**rhinoTypeR - An R package for rhinovirus genotyping**

**Abstract:** Rhinoviruses (RV), common respiratory pathogens, are positive-sense, single-stranded RNA viruses characterized by a high antigenic diversity and mutation rate. With their genome approximately 7.2 kb in length, RVs exhibit mutation rates between 10^-3 and 10^-5 mutations per nucleotide per replication event. These viruses are classified into 169 types across three species: RV-A, RV-B, and RV-C. Genotype assignment, a critical aspect of RV research, is based on pairwise genetic distances and phylogenetic clustering with prototype strains, a process currently executed through multiple steps, making it laborious. Our project aims to develop an R package to streamline RV genotype assignment, facilitating genomic scientists in efficiently genotyping RV infections.

**Data:** The project will utilize VP4/2 sequences available in the public domain from GenBank and reference prototype strains from www.picornaviridae.com

**Tools and Analysis:** The development will be carried out using R and RStudio. The package will encompass functions to compute genetic distances, perform phylogenetic clustering, and compare sequences against RV prototype strains. These functionalities will be designed to be user-friendly and adaptable to various research needs.

Other tools necessary may be `AliView` for sequence visualization, and command-line based `mafft` for sequence alignment.

**Proposed Methods:**

Our methodology will involve:

1. Parsing and preprocessing of VP4/2 sequence data.
2. Implementation of algorithms to calculate pairwise genetic distances.
3. Integration of methods for constructing Maximum Likelihood phylogenetic trees.
4. Automated comparison of sequences with prototype strains for genotype assignment.

**Skillset:** Expertise in bioinformatics, specifically in genomic data processing and phylogenetics. Proficiency in R programming and use of Git and GitHub is essential. Experience in package development and maintenance is desirable. Familiarity with virology/ genomics will be an asset to understand the nuances of RV genotype variability.

**End product:** The culmination of this project will be an R package that streamlines the genotyping of RVs. It will enable genomic researchers to assign RV VP4/2 sequences to appropriate genotypes accurately and efficiently. This tool aims to expedite RV research by simplifying the genotype assignment process, contributing to broader insights into RV epidemiology and evolution.

**References:**

*RV epidemiology*

1. <https://www.nature.com/articles/s44298-023-00008-y>

*RV genotype assignment*

1. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749525/>
2. <https://pubmed.ncbi.nlm.nih.gov/20610666/>

*Models of molecular evolution*

1. <https://genome.cshlp.org/content/8/12/1233.full>

- Ruth take us through her R package & introduce the course

- Martha present on rhinoviruses, diversity, currently classification and aim/objectives

Coding:

- create prototype db. Any curation needed?

- algorithm

- alignments (if not aligned?)

- handling gaps/ short sequences

- evolution models: p-distance (JC), Max composite,

- handling unknown types

- handling ‘near misses’ ie doesn’t match another type but very close to known type

- overlapping genotypes i.e. p-distance <10.5 to 2 protypes

- outputs:

i) csv/txt of `seq\_id, type` , p-distance

ii) circular barplot of frequencies

iii) overall diversity – think around this e.g. x axis intra-type diversity, y axis type; histograms of p-distances..hmm

iv) phylogeny with typed sequences and only matched prototypes

- documentation!! & a fine GitHub page too

Which model to use?

Determining the best model for estimating genetic distances between newly generated rhinovirus sequences and prototype strains involves considering the evolutionary characteristics of the virus, such as the rates of mutation, the prevalence of transitions over transversions, and the base composition across the genome. Rhinoviruses, being part of the Picornaviridae family, have single-stranded RNA genomes that can exhibit high mutation rates. This implies that models accounting for varying substitution rates and base composition bias could be more appropriate.

**Recommended Models:**

1. **Tamura-Nei Model (TN93):** This model is suitable for sequences with unequal base frequencies and different transition/transversion ratios, which is often the case in RNA viruses like rhinoviruses. The TN93 model provides a more accurate estimation of genetic distances by accounting for these factors.
   * Citation: Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution, 10*(3), 512-526.
2. **General Time Reversible (GTR) Model:** The GTR model is one of the most general and flexible models, allowing for different rates of substitution between each pair of nucleotides and different base frequencies. This model is particularly useful when the evolutionary history of the sequences is complex.
   * Citation: Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences (American Mathematical Society), 17*, 57-86.
3. **Maximum Composite Likelihood (MCL) Method:** While not a substitution model per se, the MCL method is a technique used in software like MEGA to estimate genetic distances across a wide range of models, providing flexibility in handling various evolutionary scenarios typical of rhinovirus diversity.
   * Citation: Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution, 30*(12), 2725-2729.