Project Proposal

John Palmer

**Background:**

The hepatitis C virus (HCV) is a single-stranded RNA virus from the *Flaviviridae* family that is currently causing a pandemic affecting approximately 185 million people worldwide (Messina et al., 2015). In 2017 alone, roughly 400,000 infected individuals were killed due to diseases induced by HCV (WHO). Chronic infection with HCV is linked to the development of severe liver diseases, which include acute and chronic hepatitis, cirrhosis, and hepatocellular carcinoma (Messina et al., 2015).

The diversity of circulating hepatitis C viruses throughout the world has been classified into seven main genotypes and additional more closely related subtypes (Messina et al., 2015). Past studies have demonstrated that patient prognoses and overall progression of disease can differ between HCV genotypes and subtypes (Schröter et al., 1999). It is therefore, important to track the prevalence of different HCV subtypes in all regions of the world so that public health systems can respond accordingly.

The Los Alamos National Laboratory has curated a database of HCV sequences collected from infected patients around the world (https://hcv.lanl.gov/content/index). For this project, I plan to query this database for patient HCV sequence data to analyze the prevalence and relatedness of infections found on two different continents: Asia and South America. I will focus this analysis on the core protein of HCV as it is a well populated genetic region within the database and plays an essential functional role in the formation of new HCV particles (Gawlik and Gallay, 2014).

**Data:**

* **Source:** HCV Database at Los Alamos National Laboratory
* **Format:** Nucleotide sequences in FASTA format
* **Header Metadata:** Subtype, patient ID, accession number
* **Region / Gene:** Core protein
* **Genotype / Subtype:** All
* **Geographical Regions:**
  + South America (~500 sequences)
  + Asia (~1900 sequences)
* **Additional Information:** 
  + Reference sequence of HCV (NC\_004102)

**Objectives:**

1. To sort sequences by country and find the most prevalent HCV subtype in each country and continent
2. To sort sequences by subtype and compute genetic distances between these different subtypes by comparing random samples of sequences to each other
3. To translate nucleotide sequences into amino acid sequences and determine whether there are differences in hydrophobicity between subtypes
4. To identify those patients who have been sequenced more than once

**References:**

Gawlik, K., and Gallay, P.A. (2014). HCV core protein and virus assembly: what we know without structures. Immunol. Res. *60*, 1–10.

Los Alamos National Laboratory Hepatitis C Virus Databases.

Messina, J.P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G.S., Pybus, O.G., and Barnes, E. (2015). Global distribution and prevalence of hepatitis C virus genotypes. Hepatology *61*, 77–87.

Schröter, M., Feucht, H.-H., Schäfer, P., Zöllner, B., and Laufs, R. (1999). Serological Determination of Hepatitis C Virus Subtypes 1a, 1b, 2a, 2b, 3a, and 4a by a Recombinant Immunoblot Assay. J. Clin. Microbiol. *37*, 2576–2580.

WHO Hepatitis C.