# **es.dhaka**: an R package for processing environmental sampling data

This package provides reproducible functions for parsing and compiling data sets that measure infectious disease burden using Environmental Sampling (ES) methods. The tools here were developed specifically for ES studies in Dhaka, Bangladesh and are designed to provide data outputs that can be used with the open-source companion package ‘es’ which can be found [HERE](https://gilesjohnr.github.io/es/). An overview of data and methods are outlined below, with more detailed vignettes available HERE.

## Data

The tools here are intended to do the heavy lifting when combining data from multiple sources such as:

* Data in .eds and .xls format exported from [Applied Biosystems qPCR Systems](https://www.thermofisher.com/us/en/home/life-science/pcr/real-time-pcr/real-time-pcr-instruments.html?gclid=CjwKCAiAi6uvBhADEiwAWiyRdo3P55Qq1NO8sErSvCmVGT2Cs4-5NCzYcgqYhCC6KemKTn6sw7850BoCqbkQAvD_BwE&ef_id=CjwKCAiAi6uvBhADEiwAWiyRdo3P55Qq1NO8sErSvCmVGT2Cs4-5NCzYcgqYhCC6KemKTn6sw7850BoCqbkQAvD_BwE:G:s&s_kwcid=AL!3652!3!606132911219!p!!g!!taqman%20applied%20biosystems!17574808700!139287686778&cid=gsd_pcr_sbu_r02_co_cp1491_pjt9623_gsd00000_0se_gaw_rs_lgn_&gad_source=1) using [QuantStudio software](https://www.thermofisher.com/us/en/home/global/forms/life-science/quantstudio-6-7-flex-software.html).
* Data in .xlsx format containing measurements from [Aquaread water monitoring instruments](https://www.aquaread.com/)

## Methods

1. Parse card-level data output produced by QuantStudio software and extract its contents into individual .csv files containing raw data.
2. Compile card-level .csv files into a single data set. When compiling, several QC operations are performed:
   1. Excludes sensitive targets based on the data in the key data dictionary included in this package.
   2. Parse sample IDs and edit duplicated target names based on the key data dictionary.
   3. Adjustments to the Ct values are made using blank samples to control for contamination. Specifically, if the corresponding blank is not “Undetermined” for a pathogen in a particular sample, then that pathogen’s Ct value is set to NA for all samples on the same card as the blank and for the following two cards as well.
   4. Adjustment to Ct values based on amplification controls. Each pathogen is mapped to its corresponding control (MS2 for RNA targets and PhHV for DNA targets) as defined in the key data dictionary. When an amplification control in a sample is positive (Ct < threshold of 40), then all relevant targets in that sample that are “Undetermined” are set to the threshold value (40 in this example).

## Installation

#### 1) Check dependencies

The data parsing functions in this package depend on Python 3 and Java. To check and install these you can try the following:

Download Python 3 [HERE](https://www.python.org/downloads/) or check current installation with:

user@computer:~$ python3 --version  
Python 3.11.5

Download Java [HERE](https://www.oracle.com/java/technologies/downloads/) or check current installation with:

user@computer:~$ java --version  
java 21.0.2 2024-01-16 LTS  
Java(TM) SE Runtime Environment (build 21.0.2+13-LTS-58)  
Java HotSpot(TM) 64-Bit Server VM (build 21.0.2+13-LTS-58, mixed mode, sharing)

#### 2) Install from Github

Use the devtools package to install the development version of es.dhaka from the GitHub repository. R version >= 3.5.0 recommended.

install.packages('devtools')  
devtools::install\_github("gilesjohnr/es.dhaka", dependencies=TRUE)

## Troubleshooting

This package is currently in development and maintained by John Giles ([@gilesjohnr](https://github.com/gilesjohnr)). For general questions, contact John Giles (john.giles@gatesfoundation.org) and/or Jillian Gauld (jillian.gauld@gatesfoundation.org). Note that this software is constructed under Copyright 2024 Bill & Melinda Gates Foundation and was developed for specific environmental sampling applications and not intended to generalize perfectly to all settings.

## Funding

This work was developed at the [Institute for Disease Modeling](https://www.idmod.org/) in support of funded research grants made by the [Bill & Melinda Gates Foundation](https://www.gatesfoundation.org/).