**CDMAP\_SO\_Main.R:**

**Purpose:** This is the Main executable script for the CDMAP Package that runs the Single Organism Analysis Pipeline. In this script, CDMAP first checks for, and if necessary installs all required packages. Then CDMAP prompts the user for relevant input information such as the cleaned VCF file, GBK file, and reference FASTA file and then dynamically creates all the relevant child directories needed to store all output files. Then CDMAP proceeds to calculate the counts and rates for all 64 site-specific contexts, codon usage, both with respect to the leading and lagging strand.

**Required Packages:**

* **Seqinr** – Multi-use bioinformatics sequence analysis package for parsing and searching FASTA files, contains **ORILOC** for ORI/TERM determination
* **BiocManager** – multi-use bioinformatics package containing **genbankr**, a package for parsing and manipulation of GBK files
* **Pracma –** Mathematics package for R containing a library of numerical analysis functions
* **Lattice** - Lightweight graphics visualization package used for generation of heatmaps
* **Beepr** – Notification package that sends audio notifications during certain steps of the analysis process.

**Variables:**

* **Organism** – String variable containing the name of the organism being analyzed.
* **DirCheck** – String variable that acts as a flag parameter to check if you are working from home/work/ or as a guest (depreciated).
* **Pathwd** – Working directory in which the executable scripts are contained.
* **Path\_outhput** – File path of where the user desires output data generated**.**
* **Path\_RefFile** – Reference FASTA sequence file path.
* **Path\_GBFile** – Reference genbank GBK file path.
* **Path\_correlate\_repo** – Path of output generated multi-organism correlation repository.
* **generations** – Number of mutation accumulation generations carried out during the MA experiment
* **param\_flag** – Visualization output scaling flag.
* **RefFile** – Raw reference FASTA file parsed by seqinr
* **RefSeq\_arr** – Munged and wrangled version of **RefFile**, which has been converted to an uppercase character array.
* **len\_refseq –** Integer value for that references the length of **RefSeq\_arr.**
* **MutBaseCalls** – Data frame that holds the position (numeric), the reference (character), and mutant (character) nucleotide values.
* **ori\_ref** - Generates the oriloc data object, this is a test variable.
* **ori\_pos –** Numeric that finds the position of the origin of replication (in KB)
* **ori\_bp** – Converted version of **ori\_pos** that has been adjusted for relative nucleotide position length
* **ori\_value** – Skew value of the ORI position
* **term\_pos -** Numeric that finds the position of the terminus of replication (in KB)
* **term\_value –** Skew value of the term position
* **term\_bp -** Converted version of **term\_pos** that has been adjusted for relative nucleotide length

**createDirectories.R:**

**Purpose**: The primary objective of this R script is to generate all of the ordered output directories dynamically for a given organisms triplet counts and context dependent mutation rates with respect to the leading and lagging strand. For reach directory, CDMAP checks if the directory exists, then proceeds to create it if it does not exist.

**generateContextCounts.R:**

**Purpose:** The objective of this Script is to generate the context mutation counts for a given chromosome of a single organism. This file parses the MutBaseCalls data frame and then for each location in MutBaseCalls it grabs the 5mer fragment for each base pair substitution and calculates the linear distance each BPS from the given ORI and TERM locations.

**partitionReplichores.R:**

**Purpose:** This script is responsible for taking the output context\_output\_full, context\_output\_upstream, and context\_output\_downstream matrices and then to partition them accordingly into left and right replichore outputs for each matrix.