Student: Lucy de Jong

Project: Nanopore sequencing of repeat sequences in human DNA

Supervisor: Professor Martin Kennedy

Sponsor: Canterbury Medical Research Foundation

Introduction:

This project was all about exploring the capabilities of a brand new piece of technology, called the MinION, to study parts of the human genome that are traditionally very difficult to examine: repetitive regions of DNA. DNA is the molecule that encodes everything that makes your body function. DNA sequencing is the process of determining the order of the bases A, T, C and G (the "code") on the DNA strand. Current technologies do this by fragmenting the DNA, sequencing all the pieces, then using software to put the sequence together.

Unfortunately, this method doesn't work very well on repeat regions of DNA because it can't put these pieces back together easily. Repetitive regions are scattered through the genomes of many species and are of considerable interest in areas such as human forensic investigations and animal conservation. They are also implicated in gene regulation and neurological disorders (such as Huntington's disease), however the full effects of repetitive DNA on genome function is largely unknown. In fact, repetitive regions of DNA are amongst the most poorly explored regions of the human genome, because they are difficult to study. The MinION is a device about half the size of an iPhone 6 and works by threading single strands of DNA through a tiny pore set into a membrane. The membrane has an electrical current across it and as the DNA goes through the pore, the bases A, T, C and G disrupt the current in such a way that the sequence can be read by analysing the electrical signal generated. The MinION generates long DNA reads with no fragmentation, tens of thousands more bases than the original method of Sanger sequencing and a lot more data in each run. Over summer I worked to see how well this long read nanopore sequencer could read repetitive DNA. The sequence I initially looked at is the CAG repeat implicated in Huntington's disease, repeated fifty times. With this particular repeat, the larger the number of repeats present the greater the likelihood of a patient getting the disease at an earlier age.

Aim:

To explore and evaluate the ability of the MinION nanopore sequencer to correctly read and quantify a variety of human repeat DNA tracts.

Method:

1. Amplified the CAG repeat sequence from the human genome using bacterial plasmids.
2. Sanger sequenced the DNA from the plasmid in order to get a reference sequence.
3. MinION sequencing of the plasmid.
4. Compared and analysed sequencing reads between the Sanger sequencer and the MinION. It became apparent throughout the project that we would have to look at the raw data output from the MinION as well as the sequence.

Results:
Many different methods and software packages were tried, with the aim of finding one method that was able to identify both the sequence AND the number of repeats. The best software turned out to be the very first one tried, called LAST. This generated an average result; the sequence was easily identified, however it was not so good at identifying the exact number of repeats. By looking at the raw data (generating a "squiggle plot") we can easily count the number of repeats in a sequence. There has been a computer algorithm designed which can do this for a single read and it would be fairly straightforward to expand this to include multiple reads for accuracy.

Conclusion:
The MinION can correctly read and quantify a CAG repeat, however at this stage different methods are required to achieve each goal of identifying the sequence and counting the number of repeats present. Further investigation into how the MinION can handle different types of repeats is ongoing. We will be looking at another type of 3 base repeat and two each of 4, 5, and 6 base repeats.

This project has confirmed that it is possible to accurately identify repeats, and has confirmed that further work in the software area is required to create a seamless analysis of repetitive regions of DNA.