and c-MET gene. (e-f) Percentage of cleaved caspase-3 in target tumor cells. (g) IGV snapshot of aggregate H3K27ac profiles around TGFb2, XIST, SPATA2, RFPL2 and MAMDC2 genes. (h) Box plot showing the gene expression level of TGFb2, XIST, SPATA2, RFPL2 and MAMDC2 genes. (i-k) IGV snapshot of aggregate H3K27ac profiles around CEACAM-1(i), LAG-3(j) and BTLA(k).

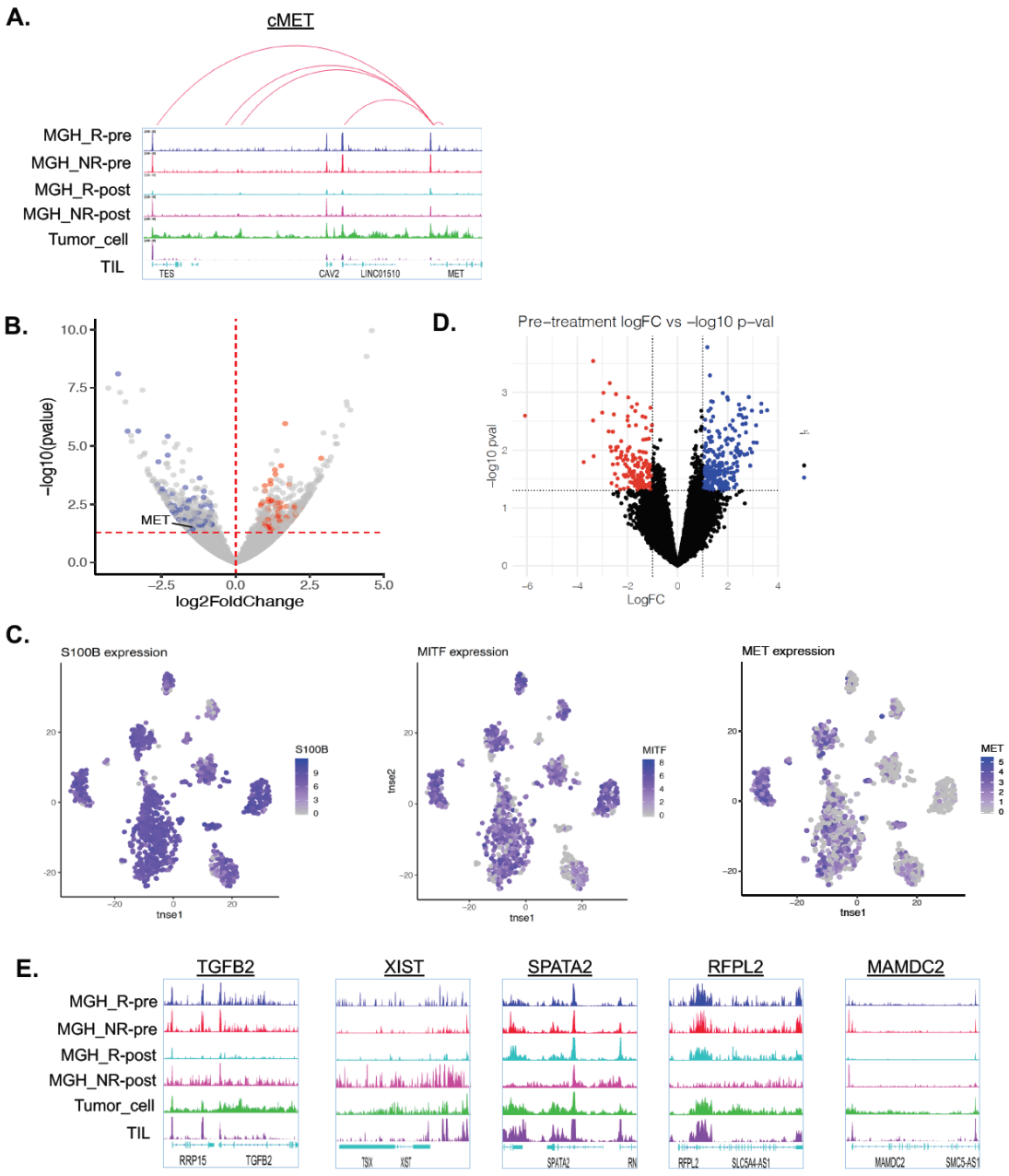


Figure 4-5: **Identification of gene targets for enhancers enriched in ICB non-responder melanoma samples**. (a) IGV snapshot of aggregate H3K27ac profiles around c-MET (left) and TGFβ2 (right) in non-responder samples, responder samples, isolated melanoma short term cultures (STCs) or isolated TILs (b) Volcano plot showing MDACC cohort differentially expressed genes (gray dots) and differentially enriched enhancers targeted genes. (c) Distribution of expression of S100B, MITF and MET genes in 2-dimensional embedding obtained by tSNE. (d) Volcano plot showing MDACC and MGH cohort combined expression data differentially expressed genes (blue and red) in responder vs non-responders. (e) IGV snapshot of aggregate H3K27ac profiles around XIST, SPATA2, RFPL2 and MAMDC2.

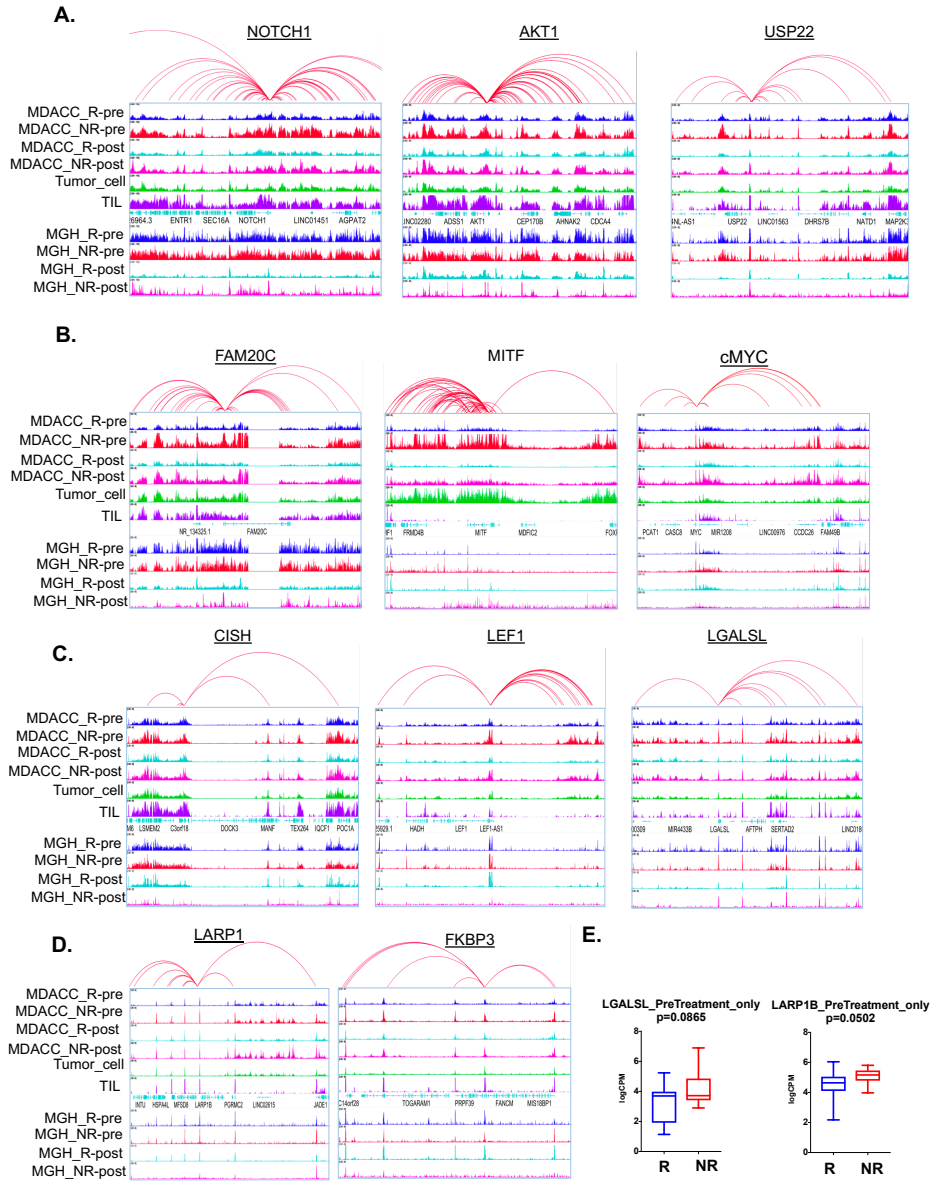


Figure 4-6: **Enhancers enriched in ICB non-responders activate important resistance-causing genes in melanoma cells**. (a-d) IGV snapshot of aggregate H3K27ac profiles around NOTCH1, AKT1 and USP22 (a); FAM20C, MITF and cMYC (b); CISH, LEF1 and LGALSL (c); LARP1 and FKBP3 (d) in non-responder samples, responder samples, isolated melanoma short term cultures (STCs) or isolated TILs. The top track shows the HiChIP data derived from STC2765 cells. (e) Box plot showing the gene expression level of LGALSL and LARP1 genes in non-responder and responder pretreatment samples.

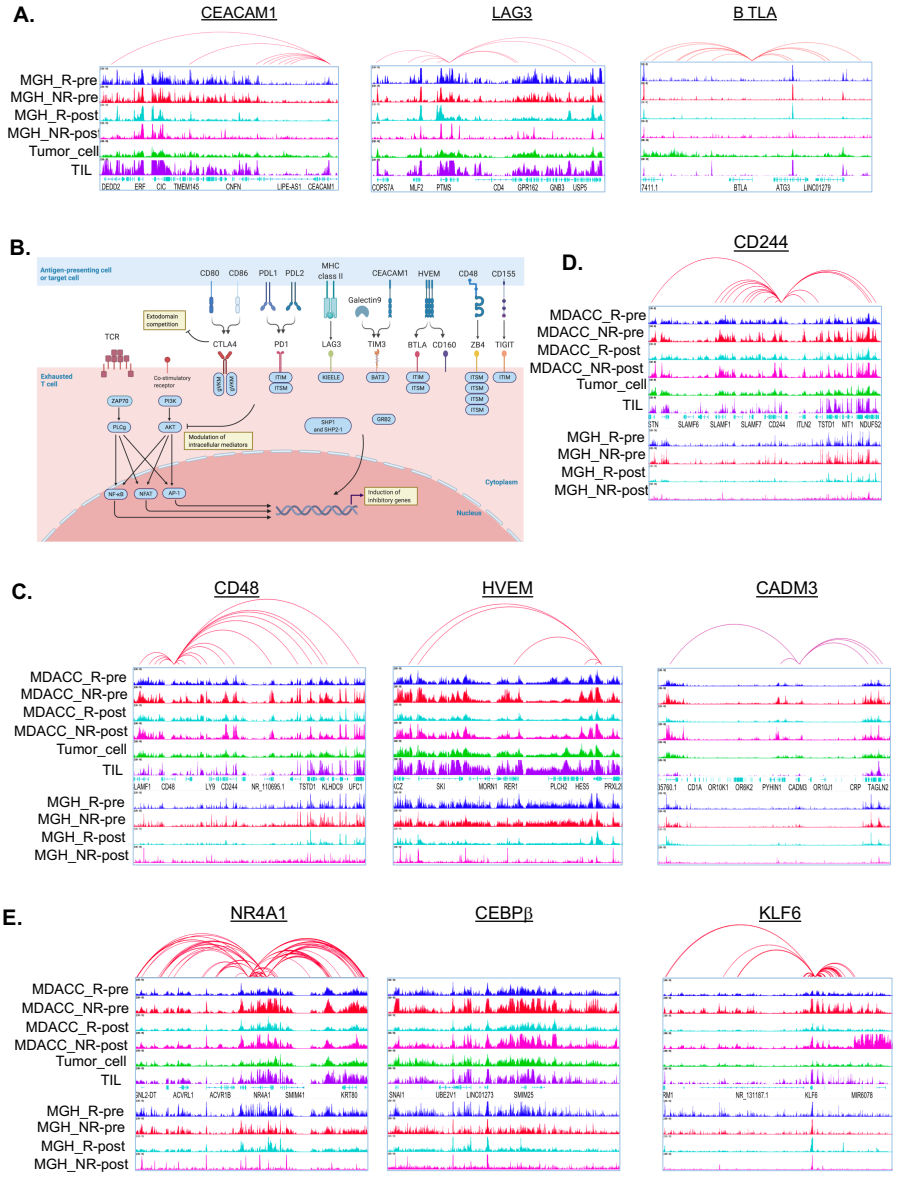


Figure 4-7: **Enhancers enriched in ICB non-responders activate important checkpoint receptors in TILs** . (a) IGV snapshot of aggregate H3K27ac profiles around CEACAM1, LAG3 and BTLA in non-responder samples, responder samples, isolated melanoma short term cultures (STCs) or isolated TILs. The top track shows the HiChIP data derived from STC2765 cells. (b) Schematic showing the key Immune checkpoint receptors on exhausted tumor-infiltrating T cells. (c-e) IGV snapshot of aggregate H3K27ac profiles around CD48, HVEM and CADM3 (c); CD244 (d); NR4A1, CEBPB and KLF6 (e) in non-responder samples, responder samples, isolated melanoma short term cultures (STCs) or isolated TILs.

4.3.5 Enhancer Reprogramming During ICB treatment

Chromatin state transition between pre- and post-treatment samples showed massive transitions from active states to repressed states in the responder samples, whereas those in non-responder samples were distributed more evenly between repressive and active states (Figure 4-8a). To determine the reprogramming of active enhancers during the treatment stage, we computed the chromatin state transition of active enhancer state E7 between post-treatment and pre-treatment samples (Figure 4-8b). The clustering of these states based on the transition revealed 4 clusters of which Cluster 1 enhancers gained repressive states or lost the active enhancer marking, whereas Cluster 4 enhancers remained in active enhancer state even at the post-treatment stage. Cluster 1 enhancers were enriched in VEGFA, autophagy, HIF1 signaling including VEGFA, RUNX3, AKT2 genes (Figure 4-8c-d). Unaffected Cluster 4 enhancers were enriched in TGFβ, PI3K/AKT/mTOR signaling pathways, AHR and oxidative stress pathways, including on genes such as TGFβ and LOXL4 (Figure 4-8e-f).

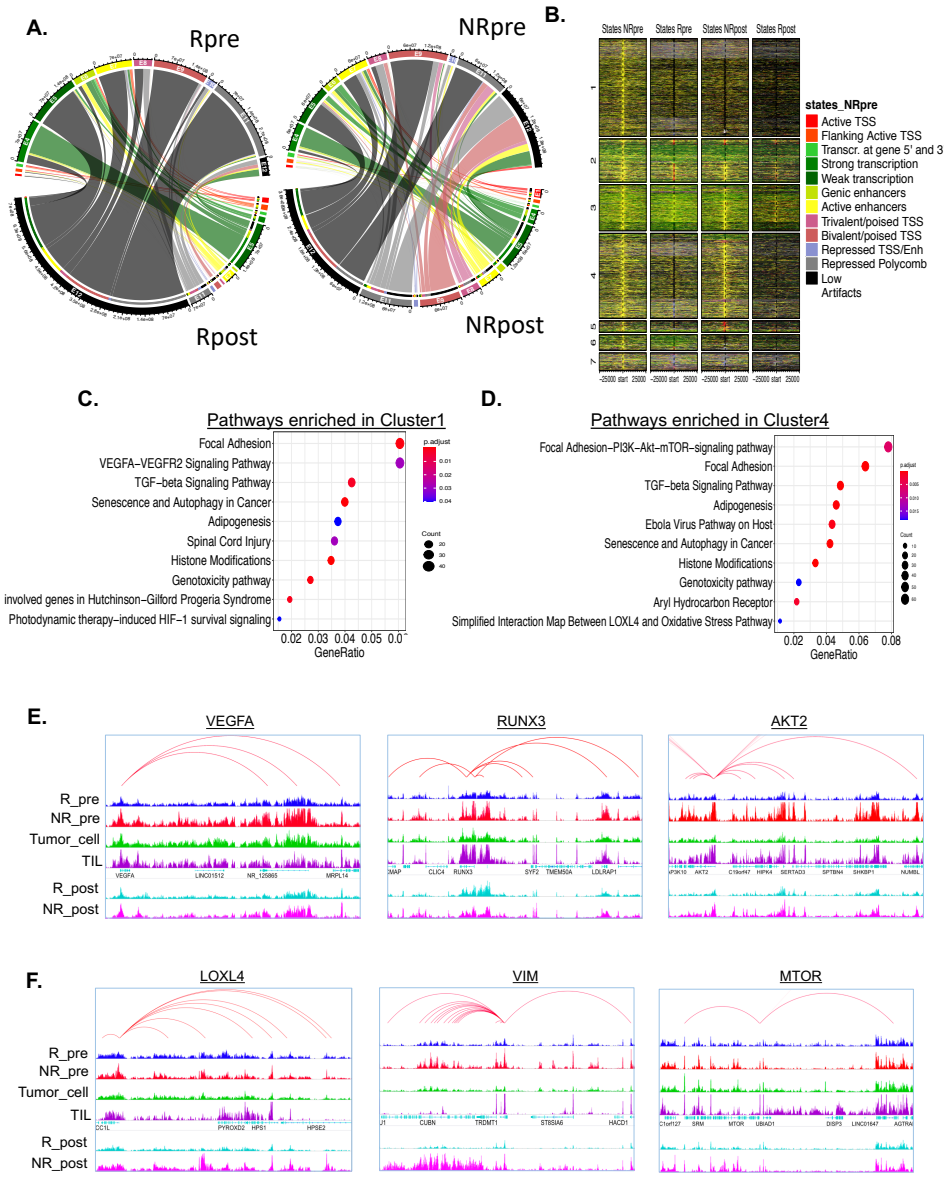


Figure 4-8: **Analysis of enhancer reprogramming between pre- and post-treatment ICB-treated melanoma samples**. (a)Circos plot showing chromatin state switches between responder pre-treatment and responder post-treatment samples (left) or non-responder pre-treatment and non-responder post-treatment samples (right) based on epilogos-derived transitions. (b) Heatmap of chromatin state intensities for 31,155 loci that show switch from E7 in non- responder pre-treatment samples to any other state (c-d) Dot plot showing the significantly enriched pathways in genes targeted by non-responder E7 state cluster1 (c) and cluster4 (d) enhancers (from panel A). Dot size represents the gene counts, adjusted p values are shown and are color-coded based on the level of significance (e-f)IGV snapshot of aggregate H3K27ac profiles around VEGFA, RUNX3 and AKT2 (e); LOXL4, VIM and MTOR (f).

4.3.6 Combination of BRD inhibitors with anti-PD-1 enhances the response in mouse melanoma models

Since enhancer activation marks multiple genes that regulate resistance to anti-PD-1 antibodies, we reasoned that inhibitors of acetylation-reader bromodomain could be used as an umbrella approach to target many resistance mechanisms at once along with anti-PD-1 therapy to enhance its efficacy. Consistently, we noted higher BRD4 levels in metastatic melanoma in the TCGA SKCM dataset (Figure 4-9a). The tumors harboring high levels of BRD4 survived poorly in comparison to those harboring lower levels of BRD4 (Figure 4-9b). Treatment of tumors generated by transplantation of murine melanoma cells, BP (from Bosenberg model [104]) and B16F10, with combination of iBET-762 with anti-PD-1 antibody significantly reduced tumor growth at doses which failed to generate much response when used as monotherapy (Figure 4-9c-d). Profiling of infiltrating CD8+ T-cells in these experiments suggested increased infiltration of these cells upon combination treatment in comparison to monotherapy (Figure 4-9e, 4-10a). In addition, treatment of STC2765 cells with bromodomain inhibitors increased the TIL2765 mediated killing in a co-culture assay (Figure 4-9f), also increased the MHC class I expression on tumor cells (Figure 4-9g). ChIP-Seq profiling of the tumors from mice treated with these agents showed a significant decrease of intensities of enhancers on c-MET, TGFβ, XIST, LAG3, MAMDC2, FKBP3, AKT1, MYC, SPATA2, NOTCH1, HES1, and SIK1 among others in the combination treatment but not in monotherapy suggesting that enhancer depletion may contribute to observed decrease of tumors growth in combination treatment (Figure 4-9h and Figure 4-10b-c). Indeed, RNA-Seq profiling revealed a loss of expression of a large number of genes in the combo treatment in comparison to the monotherapy or control IgG samples which coincided with loss of binding of BRD4 and modest loss of active enhancer marks, H3K27ac (Figure 4-9i). These genes were enriched in WNT, TGFβ, epithelial-to-mesenchymal transition (EMT) and UV response pathways (Figure 4-9j, 4-10d). Overall, this data provide evidence toward enhancer mediated activation of key resistance-driving genes/pathways as an epigenetic mechanism for resistance

to ICB and need for clinical studies focused on the combination of enhancer blocking agents in combination with ICB to improve the response rate in melanoma and potentially other malignancies.

4.4 Discussion

Our data help address two major clinical needs regarding ICB therapy in metastatic melanoma: 1) biomarkers that predict response and 2) combination therapy strategies to improve the response to ICB. We observed that gains of enhancer activity on a set of genomic loci are associated with response to ICB, which could potentially act as a predictive biomarker of response to ICB in metastatic melanoma. Our data also suggests causative roles for enhancers gains in non-response to ICB and supports the use of enhancer blocking clinical agents with anti-PD1 as a potential strategy that can be tested in future clinical trials. Importantly, the overlap between the MDACC and MGH cohorts was highly significant at the pre-treatment time point, although marginal at on-treatment time points, suggesting that baseline chromatin states of the tumor are likely important drivers of ICB response. Our results also imply that the impact of the ICB therapy on enhancer patterns of melanoma tumors significantly varies between different patients. Indeed, pre-existing chromatin states of the tumor or T cells are likely to be deterministic towards activation of pathways increasing immunogenicity or facilitating T cell mediated killing (e.g. GAS/STING, IFN pathways, or MHC expression) or repressing those that prevent immune recognition (e.g. checkpoint receptors/ligands such as PD-L1) or resist T cell mediated killing (e.g. EMT). Our studies support further prospective investigation into the utility of these enhancer signatures in predicting response to immunotherapy, offering prognostic information, and informing combinatorial clinical trials facilitated by cutting-edge epigenomic tools.

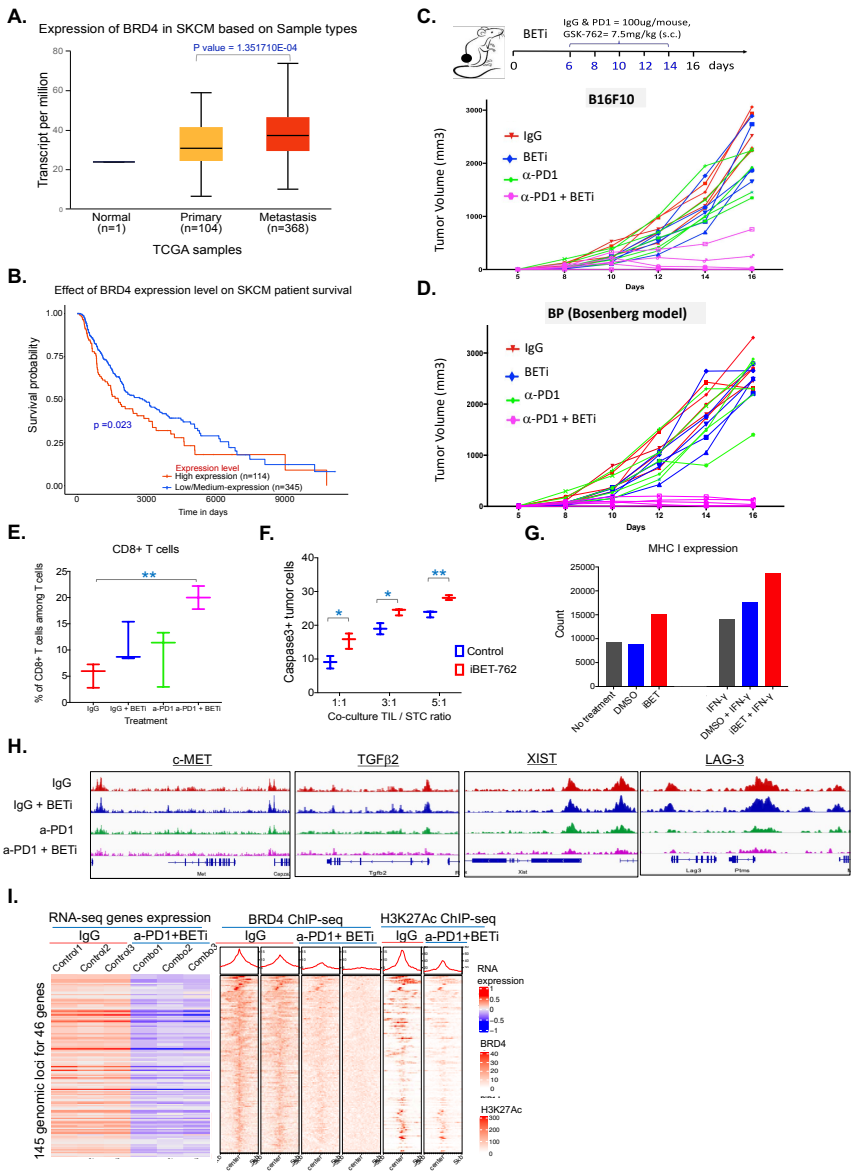


Figure 4-9: **Targeting enhancers using bromodomain inhibitors in combination with anti-PD-1 antibody confers synergistic growth reduction**. (a) Box plot showing the gene expression level of BRD4 in TCGA. (b) Difference in overall survival between BRD4 high expression and BRD low expression groups in TCGA melanoma. (c) Schematic for mouse treatments. (d) Tumor growth curves for BP cells upon treatment with four different strategies as shown in panel a (e) Graph showing infiltrated CD8+ T-cell percentages in B16F10 mice treated with IgG or anti-PD-1 alone or in combination with iBET-762. (f) Percentage of cleaved caspase-3 in target tumor cells. (g) Control, DMSO and iBET treated alone or along with IFN-g treated tumor cell 2765 were analyzed by flow cytometry for the expression of MHC class I molecules. (h) IGV snapshot of aggregate BRD4 profiles around genes in c-MET and TGFβ. (i) Heatmaps for differentially expressed genes between IgG alone and iBET-762 with anti-PD-1 combo treatment mouse B16F10 melanoma tumors. (j) Pathway analysis (Hallmark) of the genes from the panel f.

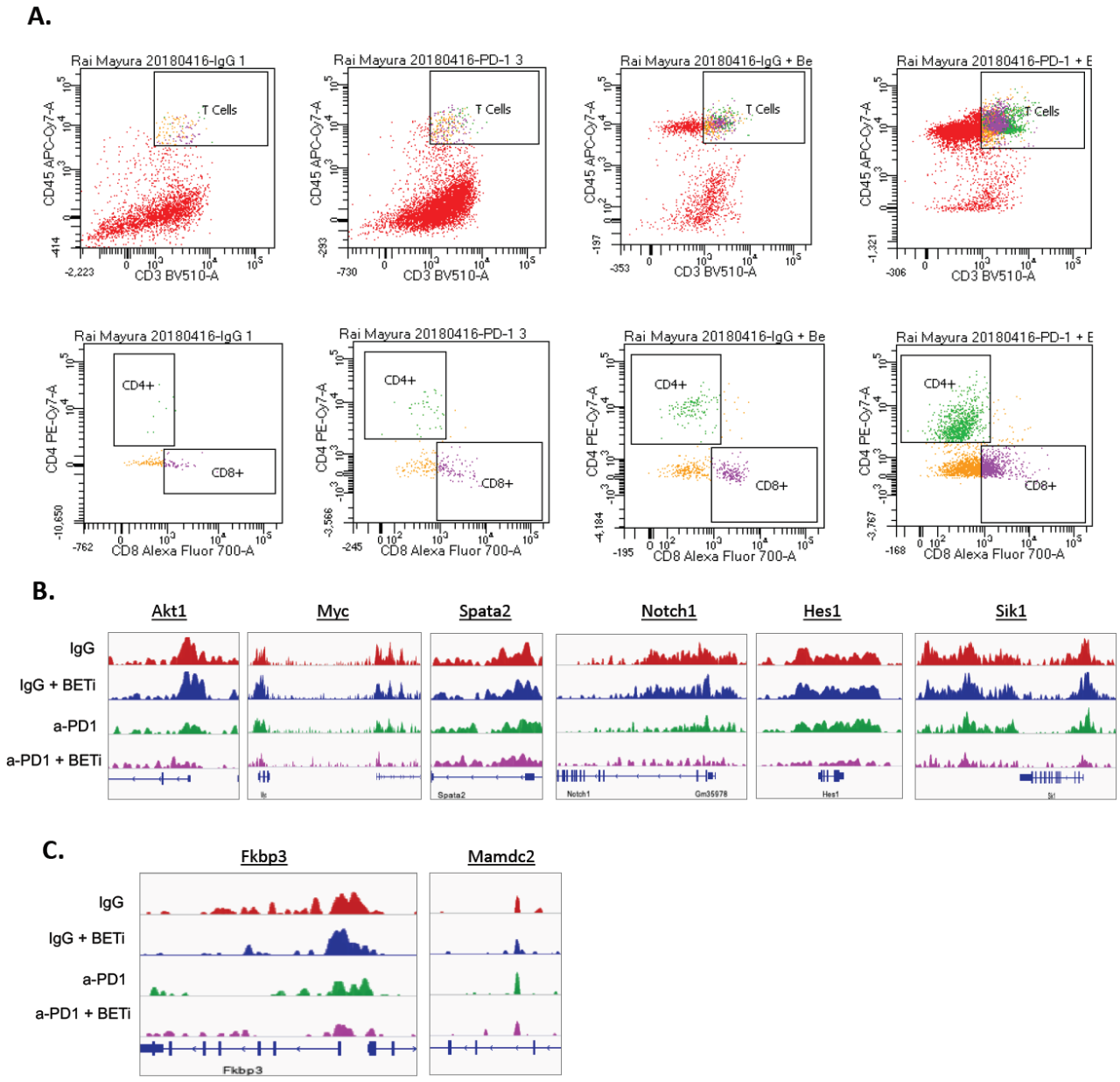


Figure 4-10: **Combination of bromodomain inhibitors and anti-PD1 therapy significantly alters the immune and enhancer landscape in treated murine tumors**. (a) FACS analysis of TILs from four different treatment groups (from Figure 4-9c. The boxes indicate CD31+CD45+CD4+/CD8+ T cells. (b-c) Genome browser view of H3K27Ac ChIP-Seq signal track surrounding the FKBP3 and MAMDC2 (c); AKT1, MYC, SPATA2, NOTCH1, HES1 and SIK1 (b) in four different treatment groups shown in Fig 4-9c. (d) Pathway network analysis (from Fig 4-9i) showing the genes downregulated in iBET-762 with anti-PD-1 combo treatment tumors compare to IgG alone tumors.

Chapter 5

Epigenetic predictors of immunotherapy response

5.1 Introduction

Given the lower overall response to checkpoint blockade immunotherapies, it is critical for the field to develop prognostic predictors of immunotherapy response [105, 56]. Prior developmenets in the space have primarily revolved around tumor mutational burden (TMB), especially for urothelial carcinoma, small cell lung cancers, NSCLC, and melanoma [12]. A previous meta-analysis of 27 cancer types demonstrated that response rate was positively correlated with log TMB [6]. There additional studies examining epigenetic biomarkers in conjunction with TMB. Cai et al. demonstrated that high TMB NSCLCs had more DNA methylation aberrations and copy number variations, showing predictive efficacy [8]. This opens the avenue for exploration of epigenetic liquid biopsies in the future.

In this chapter, we will further the development of solid-tumor-based epigenetic biomarkers. We will look at the direct prediction of immunotherapy response from H3K27ac ChIP-seq signal io a cohort of responders and non-responders. We hope to demonstrate efficacy of active enhancer marks as a power predictive tool in predicting immunotherapy resistance.

5.2 Methods

5.2.1 M-value processing and IDR calculations

In order to derive the M-values, we first used .bam files from both the ChIP and WCE files, along with a common peak file of 244,472 peaks, as inputs to MANorm using default arguments. The common peak file was generated using MACS2 'bdgdiff' function between combined pileups of responder and non-responder samples across the MDACC and MGH cohorts. The resulting normalized outputs from MAnorm [106] were first used to filter samples by imposing a $M>0$ and $p<0.05$ filter. All samples must have 20% of peaks bypassing the threshold or else it was discarded from the analysis. 30 samples from the MGH cohort passed this filter, while 27 MDACC samples passed this filter. Next, we subjected the samples to the IDR algorithm. In this case, average M-values for all peaks were calculated for both cohorts, and the two average M-value vectors were utilized as inputs to the IDR algorithm with default arguments. The resulting 77356 peaks was considered the final replicated peak set used for all downstream analyses.

5.2.2 Differential H3K27ac ChIP activity calling

By leveraging the M-values, we can easily compare responder vs. non-responder differential response. We first batch normalized the two cohorts' M-values using the ComBat algorithm from the R package 'sva' [107]. Next, we used limma's [40] empirical Bayes modeling framework to construct a linear model regressing response and treatment time against M-values. We modeled patient identity - for patients with more than one sample analyzed - as a random effect.

5.2.3 Global test for groups of peaks

To run the global test for genes, we first associated each of the peaks with a gene in the common peak set with a gene via HOMER10 annotatePeak function. Each gene's associated peaks were organized as a group for the global test. The global test

was conducted using the function 'gt' with default parameters using the 'globaltest' R package [108].

5.3 Results

5.3.1 Epigenetic predictors of anti-PD1 resistance

In order to derive a concrete set of epigenomic features that had multiple, independent lines of evidence as epigenomic correlates of ICB resistance, we collected an independent cohort of H3K27ac ChIP-seq samples from the MGH melanoma biobank (Figure 4-1a). In order to make our previously presented dataset from MD Anderson cancer center (MDACC) jointly analyzable with the MGH data, we defined a common metric that can be used across both cohorts by using MAnorm16 to calculate a log2 ratio of read densities (M-value) between ChIP and whole-cell extract (WCE) control that is adjusted for the average log2 read density at all peaks (Figure 5-1a-b). This allows any two peak regions to be compared on the same scale across the two distinct cohorts by accounting for variable total read depth at peak regions of interest. Using IDR analysis (see section 5.2.1), we identified a subset of 86,226 of 244,472 peaks as reproducible peaks betwene the MDACC and MGH cohorts (Figure 5-2a) which enriched in different functional classes including promoter, intron, and TSS (Figure 5-1c).

Next, we subjected these 86k peaks' M-values to differential peak calling via limma [23] independently in each cohort. We identified 3008 MGH and 7984 MDACC pre-treatment peaks whose activity was significantly different between responders and non-responders at nominal $p<0.05$. To identify a replicated peak-set, we took the intersection of MDACC and MGH significant peaks to determine whether the doubly significant set exhibited statistically significant enrichment above null expectation. Only the pre-treatment comparisons exhibited a significant enrichment in the number of replicated peaks relative to the null expectation ($p<4.193e-05$, One-sided Exact Binomial Test) and 189 peaks were doubly significant in both the MDACC and MGH

pre-treatment comparisons. We note an excess enrichment in the signal from the MDACC cohorts in both the pre-treatment (Figure 5-1d) and on-treatment (Figure 5-1e).

Next, to identify a subset of enhancers with predictive ability toward patient response, we concentrated on the pre-treatment overlapping significant peak-set. We utilized the the replicated 189 peaks as a feature set in a cross-validation setting and trained two random forest models as follows: one in which the MDACC cohort was designated as the training set and the MGH cohort the testing set, and vice versa. The results were combined into a single receiver operator characteristic (ROC) for evaluation. We also evaluated the area under the ROC (auROC) as a measure of model performance (Figure 5-2b). Using the 189 peaks, we were able to achieve a performant AUC of 0.91, suggesting that epigenomic features are just as performant as those from transcriptome (AUC range 0.8-0.9)[61].

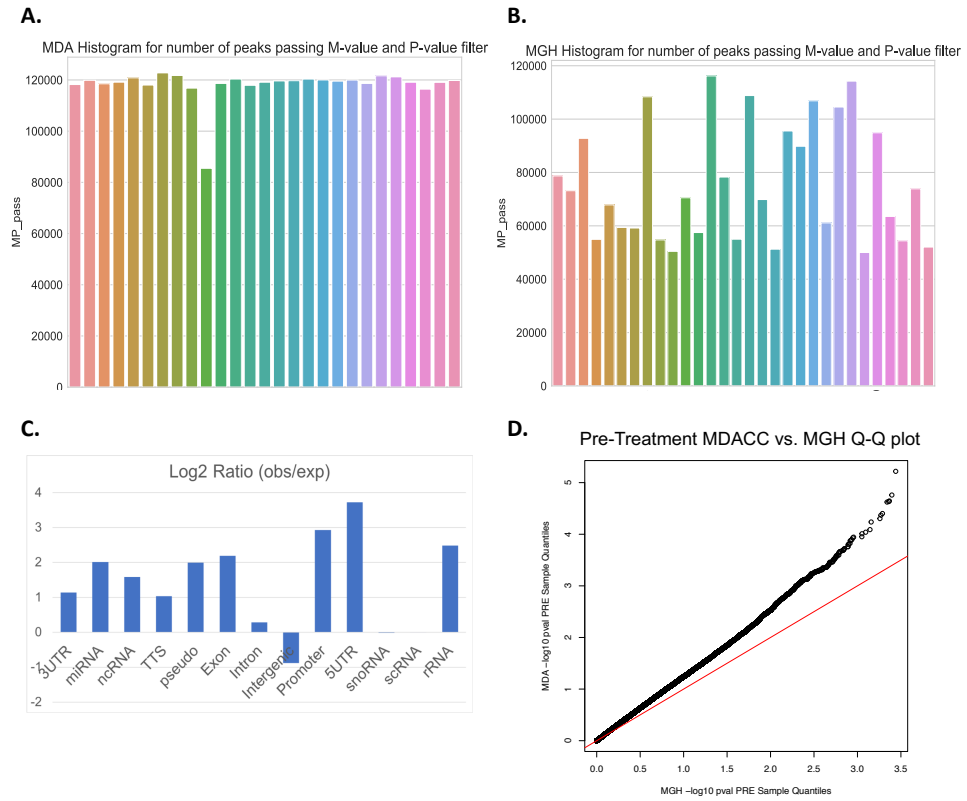
We next assayed to what extent these 189 peaks stratified overall survival (OS) and progression-free survival (PFS) in our clinical cohort. To do so, we performed Cox proportional hazards regression with M-values as the design matrix. Our analysis showed that 4 out of the 189 peaks significantly stratified survival at $p<0.05$ via Cox regression in both the MGH and MDACC cohorts. To visualize the survival differences, we first clustered the peaks by their sample M-values. We noted the presence of two distinct clusters (Figure 5-2c, 5-3a-b) corresponding to two opposite directionalities by which increased epigenomic activity at these peaks modulate survival time (Figure 5-2d-e, 5-3c-d). Our results show that a distinct set of epigenomic peaks which are significantly associated with treatment response and survival stratification in two independent cohorts, making these ideal targets for follow-up studies.

Since a gene could be activated by multiple enhancers, we sought to converge the enhancer dysregulation into the level of genes which might impact response to ICB. To do so, we employed the global test, a testing procedure designed to test whether a group of peaks is significantly associated with the clinical outcome as a unit [108], to systematically test whether groups of peaks associated with individual genes (see section 5.2.3) are different between responders and non-responders across both the

MGH and MDACC cohorts. We designated all peaks associated with a particular gene after annotation as a group and conducted a total of 19,212 global tests for genes with peak activity levels as the independent variable and binary clinical response as the response variable. Of these tests, 922 genes had a global test p-value of $p<0.05$. The top-ranked hits from this analysis are displayed in Figure 5-2f, includes MIR-4492, whose downstream targets includes IL-10 and TNF-α and lncRNA TLR8-AS1, which modulates TLR8 expression levels via stabilizing TLR8 mRNA21, as well as FOXP3, a master regulator of Treg differentiation. GO enrichment analysis of the 922 global test significant genes revealed enrichment in a diverse set of immune and metabolic related pathways, including NK T cell activation, neutrophil activation, and macrophage proliferation (Figure 5-2g).

5.4 Discussion

We identify for the first time an enhancer-based signature that could be potentially used as an epigenomic biomarker for non-response to ICB therapy in melanoma. Of these, 147 enhancers further predicted OS and PFS suggesting the potential use of this epigenomic signature as a prognostic indicator. These signatures have the potential to be utilized alone, or in combination with other genomic, transcriptomic, or immune features to generate a multi-omic signature to predict response or survival in patients on ICB. Indeed, other features such as tumor mutation burden [60] (TMB), specific genetic features (PTEN deletion [110], IFNG-R deletions [111], PBRM1 mutations, KMT2D mutations [77]) have been shown to be associated with response to ICB.



auROC	189 peak set AUC
Hugo Stratified CV	0.62
Hugo LOO CV	0.57
Riaz Stratified CV	0.64
Riaz LOO CV	0.69
Liu Stratified CV	0.60
Liu LOO CV	0.63
Van Allen Stratified CV	0.65
Van Allen LOO CV	0.64

Figure 5-1: **Validation of an enhancer signature of non-response in an independent cohort**. (a) Number of peaks from MDACC cohort sample peaks passing quality threshold M-value>0 and MANorm p>0.1 (b) Number of peaks from MGH cohort sample peaks passing quality threshold M-value>0 and MANorm p>0.1 (c) Functional enrichments for the 86,226 peaks passing the IDR threshold (d) QQ-plot between MGH (x-axis) and MDACC (y-axis) sample quantiles from the pre-treatment comparison. (e) QQ-plot between MGH (x-axis) and MDACC (y-axis) sample quantiles from the on-treatment comparison. (f) Receiver operating characteristic (ROC) of random forest trained predictive models utilizing the 189 replicated pre-treatment peaks ($p < 0.05$) or 410 replicated pre-treatment peaks ($p < 0.1$). The ROC values were calculated for the published anti-PD-1 treated melanoma patient datasets[109, 28, 105, 56].

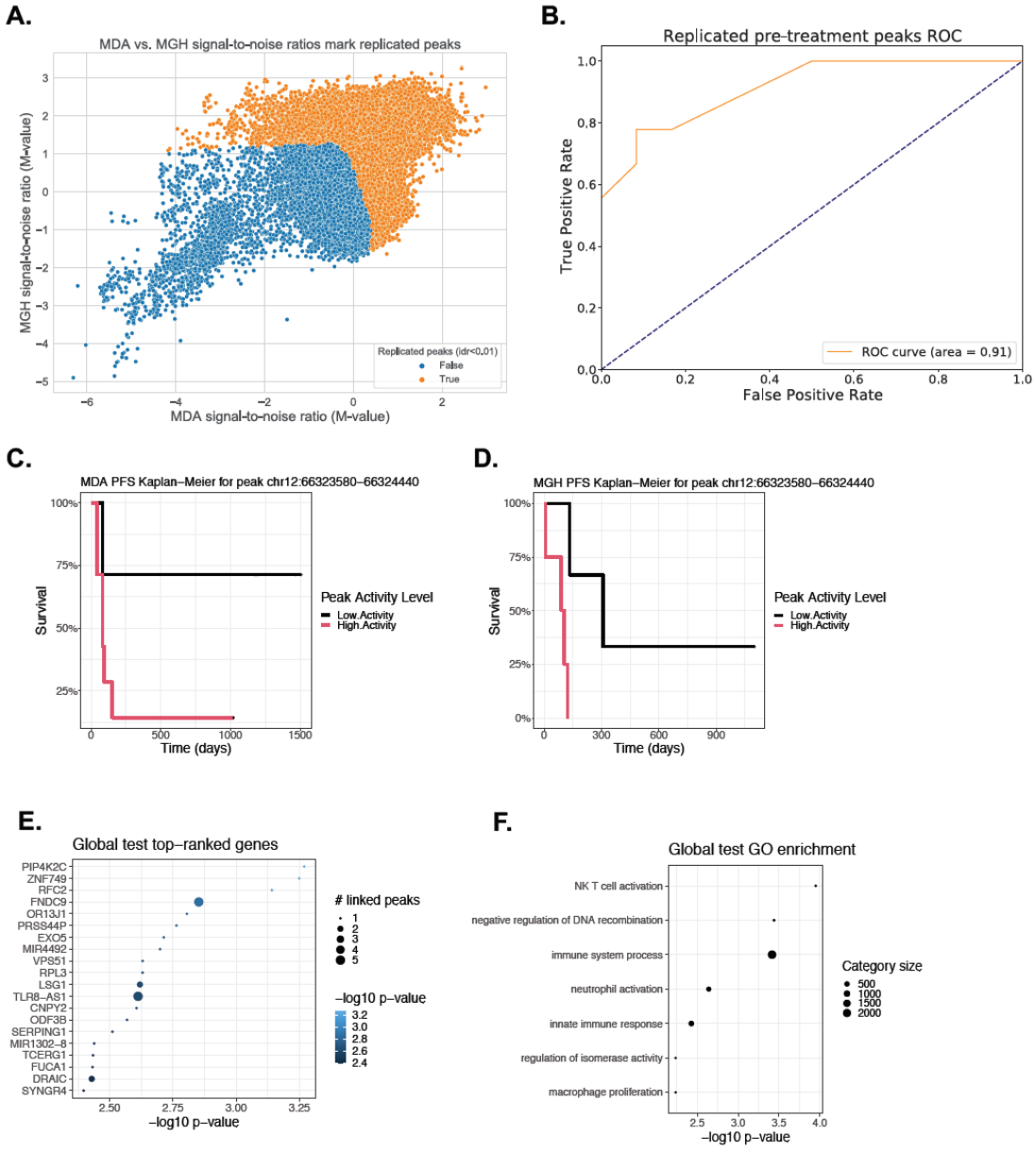


Figure 5-2: **Validation of enhancer signature prediction of response in an independent cohort**. (a) MGH vs MDACC cohort average M-value with IDR status <0.01. Individual points represent averaged M-value across the MDACC cohort (X-axis) and across the MGH cohort (Y-axis). (b) Receiver operating characteristic (ROC) of random forest trained predictive models utilizing the 189 replicated pre-treatment peaks. (c) Progression-free survival Kaplan-Meier plots in MDACC for peak Chr12:65,929,800-65,930,660. (d) Progression-free survival Kaplan-Meier plots in MGH for peak Chr12:65,929,800-65,930,660. (e) Top genes from Global Test for peaks. (f) Selected GO enrichment from the 922 gene significant from the Global Test.

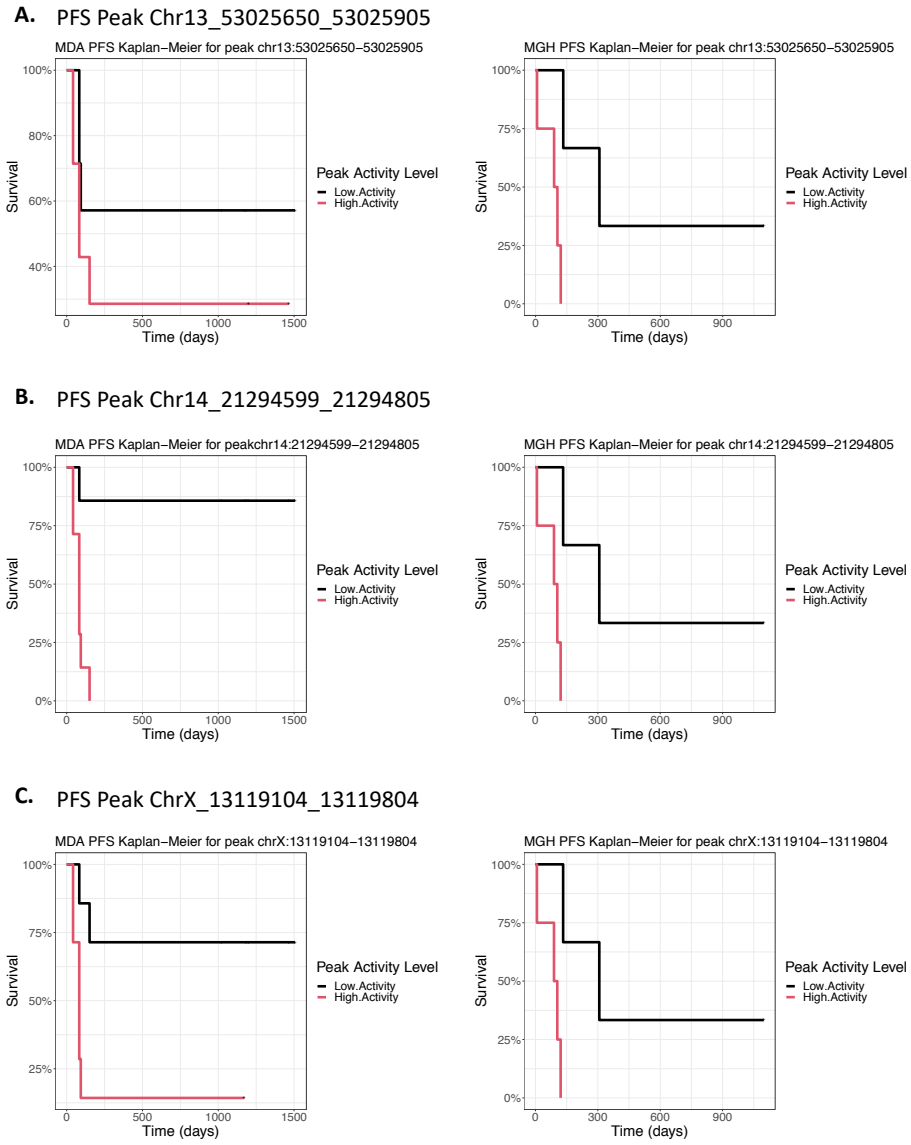


Figure 5-3: **Predictive power of enhancer peaks for progression free survival**. (a) PFS Kaplan-Meier plot of MDACC (left) and MGH (right) for peak Chr13:53,025,650-53,025,905 (b) PFS Kaplan-Meier plot of MDACC (left) and MGH (right)) for peak Chr14:21,294,599-21,294,805. (c) PFS Kaplan-Meier plot of MDACC (left) and MGH (right) for peak ChrX:13,119,104-13,119,804.

Chapter 6

TCR repertoire of thymic conventional and regulatory T-cells

6.1 Introduction

The ability to generate a rapid and sustained T cell response to external pathogens and transformed malignant cells is essential for the protection of the host. At the same time, deletion of autoreactive T cells during development and repression of excessive or autoreactive responses in peripheral tissues is essential to proper T cell protective function ([112]). This duality in the regulation of T cell function is accomplished by two types of T cells: conventional T cells that provide "helper" (CD4+) and "killer" (CD8+) functions and regulatory T cells (Treg) that suppress conventional T cell–dependent responses. Treg have been assigned to two subsets based on the origin of their generation: thymic Treg (tTreg) (or natural Treg) that develop in the thymus [113, 114, 115, 116] and peripheral Treg generated in the periphery from thymic conventional CD4+ T cells (Tconv) under specific conditions [117, 118, 119]. The development of tTreg appears to require both TCR signals and other factors, such as costimulatory signaling and cytokines, but the precise mechanisms of tTreg generation have not been fully elucidated.

A key factor in tTreg generation is the specificity of the TCR whose interaction with self-antigen/MHC plays a critical role in tTreg differentiation. Several studies

using transgenic (Tg) mouse models suggest that the signal strength of TCR recognition of self-antigen/MHC ligand differs in tTreg and Tconv, with tTreg differentiation involving a higher level of signal strength [120, 121]. Indeed, the disruption of normal self-antigen/MHC ligand expression in the thymus because of Aire deficiency causes a change in the fate of self-antigen–specific T cells from tTreg to Tconv [122, 123]. Subsequent studies suggest that Ag presentation by different APCs (classical and plasmacytoid dendritic cells, cortical and medullary thymic epithelial cells, and B cells) at different thymic locations (cortex and medulla) influences the deletion of autoreactive thymocytes and the differentiation of tTreg [114, 124]. In addition, factors such as cytokines (including IL-2 and TGF-b) [125, 126, 127] and costimulatory receptors (CD28) [128] have also been implicated in the development of tTreg. Collectively, it is clear that no single factor alone determines differentiation to the tTreg fate, but precisely how these factors act in combination remains to be determined.

Initial analyses of TCR sequences in Treg and Tconv of TCRa or TCRb Tg mice reported that CDR3a and CDR3b sequence repertoires of Treg and Tconv are different either by exclusive appearance in only one of these lineages or by their relative abundance in Treg or Tconv when sequences were found in both cell types [129, 130]. Subsequent studies using high-throughput sequencing generated larger numbers of TCR sequences in Treg and Tconv. Studies using TCRb Tg mice to compare TCRa sequences between tTreg and Tconv reported little overlap of TCRa sequence between tTreg and effector T cells [131] or between tTreg and Tconv that recognize the same foreign Ag [132]. Study of a TCRa Tg mouse to compare TCRb sequences between Treg and Tconv from spleen and peripheral lymph nodes found that 12% of TCRb sequences are shared by peripheral Treg and Tconv and are thus presumed to be derived from common progenitors [133]. However, there has been no reported deep sequencing analysis examining endogenous TCRa and TCRb of tTreg and Tconv from the thymus of non-TCR Tg mice. It is therefore unclear what degree of TCR sequence uniqueness and similarity exists overall between tTreg and Tconv or, importantly, whether there are general sequence features that distinguish ab TCR of tTreg from those of Tconv.

In this study, we addressed the role of TCR sequence in determining whether T cells develop into tTreg or Tconv lineages. We report a comprehensive comparison of TCRa and TCRb sequences of tTreg and Tconv using a Unique Molecular Identifier (UMI) methodology incorporating a 59 single universal primer for PCR amplification of all V genes, significantly reducing PCR bias of amplification and sequencing errors affecting the quantitation of TCR frequency [134, 135]. Comparison of TCRa and TCRb sequences between tTreg and Tconv from two normal mouse strains revealed that, although many sequences were unique to either Treg or Tconv, a substantial proportion of TCRa (21–30%) and TCRb (5–20%) sequences from tTreg were also found in Tconv. Analysis of a TCRb Tg mouse line revealed an even higher proportion (71%) of TCRa sequences found in tTreg that were also found in Tconv. Interestingly, these shared TCRa clonotypes that were common to tTreg and Tconv were significantly more abundant than nonshared TCRa sequences of tTreg and Tconv. Finally, we used machine learning (ML) to develop an algorithm that was capable of distinguishing nonshared TCRa and TCRb sequences expressed by tTreg from those of Tconv and, in addition, found that specific amino acid trimers were differentially expressed in either tTreg or Tconv. When we applied the same ML algorithm to an analysis of those TCR sequences that were shared by tTreg and Tconv, the vast majority of these sequences were classified as characteristic of Tconv and not tTreg. Taken together, our findings identify TCR sequence characteristics that bias to tTreg or Tconv fate, in addition to the presence of factors that can drive cells with an identical TCR sequence into either Tconv or tTreg lineages.

6.2 Methods

6.2.1 Isolation of tTreg and Tconv from thymus

tTreg and Tconv were isolated from 4- to 8-wk-old mice of three strains, all on a C57BL/6 background: 1) Rag-GFP-Foxp3-RFP [136, 137]: tTreg and Tconv were isolated from three individual mice based on GFP and Foxp3-RFP expression (Sup-

plemental Fig. 1) and were sequenced independently; 2) TcrdCreERZsGreen-Foxp3-RFP: TcrdCreER knock-in with tamoxifeninduced ZsGreen reporter for TCRd expression (three doses of 1 mg tamoxifen i.p. every other day, cells isolated 2 wk after last injection) (27): tTreg and Tconv cells were isolated from nine mice based on expressions of ZsGreen, Foxp3-RFP, and CD25 (Fig. 6-1), and sequencing was performed on three pools (three mice pooled in one sample); and 3) Foxp3-GFP TCRa+/2 TCRb-Tg mice carrying the DO11.10 TCRb [138, 139]: tTreg and Tconv samples from three individual mice were sorted based on Foxp3 reporter + and 2, respectively, and individual samples were sequenced independently. Foxp3-GFP mice were provided by Vijay Kuchroo [140]. All mice were maintained under specific pathogen free conditions at the animal facility of National Cancer Institute and Duke University. Animal procedures were reviewed and approved by National Institutes of Health or Duke Institutional Animal Care and Use Committee.

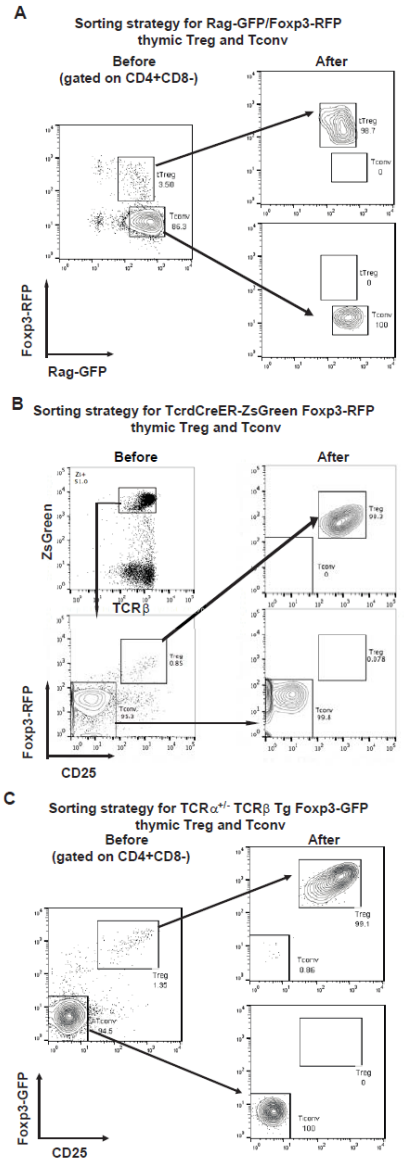


Figure 6-1: **Gating strategy for isolation of tTreg and Tconv cells from thymus of three strains of mice**. (a) Gating strategy for isolating tTreg and Tconv cells from Rag-GFP/Foxp3-RFP mice. Gating was first on single positive CD4+CD8- followed by the gating shown (b) Gating strategy for isolating tTreg and Tconv cells from TcrdCreERZsGreen-Foxp3-RFP mice. Cells were gated on ZsGreen and TCRb positive first and then gated on Foxp3-RFP and CD25. (c) Gating strategy for isolating tTreg and Tconv cells from TCRa +/- TCRb Tg-Foxp3-GFP mice.

6.2.2 Library construction and sequencing strategy

cDNA library construction was previously described (24). Briefly, total RNA was isolated from tTreg and Tconv using a QIAGEN RNeasy Micro Kit. Isolated total RNA

(50–500 ng) was used for cDNA synthesis using TCRa and TCRb C region–specific primers mTRAC1 (59-GGCGTTGGTCTCTTTGAAG-39) and mTRBC1 (5'-CACTTGTCCTCCTCT 3') (all oligos made by Eurofins USA), SMARTScribe Reverse Transcriptase (Takara Bio), and SmartN oligos (5'-AAGCAGUGGTAUCAACGCAGAGUNNNNUNNNNUNNNNUCTTrGrG 3') for template switching at the 5' end to incorporate a UMI and M1SS sequence for PCR. The cDNA products were treated with uracil-DNA glycosylase (New England BioLabs) at 37°C for 30 min to remove SmartN oligos.

Three rounds of PCR using Super Fidelity Platinum Taq DNA Polymerase (Thermo Fisher Scientific) were performed to prepare libraries for sequencing. The first PCR (18–24 cycles) was used to enrich TCRs using M1SS (5'-AAGCAGTGGTATCAACGCA-3', part of SmartN used as a 5' PCR primer) and TCR C region primers (mTRAC2: 5'-CGGCACATTGATTTGGGAG-3' and mTRBC2: 5'-TGTGGACCTCCTTGCCATTC-3'), and primers were removed by the QIAquick PCR Purification Kit (QIAGEN). The second PCR (20–32 cycles) was used to add an 8-bp sample barcode to each sample at 59 end (P7M1S-n: 5'-CGTGTGCTCTTCCGATC(N)1–2-NNNNNNNN(8 bp barcode)-CAGTGGTATCAACGCAGAG-3') and internal C region primers (mTRAC3: 5'-AGGTTCTGGGTTCTGGATG-3' and mTRBC3: 5'-GGTGGAGTCACATTTCTCAG-3'). PCR products were separated by 2% agarose (UltraPure; Thermo Fisher Scientific) gel electrophoresis, and DNA fragments (400–800 bp) were further purified by a QIAquick Gel Extraction Kit (QIAGEN). Purified DNAs of each sample were quantitated by a BioAnalyzer (Agilent Technologies) and combined for the third round of PCR (10 cycles), which incorporates the Illumina adaptor (P7: 5'-CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3', mP5TA: 5'-AATGATACGGCGACCACCGATCGTCGAGGTTCTGGGTTCTGGATG-3' and mP5TB: 5'-AATGATACGGCGACCACCGATCGTCGGGTGGAGTCACATTTCTC-3'). Amplified DNA was separated by 2% agarose gel electrophoresis and further purified by QIAquick Gel Extraction and followed by the PCR purification kits. The amount of purified DNA was measured using a Qubit Fluorometer (Thermo Fisher Scientific). Fifty picomoles of DNA were used for sequencing on an Illumina HiSeq 2500 system. A modified paired end sequencing protocol was used: TCR-specific

sequencing primers TRA (5'-TCGTCGAGGTTCTGGGTTCTGGATG-3') and TRB (5'-TCGTCGGGTGGAGTCACATTTCTCAG- 3') were used for first round sequencing of 150 bps. Illumina RD2 primer (5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3') was used for second round sequencing of 50 bp, covering the sample barcode and UMI.

6.2.3 Analysis of overlapping TCRα and TCRβ sequences

Overlap of TCRa and TCRb sequences between different mice/samples of the same strain of mice or between two strains of mice was analyzed at two levels: 1) overlapping distinct TCR sequences (Eq. 6.1) 2) and overlapping total sequences based on UMI counts (Eq. 6.2). The overlapping sequences are presented as percentages for each pair of comparison. Eq. 6.1 was used for the calculation of the overlapping percentage of distinct TCR sequences, as follows:

$$\text{Overlap}(\%) = \frac{\text{TCRs found in both S1 and S2}}{\text{All distinct TCR in S1} + \text{All distinct TCRs in S2 - TCRs found in both}} \tag{6.1}$$

Eq. 6.2 was used for the calculation of the overlapping percentage of total TCR sequences, as follows

$$\text{Overlap}(\%) = \frac{\text{Sum of UMI counts of TCRs found in both S1 and S2}}{\text{UMI counts of all TCRs in S1} + \text{UMI counts of all TCRs in S2}} \tag{6.2}$$

The sum of UMI counts of TCRs found in both S1 and S2 is calculated as UMI counts of TCRs found in both S1 and S2 in S1 plus UMI counts of TCRs found in both S1 and S2 in S2.

6.2.4 Identification of distinct TCR sequences between tTreg and Tconv by ML algorithm

To build an ML classifier to analyze ab TCR sequences from tTreg and Tconv, we converted the TCR CDR3 sequences into a matrix of length 3 aa (3-mers or trimers)

from all three strains of mice used in these studies. We first enumerated all possible 3-mers from all CDR3s (3-mer library) and then embedded each CDR3 into a length L vector that supposed there are L possible 3-mers observed in our data, in which each entry is the number of times that 3-mer appears in a CDR3; all other 3-mers in the library but not found in the analyzed TCR were labeled as 0. To calculate the relative starting location of each 3-mer, we recorded where the 3-mer appeared in CDR3 relative to the 3-mer starting locations that are possible for each CDR3 length (e.g., a length 12 CDR3 has 10 possible 3-mer locations). To determine multiple copies of the same 3-mer in a TCR, we used a Python dictionary to keep track of the number of occurrences of a 3-mer in a particular TCR. These are then combined into an N by L matrix, where N is the number of CDR3s. Furthermore, we vectorized the V/J information and concatenated this with the $N \times L$ 3-mer matrix. Suppose there are M possible V genes and K possible J genes; each CDR3's V gene information is embedded into a length M vector, where the V gene entry that corresponds to the V gene in the TCR is labeled as 1 and the rest as 0 to produce an $N \times M$ matrix. The same procedure was done for the length K J gene vector to produce an $N \times K$ matrix. To generate the final matrix for ML, we concatenated the three matrices to produce a matrix of size $N \times (L+M+K)$ matrix. For two-class classification, we train a random forest binary classifier, which takes in a L vector and predicts the compartment to which a CDR3 belongs. We used 70% of the distinct TCR sequences as a training set and 30% of the sequences as a testing set. Training was performed using the random forest classifier from scikit-learn using 150 trees and default settings for other parameters [141]. Model performance was evaluated by calculating the area under the receiver operating characteristic (ROC) using the scikit-learn metrics package. The ROC was plotted using Python's matplotlib library.

6.2.5 Analysis of enriched amino acid trimers in tTreg and Tconv

To compare trimer enrichment in the non-V/J portions of the CDR3, we first removed 3 aa from either ends of CDR3 and used the central CDR3 sequences for trimer analysis. Next, to compare the relative abundance of a particular amino acid trimer in tTreg and Tconv, we created a 2×2 contingency table for each k-mer and then computed the p value using a χ^2 test using the Python library SciPy (34). We corrected for multiple comparisons using the Benjamini-Hochberg procedure. The significantly enriched trimers in tTreg were defined as those that met the criteria tTreg/Tconv ratio ≥ 1.5 and false discovery rate (FDR) ≤ 0.05. The percentages of amino acid usage in the enriched trimers were calculated by the sum of each amino acid in the enriched trimers multiplied by their respective UMI counts divided by the total number of amino acids based on UMI counts in these enriched trimers. The percentages of amino acid usage in all trimers were calculated by sum of each amino acid in all the trimers multiplied by their respective UMI counts divided by the total number of amino acids based on UMI counts in all trimers unique to either tTreg or Tconv.

6.2.6 Statistical analysis

The Mann–Whitney U test was used to calculate the significant difference of UMI/TCR ratios between shared and nonshared TCRs with tTreg and Tconv. A p value <0.05 was considered significant.

6.3 Results

6.3.1 TCRα and TCRβ repertoires of tTreg and Tconv are comparably diverse

To analyze ab TCR repertoires of tTreg and Tconv, we isolated similar numbers of recently generated tTreg and Tconv from the thymus of Rag-GFP/Foxp3-RFP (25, 26) and TcrdCreERZsGreen-Foxp3-RFP mice [142] by cell sorting and applied a high-throughput sequencing method with UMI labeling of TCR mRNA. The Rag-GFP/Foxp3-RFP strain marked newly generated tTreg and Tconv in green fluorescent dye–GFP [136, 137], whereas the TcrdCreERZsGreen-Foxp3-RFP strain labeled newly produced tTreg and Tconv in fluorescent dye–ZsGreen after tamoxifen induction [142]. These fluorescent markers were used in flow cytometric isolation of tTreg and Tconv to ensure the thymic origin of tTreg and Tconv by excluding the potential contamination of recirculating T cells (Fig. 6-1a-b). We analyzed a total of 3.7×10^5 tTreg and 3.0×10^5 Tconv from three Rag-GFP/Foxp3-RFP mice and found that the estimated size of TCRa and TCRb repertoires, identified as the number of distinct sequences, was comparably diverse between tTreg and Tconv when similar cell numbers were analyzed (Table 6.1). We also isolated tTreg (6.0×10^4 cells) and Tconv (7.2×10^4 cells) from TcrdCreERZsGreen-Foxp3-RFP mice by cell sorting and determined their TCRa and TCRb repertoires. Again, we found that the sizes of estimated TCRa and TCRb repertoires were comparably diverse in similar numbers of Tconv and tTreg in TcrdCreERZsGreen-Foxp3-RFP mice (Table 6.1). However, it should be noted that the total TCR repertoire size of Tconv in a mouse is likely larger than that of tTreg when the actual number of cells in the thymus is adjusted. Consistent with this is the observation that there were higher percentages of overlap in tTreg TCRs (3.4–5.2%) than in Tconv TCRs (1.8–2.1%) between individual mice (Rag-GFP/Foxp3-RFP) or different samples (TcrdCreERZsGreen-Foxp3-RFP) (Fig. 6-2a-b). Because we analyzed TCRa and TCRb repertoires separately, it could not be determined from this analysis whether the ab combinatorial TCR repertoire was also comparable between tTreg and Tconv. V gene usage and CDR3 length distributions of TCRa and TCRb were also not substantially different between tTreg and Tconv

(Fig. 6-3a-d).

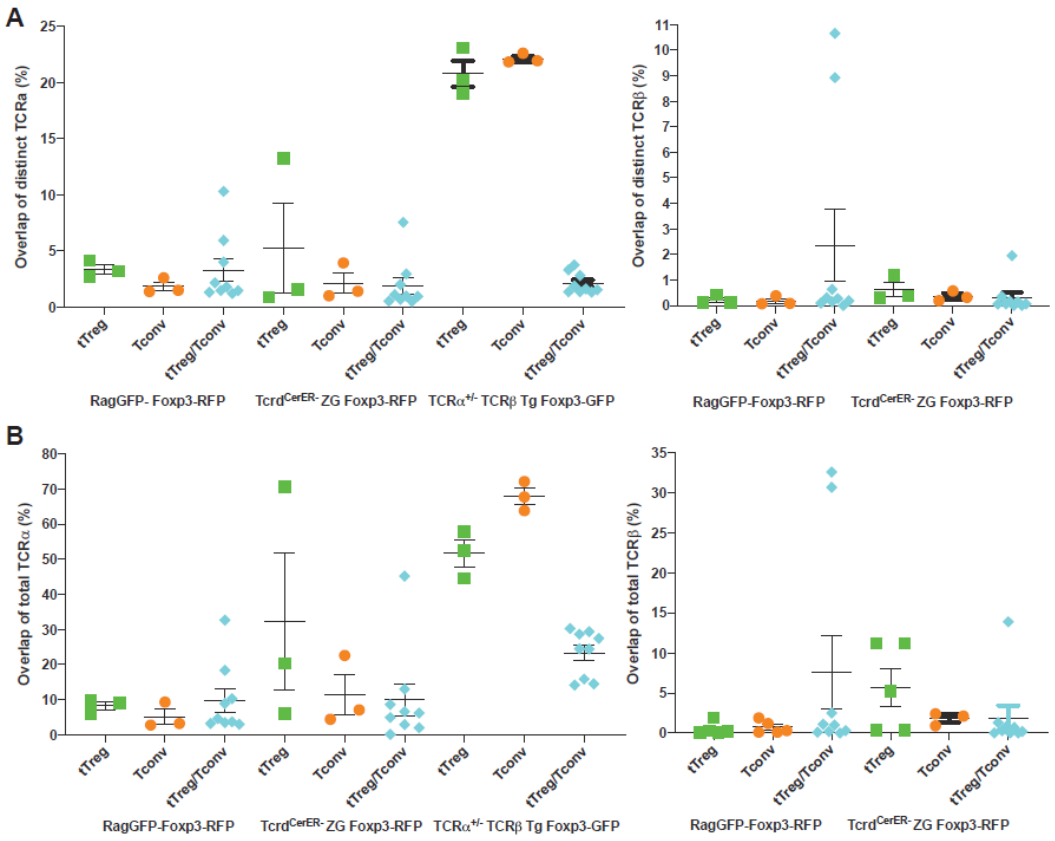


Figure 6-2: **Overlap of TCRα and TCRβ sequences of tTreg and Tconv cells** among different samples or mice within the same strain in three categories: tTreg-specific, Tconv-specific, and tTreg/Tconv shared. (a) Overlap percentages of distinct TCRα and TCRβ sequences among different samples or mice within the same strain of mice. Each dot represents the value of two mice compared. (b) Overlap percentages of total TCRα and TCRβ sequences (based on UMI counts) among different samples or mice within the same strain of mice. The UMI percentages of corresponding distinct TCR sequences are presented.

Cell type	Cell No.	Distinct TCRα	UMI	Estimated TCRα	Distinct TCRβ	UMI	Estimated TCRβ
Treg (GFP+RFP+)	370,000	4,592	439,728	13,747	2,532	164,510	8,244
Tconv (GFP+RFP-)	300,000	3,899	246,106	14,350	2,652	185,677	7,310
Treg (ZsGreen+FoxP3+CD25+)	59,990	1,716	216,004	9,919	1,837	145,782	2,817
Tconv (ZsGreen+Foxp3-CD25-)	71,760	4,335	291,632	10,219	2,693	240,464	3,547
Treg (Foxp3+CD25+)	317,700	971	79,329	3,021			
Tconv (Foxp3-CD25-)	1,500,000	14,741	215,539	19,099			

Table 6.1: **Summary of TCRα and TCRβ repertoire of tTreg and Tconv**

.

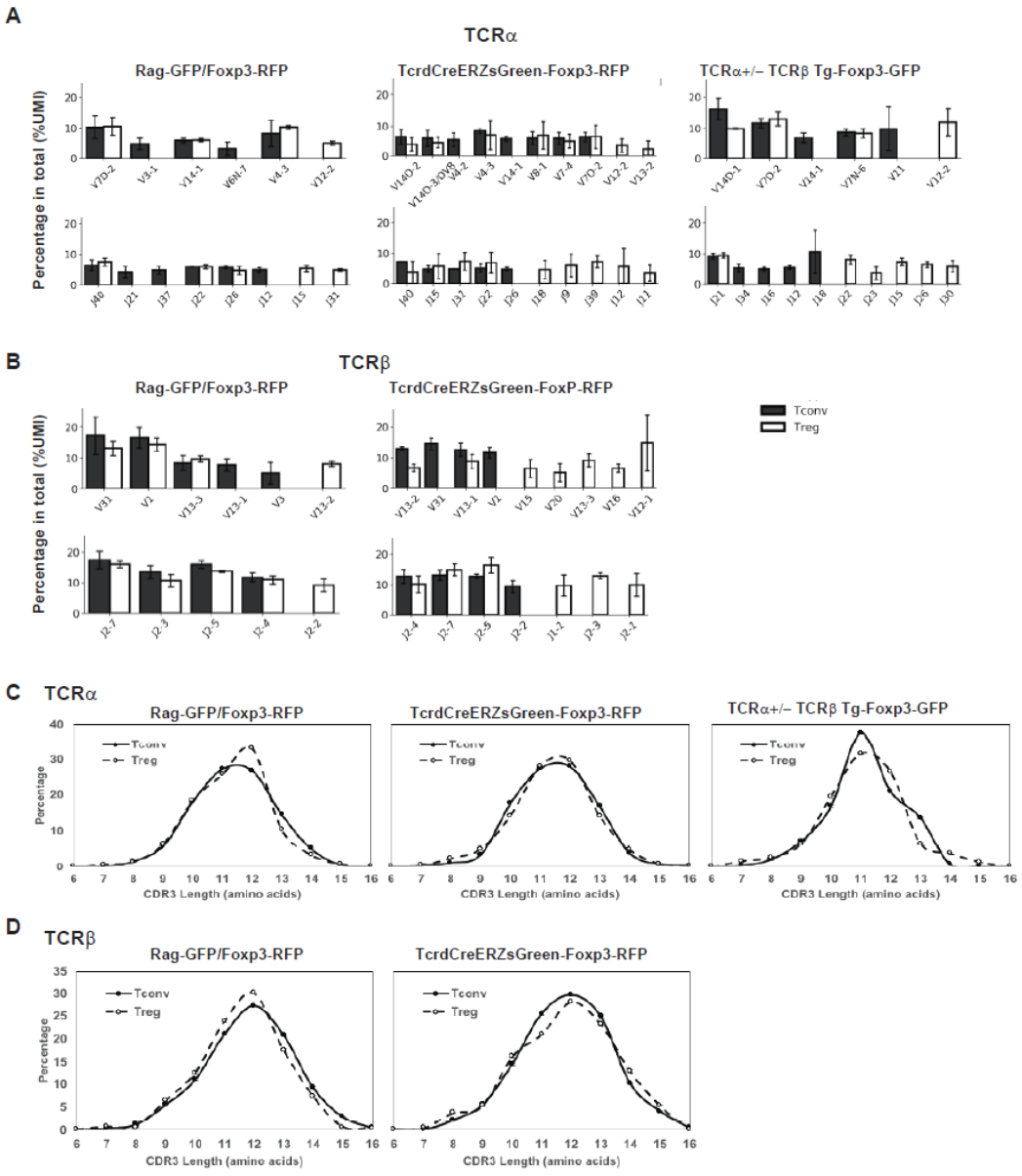


Figure 6-3: **V gene usage and CDR3 length of tTreg and Tconv cells** (a) Top 3 most used TRAV and TRAJ genes in tTreg and Tconv cells from three strains of mice. (b) Top 5 most used TRBV and TRBJ genes in tTreg and Tconv cells from three strains of mice. (c) CDR3 length of TCRα of three strains of mice. The average values and standard deviation of three samples of each strain of mice are presented. (d) CDR3 length of TCRβ of three strains of mice.

6.3.2 Abundance of TCRα and TCRβ sequences distinct to tTreg or Tconv or shared between lineages

Next, we determined the degree to which ab TCR sequences of tTreg and Tconv of normal non-TCR Tg mice are similar or distinct. To overcome the limited numbers of cells available from each individual mouse, in particular for tTreg populations, we combined TCRa and TCRb sequences of tTreg and Tconv from three samples of each strain of mice and then compared these pooled sequence sets (Fig. 6-4a). In Rag-GFP/Foxp3-RFP mice, we found that 12% of distinct tTreg TCRa sequences (580 out of total 4906 TCRa sequences from tTreg) were found in Tconv and accounted for 14% of distinct TCRa Tconv sequences. Those shared TCRa sequences were more abundant than nonshared sequences as they accounted for 21 and 25% of total tTreg and Tconv sequences (based on UMI counts), respectively (Fig. 6-4a). Compared with TCRa, the overlap in TCRb between tTreg and Tconv was slightly lower, accounting for 11 and 10% of distinct TCRb in Treg and Tconv, respectively (Fig. 6-4c). These overlapping TCRb sequences were also more abundant, accounting for 20 and 26% of total Treg and Tconv (Fig. 6-4c). TCRa and TCRb sequences found in both tTreg and Tconv were also observed in TcrdCreERZsGreen-Foxp3-RFP mice (Fig. 6-4b, 6-4d). Collectively these findings showed that 9–12% TCRa and 2–11% TCRb sequences of tTreg were found in thymic Tconv, and they accounted for 21–30% of TCRa and 5–20% of TCRb in total tTreg. With the caveat that these data derived from individual TCRa and TCRb sequences do not directly measure ab pairing, these findings indicate that TCR sequence is not the only factor determining tTreg generation in thymus.

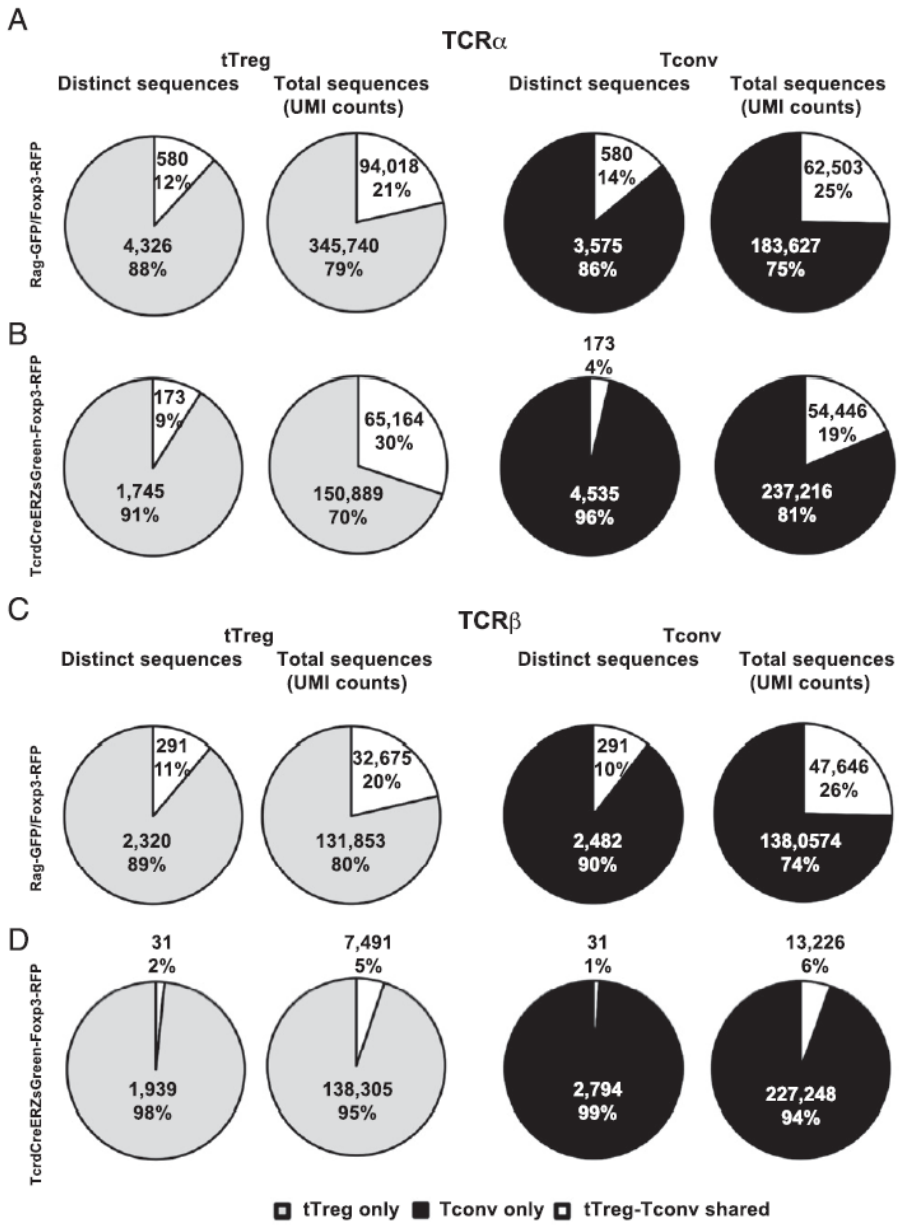


Figure 6-4: **Number and percentages of distinct and shared TCRα and TCRβ sequences between tTreg and Tconv from two lines of normal mice**. (a) Shared TCRa clonotypes in tTreg and Tconv of Rag-GFP/Foxp3-RFP mice. (b) Shared TCRa sequences in tTreg and Tconv of TcrdCreERZsGreen-Foxp3-RFP mice. (c) Shared TCRb sequences in tTreg and Tconv of Rag-GFP/Foxp3-RFP mice. (d) Shared TCRb sequences in tTreg and Tconv of TcrdCreERZsGreen-Foxp3-RFP mice.

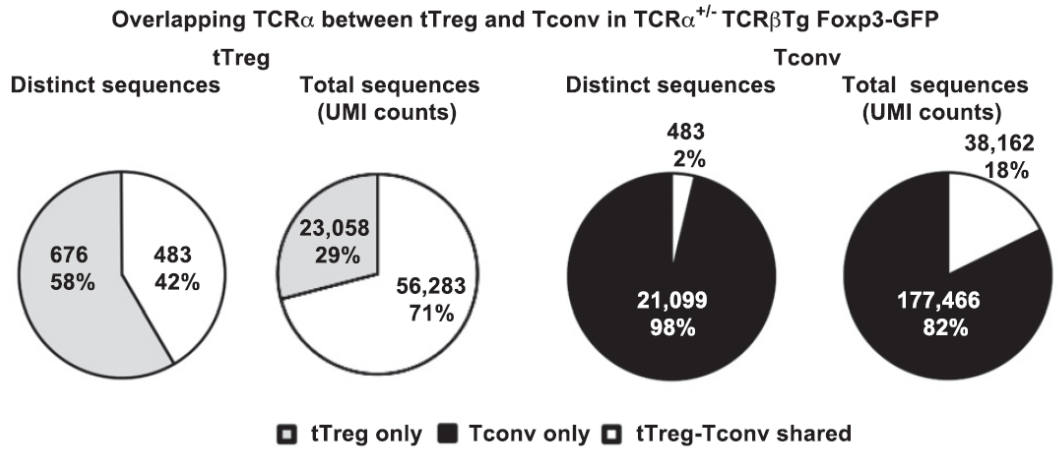


Figure 6-5: **Number and percentages of distinct and shared TCRa sequences between tTreg and Tconv of TCRα+/- TCRβ Tg-Foxp3-GFP mice**. The number and the proportion (%) of sequences found in tTreg only, in Tconv only, or in both tTreg and Tconv (overlap or shared) are presented as a function of all distinct TCRa clonotypes (headed as distinct sequences) and as a proportion of the total UMI counts corresponding to these TCRa (headed as total sequences).

6.3.3 TCRα sequences distinct to tTreg or Tconv or shared between lineages in TCRa+/2 TCRβ Tg mice

To more directly analyze ab combinatorial TCR expression in tTreg and Tconv, we analyzed TCRa sequences of tTreg and Tconv from TCRa+/- TCRb Tg-Foxp3-GFP mice (Table I). Each T cell from these mice expresses a single TCRa (because only one TCRa allele is expressed in these TCRa+/- heterozygotes) in combination with the Tg TCRb (the Tg TCRb accounted for 99.99% of TCRb sequences) so that the TCRa repertoire reflects overall ab TCR clonotype expression. Again, we pooled TCRa sequences of three mice and compared tTreg and Tconv. Strikingly, we found that 42% (483 out of total of 1159) of distinct TCRa sequences of tTreg were shared with Tconv, accounting for 71% of total tTreg TCRa sequences (Fig. 6-5). Shared TCRa sequences represented only 2% of distinct TCRa of Tconv and accounted for 18% of Tconv sequences (Fig. 6-5). These results showed that approximately half of tTreg expressed abTCR identical to those expressed by Tconv in this Tg mouse, indicating that the TCR sequence is not the sole determinant of Treg fate for this large proportion tTreg.

To further characterize the ab TCR sequences shared between tTreg and Tconv, we analyzed their relative abundance by UMI counts of each TCR in two normal mouse strains as well as in a TCRb Tg. We grouped each of the TCRa and TCRb sequences into four groups: 1) found only in tTreg or 2) only in Tconv, 3) shared sequences expressed in Treg, and 4) shared sequences expressed in Tconv. We found that shared TCRa sequences were significantly more abundant (two to nine times) than those of distinct TCRa sequences in both tTreg and Tconv in two normal stains of mice as well as in the TCRa+/- TCRb Tg-Foxp3-GFP mice (Fig. 6-6a). This suggests that those shared TCRab expressing Tconv and tTreg may either be derived from more abundant progenitor cells or might have undergone preferential expansion after differentiation. Shared TCRb sequences were significantly more abundant (three to five times) in Tconv but not in tTreg in two normal strains of mice (Fig. 6-6b).

Last, if the tTreg and Tconv TCRa sequences of TCRa+/- TCRb Tg-Foxp3-GFP are selected based on their sequences, we would expect to see conservation of these TCRa sequences in other strains of mice. To address this, we first pooled sequences from three samples of each strains and then compared tTreg and Tconv TCRa sequences of TCRa+/- TCRb Tg-Foxp3-GFP with those from the non-Tg mice (Rag-GFP/Foxp3-RFP and TcrdCreERZsGreen-Foxp3-EGFP). Indeed, we found a small overlap in TCRa sequences between TCRa+/- TCRb Tg-Foxp3-GFP mice and the two non-Tg strains: 0.14 and 0.08% for distinct and total sequences (based on UMI counts) in tTreg-specific TCRa, 2.7 and 8.3% in Tconv-specific TCRa, and 1.3 and 7.6% in tTreg/ Tconv–shared TCRa (Fig. 6-7).

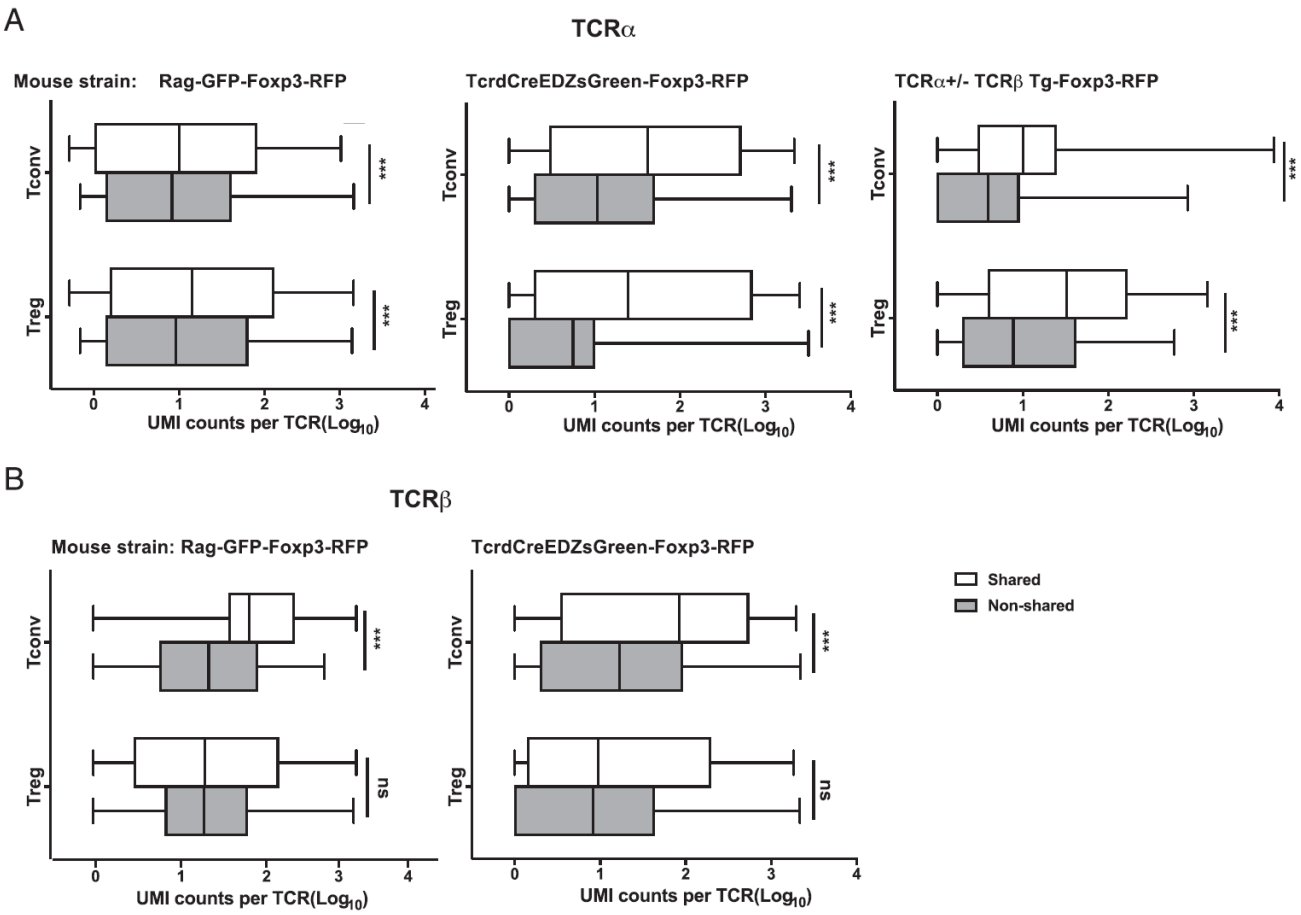


Figure 6-6: **Abundance of shared and nonshared TCRa and TCRb sequences in tTreg and Tconv**. (a) Abundance of TCRa sequences in tTreg and Tconv of three strains of mice. The number of UMI counts corresponding to those shared or nonshared TCRa sequences from all three individual samples of each strain. The data are presented as Log10 transformed values. (b) Abundance of TCRb in tTreg and Tconv of two strains of mice. The number of UMI counts corresponding to those shared or nonshared TCRb sequences from all three individual samples of each strain are presented as the box whisker plot. ***p , 0.001 using Mann–Whitney U test.

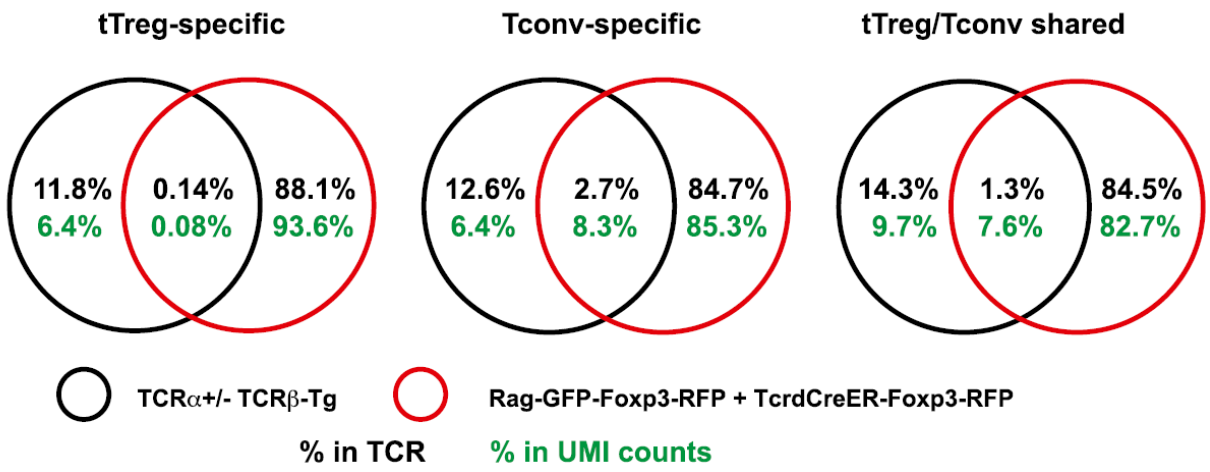


Figure 6-7: **Conservation of TCRa sequences across mouse strains**. Venn diagram displays the percentages of overlap of TCRa sequences (tTreg-specific, Tconv-specfic, and tTreg/Tconv–shared) found in TCRa+/- TCRb Tg-Foxp3-GFP mice with the pooled sequences from Rag-GFP/Foxp3-RFP and TcrdCreERZsGreen-Foxp3-RFP mice. The percentages in black refer to distinct TCRa sequences, and percentages in green refer to total TCRa sequences (based on UMI counts).

6.3.4 Nonshared αβ TCR sequences from tTreg can be distinguished from Tconv

The identification of ab TCR sequences found only in tTreg or only in Tconv could reflect true differences in the TCR repertoires of these populations or might be, at least in part, a consequence of the inability to sequence to saturation all TCR in the large Tconv population. We therefore next asked whether the ab TCR sequences that were not identified as shared between tTreg and Tconv have distinct features that distinguish tTreg and Tconv. We applied an ML algorithm to compare the nonshared ab TCR (V gene-CDR3 amino acids-J gene) of tTreg and Tconv. We first generated a continuous 3-aa motif (trimer) library found in TCR CDR3 [143] and incorporated V and J gene information. We then used 70% of the combined distinct TCRa and TCRb sequences of tTreg and Tconv as a training set and 30% as a testing set using a random forest classifier from scikit-learn [141]. The model performance on the testing set was calculated using the area under the ROC. We found that nonshared TCRa of tTreg were distinguishable from TCRa of Tconv with ROC = 0.82 (Fig. 6-8a) and that TCRb of tTreg were distinguishable from TCRb of Tconv with ROC = 0.72 (Fig. 6-8b); thus, both TCRa and TCRb nonshared sequences of tTreg were distinguished from their counterparts in Tconv.

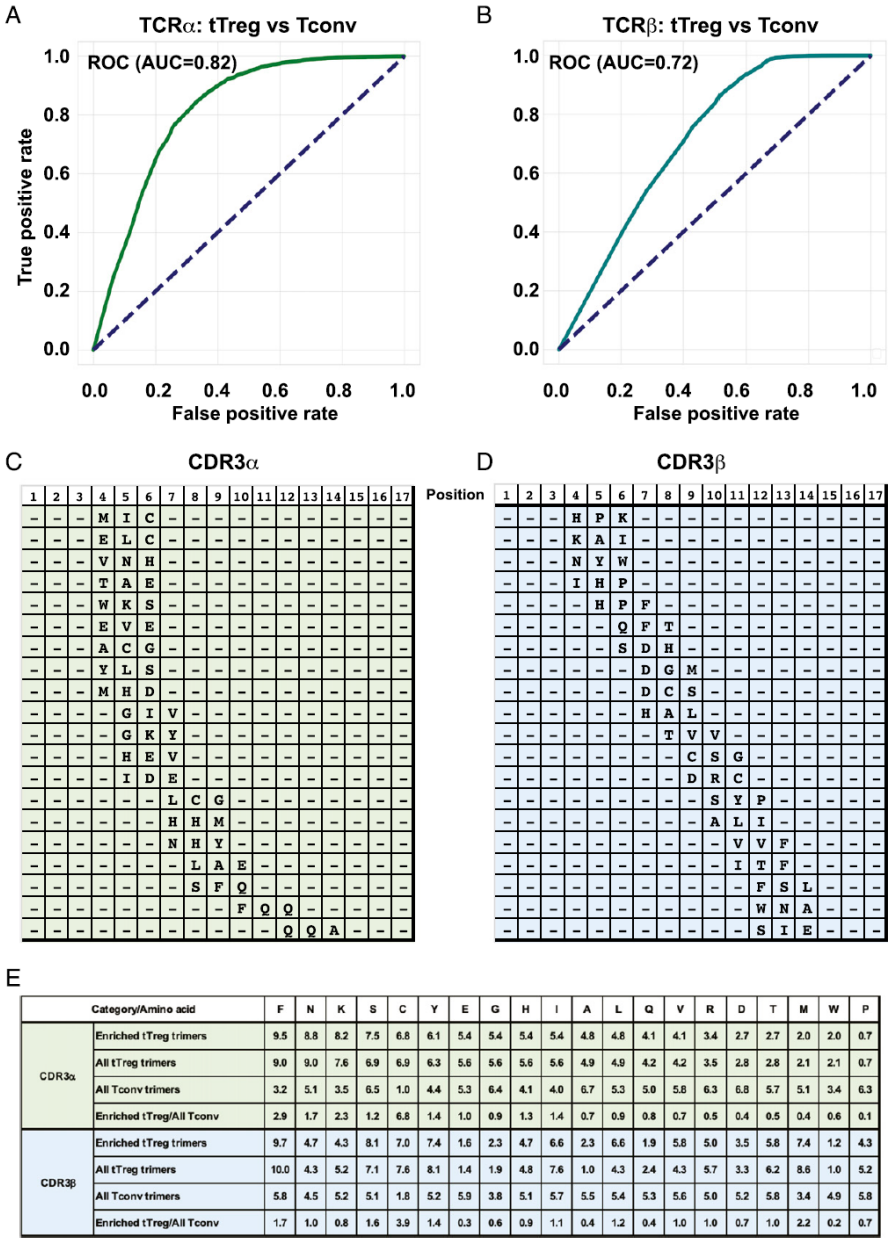
To further determine the features of tTreg-restricted CDR3 amino acid sequences, we compared the abundance in tTreg and Tconv of specific trimers in the central region of CDR3, which mediates direct contact with Ag–MHC, excluding the N-terminal amino acid and the C-terminal 3 aa of CDR3 [135]. We found that a number of trimers were significantly more abundant in tTreg than in Tconv (trimer ratio tTreg/Tconv $\geq$ 1.5, FDR *leq*0.05) in CDR3a (n = 49 found in 2.2% of total tTreg TCRa sequences) and CDR3b (n = 86 found in 1.2% of total tTreg TCRb sequences), and the 20 most abundant trimers for CDR3a and CDR3b are presented in Fig. 6-8c, 6-8d. Strikingly, 2 aa present in these trimers were highly enriched in the abundant trimers of both CDR3a and CDR3b of tTreg: cysteine (enriched by 6.8- and 3.9-fold compared with the CDR3a and CDR3b of Tconv, respectively) and phenylalanine

(enriched by 2.9- and 1.7-fold to the CDR3a and CDR3b of Tconv, respectively) (Fig. 6-8e), suggesting some common biophysical properties of tTreg TCRs. In addition, lysine was enriched by 2.3-fold in CDR3a, and methionine was enriched by 2.2-fold in CDR3b of tTreg. These findings indicate that TCRa and TCRb in tTreg express a distribution of amino acids and trimer sequences that differs from those of Tconv.

To determine the degree to which TCRa and TCRb sequences shared by Treg and Tconv resemble sequences that are found only in tTreg or only in Tconv, we applied the ML algorithm described above. Strikingly, we found that the great majority of TCRa and TCRb sequences that are shared by Treg and Tconv were classified as Tconv in origin by this algorithm (Fig. 6-9a-b). Therefore, 82.4% of TCRa and 91.9% of TCRb sequences were classified as Tconv/TCR–based on cutoffs selected from the receiver operator characteristic of the TCRa and TCRb ML classifiers, respectively. Progenitors that express these shared TCR can thus differentiate to tTreg fate despite the expressions of TCR that are classified as more similar to Tconv.

6.4 Discussion

The factors that determine the selection of Tconv or tTreg fate during thymic development are not completely understood. We designed studies to assess the degree to which TCR sequence determines this lineage choice. We conducted TCRa and TCRb sequencing of tTreg and Tconv using a UMI-based method. Comparing TCRa and TCRb sequences between tTreg and Tconv from normal strains of mice, we found that a substantial proportion of TCRa sequences and TCRb sequences were shared between tTreg and Tconv; in TCRa+/2 TCRb Tg-Foxp3-GFP mice, shared TCRa sequences or clonotypes were even more abundant. Notably, the TCRa and TCRb sequences that were not shared between tTreg and Tconv were distinct in tTreg and Tconv as recognized by an ML algorithm and by identification of amino acid trimers more commonly used in CDR3a and CDR3b of tTreg than in Tconv. Finally, ML indicated that the great majority of TCRs that are shared by tTreg and Tconv have features in common with the sequences distinct to Tconv but not with sequences dis-



E

Category/Amino acid		F	N	K	S	C	Y	E	G	H	I	A	L	Q	V	R	D	T	M	W	P
CDR3α	Enriched tTreg trimers	9.5	8.8	8.2	7.5	6.8	6.1	5.4	5.4	5.4	5.4	4.8	4.8	4.1	4.1	3.4	2.7	2.7	2.0	2.0	0.7
	All tTreg trimers	9.0	9.0	7.6	6.9	6.9	6.3	5.6	5.6	5.6	5.6	4.9	4.9	4.2	4.2	3.5	2.8	2.8	2.1	2.1	0.7
	All Tconv trimers	3.2	5.1	3.5	6.5	1.0	4.4	5.3	6.4	4.1	4.0	6.7	5.3	5.0	5.8	6.3	6.8	5.7	5.1	3.4	6.3
	Enriched tTreg/All Tconv	2.9	1.7	2.3	1.2	6.8	1.4	1.0	0.9	1.3	1.4	0.7	0.9	0.8	0.7	0.5	0.4	0.5	0.4	0.6	0.1
CDR3β	Enriched tTreg trimers	9.7	4.7	4.3	8.1	7.0	7.4	1.6	2.3	4.7	6.6	2.3	6.6	1.9	5.8	5.0	3.5	5.8	7.4	1.2	4.3
	All tTreg trimers	10.0	4.3	5.2	7.1	7.6	8.1	1.4	1.9	4.8	7.6	1.0	4.3	2.4	4.3	5.7	3.3	6.2	8.6	1.0	5.2
	All Tconv trimers	5.8	4.5	5.2	5.1	1.8	5.2	5.9	3.8	5.1	5.7	5.5	5.4	5.3	5.6	5.0	5.2	5.8	3.4	4.9	5.8
	Enriched tTreg/All Tconv	1.7	1.0	0.8	1.6	3.9	1.4	0.3	0.6	0.9	1.1	0.4	1.2	0.4	1.0	1.0	0.7	1.0	2.2	0.2	0.7

Figure 6-8: **Characterization of distinct nonshared TCRa and TCRb sequences expressed by tTreg or Tconv**. (a and b) ML classification of TCRa (a) and TCRb (b) of tTreg and Tconv. The ROC and the associated areas under the curve (AUC) are presented. (c and d) Top 20 most abundant tTreg enriched trimers and their locations in CDR3a (c) and CDR3b (d). (e) Enriched amino acids in these tTreg trimers. Each amino acid present in the enriched trimers and in all trimers are summed and then divided by the total number of amino acids in these enriched trimers or all trimers from tTreg and Tconv. The resulting percentages and the ratios of percentage in enriched tTreg trimer/percentage in Tconv are presented.

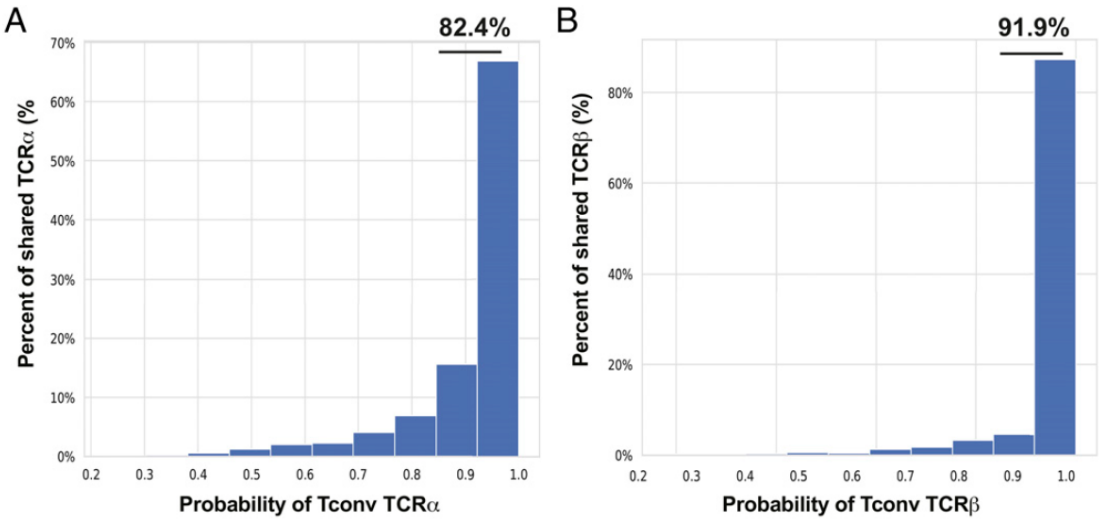


Figure 6-9: **TCRα and TCRβ sequences shared by tTreg and Tconv resemble TCR found uniquely in Tconv**. ML classification of TCRa (a) and TCRb (b) sequences shared by tTreg and Tconv. A classified TCR sequence with the $p \geq 0.8$ is considered as Tconv in origin, whereas ≤ 0.2 is considered as tTreg origin. By this definition, 82.4% of shared TCRa sequences were classified as Tconv and 91.9% shared TCRb sequences were classified as Tconv origin.

tinct to tTreg. Our TCR sequence analysis identified two populations of tTreg, one in which Treg fate is associated with the unique properties of the TCR and another with TCR properties that are characteristic of Tconv and for which tTreg fate may be therefore influenced by other factors than TCR. As previously described in an instructive model of intraclonal competition, the tTreg fate decision can be influenced by the abundance of tTreg precursors during thymic development (36). Further, the potential presence of precursor tTreg within the Tconv population could affect the analysis of the TCR contribution.

It has been reported that tTreg development can proceed through two progenitor cell pathways, with mature CD25+Foxp3+ tTreg being generated from either CD252Foxp3lo or CD25+ Foxp32 precursors (37). These two progenitor tTreg produce functionally distinct mature tTreg. tTreg derived from CD25+ tTreg progenitors are able to prevent experimental autoimmune encephalitis, whereas tTreg from Foxp3lo progenitor cells do not, and TCR of CD25+ tTreg progenitors have a higher affinity than those Foxp3lo progenitor cells. Sequence analysis of the Va2 TCR fam-

ily in that study revealed both distinct and overlapped sequences between these two progenitor cells. Whether these reported features of the Va2 TCR family generalize across the entire TCR repertoire and whether the tTreg-specific TCR and the tTreg/Tconv–shared TCR that we have identified in this study have distinct progenitor origins will require further study.

Our study provides a quantitative assessment of the degree of uniqueness and similarity of ab TCR between tTreg and Tconv. The distinct tTreg and Tconv sequences that we identified are present at a small but consistent percentage among individual mice within the same strain and between different strains, suggesting sequence conservation of tTreg- or Tconv-specific TCR characteristics. Furthermore, these conserved sequences occupy a larger fraction of total TCRa and TCRb sequences based on UMI counts than on distinct TCR sequences. Several factors could affect the estimated proportion of these shared TCR. Because the number of Tconv that could be used for sequencing was only a fraction of total Tconv in the thymus, the percentage of shared TCR sequences in tTreg may be an underestimate. In contrast, the existence of a potential precursor tTreg in the Tconv population could increase apparent sharing, although it should be noted that the Tconv (Foxp3-RFP2) analyzed for Rag-GFP/Foxp3-RFP thymocytes were only 1% CD25+ and that Tconv for other analyses were 99% Foxp3- CD252. The extensive sharing of ab TCR between tTreg and Tconv is further confirmed through the analysis of TCRa sequences in a TCRb Tg mouse, in which we found sharing of TCRa and therefore sharing of identical ab TCR clonotypes between tTreg and Tconv in 71% of total Treg. This high degree of sharing may result from the substantially reduced size of the TCRa repertoire that can be paired with the single Tg TCRb. Together, these findings suggest that TCR sequence alone does not determine the fate of tTreg or Tconv for a significant proportion of tTreg.

The UMI marking of TCR mRNA molecules allowed us to calculate the abundance of TCR sequences in tTreg and Tconv more accurately than prior approaches. This strategy led to the observation that those TCRa sequences in normal mice and ab TCR clonotypes in TCRa+/- TCRb Tg-Foxp3-GFP mice that are shared between

tTreg and Tconv are significantly more abundant than the nonshared TCRa sequences and ab TCR clonotypes. It is not clear whether the abundance of those shared TCR occurs prior to or after the fate decision of tTreg and Tconv. A study using fluorescent tracking of thymic mature CD4+ and tTreg showed only one cell division postselection (38). This suggests that a high abundance of a given preselection progenitor may simply increase the probability that some progenitors of that clonotype will differentiate to tTreg and others to Tconv, reflecting a role of additional factors that drive Treg differentiation in concert with TCR signaling. However, TCRb sequences shared by tTreg and Tconv were significantly more abundant than nonshared sequences expressed only in Tconv but not more abundant than nonshared sequences expressed only in tTreg. It remains to be determined whether there is selective expansion occurring after the fate decision in the thymus for Tconv that have shared TCR with tTreg.

ML has become an increasingly powerful tool in biological studies, in particular for those involving large datasets [144, 145]. In this study, we applied ML to analyze TCRs by first partitioning each distinct TCR into its component V-CDR3 (multiple continuous tri–amino acids)-J factors and determining whether nonshared TCRa and TCRb sequences are distinct between tTreg and Tconv. Indeed, a random forest classifier is able to distinguish those TCRa and TCRb sequences of tTreg from those of Tconv with high accuracy. The identification of specific trimers that occur at a higher frequency in tTreg CDR3a and CDR3b reveals that cysteine and phenylalanine are enriched in both CDR3a and CDR3b of tTreg, suggesting common structural features of at least some tTreg TCR. This ML algorithm can be further improved when more TCR sequences of tTreg and Tconv are generated and can also be modified to test ab paired TCRs as the potential for single cell sequencing advances. In addition, we observed that lysine is enriched in CDR3a trimers, whereas methionine is enriched in CDR3b trimers of tTreg. Although the roles of these enriched trimers and specific amino acids is currently unknown, cysteine is reported to be enriched in CDR3a and CDR3b in intraepithelial lymphocytes and type A intraepithelial lymphocytes precursors [146]. Interestingly, cysteine and phenylalanine are reported to be less

frequent in CDR3b of MHC-restricted than in MHC-independent TCR-expressing thymocytes [135], and phenylalanine was found in CDR3b of selfreactive TCR [147]. Thus, it is possible that the enriched amino acids and enriched trimers in tTreg CDR3 may be involved in the interaction with self-antigen peptide rather than MHC during tTreg differentiation. Collectively, our findings demonstrate that TCR expressed by tTreg and not by Tconv have distinct CDR3a and CDR3b sequences compared with those of Tconv, supporting the critical role of TCR in the tTreg generation in thymus. Further studies will be needed to characterize the interaction of these TCRs with potentially selecting self-peptides to provide insights into the role of TCR in tTreg generation in the thymus.

Collectively, the results reported in this study have identified features that distinguish the ab TCR sequences expressed by a significant proportion of tTreg from those expressed by Tconv in the thymus of normal mice and which therefore appear critical to determining differentiation into these lineages. The identification in tTreg of preferentially used trimers and selected amino acids in CDR3a and CDR3b provides molecular features for further understanding of TCR and Ag interaction in tTreg generation. For TCR clonotypes that are shared between tTreg and Tconv, it remains to be elucidated what are the non-TCR factors that drive the same TCR-carrying progenitors into either Treg or Tconv lineage. Candidates for such factors include differential costimulatory signaling, cytokine requirements, and other aspects of APC and thymic environment. The ML classification of tTreg and Tconv subsets by TCR sequence described in this study provides a strategy for dissecting the molecular pathways that mark these lineages at a single cell level. Combining ML and single cell analysis of ab TCR sequence with transcriptome and other molecular parameters will allow better definition of the selection, function, and activation state of these T cell subsets.

Chapter 7

Conclusion

In this thesis, we have introduced extracellular vesicles (EVs) as biomarkers for immunotherapy response, delineated the epigenetic changes that result from checkpoint blockade immunotherapy, and elucidated the role of TCR repertoires in driving CD4+ SP differentiation into Tconv or tTreg lineages.

In chapter 2, we first described the potential of EV RNA as a potential predictive biomarker for checkpoint blockade immunotherapy patients. We show a high level of correlation between both tumor cell-lines and tumor cell-line-derived EVs, as well as between patient tumors and patient plasma-derived EVs. Using concordance analysis, we show enrichment for immune related pathways and cell types in the EV population, suggesting that EV populations maybe more reflective of an immune role. We pinpoint DEGs and pathways that stratify responders and non-responders in both the pre-treatment and on-treatment EV samples. We note the presence of cancer testis antigens MAGEA1 and MAGEA3, which are uniquely expressed in melanoma cells, as on-treatment DEGs. In addition, we demonstrate that the pre-treatment samples are moderately predictive of immunotherapy response. Finally, we demonstrate that EV RNA-seq mutational information can serve as a proxy for somatic mutational load and can stratify responders and non-responders. In chapter 3, we present a Bayesian probabilistic model that explicitly models a "scaling" coefficient meant to capture the export process between tumors and EVs, as well as a mixture coefficient to denote the fraction of plasma-derived EV signal that is thought to be

patient tumor DEV-derived. We validate this model using the *in silico* deconvolution method CIBERSORTx, and we leverage the model to provide interpretations for our on-treatment and pre-treatment DEGs.

In chapter 4, we described the epigenetic state changes that differ between responders and non-responders. We first pinpointed a set of peaks that transition from enhancer states in non-responders to repressed or polycomb states in responders. We demonstrate that GO pathways such as Notch, MET, and ECM signaling are enriched for in this set of peaks. Next, we dissected the MET locus, with its 4 distal enhancers, and showed that silencing these enhancer interactions with dCas9-KRAB resulted in increased tumor killing in a co-culture experiment. We also found a number of other genes, including TGFβ2, XIST, and SPATA2, also demonstrated similar patterns as the c-MET locus of having significantly higher non-responder H3K27ac activity. Based on this, we reasoned that inhibiting acetylation reader bromodomains via BETi inhibition could potentially target multiple resistance mechanisms at once. We show both in B16F10 and the Bosenberg model that anti-PD1 in conjunction with BETi exhibits synergism and led to the largest decrease in tumor volume. We further show that BETi increased CD8% T-cell percentage in the tumor microenvironment, tumor killing by tumor infiltrating lymphocytes, as well as MHC I expression. In chapter 5, we used MANorm and IDR to isolate a set of $\sim 86,000$ peaks for downstream analysis, of which 189 peaks showed significance in both the MGH and MDA datasets. We show that these 189 peaks are able to predict response via a random forest classifier and that 4 of the peaks stratified progression free survival. Finally, we used the Global Test to test gene-peak modules and gene-pathway modules, leading to the identification of several immune related pathways as differentially regulated.

In chapter 6, we elucidated the role of TCR in driving CD4+ SP T-cells into thymic Treg (tTreg) vs. conventional T-cell (Tconv) fates. We identified αβ TCR sequences that were unique to either tTreg or Tconv and found these sequences were distinctly recognized by a random forest classifier and preferentially used amino acid trimers in αβ CDR3 of tTreg. We also found a proportion of the αβ TCR sequences expressed by tTreg were also found in Tconv, and machine learning classified the great

majority of these shared αβ TCR sequences as characteristic of Tconv and not tTreg. These findings identify two populations of tTreg, one in which regulatory T-cell fate is associated with unique properties of the TCR and another for which tTreg fate is determined by factors beyond TCR sequence.

7.1 Looking to the future

Additional research into the potential of extravesicular RNA as a predictive biomarker for immunotherapy response is still needed. Targeted enrichment and profiling of tumor-derived EV is necessary to pinpoint which of the observed signals in our study is actually enriched for in patients. Performing targeted enrichment will also allow us to validate aspects of our *in silico* deconvolution model. Furthermore, profiling of patients outside of metastatic melanoma could also be useful to see if our findings transfer over to additional tumor types. Ultimately, the bar is a prospective clinical trial in which EVs are explicitly tested for their ability to guide clinical decision making and provide better patient stratification in the context of immunotherapy treatment. In this goal, EVs face a similar developmental process as that of tumor mutational burden, which took several trials in order establish as a gold standard biomarker in the context of immunotherapy response.

In the context of epigenetic profiling, the obvious next steps would be to investigate whether the observed synergism between BETi and anti-PD1 is relevant in humans. If so, this combinatorial drugging regime can potentially improve immunotherapy efficacy and enable overall longer survival times in patients. Beyond this, confirmation of the non-responder enhancer signature in other tumor types besides metastatic melanoma would provide additional evidence that this is a universal signature of non-response. As for the predictive enhancer signals isolated in chapter 5, these signatures should be developed into a distinct assay utilizing ChIP-qPCR or another sensitive detector of H3K27ac signal in a novel cohort to confirm whether these signatures are truly predictive.

Finally, in the context of TCR profiling, studies leveraging paired αβ TCR se-

quencing should be employed, as this would increase confidence in the applicability of the findings to the total αβ TCR repertoire, instead of individual α or β repertoires by themselves. Additional studies should be undertaken to determine the biological factors responsible for directing CD4+ SP T-cells down tTreg vs. Tconv fates. Specific improvements to individual chapters are listed below.

7.1.1 Extracellular vesicles as biomarkers for immunotherapy resistance

In the work presented in chapter 2, there are a number of areas for potential improvement. The first is the overall approach of capturing mixed, plasma-derived EVs in which tumor-derived EVs are an unknown fraction of the captured pool. To improve upon this, we can use microfluidics and nanoengineering to explicitly filter for tumor-derived EVs and profile them separately [148]. This will allow us to directly access the tumor-derived EV RNA profile, without the need to impute for them computationally, as well as specifically analyze tumor-associated changes instead of focusing on both immune and tumor related changes as we have done in our work. Additionally, more complete immunophenotyping of circulating EVs is also possible via the work done by Zhang et al. [149].

7.1.2 Deconvolution of extracellular cargo

In the work presented in chapter 3, the limitations on the work are the lack of validation data and the simplistic model for the scaling coefficient s. The deconvolution model should ideally be confirmed with *in vitro* or *in vivo* experimental model. The current validation with CIBERSORTx uses another *in silico* model to confirm an existing *in silico* model, leading to issues with how well the model realizes biological reality. Ideally, this would involve experimentation in which tumor-derived EV profiles are isolated, for example via [148], and those profiles are compared with the *in silico* predicted ones generated by the model. An additional issue is the simplistic linear model for the scaling coefficient s, that models the export process between

tumors and tumor-derived EVs. Ideally, this export process should be non-linear in nature, to capture more complex interactions between RNA concentration within the tumors and the RNA concentration as exported to tumor-derived EVs.

7.1.3 Epigenetic changes during immunotherapy resistance

A major limitation of the work presented in chapter 4 is the focus on the non-responder E7 state as the key signature of immunotherapy response. Although we were able to detect an actionable epigenetic signature from peaks residing in the non-responder E7 state, this does not preclude the presence of other epigenetic signatures residing in other chromatin states. A significance test for whether read counts differed significantly between each of the ChromHMM states would be able to identify other epigenetic signatures that stratify responders and non-responders.

7.1.4 Epigenomic predictors of immunotherapy resistance

One limitation of the work presented in chapter 5 is the limited pool of peaks that we were drawing from in order to create the feature set for input into the random forest classifier. Using only the 189 doubly significant peaks ensured that the individual peaks were predictive; however, it is potentially missing predictive features from the $\sim$ 86000 other peaks that could potentially stratify response. An alternative strategy for peak discovery is perhaps to use a Bayesian spike and slab regression [150] to select for the predictive peaks among the 86,000 peak pool prior to running a random forest classifier. A regularized regression framework may also work. This would enable the selection of peaks outside of doubly significant pool and potentially expand the predictive power of the pre-treatment epigenetic peak set.

7.1.5 TCR profiling

The major limitation in the study is the lack of paired αβ TCR sequencing. This would allow direct addressing of whether the paired αβ TCR repertoire influences tTreg vs. Tconv development. Paired αβ TCR sequencing can be performed by the protocol

presented by Howie et al. [151], and the paired αβ sequences can be analyzed using the existing analysis algorithms. This would complement the analysis with the TCRα +/- TCRβ-Tg-Foxp3-GFP mice, which synthetically constrains TCR α by using only a single TCR β gene. Paired sequences would ensure that the overlapping TCRα and TCRβ sequences between tTreg and Tconv are actually overlapping as fully formed αβ TCR.

Bibliography

[1] L. E. Davis, S. C. Shalin, and A. J. Tackett, "Current state of melanoma diagnosis and treatment," *Cancer Biol. Ther.*, vol. 20, pp. 1366–1379, Aug. 2019.

[2] D. Schadendorf, D. E. Fisher, C. Garbe, J. E. Gershenwald, J.-J. Grob, A. Halpern, M. Herlyn, M. A. Marchetti, G. McArthur, A. Ribas, A. Roesch, and A. Hauschild, "Melanoma," *Nature Reviews Disease Primers*, vol. 1, pp. 1–20, Apr. 2015.

[3] I. Mellman, G. Coukos, and G. Dranoff, "Cancer immunotherapy comes of age," *Nature*, vol. 480, pp. 480–489, Dec. 2011.

[4] F. S. Hodi, S. J. O'Day, D. F. McDermott, R. W. Weber, J. A. Sosman, J. B. Haanen, R. Gonzalez, C. Robert, D. Schadendorf, J. C. Hassel, W. Akerley, A. J. M. van den Eertwegh, J. Lutzky, P. Lorigan, J. M. Vaubel, G. P. Linette, D. Hogg, C. H. Ottensmeier, C. Lebbé, C. Peschel, I. Quirt, J. I. Clark, J. D. Wolchok, J. S. Weber, J. Tian, M. J. Yellin, G. M. Nichol, A. Hoos, and W. J. Urba, "Improved survival with ipilimumab in patients with metastatic melanoma," *N. Engl. J. Med.*, vol. 363, pp. 711–723, Aug. 2010.

[5] C. Robert, "A decade of immune-checkpoint inhibitors in cancer therapy," *Nat. Commun.*, vol. 11, p. 3801, July 2020.

[6] M. Yarchoan, A. Hopkins, and E. M. Jaffee, "Tumor mutational burden and response rate to PD-1 inhibition," *N. Engl. J. Med.*, vol. 377, pp. 2500–2501, Dec. 2017.

[7] C. S. Fuchs, M. Özgüroğlu, Y.-J. Bang, M. Di Bartolomeo, M. Mandalà, M.-H. Ryu, C. Vivaldi, T. Olesinski, C. Caglevic, H. C. Chung, K. Muro, E. Van Cutsem, J. Kobie, R. Cristescu, D. Aurora-Garg, J. Lu, C.-S. Shih, D. Adelberg, Z. A. Cao, and K. Shitara, "The association of molecular biomarkers with efficacy of pembrolizumab versus paclitaxel in patients with gastric cancer (GC) from KEYNOTE-061," *J. Clin. Orthod.*, vol. 38, pp. 4512–4512, May 2020.

[8] L. Cai, H. Bai, J. Duan, Z. Wang, S. Gao, D. Wang, S. Wang, J. Jiang, J. Han, Y. Tian, X. Zhang, H. Ye, M. Li, B. Huang, J. He, and J. Wang, "Epigenetic alterations are associated with tumor mutation burden in non-small cell lung cancer," *Journal for ImmunoTherapy of Cancer*, vol. 7, pp. 1–11, July 2019.

[9] N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, and T. A. Chan, "Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer," *Science*, vol. 348, pp. 124–128, Apr. 2015.

[10] E. Ghorani, R. Rosenthal, N. McGranahan, J. L. Reading, M. Lynch, K. S. Peggs, C. Swanton, and S. A. Quezada, "Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma," *Ann. Oncol.*, vol. 29, pp. 271–279, Jan. 2018.

[11] A. J. Rech, D. Balli, A. Mantero, H. Ishwaran, K. L. Nathanson, B. Z. Stanger, and R. H. Vonderheide, "Tumor immunity and survival as a function of alternative neopeptides in human cancer," *Cancer Immunol Res*, vol. 6, pp. 276–287, Mar. 2018.

[12] R. Bai, Z. Lv, D. Xu, and J. Cui, "Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors," *Biomarker Research*, vol. 8, pp. 1–17, Aug. 2020.

[13] M. Mathew, M. Zade, N. Mezghani, R. Patel, Y. Wang, and F. Momen-Heravi, "Extracellular vesicles as biomarkers in cancer immunotherapy," *Cancers*, vol. 12, Sept. 2020.

[14] G. Chen, A. C. Huang, W. Zhang, G. Zhang, M. Wu, W. Xu, Z. Yu, J. Yang, B. Wang, H. Sun, H. Xia, Q. Man, W. Zhong, L. F. Antelo, B. Wu, X. Xiong, X. Liu, L. Guan, T. Li, S. Liu, R. Yang, Y. Lu, L. Dong, S. McGettigan, R. Somasundaram, R. Radhakrishnan, G. Mills, Y. Lu, J. Kim, Y. H. Chen, H. Dong, Y. Zhao, G. C. Karakousis, T. C. Mitchell, L. M. Schuchter, M. Herlyn, E. J. Wherry, X. Xu, and W. Guo, "Exosomal PD-L1 contributes to immunosuppression and is associated with anti-PD-1 response," *Nature*, vol. 560, pp. 382–386, Aug. 2018.

[15] S. N. Hurwitz and D. G. Meckes, Jr, "Extracellular vesicle integrins distinguish unique cancers," *Proteomes*, vol. 7, Apr. 2019.

[16] H. Y. Zoghbi and A. L. Beaudet, "Epigenetics and human disease," *Cold Spring Harb. Perspect. Biol.*, vol. 8, p. a019497, Feb. 2016.

[17] S. Virani, J. A. Colacino, J. H. Kim, and L. S. Rozek, "Cancer epigenetics: a brief review," *ILAR J.*, vol. 53, no. 3-4, pp. 359–369, 2012.

[18] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein,

E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, p. 317, Feb. 2015.

[19] J. Ernst and M. Kellis, "Chromatin-state discovery and genome annotation with ChromHMM," *Nat. Protoc.*, vol. 12, pp. 2478–2492, Nov. 2017.

[20] E. Rosati, C. M. Dowds, E. Liaskou, E. K. K. Henriksen, T. H. Karlsen, and A. Franke, "Overview of methodologies for t-cell receptor repertoire analysis," *BMC Biotechnol.*, vol. 17, p. 61, July 2017.

[21] X. Bai, Q. Zhang, S. Wu, X. Zhang, M. Wang, F. He, T. Wei, J. Yang, Y. Lou, Z. Cai, and T. Liang, "Characteristics of tumor infiltrating lymphocyte and circulating lymphocyte repertoires in pancreatic cancer by the sequencing of T cell receptors," *Sci. Rep.*, vol. 5, p. 13664, Sept. 2015.

[22] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, pp. 1–26, Jan. 2021.

[23] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, pp. e47–e47, Jan. 2015.

[24] A. Shi, G. G. Kasumova, W. A. Michaud, J. Cintolo-Gonzalez, M. Díaz-Martínez, J. Ohmura, A. Mehta, I. Chien, D. T. Frederick, S. Cohen, D. Plana, D. Johnson, K. T. Flaherty, R. J. Sullivan, M. Kellis, and G. M. Boland, "Plasma-derived extracellular vesicle analysis and deconvolution enable prediction and tracking of melanoma checkpoint blockade outcome," *Science Advances*, vol. 6, p. eabb3461, Nov. 2020.

[25] A. Ko, M. Watanabe, T. Nguyen, A. Shi, A. Achour, B. Zhang, X. Sun, Q. Wang, Y. Zhuang, N.-P. Weng, and R. J. Hodes, "TCR repertoires of thymic conventional and regulatory T cells: Identification and characterization of both unique and shared TCR sequences," *J. Immunol.*, Jan. 2020.

[26] A. Ashida, K. Sakaizawa, H. Uhara, and R. Okuyama, "Circulating tumour DNA for monitoring treatment response to Anti-PD-1 immunotherapy in melanoma patients," *Acta Derm. Venereol.*, vol. 97, pp. 1212–1218, Nov. 2017.

[27] X. Hong, R. J. Sullivan, M. Kalinich, T. T. Kwan, A. Giobbie-Hurder, S. Pan, J. A. LiCausi, J. D. Milner, L. T. Nieman, B. S. Wittner, U. Ho, T. Chen, R. Kapur, D. P. Lawrence, K. T. Flaherty, L. V. Sequist, S. Ramaswamy, D. T. Miyamoto, M. Lawrence, M. Toner, K. J. Isselbacher, S. Maheswaran, and D. A. Haber, "Molecular signatures of circulating melanoma cells for monitoring early response to immune checkpoint therapy," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, pp. 2467–2472, Mar. 2018.

[28] N. Riaz, J. J. Havel, V. Makarov, A. Desrichard, W. J. Urba, J. S. Sims, F. Stephen Hodi, S. Martín-Algarra, R. Mandal, W. H. Sharfman, S. Bhatia, W.-J. Hwu, T. F. Gajewski, C. L. Slingluff, D. Chowell, S. M. Kendall, H. Chang, R. Shah, F. Kuo, L. G. T. Morris, J.-W. Sidhom, J. P. Schneck, C. E. Horak, N. Weinhold, and T. A. Chan, "Tumor and microenvironment evolution during immunotherapy with nivolumab," *Cell*, vol. 171, pp. 934–949.e15, Nov. 2017.

[29] S. D. Alipoor, E. Mortaz, M. Varahram, M. Movassaghi, A. D. Kraneveld, J. Garssen, and I. M. Adcock, "The potential biomarkers and immunological effects of Tumor-Derived exosomes in lung cancer," *Front. Immunol.*, vol. 9, p. 819, Apr. 2018.

[30] L. Muller, S. Muller-Haegele, M. Mitsuhashi, W. Gooding, H. Okada, and T. L. Whiteside, "Exosomes isolated from plasma of glioma patients enrolled in a vaccination trial reflect antitumor immune activity and might predict survival," *Oncoimmunology*, vol. 4, p. e1008347, June 2015.

[31] M. K. McDonald, Y. Tian, R. A. Qureshi, M. Gormley, A. Ertel, R. Gao, E. Aradillas Lopez, G. M. Alexander, A. Sacan, P. Fortina, and S. K. Ajit, "Functional significance of macrophage-derived exosomes in inflammation and pain," *Pain*, vol. 155, pp. 1527–1539, Aug. 2014.

[32] Z. Cai, F. Yang, L. Yu, Z. Yu, L. Jiang, Q. Wang, Y. Yang, L. Wang, X. Cao, and J. Wang, "Activated T cell exosomes promote tumor invasion via fas signaling pathway," *J. Immunol.*, vol. 188, pp. 5954–5961, June 2012.

[33] D. W. Greening, S. K. Gopal, R. Xu, R. J. Simpson, and W. Chen, "Exosomes and their roles in immune regulation and cancer," *Semin. Cell Dev. Biol.*, vol. 40, pp. 72–81, Apr. 2015.

[34] T. A. Chatila and C. B. Williams, "Regulatory T cells: exosomes deliver tolerance," *Immunity*, vol. 41, pp. 3–5, July 2014.

[35] C. Théry, S. Amigorena, G. Raposo, and A. Clayton, "Isolation and characterization of exosomes from cell culture supernatants and biological fluids," *Curr. Protoc. Cell Biol.*, vol. Chapter 3, p. Unit 3.22, Apr. 2006.

[36] B. S. Carvalho and R. A. Irizarry, "A framework for oligonucleotide microarray preprocessing," *Bioinformatics*, vol. 26, pp. 2363–2367, Oct. 2010.

[37] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, pp. 118–127, Jan. 2007.

[38] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, pp. i884–i890, Sept. 2018.

[39] Y. Liao, G. K. Smyth, and W. Shi, "The R package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads," *Nucleic Acids Res.*, vol. 47, p. e47, May 2019.

[40] G. K. Smith, "limma: Linear models for microarray data," *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, no. 2005, pp. 397–420, 2005.

[41] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, pp. 1739–1740, June 2011.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[43] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, and S. A. Forbes, "COSMIC: the catalogue of somatic mutations in cancer," *Nucleic Acids Res.*, vol. 47, pp. D941–D947, Jan. 2019.

[44] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. a. Alizadeh, "Robust enumeration of cell subsets from tissue expression profiles," *Nat. Methods*, vol. 12, pp. 453–457, Mar. 2015.

[45] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh, "Determining cell type abundance and expression from bulk tissues with digital cytometry," *Nat. Biotechnol.*, vol. 37, pp. 773–782, July 2019.

[46] W. Roh, P.-L. Chen, A. Reuben, C. N. Spencer, P. A. Prieto, J. P. Miller, V. Gopalakrishnan, F. Wang, Z. A. Cooper, S. M. Reddy, C. Gumbs, L. Little, Q. Chang, W.-S. Chen, K. Wani, M. P. De Macedo, E. Chen, J. L. Austin-Breneman, H. Jiang, J. Roszik, M. T. Tetzlaff, M. A. Davies, J. E. Gershenwald, H. Tawbi, A. J. Lazar, P. Hwu, W.-J. Hwu, A. Diab, I. C. Glitza, S. P. Patel, S. E. Woodman, R. N. Amaria, V. G. Prieto, J. Hu, P. Sharma, J. P. Allison, L. Chin, J. Zhang, J. A. Wargo, and P. A. Futreal, "Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance," *Sci. Transl. Med.*, vol. 9, Mar. 2017.

[47] S. Hänzelmann, R. Castelo, and J. Guinney, "GSVA: gene set variation analysis for microarray and RNA-seq data," *BMC Bioinformatics*, vol. 14, p. 7, Jan. 2013.

[48] R. W. Jenkins, A. R. Aref, P. H. Lizotte, E. Ivanova, S. Stinson, C. W. Zhou, M. Bowden, J. Deng, H. Liu, D. Miao, M. X. He, W. Walker, G. Zhang, T. Tian, C. Cheng, Z. Wei, S. Palakurthi, M. Bittinger, H. Vitzthum, J. W. Kim, A. Merlino, M. Quinn, C. Venkataramani, J. A. Kaplan, A. Portell, P. C. Gokhale, B. Phillips, A. Smart, A. Rotem, R. E. Jones, L. Keogh, M. Anguiano, L. Stapleton, Z. Jia, M. Barzily-Rokni, I. Cañadas, T. C. Thai, M. R. Hammond, R. Vlahos, E. S. Wang, H. Zhang, S. Li, G. J. Hanna, W. Huang, M. P. Hoang, A. Piris, J.-P. Eliane, A. O. Stemmer-Rachamimov, L. Cameron, M.-J. Su, P. Shah, B. Izar, M. Thakuria, N. R. LeBoeuf, G. Rabinowits, V. Gunda, S. Parangi, J. M. Cleary, B. C. Miller, S. Kitajima, R. Thummalapalli, B. Miao, T. U. Barbie, V. Sivathanu, J. Wong, W. G. Richards, R. Bueno, C. H. Yoon, J. Miret, M. Herlyn, L. A. Garraway, E. M. Van Allen, G. J. Freeman, P. T. Kirschmeier, J. H. Lorch, P. A. Ott, F. S. Hodi, K. T. Flaherty, R. D. Kamm, G. M. Boland, K.-K. Wong, D. Dornan, C. P. Paweletz, and D. A. Barbie, "Ex vivo profiling of PD-1 blockade using organotypic tumor spheroids," *Cancer Discov.*, vol. 8, pp. 196–215, Feb. 2018.

[49] M. Janghorban, L. Xin, J. M. Rosen, and X. H.-F. Zhang, "Notch signaling as a regulator of the tumor immune response: To target or not to target?," *Front. Immunol.*, vol. 9, p. 1649, July 2018.

[50] T. Sinnberg, M. P. Levesque, J. Krochmann, P. F. Cheng, K. Ikenberg, F. Meraz-Torres, H. Niessner, C. Garbe, and C. Busch, "Wnt-signaling enhances neural crest migration of melanoma cells and induces an invasive phenotype," *Mol. Cancer*, vol. 17, p. 59, Feb. 2018.

[51] E. I. Buchbinder and A. Desai, "CTLA-4 and PD-1 pathways: Similarities, differences, and implications of their inhibition," *Am. J. Clin. Oncol.*, vol. 39, pp. 98–106, Feb. 2016.

[52] A. Memon and W. K. Lee, "KLF10 as a tumor suppressor gene and its TGF-β signaling," *Cancers*, vol. 10, May 2018.

[53] L. Gattinoni, X.-S. Zhong, D. C. Palmer, Y. Ji, C. S. Hinrichs, Z. Yu, C. Wrzesinski, A. Boni, L. Cassard, L. M. Garvin, C. M. Paulos, P. Muranski, and N. P. Restifo, "Wnt signaling arrests effector T cell differentiation and generates CD8+ memory stem cells," *Nat. Med.*, vol. 15, pp. 808–813, July 2009.

[54] K. Ohman Forslund and K. Nordqvist, "The melanoma antigen genes–any clues to their functions in normal tissues?," *Exp. Cell Res.*, vol. 265, pp. 185–194, May 2001.

[55] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, "GAGE: generally applicable gene set enrichment for pathway analysis," *BMC Bioinformatics*, vol. 10, p. 161, May 2009.

[56] W. Hugo, J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, E. Seja, S. Lomeli, X. Kong, M. C. Kelley, J. A. Sosman, D. B. Johnson, A. Ribas, and R. S. Lo, "Genomic and transcriptomic features of response to Anti-PD-1 therapy in metastatic melanoma," *Cell*, vol. 165, no. 1, pp. 35–44, 2016.

[57] J. Deng, E. S. Wang, R. W. Jenkins, S. Li, R. Dries, K. Yates, S. Chhabra, W. Huang, H. Liu, A. R. Aref, E. Ivanova, C. P. Paweletz, M. Bowden, C. W. Zhou, G. S. Herter-Sprie, J. A. Sorrentino, J. E. Bisi, P. H. Lizotte, A. A. Merlino, M. M. Quinn, L. E. Bufe, A. Yang, Y. Zhang, H. Zhang, P. Gao, T. Chen, M. E. Cavanaugh, A. J. Rode, E. Haines, P. J. Roberts, J. C. Strum, W. G. Richards, J. H. Lorch, S. Parangi, V. Gunda, G. M. Boland, R. Bueno, S. Palakurthi, G. J. Freeman, J. Ritz, W. N. Haining, N. E. Sharpless, H. Arthanari, G. I. Shapiro, D. A. Barbie, N. S. Gray, and K.-K. Wong, "CDK4/6 inhibition augments antitumor immunity by enhancing t-cell activation," *Cancer Discov.*, vol. 8, pp. 216–233, Feb. 2018.

[58] L. Zhang and Z. Zhang, "Recharacterizing Tumor-Infiltrating lymphocytes by Single-Cell RNA sequencing," *Cancer Immunol Res*, vol. 7, pp. 1040–1046, July 2019.

[59] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, pp. 1145–1159, July 1997.

[60] A. M. Goodman, S. Kato, L. Bazhenova, S. P. Patel, G. M. Frampton, V. Miller, P. J. Stephens, G. A. Daniels, and R. Kurzrock, "Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers," *Mol. Cancer Ther.*, vol. 16, pp. 2598–2608, Nov. 2017.

[61] N. Auslander, G. Zhang, J. S. Lee, D. T. Frederick, B. Miao, T. Moll, T. Tian, Z. Wei, S. Madan, R. J. Sullivan, G. Boland, K. Flaherty, M. Herlyn, and E. Ruppin, "Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma," *Nat. Med.*, vol. 24, pp. 1545–1549, Oct. 2018.

[62] L. Han, E. W.-F. Lam, and Y. Sun, "Extracellular vesicles in the tumor microenvironment: old stories, but new tales," *Mol. Cancer*, vol. 18, pp. 1–14, Mar. 2019.

[63] T. L. Whiteside, "The potential of tumor-derived exosomes for noninvasive cancer monitoring," *Expert Rev. Mol. Diagn.*, vol. 15, no. 10, pp. 1293–1310, 2015.

[64] T. L. Whiteside, "Exosomes and tumor-mediated immune suppression," *J. Clin. Invest.*, vol. 126, no. 4, pp. 1216–1223, 2016.

[65] M. D. Hoffman and A. Gelman, "The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian monte carlo," *J. Mach. Learn. Res.*, vol. 15, no. April, pp. 1593–1623, 2014.

[66] A. Gelman, D. Lee, and J. Guo, "Stan: A probabilistic programming language for bayesian inference and optimization," *J. Educ. Behav. Stat.*, vol. 40, no. 5, pp. 530–543, 2015.

[67] J. Lu, J. Li, S. Liu, T. Wang, A. Ianni, E. Bober, T. Braun, R. Xiang, and S. Yue, "Exosomal tetraspanins mediate cancer metastasis by altering host microenvironment," *Oncotarget*, vol. 8, pp. 62803–62815, Sept. 2017.

[68] K. A. Papadakis, J. Krempski, J. Reiter, P. Svingen, Y. Xiong, O. F. Sarmento, A. Huseby, A. J. Johnson, G. A. Lomberk, R. A. Urrutia, and W. A. Faubion, "Krüppel-like factor KLF10 regulates transforming growth factor receptor II expression and TGF-β signaling in CD8+ T lymphocytes," *Am. J. Physiol. Cell Physiol.*, vol. 308, pp. C362–71, Mar. 2015.

[69] F. S. Hodi, S. J. O'Day, D. F. McDermott, R. W. Weber, J. A. Sosman, J. B. Haanen, R. Gonzalez, C. Robert, D. Schadendorf, J. C. Hassel, W. Akerley, A. J. M. van den Eertwegh, J. Lutzky, P. Lorigan, J. M. Vaubel, G. P. Linette, D. Hogg, C. H. Ottensmeier, C. Lebbé, C. Peschel, I. Quirt, J. I. Clark, J. D. Wolchok, J. S. Weber, J. Tian, M. J. Yellin, G. M. Nichol, A. Hoos, and W. J. Urba, "Improved survival with ipilimumab in patients with metastatic melanoma," *N. Engl. J. Med.*, vol. 363, pp. 711–723, Aug. 2010.

[70] D. Schadendorf, G. V. Long, D. Stroiakovski, B. Karaszewska, A. Hauschild, E. Levchenko, V. Chiarion-Sileni, J. Schachter, C. Garbe, C. Dutriaux, H. Gogas, M. Mandalà, J. B. A. G. Haanen, C. Lebbé, A. Mackiewicz, P. Rutkowski, J.-J. Grob, P. Nathan, A. Ribas, M. A. Davies, Y. Zhang, M. Kaper, B. Mookerjee, J. J. Legos, K. T. Flaherty, and C. Robert, "Three-year pooled analysis of factors associated with clinical outcomes across dabrafenib and trametinib combination therapy phase 3 randomised trials," *Eur. J. Cancer*, vol. 82, pp. 45–55, Sept. 2017.

[71] J. R. Brahmer, S. S. Tykodi, L. Q. M. Chow, W.-J. Hwu, S. L. Topalian, P. Hwu, C. G. Drake, L. H. Camacho, J. Kauh, K. Odunsi, H. C. Pitot, O. Hamid, S. Bhatia, R. Martins, K. Eaton, S. Chen, T. M. Salay, S. Alaparthy, J. F.

Grosso, A. J. Korman, S. M. Parker, S. Agrawal, S. M. Goldberg, D. M. Pardoll, A. Gupta, and J. M. Wigginton, "Safety and activity of Anti–PD-L1 antibody in patients with advanced cancer," *N. Engl. J. Med.*, vol. 366, pp. 2455–2465, June 2012.

[72] S. L. Topalian, F. S. Hodi, J. R. Brahmer, S. N. Gettinger, D. C. Smith, D. F. McDermott, J. D. Powderly, R. D. Carvajal, J. A. Sosman, M. B. Atkins, P. D. Leming, D. R. Spigel, S. J. Antonia, L. Horn, C. G. Drake, D. M. Pardoll, L. Chen, W. H. Sharfman, R. A. Anders, J. M. Taube, T. L. McMiller, H. Xu, A. J. Korman, M. Jure-Kunkel, S. Agrawal, D. McDonald, G. D. Kollia, A. Gupta, J. M. Wigginton, and M. Sznol, "Safety, activity, and immune correlates of Anti–PD-1 antibody in cancer," *N. Engl. J. Med.*, vol. 366, pp. 2443–2454, June 2012.

[73] S. L. Topalian, M. Sznol, D. F. McDermott, H. M. Kluger, R. D. Carvajal, W. H. Sharfman, J. R. Brahmer, D. P. Lawrence, M. B. Atkins, J. D. Powderly, P. D. Leming, E. J. Lipson, I. Puzanov, D. C. Smith, J. M. Taube, J. M. Wigginton, G. D. Kollia, A. Gupta, D. M. Pardoll, J. A. Sosman, and F. S. Hodi, "Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab," *J. Clin. Oncol.*, vol. 32, pp. 1020–1030, Apr. 2014.

[74] M. A. Postow, J. Chesney, A. C. Pavlick, C. Robert, K. Grossmann, D. McDermott, G. P. Linette, N. Meyer, J. K. Giguere, S. S. Agarwala, M. Shaheen, M. S. Ernstoff, D. Minor, A. K. Salama, M. Taylor, P. A. Ott, L. M. Rollin, C. Horak, P. Gagnier, J. D. Wolchok, and F. S. Hodi, "Nivolumab and ipilimumab versus ipilimumab in untreated melanoma," *N. Engl. J. Med.*, vol. 372, pp. 2006–2017, May 2015.

[75] T. I. Lee and R. A. Young, "Transcriptional regulation and its misregulation in disease," *Cell*, vol. 152, pp. 1237–1251, Mar. 2013.

[76] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-Resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, pp. 823–837, May 2007.

[77] G. Wang, R. D. Chow, L. Zhu, Z. Bai, L. Ye, F. Zhang, P. A. Renauer, M. B. Dong, X. Dai, X. Zhang, Y. Du, Y. Cheng, L. Niu, Z. Chu, K. Kim, C. Liao, P. Clark, Y. Errami, and S. Chen, "CRISPR-GEMM pooled mutagenic screening identifies KMT2D as a major modulator of immune checkpoint blockade," *Cancer Discov.*, vol. 10, pp. 1912–1933, Dec. 2020.

[78] D. Peng, I. Kryczek, N. Nagarsheth, L. Zhao, S. Wei, W. Wang, Y. Sun, E. Zhao, L. Vatan, W. Szeliga, J. Kotarski, R. Tarkowski, Y. Dou, K. Cho, S. Hensley-Alford, A. Munkarah, R. Liu, and W. Zou, "Epigenetic silencing of TH1-type chemokines shapes tumour immunity and immunotherapy," *Nature*, vol. 527, pp. 249–253, Nov. 2015.

[79] D. M. Woods, A. L. Sodré, A. Villagra, A. Sarnaik, E. M. Sotomayor, and J. Weber, "HDAC inhibition upregulates PD-1 ligands in melanoma and augments immunotherapy with PD-1 blockade," *Cancer Immunol Res*, vol. 3, pp. 1375–1385, Dec. 2015.

[80] S. Mulero-Navarro and M. Esteller, "Epigenetic biomarkers for human cancer: the time is now," *Crit. Rev. Oncol. Hematol.*, vol. 68, pp. 1–11, Oct. 2008.

[81] M. Tang, "pyflow-ChIPseq: a snakemake based ChIP-seq pipeline," June 2017.

[82] J. Köster and S. Rahmann, "Snakemake—a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, pp. 2520–2522, Aug. 2012.

[83] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, pp. 1–10, Mar. 2009.

[84] F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, and T. Manke, "deeptools2: a next generation web server for deep-sequencing data analysis," *Nucleic Acids Res.*, vol. 44, pp. W160–5, July 2016.

[85] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nat. Biotechnol.*, vol. 29, pp. 24–26, Jan. 2011.

[86] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol.*, vol. 9, p. R137, Sept. 2008.

[87] Z. Gu, R. Eils, and M. Schlesner, "Complex heatmaps reveal patterns and correlations in multidimensional genomic data," *Bioinformatics*, vol. 32, pp. 2847–2849, Sept. 2016.

[88] G. Yu, L.-G. Wang, and Q.-Y. He, "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization," *Bioinformatics*, vol. 31, pp. 2382–2383, Mar. 2015.

[89] J. Lovén, H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. A. Young, "Selective inhibition of tumor oncogenes by disruption of super-enhancers," *Cell*, vol. 153, pp. 320–334, Apr. 2013.

[90] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Oct. 2012.

[91] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, pp. 1–21, Dec. 2014.

[92] D. S. Chandrashekar, B. Bashel, S. A. H. Balasubramanya, C. J. Creighton, I. Ponce-Rodriguez, B. V. S. K. Chakravarthi, and S. Varambally, "UALCAN: A portal for facilitating tumor subgroup gene expression and survival analyses," *Neoplasia*, vol. 19, pp. 649–658, Aug. 2017.

[93] M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang, "HiChIP: efficient and sensitive analysis of protein-directed genome architecture," *Nat. Methods*, vol. 13, pp. 919–922, Nov. 2016.

[94] N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C.-J. Chen, J.-P. Vert, E. Heard, J. Dekker, and E. Barillot, "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing," *Genome Biol.*, vol. 16, pp. 1–11, Dec. 2015.

[95] C. A. Lareau and M. J. Aryee, "hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data," *Nat. Methods*, vol. 15, pp. 155–156, Feb. 2018.

[96] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an R package for comparing biological themes among gene clusters," *OMICS*, vol. 16, pp. 284–287, May 2012.

[97] M. Garber, N. Yosef, A. Goren, R. Raychowdhury, A. Thielke, M. Guttman, J. Robinson, B. Minie, N. Chevrier, Z. Itzhaki, R. Blecher-Gonen, C. Bornstein, D. Amann-Zalcenstein, A. Weiner, D. Friedrich, J. Meldrim, O. Ram, C. Cheng, A. Gnirke, S. Fisher, N. Friedman, B. Wong, B. E. Bernstein, C. Nusbaum, N. Hacohen, A. Regev, and I. Amit, "A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals," *Mol. Cell*, vol. 47, pp. 810–822, Sept. 2012.

[98] K. Rai, K. C. Akdemir, L. N. Kwong, P. Fiziev, C.-J. Wu, E. Z. Keung, S. Sharma, N. S. Samant, M. Williams, J. B. Axelrad, A. Shah, D. Yang, E. A. Grimm, M. C. Barton, D. R. Milton, T. P. Heffernan, J. W. Horner, S. Ekmekcioglu, A. J. Lazar, J. Ernst, and L. Chin, "Dual roles of RNF2 in melanoma progression," *Cancer Discov.*, vol. 5, pp. 1314–1327, Dec. 2015.

[99] Q. Cao, C. Anyansi, X. Hu, L. Xu, L. Xiong, W. Tang, M. T. S. Mok, C. Cheng, X. Fan, M. Gerstein, A. S. L. Cheng, and K. Y. Yip, "Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines," *Nat. Genet.*, vol. 49, pp. 1428–1436, Oct. 2017.

[100] N. Joller and V. K. Kuchroo, "Tim-3, lag-3, and TIGIT," *Curr. Top. Microbiol. Immunol.*, vol. 410, pp. 127–156, 2017.

[101] N. Watanabe, M. Gavrieli, J. R. Sedy, J. Yang, F. Fallarino, S. K. Loftin, M. A. Hurchla, N. Zimmerman, J. Sim, X. Zang, T. L. Murphy, J. H. Russell, J. P. Allison, and K. M. Murphy, "BTLA is a lymphocyte inhibitory receptor with similarities to CTLA-4 and PD-1," *Nat. Immunol.*, vol. 4, pp. 670–679, June 2003.

[102] Y.-H. Huang, C. Zhu, Y. Kondo, A. C. Anderson, A. Gandhi, A. Russell, S. K. Dougan, B.-S. Petersen, E. Melum, T. Pertel, K. L. Clayton, M. Raab, Q. Chen, N. Beauchemin, P. J. Yazaki, M. Pyzik, M. A. Ostrowski, J. N. Glickman, C. E. Rudd, H. L. Ploegh, A. Franke, G. A. Petsko, V. K. Kuchroo, and R. S. Blumberg, "CEACAM1 regulates TIM-3-mediated tolerance and exhaustion," *Nature*, vol. 517, pp. 386–390, Jan. 2015.

[103] E. J. Wherry and M. Kurachi, "Molecular and cellular insights into T cell exhaustion," *Nat. Rev. Immunol.*, vol. 15, pp. 486–499, Aug. 2015.

[104] D. Dankort, D. P. Curley, R. A. Cartlidge, B. Nelson, A. N. Karnezis, W. E. Damsky, Jr, M. J. You, R. A. DePinho, M. McMahon, and M. Bosenberg, "Braf(V600E) cooperates with pten loss to induce metastatic melanoma," *Nat. Genet.*, vol. 41, pp. 544–552, May 2009.

[105] E. M. Van Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. G. Foppen, S. M. Goldinger, J. Utikal, J. C. Hassel, B. Weide, K. C. Kaehler, C. Loquai, P. Mohr, R. Gutzmer, R. Dummer, S. Gabriel, C. J. Wu, D. Schadendorf, and L. A. Garraway, "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma," *Science*, vol. 350, pp. 207–211, Oct. 2015.

[106] Z. Shao, Y. Zhang, G.-C. Yuan, S. H. Orkin, and D. J. Waxman, "MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets," *Genome Biol.*, vol. 13, p. R16, Mar. 2012.

[107] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, pp. 882–883, Mar. 2012.

[108] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, pp. 93–99, Jan. 2004.

[109] D. Liu, B. Schilling, D. Liu, A. Sucker, E. Livingstone, L. Jerby-Arnon, L. Zimmer, R. Gutzmer, I. Satzger, C. Loquai, S. Grabbe, N. Vokes, C. A. Margolis, J. Conway, M. X. He, H. Elmarakeby, F. Dietlein, D. Miao, A. Tracy, H. Gogas, S. M. Goldinger, J. Utikal, C. U. Blank, R. Rauschenberg, D. von Bubnoff, A. Krackhardt, B. Weide, S. Haferkamp, F. Kiecker, B. Izar, L. Garraway, A. Regev, K. Flaherty, A. Paschen, E. M. Van Allen, and D. Schadendorf, "Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma," *Nat. Med.*, vol. 25, pp. 1916–1927, Dec. 2019.

[110] W. Peng, J. Q. Chen, C. Liu, S. Malu, C. Creasy, M. T. Tetzlaff, C. Xu, J. A. McKenzie, C. Zhang, X. Liang, L. J. Williams, W. Deng, G. Chen, R. Mbofung, A. J. Lazar, C. A. Torres-Cabala, Z. A. Cooper, P.-L. Chen, T. N. Tieu,

S. Spranger, X. Yu, C. Bernatchez, M.-A. Forget, C. Haymaker, R. Amaria, J. L. McQuade, I. C. Glitza, T. Cascone, H. S. Li, L. N. Kwong, T. P. Heffernan, J. Hu, R. L. Bassett, Jr, M. W. Bosenberg, S. E. Woodman, W. W. Overwijk, G. Lizée, J. Roszik, T. F. Gajewski, J. A. Wargo, J. E. Gershenwald, L. Radvanyi, M. A. Davies, and P. Hwu, "Loss of PTEN promotes resistance to T Cell-Mediated immunotherapy," *Cancer Discov.*, vol. 6, pp. 202–216, Feb. 2016.

[111] J. Gao, L. Z. Shi, H. Zhao, J. Chen, L. Xiong, Q. He, T. Chen, J. Roszik, C. Bernatchez, S. E. Woodman, P.-L. Chen, P. Hwu, J. P. Allison, A. Futreal, J. A. Wargo, and P. Sharma, "Loss of IFN-γ pathway genes in tumor cells as a mechanism of resistance to Anti-CTLA-4 therapy," *Cell*, vol. 167, pp. 397–404.e9, Oct. 2016.

[112] S. Z. Josefowicz, L.-F. Lu, and A. Y. Rudensky, "Regulatory T cells: mechanisms of differentiation and function," *Annu. Rev. Immunol.*, vol. 30, pp. 531–564, Jan. 2012.

[113] S. Z. Josefowicz and A. Rudensky, "Control of regulatory T cell lineage commitment and maintenance," *Immunity*, vol. 30, pp. 616–625, May 2009.

[114] L. Klein, E. A. Robey, and C.-S. Hsieh, "Central CD4+ T cell tolerance: deletion versus regulatory T cell differentiation," *Nat. Rev. Immunol.*, vol. 19, pp. 7–18, Jan. 2019.

[115] W. Chen and J. E. Konkel, "Development of thymic foxp3(+) regulatory T cells: TGF-β matters," *Eur. J. Immunol.*, vol. 45, pp. 958–965, Apr. 2015.

[116] D. Malhotra and M. K. Jenkins, "Regulatory T cells: A crisis averted," *Immunity*, vol. 44, pp. 1079–1081, May 2016.

[117] A. M. Bilate and J. J. Lafaille, "Induced CD4+Foxp3+ regulatory T cells in immune tolerance," *Annu. Rev. Immunol.*, vol. 30, pp. 733–758, Jan. 2012.

[118] G. Plitas and A. Y. Rudensky, "Regulatory T cells: Differentiation and function," *Cancer Immunol Res*, vol. 4, pp. 721–725, Sept. 2016.

[119] E. M. Shevach and A. M. Thornton, "ttregs, ptregs, and itregs: similarities and differences," *Immunol. Rev.*, vol. 259, pp. 88–102, May 2014.

[120] H.-M. Lee, J. L. Bautista, J. Scott-Browne, J. F. Mohan, and C.-S. Hsieh, "A broad range of self-reactivity drives thymic regulatory T cell selection to limit responses to self," *Immunity*, vol. 37, pp. 475–486, Sept. 2012.

[121] G. L. Stritesky, S. C. Jameson, and K. A. Hogquist, "Selection of self-reactive T cells in the thymus," *Annu. Rev. Immunol.*, vol. 30, pp. 95–114, 2012.

[122] S. Malchow, D. S. Leventhal, V. Lee, S. Nishi, N. D. Socci, and P. A. Savage, "Aire enforces immune tolerance by directing autoreactive T cells into the regulatory T cell lineage," *Immunity*, vol. 44, pp. 1102–1113, May 2016.

[123] J. S. A. Perry, C.-W. J. Lio, A. L. Kau, K. Nutsch, Z. Yang, J. I. Gordon, K. M. Murphy, and C.-S. Hsieh, "Distinct contributions of aire and antigen-presenting-cell subsets to the generation of self-tolerance in the thymus," *Immunity*, vol. 41, pp. 414–426, Sept. 2014.

[124] L. Klein, B. Kyewski, P. M. Allen, and K. A. Hogquist, "Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see)," *Nat. Rev. Immunol.*, vol. 14, pp. 377–391, June 2014.

[125] A. L. Bayer, A. Yu, D. Adeegbe, and T. R. Malek, "Essential role for interleukin-2 for CD4(+)CD25(+) T regulatory cell development during the neonatal period," *J. Exp. Med.*, vol. 201, pp. 769–777, Mar. 2005.

[126] W. Ouyang, O. Beckett, Q. Ma, and M. O. Li, "Transforming growth factor-beta signaling curbs thymic negative selection promoting regulatory T cell development," *Immunity*, vol. 32, pp. 642–653, May 2010.

[127] Y. Liu, P. Zhang, J. Li, A. B. Kulkarni, S. Perruche, and W. Chen, "A critical function for TGF-beta signaling in the development of natural CD4+CD25+Foxp3+ regulatory T cells," *Nat. Immunol.*, vol. 9, pp. 632–640, June 2008.

[128] B. Salomon, D. J. Lenschow, L. Rhee, N. Ashourian, B. Singh, A. Sharpe, and J. A. Bluestone, "B7/CD28 costimulation is essential for the homeostasis of the CD4+CD25+ immunoregulatory T cells that control autoimmune diabetes," *Immunity*, vol. 12, pp. 431–440, Apr. 2000.

[129] C.-S. Hsieh, Y. Zheng, Y. Liang, J. D. Fontenot, and A. Y. Rudensky, "An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires," *Nat. Immunol.*, vol. 7, pp. 401–410, Apr. 2006.

[130] R. Pacholczyk, H. Ignatowicz, P. Kraj, and L. Ignatowicz, "Origin and T cell receptor diversity of Foxp3+CD4+CD25+ T cells," *Immunity*, vol. 25, pp. 249–259, Aug. 2006.

[131] A.-S. Bergot, W. Chaara, E. Ruggiero, E. Mariotti-Ferrandiz, S. Dulauroy, M. Schmidt, C. von Kalle, A. Six, and D. Klatzmann, "TCR sequences and tissue distribution discriminate the subsets of naïve and activated/memory treg cells in mice," *Eur. J. Immunol.*, vol. 45, pp. 1524–1534, May 2015.

[132] L. M. Relland, J. B. Williams, G. N. Relland, D. Haribhai, J. Ziegelbauer, M. Yassai, J. Gorski, and C. B. Williams, "The TCR repertoires of regulatory and conventional T cells specific for the same foreign antigen are distinct," *J. Immunol.*, vol. 189, pp. 3566–3574, Oct. 2012.

[133] K. J. Wolf, R. O. Emerson, J. Pingel, R. Mark Buller, and R. J. DiPaolo, "Conventional and regulatory CD4+ T cells that share identical TCRs are derived from common clones," *PLoS One*, vol. 11, p. e0153705, Apr. 2016.

[134] M. Shugay, O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov, "Towards error-free profiling of immune repertoires," *Nat. Methods*, vol. 11, pp. 653–655, May 2014.

[135] J. Lu, F. Van Laethem, A. Bhattacharya, M. Craveiro, I. Saba, J. Chu, N. C. Love, A. Tikhonova, S. Radaev, X. Sun, A. Ko, T. Arnon, E. Shifrut, N. Friedman, N.-P. Weng, A. Singer, and P. D. Sun, "Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire," *Nat. Commun.*, vol. 10, p. 1019, Mar. 2019.

[136] W. Yu, H. Nagaoka, Z. Misulovin, E. Meffre, H. Suh, M. Jankovic, N. Yannoutsos, R. Casellas, E. Besmer, F. Papavasiliou, X. Qin, and M. C. Nussenzweig, "RAG expression in B cells in secondary lymphoid tissues," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 64, pp. 207–210, 1999.

[137] Y. Y. Wan and R. A. Flavell, "Identifying foxp3-expressing suppressor T cells with a bicistronic reporter," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, pp. 5126–5131, Apr. 2005.

[138] Y. Shinkai, S. Koyasu, K. Nakayama, K. M. Murphy, D. Y. Loh, E. L. Reinherz, and F. W. Alt, "Restoration of T cell development in RAG-2-deficient mice by functional TCR transgenes," *Science*, vol. 259, pp. 822–825, Feb. 1993.

[139] K. S. Hathcock, S. Bowen, F. Livak, and R. J. Hodes, "ATM influences the efficiency of TCRβ rearrangement, subsequent TCRβ-Dependent T cell development, and generation of the Pre-Selection TCRβ CDR3 repertoire," *PLoS One*, vol. 8, p. e62188, Apr. 2013.

[140] T. Korn, J. Reddy, W. Gao, E. Bettelli, A. Awasthi, T. R. Petersen, B. T. Bäckström, R. A. Sobel, K. W. Wucherpfennig, T. B. Strom, M. Oukka, and V. K. Kuchroo, "Myelin-specific regulatory T cells accumulate in the CNS but fail to control autoimmune inflammation," *Nat. Med.*, vol. 13, pp. 423–431, Apr. 2007.

[141] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[142] B. Zhang, Q. Jia, C. Bock, G. Chen, H. Yu, Q. Ni, Y. Wan, Q. Li, and Y. Zhuang, "Glimpse of natural selection of long-lived t-cell clones in healthy life," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, pp. 9858–9863, Aug. 2016.

[143] J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, and M. M. Davis, "Identifying specificity groups in the T cell receptor repertoire," *Nature*, vol. 547, pp. 94–98, July 2017.

[144] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins, "Next-Generation machine learning for biological networks," *Cell*, vol. 173, pp. 1581–1592, June 2018.

[145] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nat. Genet.*, vol. 51, pp. 12–18, Jan. 2019.

[146] R. C. Wirasinha, M. Singh, S. K. Archer, A. Chan, P. F. Harrison, C. C. Goodnow, and S. R. Daley, "$\alpha\beta$ t-cell receptors with a central CDR3 cysteine are enriched in CD8$\alpha\alpha$ intraepithelial lymphocytes and their thymic precursors," *Immunol. Cell Biol.*, vol. 96, pp. 553–561, July 2018.

[147] B. D. Stadinski, K. Shekhar, I. Gómez-Touriño, J. Jung, K. Sasaki, A. K. Sewell, M. Peakman, A. K. Chakraborty, and E. S. Huseby, "Hydrophobic CDR3 residues promote the development of self-reactive T cells," *Nat. Immunol.*, vol. 17, pp. 946–955, Aug. 2016.

[148] E. Reátegui, K. E. van der Vos, C. P. Lai, M. Zeinali, N. A. Atai, B. Aldikacti, F. P. Floyd, Jr, A. H Khankhel, V. Thapar, F. H. Hochberg, L. V. Sequist, B. V. Nahed, B. S Carter, M. Toner, L. Balaj, D. T Ting, X. O. Breakefield, and S. L. Stott, "Engineered nanointerfaces for microfluidic isolation and molecular profiling of tumor-specific extracellular vesicles," *Nat. Commun.*, vol. 9, p. 175, Jan. 2018.

[149] P. Zhang, X. Zhou, and Y. Zeng, "Multiplexed immunophenotyping of circulating exosomes on nano-engineered ExoProfile chip towards early diagnosis of cancer," *Chem. Sci.*, vol. 10, pp. 5495–5504, June 2019.

[150] H. Ishwaran and J. Sunil Rao, "Spike and slab variable selection: Frequentist and bayesian strategies," *aos*, vol. 33, pp. 730–773, Apr. 2005.

[151] B. Howie, A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, and H. S. Robins, "High-throughput pairing of T cell receptor α and β sequences," *Sci. Transl. Med.*, vol. 7, p. 301ra131, Aug. 2015.