Zachary Chiang
Biophysics 205 Final Project Proposal
April 24th, 2017

Lineage Tracing of Stem-like Tumor-Propagating Cells via CRISPR Evolution

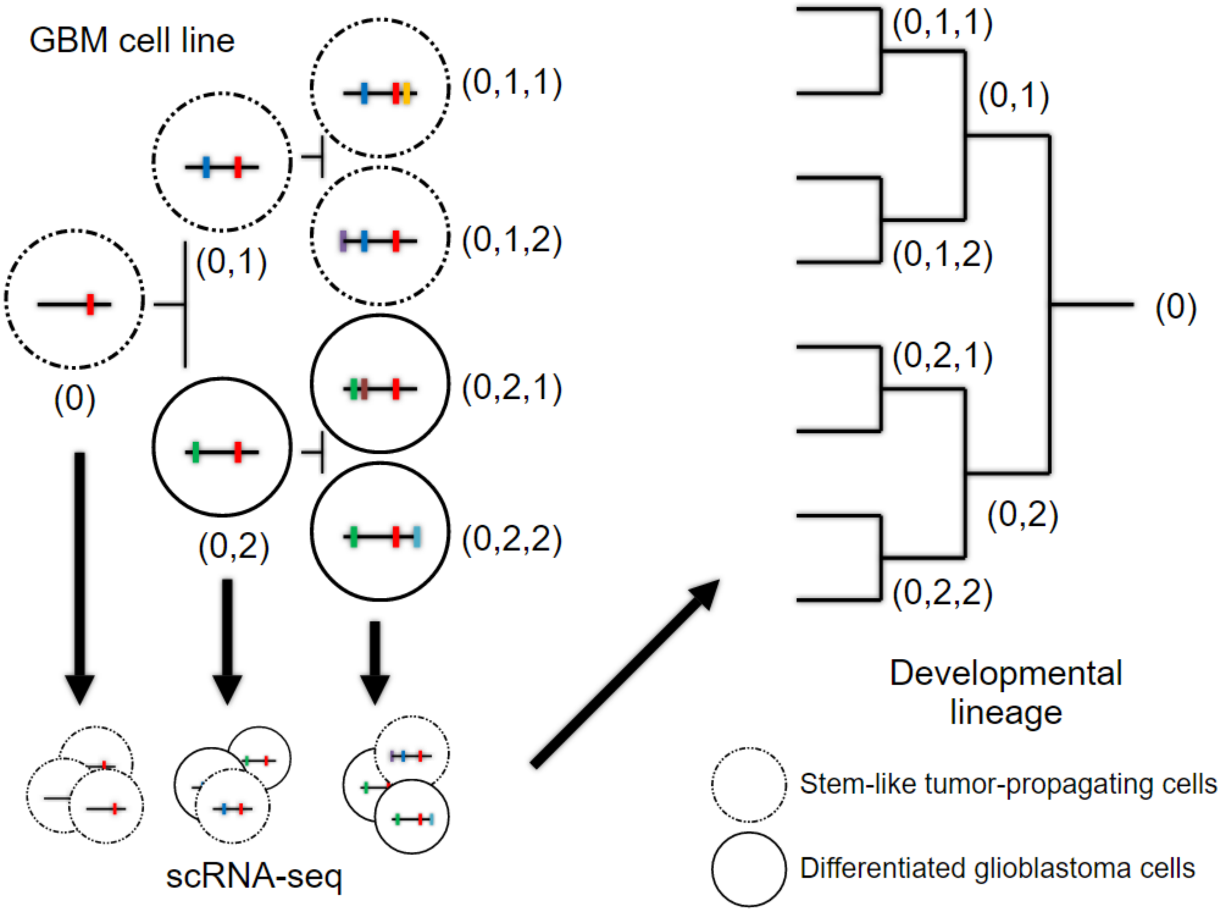
Introduction and Motivation

Glioblastoma (GBM) is the most common form of brain cancer in adults. It is highly malignant, and generally has a poor prognosis despite aggressive treatment[1]. Targeted therapy has proved to be challenging because GBM displays remarkable tumor heterogeneity, both between patients with the same tumor type, and between cells within a tumor[2]. The first type of heterogeneity is starting to be addressed, as whole-genome sequencing and transcriptional profiling have revealed several clinically-distinct GBM subtypes[3]. The second type is more difficult due to current limitations of single-cell technologies. However, a recent study suggests that all subtypes share the same basic developmental hierarchy, meaning that most differences can be attributed to signature genetic events and the tumor microenvironment[4]. Therefore, developing methods to characterize this shared hierarchy and determine how certain cell subpopulations contribute to tumor development is of great interest.

One of the primary goals of this proposal is to understand the mechanism of tumor propagation at single-cell resolution. There is strong evidence that glioblastomas contain a subpopulation of cells that resemble neural stem cells[5]. Since higher-grade tumors have larger pools of these undifferentiated cells and exhibit enhanced proliferation, these stem-like cells are thought to drive tumor propagation and contribute to therapeutic resistance[6]. Previous studies have used next-generation sequencing techniques to compare these stem-like tumor-propagating cells (TPCs) to differentiated glioblastoma cells (DGCs). This led to the identification of a set of transcription factors (TFs) that reprogram differentiated cells into a stem-like state when induced[7]. A significant limitation of this approach is that both types of cells were isolated and grown within a specific medium as to preserve their current state. More work is necessary to understand how these subpopulations interact and change within a heterogeneous population over the course of development.

Cell lineage tracing is a method of identifying the relationships between individual cells in a population at different stages of development. Homing guide RNAs (hgRNAs) are the latest advancement in using genome editing technologies such as CRISPR/Cas9 to reconstruct such lineages[8,9]. The basic idea is to generate a unique barcode in each cell that it will pass on to all its descendants by introducing diverse mutations in small regions of DNA that do not affect cell viability. These barcodes can then be computationally aligned and clustered to reconstruct a developmental hierarchy. The advantage of using hgRNAs over other modern genome editing-based techniques is that by introducing a PAM sequence into the endogenous locus of your guide RNA, the Cas9:gRNA complex will continually target its own locus. This vastly increases the number and diversity of new mutations that can be introduced in one barcode, making it feasible to track an entire population of cells over several weeks.

Here I propose to combine a hgRNA lineage tracing system and single-cell RNA-seq at regularly spaced time points to characterize the role of TPCs and DGCs within GBM spherical cultures. Although it is unclear how well gliomaspheres recapitulate the heterogenous tumor environment, this experiment will establish a basic framework for how these subpopulations of cells interact and change over time. Being able to link transcriptional signatures to a cell's role in the developmental hierarchy may provide clues about the mechanisms responsible for unchecked proliferation and tumor development. If there is some feature that is predictive of how a cell's descendants will behave, then that may be a potential marker for new therapeutic strategies to eliminate TPCs in glioblastoma. Lastly, the success of this experiment will have resulted in the creation of a robust system that can be adapted to investigate other healthy and diseased cell populations, possibly under various drug conditions or using different types of single-cell sequencing assays.



Specific Aim 1: Introduce a functional hgRNA system into GBM cells.

The purpose of the first aim is to construct a system that will enable lineage tracing in a heterogenous population of glioblastoma cells. To this end, I will introduce a functional hgRNA locus into GBM cells via lentiviral integration. Like a typical single guide RNA (sgRNA), a hgRNA requires a 16-18 nucleotide spacer and a scaffold in order for Cas9 to induce a double-stranded break at its target site. However, unlike most sgRNAs, a functional hgRNA must also contain a protospacer adjacent motif (PAM) sequence that directs Cas9 to cleave its own locus, in addition to any separate target sites[8,9]. To engineer this, I will mutate the sequence just downstream of the of the spacer to the required 'NGG' motif, as well the nucleotides that hybridize to this sequence in the scaffold, as to not affect the secondary structure.

It will be very important to demonstrate that the hgRNA system is functional. To show that the hgRNA can target its own locus, I will introduce a Cas9 under the control of an inducible promoter into cells along with a hgRNA locus and a separate target site. After inducing Cas9 and sequencing these regions, both sites should show evidence of cutting and repair in the form of NHEJ-mediated insertions and deletions. An important control for this experiment is to repeat the procedure with a normal sgRNA in place of the hgRNA to ensure that there is no homing activity without the introduced PAM sequence.

Previous studies have shown that a hgRNA will cease targeting its own loci when its spacer is truncated to a length shorter than 16-18 nucleotides[8]. Since it is crucial that this system can generate sequence diversity across an experimental time course of several weeks, I will add a 50-nucleotide stuffer sequence between the TSS and the spacer of the hgRNA locus to prevent this type of truncation. As the hgRNA locus goes through multiple rounds of targeted double-stranded breaks and accumulates sequence deletions, the stuffer will slowly become a part of the spacer. There is evidence that adding a 50-nucleotide stuffer somehow slows down the rate of the entire system, allowing the hgRNA to generate sequence diversity at least 14 days following Cas9 induction, unless the PAM is mutated[8].

The final requirement for a hgRNA system is that it must be possible to consistently read out the spacer sequence as a cellular barcode in the primary assay, which is single-cell RNA-seq in this case. To do this, I will design reverse transcription (RT) primers for the hgRNA locus regions flanking the spacer, as to not interfere with the homing activity. The left primer will be located downstream of the TSS inside the stuffer sequence, which be preserved as is unless an extreme amount of spacer truncation occurs. Because it is not feasible to amplify the entire hgRNA locus, the right primer will be located within the hgRNA scaffold sequence, in a region shown to be tolerant of insertions[8].

Specific Aim 2: Generate transcriptional profiles of individual GBM cells.

The overarching goal of the second aim is to characterize individual glioblastoma cells at specific timepoints in the developmental hierarchy. Although there are many experimental options, my approach will be to generate transcriptional profiles of GBM cells using single-cell RNA-seq. I am planning to do scRNA-seq as opposed to other modern single-cell assays, such as BS-seq or ATAC-seq, because the transcriptome arguably gives the most complete picture of all the processes occurring in the cell. However, because epigenetic marks and open/closed chromatin regions are believed to play key roles in maintaining cell lineages, it would make sense to combine the hgRNA system with these assays in future studies. Another important consideration is the selection of timepoints. Since GBM cells are typically passaged every three days, that will also be when I take samples for scRNA-seq. The procedure will be performed over a period of two weeks, throughout which the hgRNA system should be active.

At each timepoint, I will use the Smart-seq2 protocol as previously described[10] to generate transcriptional profiles for 96 individual cells. I plan to use Smart-seq2 as opposed to other widely-used scRNA-seq protocols because it currently results in the best coverage of full-length transcripts due to its template switching strategy. I don't anticipate needing to profile thousands of cells to reconstruct a developmental hierarchy, but if there is evidence that rare cell subpopulations make unique contributions to tumor propagation, I will switch to the Drop-seq protocol to increase the number of cells[11]. To ensure that the scRNA-seq protocol does not result in any significant biases, I will also perform bulk RNA-seq on the same samples and run a basic correlation analysis to compare it to pooled scRNA-seq data.

Making sense of single-cell RNA-seq data can be more challenging than it is with bulk RNA-seq due to cell-to-cell variability and the sparse nature of each transcriptome. To account for some of these known confounders, I will use the PAGODA package[12] to normalize variance based on a negative binomial model and adjust for the dropout of low-abundance transcripts. After processing the data, I will identify cell subpopulations through unsupervised clustering. One commonly-used approach to cluster single cells is hierarchical agglomerative clustering because it doesn't require a predefined number of clusters or depend on random initializations like k-means clustering. Using the implementation from the python scikit-learn library, I will run hierarchical clustering on the set of cells from each timepoint.

Since I expect to see two major cell subpopulations in my data -- TPCs and DGCs -- I will use the resulting dendrograms to identify the two largest clusters at each time point, removing any major outliers. There are a few ways to validate that these two clusters correspond to TPCs and DGCs using transcriptional profiles. One approach would be to check if the set of TFs that reprogram cells into a stem-like state are upregulated in one cluster. Another method would be to perform differential gene expression analysis using PAGODA[12] and looking for enriched GO terms that obviously correspond to one of the two cell types. Assuming that I am able to label the clusters at each timepoint, it will also be important to check that the clusters are consistent across development. One way to do this would be to calculate UPGMA (Unweighted Pair Group Method with Arithmetic Mean) values for each cluster compared to a pooled set of the data. A high value would indicate that a cluster at a particular timepoint has a unique transcriptional signature, indicative of either experimental error or some biologically significant phenomenon.

Specific Aim 3: Reconstruct a GBM developmental hierarchy from barcode sequences.

The purpose of the final aim is to reconstruct a general developmental hierarchy for glioblastoma using the barcode sequences retrieved from each individual cell. If the hgRNA system remains active for two weeks, every GBM cell sequenced should have a spacer region with a unique combination of insertions and deletions due to the randomness of the NHEJ repair pathway. Since these indels are passed down all a cell's descendants, it is possible to use temporal information to pinpoint when an event occurred, which can then be used to place each cell at the correct location in the developmental lineage.

Because the experimental procedure calls for profiling 96 cells at all eight of the developmental timepoints, a method to systematically determine the insertions and deletions in each barcode is necessary. I will write a simple script that compares each barcode to the original spacer sequence using a local alignment with gap penalty strategy. On average, I would expect that barcodes from later timepoints are further from the original sequence. There may be some cases where the barcode is so far from the original sequence that it is impossible to deconvolute the mixture of insertions and deletions that occurred. If this is the case, I will write a script that iterates through each timepoint to find intermediate barcodes that provide information about the order and location of events. Finally, I will write a script that creates the hierarchy by linking barcodes from adjacent timepoints with the greatest similarity, and removes any ambiguous cases.

Once the developmental hierarchy is complete, I will start to generate statistics to find biologically significant trends. Relevant questions include: how the sizes of subpopulations change over time, which subpopulations tend to have more descendants, how often descendants become a different cell type, etc. After that, I can begin to look at the associated transcriptional profiles to determine if expression of certain genes correspond to a cell's role in the hierarchy, or how expression of key transcription factors fluctuates. However, the most important question I hope to answer through this experiment is whether there are factors that predict the proliferative activity of a stem-like TPC and its descendants. To do this, I will identify the TPCs with the most stem-like descendants in the hierarchy and do differential gene expression analysis to compare these cells to all other TPCs. I hypothesize that there may be some transcriptional signature that causes some TPC to drive tumor propagation, and if so, this could be used as a potential biomarker for glioblastoma therapeutics.

References

1. Jansen M, Yip S, Louis DN. Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers. Lancet Neurol. 2010;9(7):717-26.
2. Sottoriva A, Spiteri I, Piccirillo SG, et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc Natl Acad Sci USA. 2013;110(10):4009-14.
3. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17(1):98-110.
4. Venteicher AS, Tirosh I, Hebert C, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science. 2017;355(6332)
5. Bao S, Wu Q, Mclendon RE, et al. Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. Nature. 2006;444(7120):756-60.
6. Chen J, Li Y, Yu TS, et al. A restricted cell population propagates glioblastoma growth after chemotherapy. Nature. 2012;488(7412):522-6.
7. Suvà ML, Rheinbay E, Gillespie SM, et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. Cell. 2014;157(3):580-94.
8. Kalhor R, Mali P, Church GM. Rapidly evolving homing CRISPR barcodes. Nat Methods. 2017;14(2):195-200.
9. Perli SD, Cui CH, Lu TK. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. Science. 2016;353(6304)
10. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171-81.
11. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161(5):1202-14.
12. Fan J, Salathia N, Liu R, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016;13(3):241-4.