Section 1: Personal

How I Became Interested in Cancer Systems Biology

It began with a serendipitous encounter with a *Scientific American* magazine during science class in eighth grade that discussed the idea of parallel universes. As a shocked thirteen year-old, I had to learn more, so I watched a documentary about the underpinnings of our universe hosted by Dr. Brian Greene, which amplified my curiosity even more. I went on to read a myriad of books in the field by authors such as Stephen Hawking. This exploration sparked in me an unquenchable fascination with the universe, rapidly replacing all feelings of contempt I previously held for the field as a child. Realizing that science could unleash unbelievable wonders filled me with a sense of purpose and a newfound appreciation for the study, which I had previously dismissed as mundane and full of extraneous details. **No longer was "science" just a dry term describing a body of knowledge. It became a way of life that opened doors to exploration, adventure, and altruism in unimaginable ways.**

With my exploration of theoretical physics, I learned that the mysteries of the universe must be addressed with mathematics. If we want any kind of structure, precision, and the ability to predict phenomena, we must implement mathematics to define and explain what we see in nature. Many of the peers around me felt that mathematics had no role in daily life, and was an absolutely useless subject. I could see where they were coming from, but after delving into theoretical physics, I gained a unique perspective as to why it's important: it describes everything.

At the same time, I started to become interested in cancer research. After hearing about other teenagers' successful science fair projects, such as those of Jack Andraka, Shree Bose, Brittany Wenger, Eric Chen, and many others, I realized that we can make progress on biomedical challenges no matter what our age. After all, a great number of singers and actors and artists begin their careers while they are just in middle school. So why is it shocking when someone involves themselves in research at a young age too? I didn't have a PhD (and didn't even know what that was), but luckily, I was ignorant enough to believe I could do something. (I'll explain in a bit how I went about getting involved!)

Furthermore, cancer was (and still is) a **major** cause of suffering in the world today. And though millions of researchers are tirelessly working on understanding and treating

it, there is still a lot more to be done. Torn between wanting to go into physics and biomedical research, I was beyond thrilled when I realized that my deep interest in theoretical physics could actually be *applied* to biomedical research. In fact, it was needed.

Let me explain. Through reading countless magazine and journal articles on cancer, I eventually came across an interdisciplinary field of research with immense promise: systems biology, the integration of computer science, engineering, mathematics, and physics with biology to advance our understanding of and ability to treat disease.

Systems biology operates on the premise that biology should be studied in terms of complex systems. Genes do not act *in solo*, but *in concert*. We must strive to understand how networks of millions of molecular interactions give rise to cancer (and other diseases) - and when you're dealing with that much complexity, you need mathematics. There is no other way.

Wanting to involve myself in research instead of just reading about it, the summer before ninth grade, I emailed dozens of scientists from local institutions requesting an internship position. Though I received multiple rejections, I continued reaching out, and eventually got a positive response from a National Cancer Institute (NCI) researcher. I spent a year working with her on identifying heritable cancer-causing genes using computational approaches.

The next summer, I got in touch with a Johns Hopkins University (JHU) professor and helped develop a computational strategy for finding sets of cancer-causing genes from cancer genomic datasets.

And finally, last summer, at the National Institutes of Health (NIH), I developed a novel computational method for predicting cancer-driving genes and pathways, which is what this paper will go deeper into. In fact, I loved my project so much, this summer I'm back at the NIH, continuing my investigation and algorithm development!

And with each summer, my fascination and involvement with computational/systems biology has grown. Mathematics made biology more alive for me; it provided a potent lens through which I could explore the world.

Advice for High School Students

If you want to undertake a project combining science and mathematics, use your imagination! If there is any scientific problem that intrigues you, there is almost always

going to be mathematics required to understand and solve it. If there isn't currently much mathematics used, then that's even better; find a way to make it more mathematical. That will only increase your work's originality and rigor. That's what I did with biomedical research -- except there was already a cutting-edge field involving math and computation: systems biology. Look for the systems biology of your field, and if there isn't one, use your imagination. You are never too young to be innovative, so don't silence your creativity. You're more capable than you think!

Once you have identified a field or interest/problem to tackle, look for a mentor. I am endlessly grateful for my incredibly supportive mentors. Without them, this project wouldn't have become what it is. So reach out to scientists/researchers/professors in your area. Send them emails, or call them. Tell them about your curiosity for what they do, and ask them if you can work with them. You will most likely have to contact many, many people, but the most important thing is to NOT give up. Because once you find yourself a supportive mentor, they will be your most cherished source of guidance. Those who take you on will more than likely be extremely happy to help you, and will actively encourage your curiosity.

Once you're on a project and settled into a lab, or even just working on your own thing with a mentor to consult, keep yourself constantly updated to the field. If you're working at the interface of a scientific field and math, it will probably be a very dynamic and fast-changing area, so you want to keep up. This will exercise your imagination and help you become an idea-machine.

Now here is something important. As I learned the hard way, doing research is very different from reading about it. Reading a paper could take, say 45 minutes, but the paper itself could easily contain years of work. What I'm getting at here is that producing good scientific research is rigorous and requires resilience, which is something you'll have the opportunity to develop. So don't feel distraught if you are struggling; in fact, you're supposed to, or you're not learning! I initially felt very uncomfortable when confronted with math much more advanced and specialized than what I'd learned in high school. I had to ensure my calculations were backed with rigorous statistics to maintain scientific integrity. I learned that research is not glorious like they show in sci-fi movies (just smash your keyboard violently and bam, a cure for cancer! I wish.), and requires expanding your comfort zone to learn complicated math.

It's scary, but as I learned, the rigorous math is what makes it *possible* to find answers to ambitious questions.

I also learned another valuable truth: if something is your "passion," that doesn't mean you feel pleasurable bliss every moment you're engaging yourself in that activity. The satisfaction comes in the struggle, and in the consequent growth that is a result of venturing out of your comfort zone and feeling vulnerable every step of the way because you're always in unknown territory, exploring questions without answers and always, always knowing there is more to be known. The feeling doesn't go away, because this is what research is. Embrace the feeling.

I would like to take a moment to emphasize that *none* of this would have been accomplished without the support of my graciously generous mentors over the years. Taking me under their wings was more than enough, but they were also beyond enthusiastic to mentor and guide me. After experiencing firsthand what an impact such generosity can make, I am inspired to reach out to the younger generation as I progress in my own career as well. Those who take extra time out of their schedules to support others when they don't need to are the best kind of people, and will be largely responsible for the scientific progress of the future. Thus, I encourage you to keep in mind how you can support and inspire others while chasing your ambitious dreams!

Section 2: Research

Abstract

Devising effective strategies for treating cancer requires elucidating molecular mechanisms through which the disease initiates and spreads. A critical step for doing so is distinguishing driver mutations from passenger mutations, the former contributing to tumorigenesis while the latter, though abundant, being nonfunctional. An observed property of driver mutations is their mutual exclusivity; mutually exclusive driver genes often share the same functional pathway, as one driver mutation in a pathway is usually sufficient for dysregulating the pathway's function. Though this gene-gene relationship has been established, there is a lack of investigation of mutual exclusivity between mutated pathways, as well as a need for improved detection of co-mutated pathways. An accurate statistical model, balancing computational intensity and accuracy, was

developed for the evaluation of mutual exclusivity between driver genes in the human genome across twelve cancer types. With this model, driver genes were identified for each biological pathway and a novel algorithmic strategy based on the previous mutual exclusivity algorithm was devised to evaluate relationships between pathways. This strategy successfully uncovered combinations of mutually exclusive and co-occurring dysregulated pathways, including PI3K-Akt and Ras signaling, with backing from experimental studies. This algorithm is indispensable for elucidating tumorigenic mechanisms and guiding combinatorial drug targeting efforts to effectively treat cancer and mitigate resistance.

Introduction

A major challenge in cancer research is being able to distinguish driver mutations from passenger mutations, the former contributing to tumorigenesis, while the latter, though abundantly present, conferring no selective growth advantage to the cell. Recently, cancer genome sequencing projects have been able to measure genomic, transcriptomic, and proteomic levels in a vast quantity of human tumors. With the observed heterogeneity of these sequenced tumor genomes, it is necessary to characterize the variety of molecular features in order to identify drivers across and within cancer types, this knowledge ultimately guiding rational therapeutic development. Furthermore, it is known that pathways, not single genes, govern the course of tumorigenesis. This idea is strengthened by the observed heterogeneity of tumor genomes, which further suggests that driver mutations target sets of genes, not single genes. Therefore, is critical to understand how driver genes interact in pathways, and to further determine which pathways drive the initiation and progression of cancer. This information will assist with understanding the molecular mechanisms involved in tumorigenesis, consequently helping with designing rational treatments.

Recently, high-throughput measurement of molecular properties of cells has become possible and resulted in large quantities of biological data. With the power of genome analysis technologies, such as large-scale genome sequencing and microarray

measurements, the levels of thousands of different molecules can be measured simultaneously at the genomic, transcriptomic, and proteomic levels. This leads to the ability to gain a more effective characterization and understanding of tumor genomes. The Cancer Genome Atlas (TCGA) is a comprehensive effort that has profiled a large number of human tumors, including measurements of diverse characteristics of cells, such DNA, RNA, and protein levels. Such measurements yield a rich source of "big data", full of informative patterns waiting to be mined. This is both a blessing and a curse; having a rich source of information means an integrated view of tumor genomes can be constructed, but it is also a challenge to figure out where to look among all the noise. Therefore, the development of computational tools that can interrogate the commonalities, differences, and emergent themes across multiple cancer types is of high importance.

A common strategy for detecting driver mutations in cancer genomes is a simple frequency-based technique; simply select mutations in genes that are the most recurrent and compare this rate to the expected background mutation rate (BMR). However, this technique is insufficient because rare mutations would have a very weak signal, and the heterogeneity of cancer genomes deems many important drivers undetectable. This technique also does not look at driver genes in the context of the pathways to which they belong, which is why there may be so much heterogeneity in the first place. Tumors of the same type can be caused by mutations in completely different genes in two different patients.

Strategies have been developed to circumvent this problem. It has previously been observed that driver mutations belonging to the same functional pathway are mutually exclusive, due to one mutation being enough to disrupt the function of the pathway, as there is no further selective pressure placed on the cell for that particular function. In other words, it is not expected to see two driver genes belonging to the same pathway co-mutated in a given patient. The mutual exclusivity of driver mutations has been employed for the discovery of novel drivers as well as for the construction of oncogenic network modules using a human protein-interaction network. Other algorithms have

been designed to predict driver pathways from genomic profiles of cancer patients by finding highly exclusive gene sets. However, computational tools to find mutually exclusive *pathways* have not been developed, and it is unknown whether such pathways would correspond to driver pathways. If mutual exclusivity implies sufficient selective advantage, then this property may hold true for not only gene-gene relationships, but also pathway-pathway relationships. Furthermore, it is known that in carcinogenic processes, dysregulated pathways do not always act in *solo*, but can act in tandem with other mutated pathways to enhance tumorigenic effects. Therefore, detecting co-occurring driver pathways in addition to mutually exclusive driver pathways in cancer is a major challenge of interest.

First, a sufficiently accurate model for the evaluation of mutual exclusivity between cancer mutations needs to be developed. While researchers have analyzed the mutual exclusivity of driver mutations in countless studies, most of their methods have employed the hypergeometric model, but this model is not accurate enough, for it does not implement gene- and patient-specific mutation rates. The current alternative approach is the exact permutation test, but this test is extremely computationally intensive, consuming an unrealistic amount of time and space. Therefore, it is necessary to achieve a compromise between computational cost and accuracy. With this, the development of a novel algorithmic strategy for evaluation the mutual exclusivity and co-occurrence of cellular pathways, and the evaluation of its capability to uncover driver pathways and cooperating driver pathways, will be the next key challenge.

These algorithms will be developed to study The Cancer Genome Atlas (TCGA) data in the form of mutation matrices consisting of single nucleotide variations (SNV's) and copy number variations (CNV's) information of ~3,000 patients across twelve major cancer types: bladder carcinoma, breast invasive carcinoma, colorectal cancer (colon adenocarcinoma and rectum adenocarcinoma merged), glioblastoma multiforme, head and neck squamous cell cancer, kidney renal cell carcinoma, acute myeloid leukemia, lung adenocarcinoma, lung squamous cell cancer, ovarian carcinoma, and uterine corpus endometrial carcinoma.

Mathematical Model for Mutually Exclusive Mutations in Cancer

To write the algorithmic strategies devised and perform statistical tests, Python 3.4.3 and R 3.0.2 were used on the Unix Operating System (OS).

The first step was to develop an accurate model for mutual exclusivity of gene mutations across twelve different cancer types using mutation matrices incorporating SNV and CNV data.

The format for mutation matrices for each cancer types is presented in Table 1.

	gene 1	gene 2	gene 3	…	gene 23K
sample 1	1	0	0	…	1
sample 2	1	1	0	…	1
sample 3	0	0	0	…	0
…sample 3K	0	0	1	…	0

Table 1. Mutation matrix format containing SNV and SNV data, with the rows corresponding to each sample and the columns corresponding to each gene. A "1" indicates the presence of an SNV or CNV, and a "0" indicates the absence of an SNV or CNV, for a given gene in a given sample.

The SNV and CNV data was obtained from The Cancer Genome Atlas. Twelve cancer types were analyzed (with colon adenocarcinoma and rectum adenocarcinoma merged into one type, colorectal carcinoma) for ~23,000 genes across 2,867 patients.

Cancer Type	BLCA	BRCA	CRC	GBM	HNSC	KIRC	LAML	LUAD	LUSC	OV	UCEC
# of samples	74	665	199	279	287	414	194	140	140	268	207

Table 2. Number of samples for each cancer type. BLCA: bladder carcinoma; BRCA: breast invasive carcinoma; CRC: colorectal carcinoma; GBM: glioblastoma multiforme; HNSC: head and neck squamous cell carcinoma; KIRC: kidney renal cell carcinoma; LAML: acute myeloid leukemia; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian carcinoma; UCEC: uterine corpus endometrial carcinoma. For all types, there is a grand total of 2,876 samples.

The current, most widely-used approach for evaluating mutual exclusivity of gene mutations is the hypergeometric test (*i.e.,* Fisher's exact test), in which the probability that an n-trial sampling experiment without replacement results in exactly x successes from a population of N items containing k successes is calculated.

While this statistical test can be applied to computing mutual exclusivity p-values in this scenario, it does not conserve patient-specific mutation rates, thus resulting in less accurate p-values for biological scenarios, with the presence of false negatives.

An alternative approach is the permutation test. In this technique, the null model is generated by creating a large number of random permutations of the given mutational profile with the condition that each profile must preserve the observed mutation rates for each gene and sample. If mutations in a pair of genes are mutually exclusive, the number of samples in which they are mutated will be larger than expected by chance. The p-value is calculated by counting the number of random instances that have a larger number of mutated samples than the observed, original mutational profile. However, generating null profiles for each combination of genes for ~23,000 genes across 2,867 samples is impractical.

To achieve a balance between accuracy and computational cost, a variation of the permutation test was tested: weighted sampling. Instead of generating random mutational profiles via the permutation test, weighted sampling was used to generate a set of random profiles with the same expected mutation rates. This approach preserves the mutation rates for each gene and the expected mutation rate of patients, which the hypergeometric model fails to do.

Measuring Mutual Exclusivity with Weighted Sampling

Cover size was implemented as the measure of how mutually exclusive a pair of mutations in a pair of genes was. For a given pair of genes (*g1*, *g2*), cover size is defined as the union of the set of patients who have a mutation in either *g1* or *g2*. If mutations in a given pair of genes are mutually exclusive, then the number of samples in which they are collectively mutated, *i.e.,* their cover size, will be larger than expected by chance.

If *g1* was mutated in x patients, and *g2* was mutated in y patients, then mutational profiles were randomly permuted while conserving these mutation rates x and y. Likewise, each patient's mutation rate z_i was preserved in the random mutational profiles.

For each gene, 100 random mutational profiles were generated. Thus, for each pair of genes, a set of 10,000 random instances were yielded by comparing each of the 100 mutational profiles of *g1* with each of the 100 mutational profiles of *g2*. For each pair of genes, the union (*i.e.*, cover size) of samples in each pair of random profiles being compared was then recorded as a random instance, resulting in 10,000 random instances. Then, the actual observed cover size of the mutations in *g1* and *g2* was compared with the random model, and for every cover size that is greater than the observed cover size, 1/10,000 (0.0001) is added to the p-value (initially set to 0) for mutual exclusivity.

Mutual Exclusivity and Co-Occurrence of Driver Pathways in Cancer

Computation of 10,000 random cover sizes for each pair of pathways

For each gene in each pathway, 100 random permutations of its mutational profile were computed using the weighted sampling method outlined previously. Then, every gene's corresponding random mutational profile was merged, *i.e.*, the union operation, into a set of 100 random mutational profiles for the entire pathway. The same was done for the second pathway in the pair. Figure 2 is a visual example of how these random pathway mutational profiles are generated.

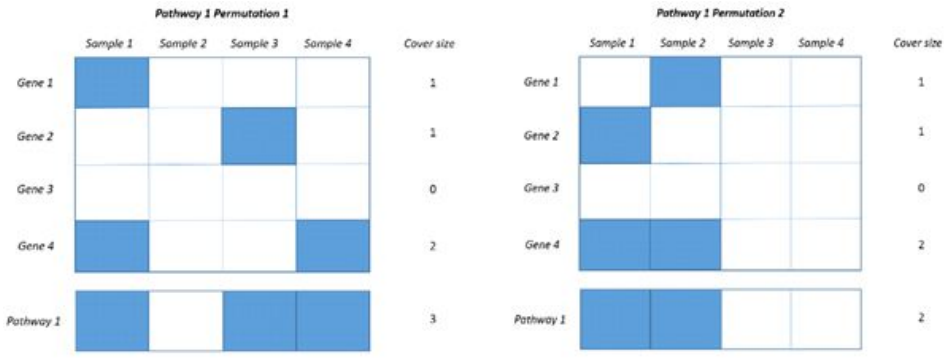

Figure 2. Above are randomly generated pathway mutation profiles composed of gene mutation profiles from each gene member of the respective pathway. For sake of simplicity in this example, there are four patients and each pathway consists of four genes. For a given pathway, a random instance is generated by merging the set of samples containing a mutation in any gene belonging to the pathway. This is repeated 100 times to generate 100 random pathway mutational profiles for each pathway.

Then, the 100 permuted profiles from each pathway in a given pair were merged with each other to produce 10,000 random profiles: for each profile belonging to Pathway 1, a cover-size computation with each of the 100 profiles of Pathway 2 was performed. It then follows that 100 x 100 = 10,000, yielding 10,000 random profiles. Figure 3 is a visual of how 10,000 random pathway mutational profiles can be obtained for each pair of pathways.

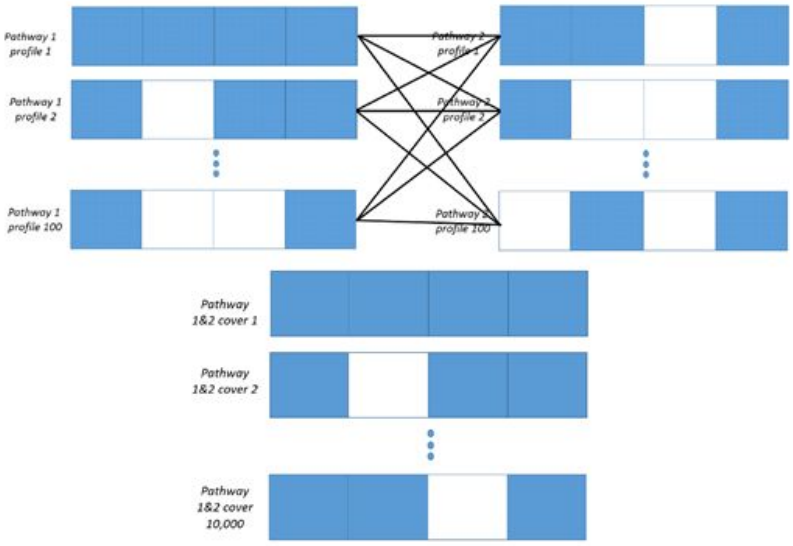

Figure 3. Generation of 10,000 random pathway mutational profiles for each pathway pair is possible through computing cover sizes for each of the 100 random instances in Pathway 1 and each of the 100 random instances in Pathway 2. This set of 10,000 random cover sizes is the null model for a given pathway pair. Mutually exclusive mutated pathways will have an observed cover size that is greater than expected, while co-occurring mutated pathways will have a smaller than expected observed cover size.

<u>Computation of observed pathway pair cover sizes</u>

The calculation of the observed cover size of a given pair of pathways is identical to the approach outlined in the previous section. In this instance, however, cover sizes of the *observed* profiles in the mutation matrices are computed, obtaining values which will be compared to the null models.

Results

P-values of mutual exclusivity of gene mutations using my weighted sampling method were significantly more correlated to the permutation test than those of the hypergeometric test. Figure 4 contains two graphs displaying the correlation coefficients of (1) hypergeometric test vs. permutation test and (2) newly devised weighted sampling method vs. permutation test.

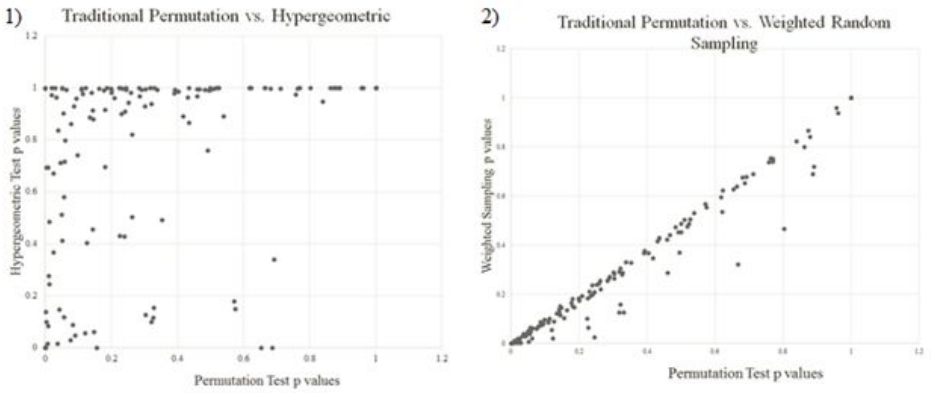

Figure 4. **Permutation Test vs. Hypergeometric Test and Permutation Test vs. Weighted Sampling**
A subset (n=298) of the gene pairs' mutual exclusivity tested in this project was computed in another study [1] using the permutation test. The graphs above show the correlation between the permutation test p-values and (1) the hypergeometric test (R = 0.558794) and (2) my weighted sampling approach (R = 0.994617). The weighted sampling test was able to obtain substantially more accurate p-values than the hypergeometric test.

Conclusion

My algorithmic procedure was able find driver pathways based on mutual exclusivity, a task for which no other computational tools exist. In addition, this computational procedure found co-occurring mutated driver pathways across multiple cancer types, a task that has needed further work on improving biological certainty. It was able to analyze all ~23,000 human genes belonging to several hundred pathways in several thousand patients, its findings having backing from experimental studies. This tool revealed combinatorial patterns of the cancer genome landscape, and knowledge of these patterns is indispensable for deciphering key tumorigenic mechanisms essential for our understanding and ability to treat cancer.

My algorithm was also capable of finding experimentally validated co-occurring driver pathways, suggesting collateral pathways that collaborate to induce cancer initiation and progression. Knowledge of dysregulated pathway combinations goes a step further and provides guidance on which drug targets combinations are promising for minimizing resistance.

Implemented in this analysis, the pathway-centric view of cancer classification provides a potentially more statistically powerful approach for clinically relevant classification than the gene-centric view. To this end, more investigation needs to be

carried out to confirm or refute the idea that pathway-centric mutual exclusivity and co-occurrence hold stronger statistical signals than gene-centric mutual exclusivity and co-occurrence.

Ultimately, this algorithm is a tool for hypothesis generation. Computational tools provide predictive capabilities, guiding future experimental research. This approach has yielded new insights into biology, which will inspire new biological questions, which will in turn further our understanding of and ability to treat cancer. The immense potential of this algorithmic tool lies in its ability to point experimental researchers in a rational direction when deciphering tumorigenic mechanisms and designing new therapeutic platforms for treating cancer.

Acknowledgements

I would like to acknowledge Dr. Teresa Przytycka and Dr. Yoo-Ah Kim for supporting me above and beyond and teaching me more about research than I fathomed I would learn. Their generous support and patience are indispensable to all that I have accomplished. Having mentors as wonderful as them is something I will cherish and be grateful for for the rest of my life.

I also want to thank Dr. Patricia Miller and Mr. Mark Curran for providing invaluable support and guidance to me during the research process. Their genuine efforts to help me make my research paper the best possible, to have a presentable poster, and to have a polished presentation, among many other things, were significantly invaluable to the success of this project. They taught me that the communication of science is just as important as the science itself.

I also express my gratitude to my mother, Inderbir Madan, for supporting me for the past 18 years. She went above and beyond to ensure I could pursue the research experiences that I did despite the hardships she was going through, supported me when I doubted myself, and without her, I'm not sure I would have accomplished anything at all.

"*If I have seen further than others, it is by standing upon the shoulders of giants.*" - Isaac Newton