There are two primary approaches to sequence the complete genome of a given species. De novo sequencing projects seek to assemble the genome of an organism from scratch using only the sequence data. Resequencing projects use a reference genome as a template to map the sequencing data. Resequencing projects are more efficient in that they use the fact that the DNA of many species are closely related. However, most current resequencing techniques are not sensitive enough to find correlations between distantly related genomes. Our group tries to overcome this by developing a tool that can profile short reads quickly, and find distant correlations by being more sensitive while not losing out on accuracy.

If we are to efficiently process all the reads we cannot afford to directly check if each read sufficiently matches with a substring in our template. What we need to do instead is, eliminate most of the incorrect reads quickly and only check those reads, which get through this initial filter, for a match. This initial filter is what my portion of the project focuses on.

I am building an indexing structure that basically indexes certain patterns and tells us where in the template they can be found. These patterns are only a portion of the complete read that we get. By using this table we save a lot of time while looking for a read-match. This is achieved as we no longer need to look through the whole template in an attempt to find the read. Instead we only look in those parts that we already know a portion of the read lies. There are a few other structures that we could have chosen to implement instead. It’s all a choice between memory speed and ease of implementation. It is sometimes hard to predict if a
certain structure fits are requirement of time and space so we may need to test out other structures in the future.

This gives us speed, but we have still not obtained sensitivity which was the main issue. To gain sensitivity we essentially allow a few mismatches between our read and the template. This is required because, as mentioned before, the genome of the target species may vary significantly from the template that we have. So we cannot afford to straight away reject any read that differs slightly from the given template. This is where a lot of complications come in.

Before I could approach this problem, I needed to read a couple papers on string matching with mismatches. This exposed me to reading research papers which is something I have not done before. It was very interesting talking to my mentor about the research papers.

In the upcoming weeks the goal is to finish the implementation and then optimize it to its limit, as speed is the primary concern. This will involve a lot of testing, finding bottlenecks and trying to use computer resources more efficiently.

Exploring this area of computer science has been interesting. After this experience I will continue to pursue undergraduate research exploring more fields so that I can better understand what I want to do in the future.