Utilizing a novel machine learning pipeline for single-cell transcriptomic characterization of a remodeled tumor microenvironment

Alan Chang

Section 1: My Personal Journey

It was just another car ride home with my brother; I was the curious freshman asking difficult questions to the knowledgeable senior. Topic of the night: viruses. It seemed almost unfair, how viruses could inject their DNA into target cells and exploit them as host cells. I inquired further, “But Kevin, what if scientists could actually reprogram these viruses to artificially alter genomes?” He paused. “Hm… never thought of that.” After I got home, I eagerly searched “genome editing viruses,” and a phrase kept popping up: “CRISPR-Cas9.” I found exactly what I was looking for: researchers were transfecting reprogrammed bacterial plasmids into target cells to selectively mutate genomes. Moreover, I was particularly interested in cancer CRISPR screening - a contemporary method for identifying tumorigenesis drivers in the tumor microenvironment. After reading more literature and discussing my interest with teachers, I realized the massive potential this form of computational analysis holds in the field of systems biology. I was determined to teach myself R language and parallel it with my enthusiasm for the CRISPR-Cas9 system to hopefully aid in the development of improved cancer immunotherapy.

After being captivated by the CRISPR-Cas9 system, in the following months, I was accepted into our high school’s three-year Authentic Science Research Program and extensively studied various textbooks and read over 20 journal articles regarding its various applications my first year. When searching for potential mentors to collaborate with, Dr. Sidi Chen at Yale

University immediately caught my attention. Working in the field of systems biology, Dr. Chen was also interested in utilizing the CRISPR-Cas9 system to study genetic tumorigenesis drivers.

Since spring of 2017, Dr. Chen and I have exchanged weekly emails, in which he answers my questions and challenges me to find gaps in the literature. In two years, I not only became fluent in CRISPR, immunology, and cancer, but also researched the biological applications of unsupervised machine learning computational analyses. In the winter of 2018, Dr. Chen challenged me to propose my own project plan. After months of research, I presented a novel detailed plan of CRISPR T-cell screening. Dr. Chen was impressed that I had surprisingly suggested a project underway in the lab. However, Dr. Chen informed me that with my time constraints, it was too ambitious. Nevertheless, my proposed ideas were promising, and Dr. Chen ultimately invited me to apply my knowledge to help co-author a manuscript for an ongoing project that summer involving another tumorigenesis driver, *Prkar1a*.

In order to prepare myself prior to arriving at the lab in June, I read countless journal articles on CRISPR, the tumor microenvironment, and systems biology, specifically regarding analyzing single-cell data. I also read two textbooks, *The Biology of Cancer,* by Robert Weinberg, and *Janeway's Immunobiology*, by Kenneth M. Murphy. In addition, I taught myself the programming language R by reading *R Cookbook,* by Paul Teetor, as well as by following various online courses (e.g. DataCamp) and watching diverse R videos (e.g. Coursera).

After arriving at the lab, I continued learning. With the goal of each objective decided when initiating the purpose of my research, I led the design of how each objective would be executed. I searched through the literature, looking for strategies that had been used in similar computational papers. After hours upon hours of researching and discussing with Mr. Renauer, a

graduate student in the lab, I decided that I not only wanted to use well-established unsupervised dimensionality reduction tools, such as PCA and tSNE, but also test the efficacy of contemporary tools, namely PHATE from 2017. Furthermore, I decided to employ the dimensionality reduction technique, NMF, to determine the optimal number of clusters for shared nearest neighbor (SNN) clustering. To use and understand these computational tools, I had to obtain a foundational knowledge in certain linear algebra topics, including dimensionality reduction, clustering, and more. I learned these advanced mathematical topics by discussing with graduate students in my lab and watching introductory YouTube videos. I also tested a recently developed tool, scmap, in addition to mapping the expression of canonical gene markers as well as performing differential gene expression to create a three-fold pipeline for identifying the immune cell clusters. Lastly, I used pathway analysis tools Ingenuity Pathway Analysis (IPA) and Database for Annotation, Visualization and Integrated Discovery (DAVID) to characterize cell activity and make final conclusions about the cell populations within the tumor microenvironment.

Throughout the course of the summer of 2018, I spent nine weeks Mon-Fri at the Sidi Chen Laboratory at Yale University, totaling more than 400 hours in the lab. I also worked extensively on my research project outside of the lab, troubleshooting my code as well as exploring different methods of analysis after returning home and on the weekends. I wrote well over 1,000 lines of R code to implement my procedure design. To teach myself additional R packages and troubleshoot my code, I researched countless online sources, such as vignettes created by the package authors and similar published studies. Oftentimes, these packages introduced a new type of object or syntax, so I had to be continuously learning new techniques.

Completing my research has not only furthered my pre-existing STEM interests, but more importantly, unveiled new ones. I originally became intrigued by this field of research due to its biological elegance; the fact that humans have managed to harness a bacterial-derived "immune system" as a precise gene editing tool fascinated me. I was eager to learn more about CRISPR's countless applications, especially CRISPR screening. After finishing this project, I am even more curious to explore the field of CRISPR editing and immunology. During my research experience, rather than sitting quietly at lunch and listening to lab members discuss recent biological breakthroughs, I found myself jumping into the conversation to discuss the information I had learned that day. Furthermore, beyond the cancer gene editing field, this research project has aroused my interests in computer science. Having only taken an introductory course freshman year, I never realized how much I loved programming until I completed my research. Needing to learn so much about complex coding in both R and Linux within nine weeks was a challenge I was excited to tackle every morning, never something I viewed as a burden. I found myself getting lost in the world of code, meticulously troubleshooting my code and optimizing my code structure. Altogether, completing my research has not only deepened my earlier STEM interests, but also introduced new ones to explore in the future.

For prospective high school students interested in undertaking a project combining science and mathematics, the most important piece of advice for them would be to remain open-minded. Especially among these interdisciplinary crossroads, the field evolves incredibly quickly, and it is important to maintain a resilient, positive mindset. I had to adapt to countless changes and challenges during my research experience, but looking back, all these obstacles made the experience that much more memorable.

Section 2: Discussing my Research

Cancer death tolls are expected to continue increasing to 13 million in 2030. Despite recent advancements in cancer research, cancer cells utilize countless genetic perturbations to resist current methods of immunotherapy. Understanding the tumor mechanisms of immune escape is imperative for designing improved immunotherapies. This study lays important groundwork for elucidating the functional roles of tumorigenesis drivers. By employing diverse machine learning approaches with a high-resolution 10X Genomics Chromium scRNA-seq dataset, this study establishes a novel pipeline to separate, identify, and characterize the remodeled cell populations within the tumor microenvironment after *Prkar1a* knockout. With this methodology, researchers can evaluate the effects of countless protumor genetic perturbations that have remained unexplored for decades. Top differentially expressed genes identified in both the tumor and immune subpopulations not only characterized the cell populations, but also reinforced current literature. In addition, this study is the first of its kind to holistically validate the efficacy of two cutting-edge tools, PHATE and scmap, via well-established methods tSNE and canonical marker expression, respectively.

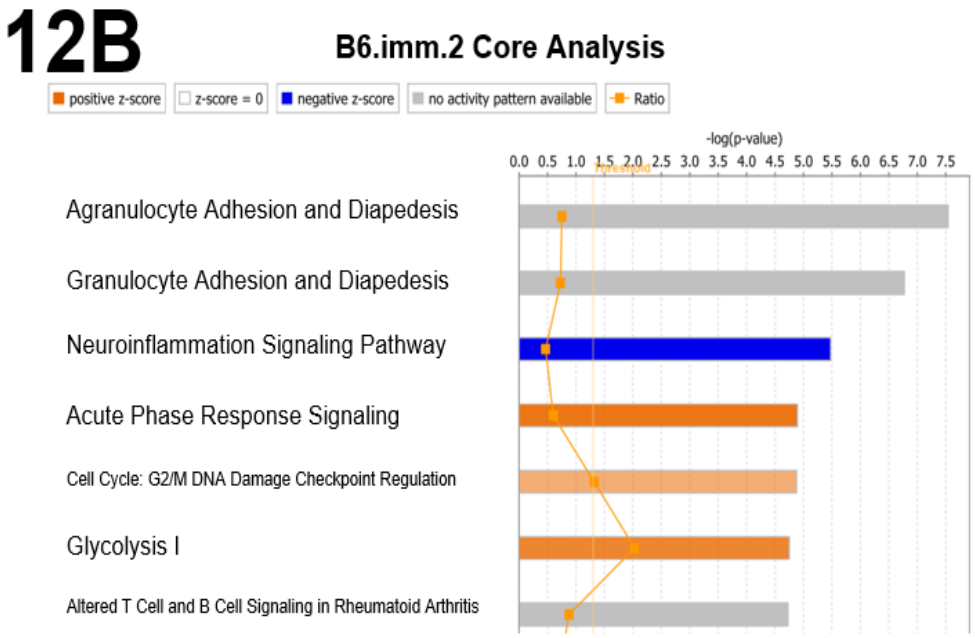
In this study, analysis of the scRNA-seq data from *Prkar1a*-mutant tumors presented an immune population predominantly composed of macrophages with a distinct antitumor M1 macrophage cluster, a protumor M2 macrophage cluster, and an ambiguous third cluster of rapidly proliferating macrophages. The immunosuppressive effects of *Prkar1a* mutation were previously unexplored, but my findings suggest convincing mechanisms by which this tumorigenesis driver facilitates both innate and adaptive immune escape. As shown by the large protumor M2 macrophage population, these macrophages are likely suppressing the

inflammatory response to promote tumor growth. Furthermore, as shown by the minuscule T-cell population and the multiple cancer cell proliferation signatures, the *Prkar1a*-mutant tumor likely grew fast enough to evade a substantial adaptive immune response. Overall, I support my alternate hypothesis that *Prkar1a* does indeed induce a myriad of immunosuppressive mechanisms that facilitate tumor growth.

Key Graphs from my Paper

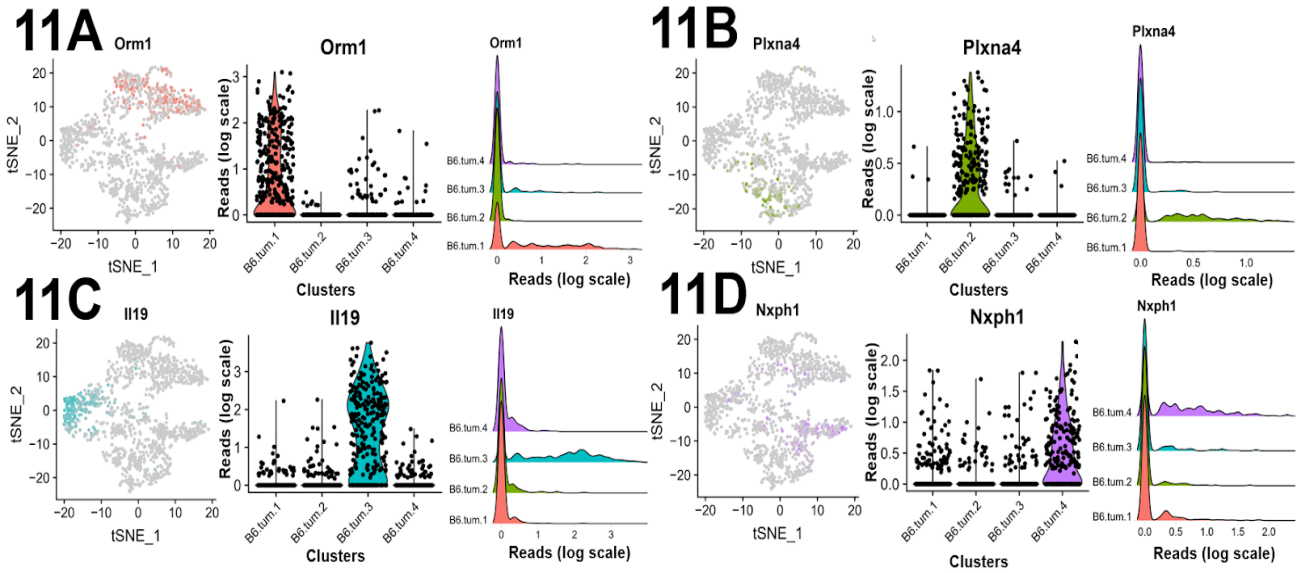
Protumor characteristics in immune subpopulation post-*Prkar1a* mutation.

As shown by Figure 12B, a cluster of immune cells (B6.imm.2) proved to be protumor, anti-inflammatory M2 macrophages, as the neuroinflammation signaling pathway was significantly downregulated (p=3.31e-6). DAVID analysis further supports the M2-macrophage phenotype, as there is a clear downregulation of inflammatory response pathways (p=1.81e-6).



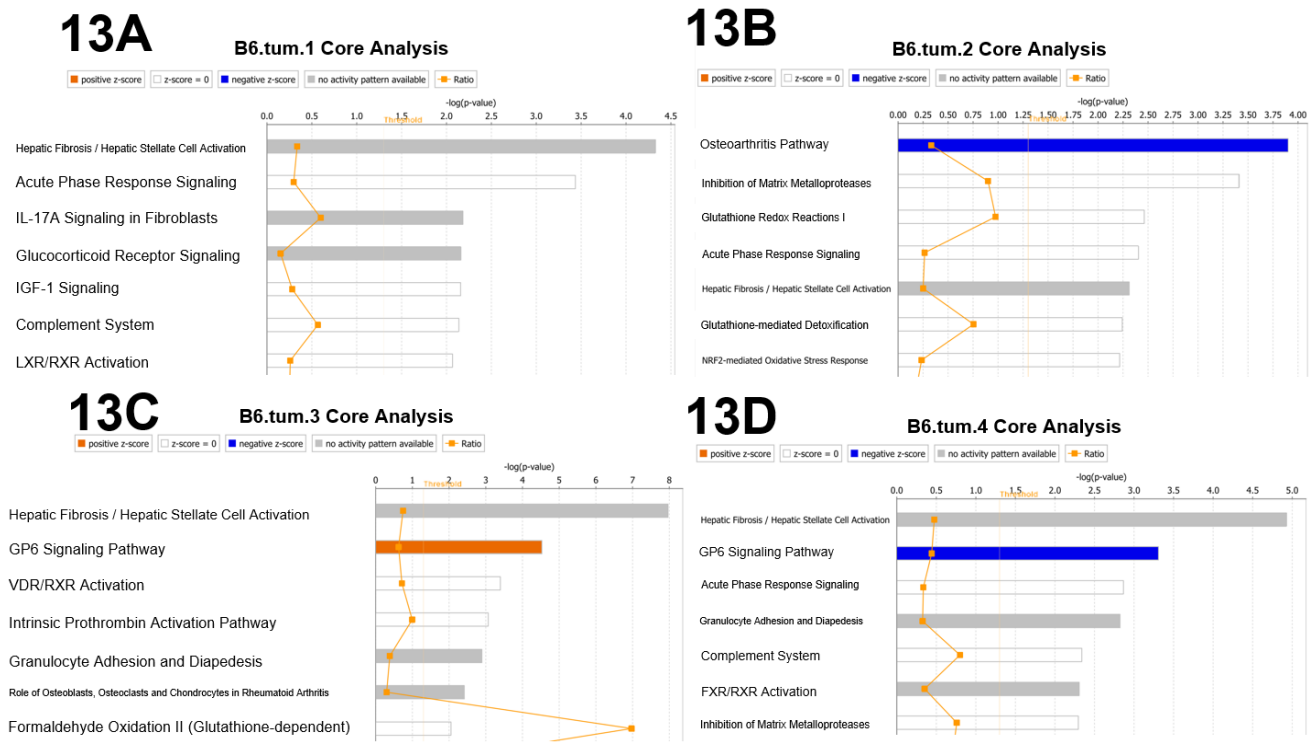
Protumor characteristics in tumor subpopulation post-*Prkar1a* mutation.

Differential gene expression analysis (Figure 11). In cancer cell cluster B6.tum.1, there was high expression of *Orm1* (Figure 11A), which is an acute-phase reactant that may be involved in aspects of immunosuppression (Fan, Stendahl, Stjernberg & Beckman, 1995). Furthermore, high expression of *Plxna4* in B6.tum.2 (Figure 11B) has been observed to enhance VEGF and bFGF signaling to promote tumor progression and angiogenesis (Kigel, Rabinowicz, Varshavsky, Kessler & Neufeld, 2011). Similarly, upregulated *Il19* gene expression in cluster B6.tum.3 (Figure 11C) promotes tumor progression by promoting cell proliferation (Hsing et al., 2012), and *Nxph1* expression in cluster B6.tum.4 (Figure 11D) suppresses the proliferation of hematopoietic progenitor cells (Kinzfogl, Hangoc & Broxmeyer, 2011).



IPA pathway analysis (Figure 13). In cluster B6.tum.1 (Figure 13A), IGF-1 signaling is significantly expressed (p=6.91e-3). IGF-1 signaling has been shown to not only play a critical role in the development and progression of various tumors, but also aid in cancer drug resistance (Denduluri et al., 2014). Furthermore, in cluster B6.tum.2 (Figure 13B), the osteoarthritis

pathway is shown to be heavily downregulated (p=1.26e-4, z =-0.45). Apoptosis occurs in osteoarthritic cartilage and may induce cartilage degeneration (Hwang & Kim, 2015). These results suggest that the tumor cluster may be downregulating similar apoptotic pathways to promote tumor progression.



All in all, various protumor mechanisms post-*Prkar1a* mutation were identified, and utilizing the novel pipeline described in this study allows for further investigation of important tumorigenesis drivers that have remained unexplored for decades.

Other interesting findings in this study regarding differential cell cluster expression lay the groundwork for future research regarding the effects of *Prkar1a* mutation within the TME. For instance, one of the immune clusters did not polarize towards an M1 or M2 phenotype; future research should focus on the potential role this cluster of radpily dividing macrophages

plays in the tumor microenvironment. Furthermore, unexplored top differentially expressed genes identified via this pipeline may be cell type-specific markers that can be used to identify cell populations. Future research should also focus on clarifying the polarization of immune cell pathway activity and elucidating the underlying reasons behind the polarization of the cancer cell clusters.

In summary, coupling high-resolution single-cell RNA sequencing with innovative unsupervised machine learning approaches has introduced powerful means of analyzing cell populations in the field of systems biology. Beyond examining the effects of TME tumorigenesis drivers, the novel pipeline designed in this study holds revolutionary potential for explicating other complex cell processes, such as cell differentiation and neuronal evolution. Further harnessing the power of such analysis enables future research to uncover key elements of intricate biological mechanisms that have remained unexplored for decades, ultimately optimizing the effectiveness of diagnoses and treatments worldwide.