Cancer resistance is an unsolved problem partly due to lack of understanding of heterogeneity within a tumor. A drug may be effective against most of the cells in a tumor, but a few survivors may cause relapse. Prevailing methods to study a complex tissue such as a tumor use identical cells, which does not accurately represent the diversity of its different, interacting cell types. Thus, I propose to study the tumor heterogeneity of chronic lymphocytic leukemia (CLL), the most common adult leukemia in western countries, at the single-cell level with Prof. Gene Yeo at University of California, San Diego (UCSD). CLL is the most common adult B-cell leukemia in Western countries, and approximately 10% of affected individuals have a mutation in the spliceosomal subunit SF3B1 associated with poor prognosis [1,12]. It is likely that there are more splicing-related mutations as known oncogenic splicing factors in other cancers [4,5]. I want to study how SF3B1 is differentially mutated between individual cells, differences between DNA and RNA at the single-cell level, and how this affects alternative splicing mechanisms. This work leverages my experience in single-cell analysis of cancer at UCSC and investigation of cancer genomics diversity at the Broad Institute.

Aim 1: Develop simultaneous RNA and DNA-sequencing from a single cell. Presently, only one of DNA or RNA can be removed from an individual cell for high-throughput sequencing, as the process destroys the cell. However, it is critical to understand the exact differences between individual cells, including how DNA affects the possible RNA. Currently, we map RNA-Sequencing (RNA-Seq) reads onto "reference" DNA that is an average of 13 individual genomes, but human-human variation is 0.01%, which could mean the difference between cancer cure and relapse. We must use the individual's DNA to inform our study of its RNA. Thus, I propose to develop simultaneous RNA and DNA sequencing (R+D-Seq) from a single cell.

I will use Yeo's Fluidigm CI single-cell sorting device to isolate cells into individual [wells] of 96-well plates. I will collaborate with single-cell genomics expert Dr. Roger Lasken from the J. Craig Venter Institute (JCVI) to optimize the experimental protocol on Craig Venter's fibroblasts. This would ensure that we get as close to exact replicates of single cells in each individual well of the 96-well plate, and with copy numbers of DNA close to one. We will use the RNA-specific Trizol reagent to extract RNA first as it is more fragile, then [other reagent] to extract DNA from individual cells. [We will avoid elution-based methods as the volumes will overwhelm the tiny amounts of genetic material.] We will likely run into issues such as within-plate heterogeneity and batch-effects, and these results will help develop algorithms for processing the individual cells from the plates, then further develop algorithms that use the sequenced DNA as a reference for RNA-Seq. We will then be prepared to analyze genomic and transcriptomic heterogeneity in CLL.

Aim 2: Determine RNA-editing events from discrepancies between RNA and DNA. RNA-editing is the controversial topic of the cell modifying individual bases of RNA to produce different versions of proteins in response to the environment. All previous studies were performed in cell populations, with separate groups for each of RNA-Seq and DNA-Seq. By performing R+D-Seq on an individual cell, we would know exactly which RNA-editing events correspond to which genomic DNA and RNA transcripts. We would use normal human fibroblast cells to develop algorithms to determine RNA-editing events, then use this software to study CLL cells.

If successful, single-cell R+D-Seq would advance the field by providing comprehensive, definitive answers of the extent of RNA editing, the pervasiveness of which is currently

controversial and unknown, bringing us closer to understanding the role of RNA editing in cancer resistance.

Aim 3: Identify alternative splicing patterns in individual cells. Alternative splicing (AS) is another method of producing variant proteins by mixing and matching different elements within an RNA transcript, a critical issue as in cancer as there are many mutations in AS sites. Armed with accurate RNA transcripts of single cells, we can determine whether AS patterns measured from a population are truly expressed by every cell, or whether they are a population average. Again, we would develop software using human fibroblasts, then apply the algorithms to study heterogeneity of AS in CLL. This will pave the way to understand how AS patterns of individual cells contribute to the response of a tumor to chemotherapy.

The intersection of single-cell genomics, alternative splicing, and RNA editing is only available here at UCSD with Prof. Yeo, who owns a prototype of a single-cell sorting device, is an expert in alternative splicing and RNA-editing, and is a member of the Moores Cancer Center, from which I will obtain clinical cancer samples in collaboration with CLL expert Dr. Thomas Kipps. Additionally, I will collaborate with single-cell genomics expert Dr. Roger Lasken at the J. Craig Venter Institute. We will use these studies to investigate the effects of SF3B1 in CLL, which will further our understanding of the roles of heterogeneity and alternative splicing in resistance.

I am excited to pursue this unique project here at UCSD. This project has the potential to revolutionize the field of genomics by providing a true reference for individual cells, uncovering the [true] [levels] of RNA-editing, and determine heterogeneity of alternative splicing events.

When I tell people that I study bioinformatics, their usual reaction is, "Bio-what?" Despite its rapid growth, bioinformatics is still a new field, and one rarely understood. I first heard of bioinformatics via the Human Genome Project in my eighth grade genetics class, and was captivated by the possibility of fully characterizing hereditary disease and personalizing medicine. I want to spread the joy and potential of bioinformatics, and for students to become as about bioinformatics excited as I was.

During my master's at University of California-Santa Cruz, I took advantage of every opportunity to teach, including bioinformatics in high school biology. As part of a team of students and a professor that pioneered this program, I spearheaded the best-received curriculum. In my unit, "Genes and Disease," students chose a gene related to a familiar disease or condition and explored its properties - protein size, biological function, chromosomal location, and so on. One creative group explored "maple syrup urine syndrome," and found that the gene associated with this condition plays a role in sugar metabolism. This bridge from genetic defect to disease causation was an anomaly, as most students struggled to find direct links between disease and genes, a realistic research experience. Pleased with the success of the pilot project, we released the curriculum on the web, where hundreds have viewed it internationally. I am still actively involved in refining and developing this curriculum through UC-San Diego's ScienceBridge, which targets underserved high schools in San Diego. I hope that in teaching underprivileged students these skills, I can empower them to pursue science.

This summer, I built on my high school teaching experience and developed a stem cell biology curriculum for minority undergraduates embarking on a summer research project. My favorite aspect of the course, though it was also the most difficult, was planning the "inquiry" part of the curriculum, where learners mirrored authentic research by developing their own unique method of investigation. Our process goal was for learners to realize that cell identity is a continuous spectrum and become more comfortable with conflicting information – important realities of research. As one of three facilitators, I taught bioinformatics as a method of exploring biology. I debated whether to clean up the data, but decided it would diminish the experience of mirroring authentic research, as real data is not clear-cut whatsoever. While initially overwhelmed, the students appreciated the realistic experience, and I was inspired by their dedication to science.

In addition to teaching, I also enjoy organizing conferences, as it is critical to share and effectively communicate scientific research. This summer I was the youngest co-chair of the Student Council Symposium (SCS) at Intelligent Systems for Molecular Biology (ISMB), the largest international computational biology conference, held in Long Beach, CA. We had 60 attendees from six continents, and I savored meeting the travel fellowship recipients and experiencing their unyielding commitment to research. The SCS was a great venue for students to present to a broad audience, and some for the first time in English. I was honored to be able to facilitate such a life-changing event for my peers.

By engaging in teaching, mentoring, and the scientific community, I have realized the importance of each unique contribution. This has strengthened my research as it has pushed me to pursue even more interdisciplinary science and teach to even broader audiences. I intend to become a professor and together with my colleagues, build a research center that integrates genomics, proteomics, metabolomics, and emergent fields to solve biological problems.

In my first research experience at Prof. Martha Bulyk's lab at Harvard Medical School, I performed molecular biology experiments and ran Perl scripts. I was amazed at how microarrays polka-dotted with tens of thousands of protein-DNA binding pairs could be transformed into eight-letter binding affinities. This experience hooked me on computation; I saw how much power there was in using mathematics and computer science to solve biological problems. During my first computational lab experience with Dr. Sean Eddy at Howard Hughes Medical Institute Janelia Farm, I developed a null model for protein sequence comparison using Hidden Markov Models. Thirsty for more research experiences, I explored two more laboratories: in Prof. David Gifford's lab at MIT's Computer Science and Artificial Intelligence Laboratory, I worked on network differences between two yeast strains, and on computationally modeling infection in the immune system; in Prof. Sebastian Seung's lab at MIT's Department of Brain and Cognitive Sciences, I studied neuron orientation in rabbit retina.

Revealed new cancer activators via integrated genomics: After college, I worked at the Broad Institute of Harvard and MIT in Cambridge, MA, with Dr. Jill Mesirov to develop REVEALER, an algorithm that integrates genomic and functional data to infer new associations. For example, a researcher may know that a certain pathways of genes are overexpressed in many cancers, and that this overexpression can be caused by a mutation in the governing oncogene. However, there are many cases where there is no mutation in that oncogene and yet the pathway is highly expressed. REVEALER finds novel candidate activators of this pathway by removing samples which already have a mutation in the original oncogene, and searching for the top genomic feature that explains the remaining samples using a mutual information (MI) metric to discern between top candidates. In a half hour, the REVEALER algorithm uncovered the relationship of resistance to knockdown of the oncogene KRAS, to an amplicon unrelated to KRAS mutation, a finding that had taken another researcher months of looking through Excel files. I am co-first author on our manuscript (in preparation) where REVEALER found additional experimentally verified novel activators. This ambitious project combines state-of-the-art computational methods with genomic analysis in the same way I hope to do in my future research.

Single-cell analysis of chemotherapy resistance in breast cancer: There are some cancers that no matter what you throw at them, they still relapse. In Prof. Nader Pourmand's laboratory at Univ. Calif.-Santa Cruz (UCSC), we were interested in breast cancer drug resistance at the single- cell level to observe how individual cells escape chemotherapy. Specifically, we used paclitaxel (taxol), which inhibits microtubule elongation, preventing proper mitotic spindle formation and, subsequently, preventing cells from dividing. Our collaborator at the Salk Institute extracted individual cells from three groups of breast cancer cell lines: untreated (6 cells), treated with paclitaxel (6 cells), and survivor (5 cells). Our lab performed whole-cell RNA-Sequencing (RNA- Seq) on these 17 samples, and one of the main technical challenges was consistently analyzing the aligned sequencing reads and producing lists of differentially expressed genes. To increase consistency between sequencing experiment interpretations, I developed an open-source RNA- Seq differential expression (RSDE) pipeline which takes in alignment files and annotation, then outputs genome-wide coverage, heatmaps, lists of differentially expressed genes, and quality control information. Using RSDE to study taxol resistance in breast cancer at the single-cell level, we found the surviving cells had differentially regulated actin cytoskeleton genes, which may indicate that these survivors have some mechanisms to compensate for improper chromosome alignment to the mitotic plate.