Statement of Purpose: Megan Le

I am excited to continue research in computational biology, which is a setting where my academic background in computer science and mathematics is directly translatable due to the generation of incredible amounts of data, as well as an area where I could have a major impact in understanding and improving human health. My computational research in genomics has not only given me experience developing and implementing analytical tools for large-scale genomic datasets, but also helped me develop strong scientific communication skills through a first-author manuscript, a conference presentation, and two invited seminar talks.

In my future research, I want to explore how we can use computing power and statistical methods to untangle the complexity behind genetic variation to ultimately improve health outcomes. I am particularly interested in exploring how machine-learning can be applied to large-scale genomic datasets for tasks such as mutation rate mapping, gene regulatory network modeling, and variant effect prediction. With my strong foundations in computer science, mathematics, and genomics, I am well-prepared to pursue a PhD in computer science and continue research in computational and statistical genomics.

Computational genomics research: As a research assistant in Dr. Vagheesh Narasimhan's lab, I grew fascinated by the way we can extract meaning about nature through analyzing raw sequences of nucleotides. In my research project, I analyzed DNA obtained from ancient human remains to examine the genomic impacts of profound cultural evolution over the last ten millennia in Europe. Ancient DNA data provides the opportunity to study natural selection by directly tracking shifts in allele frequencies between generations, but previous studies have been limited due to small sample sizes. In this study, we detected genetic variants that increased in frequency more than expected in the absence of natural selection. By using increased sample sizes and applying this method in slices of time, we resolved the timing of signals to three time periods for the first time. This allowed us to generate more informed hypotheses about the cultural transformations that may have fueled these genetic adaptations. Most of our discovered signals of selection were novel. Interestingly, they were concentrated in earlier time periods (about 6,000 to 8,000 years ago), highlighting the importance of the "time stamp" that comes with ancient DNA data.

I performed all of our analyses, which required writing efficient code to process segments of large datasets in parallel on supercomputing resources, and I helped interpret and contextualize our results after learning relevant background knowledge in evolutionary and population genetics. While my work was initially focused on implementing a method developed by my colleagues, I later participated in the development of a new statistical tool designed to detect selection acting on complex traits. This was both challenging and exhilarating, since population geneticists have previously identified many caveats and obstacles to the interpretation of signals of selection on complex traits. The statistical tool I helped develop showed increased robustness to these obstacles. I am the first author on our paper for this project, which has been preprinted and is currently under review at a journal. I also presented this work in a platform talk at the American Society of Human Genetics 2022 Annual Meeting, as well as invited seminar talks for the Variant Effect Seminar Series and UT Austin RNA & DNA Club.

Other computational research: Prior to my genomics research, I was excited to apply my computing and statistical skills to other research areas where it's important to analyze large and complex datasets. My first research experience was as a student in the Freshman Research Initiative Computational Materials stream. My independent research project investigated the use of Newton's method for finding stable states of a chemical system, which can fail due to areas of negative curvature on the potential energy surface. I implemented different solutions to this problem and compared their performances for the Lennard-Jones potential. This experience was what first excited me about performing scientific research, since I realized how much I enjoy formulating, designing, and carrying out research projects.

Continuing in the Computational Materials research stream, I developed software for a haptic device that simulates forces for chemical systems, allowing users to gain intuition about atomic interactions. I also worked on the integration of machine-learning generated potential energy surfaces with the simulation software and gave a poster presentation over this work at the 2019 Texas Advanced Computing Center Symposium for Texas Researchers. This was my first project that involved machine-learning, and I was excited by how machine-learning could both increase accuracy and decrease computation time when calculating potential energies. This sparked my interest in learning more about the machine-learning techniques that I would now like to apply to biological data.

Later, through an individual research course to complete a certificate in Computational Science, I worked on a project in physical oceanography with Dr. David Halpern and Dr. Patrick Heimbach that evaluated the accuracy of the ocean currents produced by the Estimating the Circulation and Climate of the Ocean (ECCO) consortium's predictive model. We wanted to learn how the ECCO model predictions could be improved for various current characteristics, since the model could potentially be used in place of expensive moored measurements. I wrote all of the code and performed all of our analyses, which entailed analyzing large datasets, validating my programs on test datasets, formulating hypotheses for the differences we saw between datasets, and testing those hypotheses. Ultimately, we found the ECCO model currents were not adequate substitutes for moored measurements, but we formulated equations that could supplement the model to improve the use of model currents as proxies for moored measurements. I am the second author on a manuscript that is under review.

Future goals: My deep involvement in diverse computational science research has developed my ability to think about problems in an interdisciplinary framework and synthesize information across multiple fields, as well as analyze complex datasets and formulate research questions. I am excited to apply my computational and statistical knowledge to the development of novel methods that will advance our knowledge of genomics in new ways, with the ultimate goal of improving healthcare outcomes. In addition to my research activities, I have enjoyed helping fellow students unravel complex concepts and establish strong foundations in fundamental skills during my time as a teaching assistant, tutor, and peer research mentor. For these reasons, I would like to pursue a research career in computational biology as a professor.

I hope to continue my research journey at MIT due to its interdisciplinary strength in computer science, mathematics, and biology, as well as my interest in the research being conducted by faculty in the Computer Science and Artificial Intelligence Laboratory. In particular, I am interested in Dr. Manolis Kellis' work related to studying the effects of genetic variation on disease, and I would be interested in developing methods for tasks such as causal variant identification and variant effect prediction. My experience developing methods for analyzing large-scale genomic datasets would also extend well to working in Dr. Bonnie Berger's lab, where I would be interested in developing algorithms and leveraging deep-learning methods for tasks such as mutation rate mapping and multimodal single-cell data analysis. Additionally, I would be interested in working with Dr. Caroline Uhler on developing algorithms with provable guarantees for learning gene regulatory networks. I believe there is a strong fit for my skills and interests at MIT, and that the Department of Electrical Engineering and Computer Science would be an excellent environment for me to grow as a researcher and pursue my goals.