**Thesis Objectives**

The history of science is riddled with examples where theory, technology, and serendipitous discovery drives science. The advent of Illumina short reads and continued decrease in per base sequencing cost has accelerated our understanding of human evolution and migration patterns \cite{}, identification of pathogenic mutations in patients with Mendelian diseases \cite{}, the analysis of driver mutation and transcriptomic landscape in thousands of cancer genomes \cite{}.

The inability to generate contiguous and complete reference genomes, however, with Illumina short reads and the prohibitively expensive cost of BAC clone library preparation and hierarchical shotgun sequencing has thwarted our efforts to understand genetic variation in non-model organisms \cite{}.

High-throughput and high-accuracy single-molecule sequencing technologies \cite{} overcome the limitations of the Illumina platform and propel us towards the third wave of genomic revolution where each individual will be able to have their complete and haplotype phased genome sequence, where the construction of the most complex and repetitive genomes will be possible and where the reference genomes of all organisms will be available to the scientific community.

The DToL project, for example, has generated an extraordinary public resource that comprises CCS reads, linked reads, Hi-C reads, high-quality chromosome-length scaffolds, and associated gene annotations. Comparative genomics in linear and three-dimensional space and population genetic studies with the newly assembled reference genomes will undoubtedly enhance our understanding of the process of speciation and evolution. Here, we instead aspired to better understand the mutational process operational in each species.

To determine the germline and somatic mutational process across the Tree of Life, we considered the following:

\begin{enumerate}

\item Based on the similarities between the duplex \cite{} and CCS library sequencing \cite{}, we hypothesized that CCS reads might have sufficient base accuracy for ultra-rare somatic mutation and potentially single molecule somatic mutation detection.

\item CCS reads are reported to have a predicted accuracy above Q20, but their base accuracies have not been independently examined.

\item Somatic mutation detection algorithm needs to distinguish somatic mutations from germline mutations, in addition to, sequencing, alignment and systematic bioinformatic errors. We, unfortunately, cannot differentiate somatic mutations from library errors unless there are upstream modifications to the library preparation protocol.

\item Using samples with single ongoing somatic mutational process and mutational signature analysis, we can demonstrate that CCS reads have sufficient or insufficient base accuracy for single molecule somatic mutation detection and determine the parameters that influence sensitivity and specificity.

\item If the sample in question has either high mutation rate or high mutation burden, the expected and the correct mutational spectrum will be observable from the validation and test data sets, respectively.

\end{enumerate}

In short, we aimed to measure the CCS error rate, assess whether CCS bases have sufficient base accuracy for single molecule somatic mutation detection, develop a method to detect somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome and apply the method to understand germline and somatic mutational processes across the Tree of Life.