MADE: Measuring Adaptive Distance and vaccine Efficacy using allelic barcodes

---

Hui Chen, Ivy Quek Ee Ling, Shih Chih Chuan, Zhiyan Fu, Weiwei Zhai* [2018]

Version 1

Genome Institute of Singapore

60 Biopolis Street, #02-01 Genome

Singapore 138672

Hui Chen: chenh1@gis.a-star.edu.sg

Weiwei Zhai: zhaiww1@gis.a-star.edu.sg

Summary

During vaccine production, influenza viruses are unavoidably propagated in embryonated eggs. In the culture expansion, flu viruses adapt to the egg environment, a process known as passage adaptation. In our companion study, we found that egg passage adaptation is driven by repeated substitutions (i.e. convergent evolution) over a set of codons and passage adaptation often leads to highly specific alleles in egg-passaged strains. Using a statistical analysis of these sites, we develop a metric of Adaptive Distance (AD) quantifying the strength of passage adaptation and show that there is a strong negative correlation between AD of a vaccine strain and vaccine efficacy. Based on these observations and principles, we developed MADE (Measuring Adaptive Distance and vaccine Efficacy using allelic barcodes) for vaccine developers to measure the strength of passage adaptation and predict the efficacy of a vaccine strain based on its nucleotide sequence.

Installation

In order to setup the computing environment for MADE, Docker https://www.docker.com/ is needed for subsequent installation (the version of the Docker package has to be compatible with the operational system).

MADE can be directly downloaded from the github website using:

git clone https://github.com/chenh1gis/MADE_docker_v1.git

Set up the computing environment under docker

- **Build an image from a Dockerfile**

  cat [Dockerfile] | docker build -t [a new image name] –

  For example: ***cat MADE_docker_v1/DOCKER_rmarkdown | docker build -t rmarkdown –***

  In this example, a new base image called "rmarkdown" is built.

- **Run a command in a new container (a running instance of an image) & mount the current working directory to container**

  docker run -it --rm -v [current directory]:[directory in container] [an existing image name] bash

  For example: ***docker run -it --rm -v $PWD/MADE_docker_v1:/MADE_docker_v1 rmarkdown bash***

  In this example, a container of the previous base images is running.

- **Run MADE analysis**

  With this setup, further analysis can be executed directly in the container environment.

- **Exit the container**

  exit

A quick start guide to docker

- Detach

  Ctrl+p or *Ctrl+q*

- Re-attach to an up-running container

  docker attach [container name / container ID]

- List all containers or images

  docker ps -a

  docker images

- Delete a container or image:

  docker rm [container name / container ID]

  docker rmi [image ID / image_name:image_tag]

Command line options

Given the allelic information over those key codon positions, the strength of egg passage adaptation will be measured and the vaccine efficacy will be predicted for a candidate influenza vaccine strain.

Input files

There are two different approaches user can input the allelic information to MADE.

- **Approach 1: specifying the alleles at a set of codon positions driving passage adaptation**

  allelic file [in TXT format]

  *For an example allelic file for H3N2 influenza, please refer to "**/test/file_alleles.txt**".*

  All alleles from specified codon positions should be listed into two separated columns (for different influenza viruses, the required amino acid positions will be different).

  1) For H1N1 seasonal virus, these 9 codon positions with strong egg-passage adaptation should be given:

  89, 97 129, 134, 161, 185, 186, 221, 222, 226

  2) For H1N1 pandemic virus, these 10 codon positions with strong egg-passage adaptation should be given:

  21, 127, 129, 183, 190, 191, 222, 223, 225

  3) For H3N2 virus, these 14 codon positions with strong egg-passage adaptation should be given:

  138, 145, 156, 158, 159, 160, 183, 186, 190, 193, 194, 219, 226, 246

Please note that if any allele is missing or its corresponding enrichment score is not available in our curated dataset, the analysis will be terminated immediately.

- **Approach 2: specifying the corresponding nucleotide sequence**

  nucleotide sequence file [in FASTA format]

  *For an example sequence file for H3N2 influenza, please refers to "**/test/file_sequence.fa**".*

  Alternatively, the allelic file can be generated from a sequence file.

  Please note that if any allele is missing or its corresponding enrichment score is not available in our curated dataset, the analysis will be terminated immediately.

Options

--subtype

It is **compulsory** for user to specify the subtype of the candidate influenza vaccine strain. For example, "1" denotes H1N1seasonal virus, "2" denotes H1N1pdm virus and "3" denotes H3N2 virus.

--is_allelic_file

It is **compulsory** for user to specify the type of input file. For example, "1" denotes an allelic file while "0" denotes a nucleotide sequence file.

--id

This option allows user to input the public database ID such as "NC000001" of the candidate influenza vaccine strain.

--strain

This option allows user to input the original source of the candidate influenza vaccine strain, for example, “A/Philippines/2002”.

--host

This option allows user to input the host where the candidate influenza vaccine strain sources from, for example, “human” or “embryonated egg”.

--passage

This option allows user to input the passage history of the candidate influenza vaccine strain, for example, “embryonated egg” or “Madin-Darby Canine Kidney (MDCK)”.

--input_file

It is **compulsory** to specify the input file.

Please be careful with the relative directory while specifying the file name.

Example

perl generate_report.pl --subtype 3 --is_allelic_file 0 --id NC0001 --strain A/Phllipphines/1998 --host Human --passage Egg --input_file test/H3N2_HA1_sequence.fa

or

perl generate_report.pl --subtype 3 --is_allelic_file 1 --id NC0001 --strain A/Phllipphines/1998 --host Human --passage Egg --input_file test/H3N2_14alleles.fa

Please note that ***muscle3.8.31_i86linux64*** *must to executable, e.g.* ***chmod 544 muscle3.8.31_i86linux64.***

Essential supporting documents

To perform the analysis, several supporting files are required for different virus subtypes.

I. **Reference nucleotide sequence**

Reference nucleotide sequence is provided for performing sequence alignment between the input sequence and the reference genome. After sequence alignment, we can extract the allelic status of the codons at the given positions from the input sequence.

*For H3N2 reference nucleotide sequence, please refer to the "****/data/H3N2/H3N2_HA1_sequence.fa****".*

II. **Enrichment scores of all alleles extracted from a large database of curated sequences**

For any given allele at a codon position, enrichment score is defined as the ratio of the allele frequency in the egg passaged strains (Pegg) and in the total set (Ptotal). This enrichment score file stores all the enrichment scores of the observed alleles in the database. This can be used in subsequent analysis.

*For H3N2 reference enrichment scores file, please refer to "****/data/H3N2/H3N2_enrichment_scores_329codons****".*

III. **Multi-dimensional enrichment scores across all background viral sequences**

In order to perform the PCA analysis, we need the enrichment profiles of the input sequence as well as all the background sequences from the public database.

*For H3N2 subtype, please refer to "**/data/H3N2/H3N2_background strains_14alleles**".*

We have curated the profiles of enrichment scores across all the sequences in the GISAID database. This file will be used in the PCA map.

IV. **Allele frequencies and *p* value of all alleles extracted from a large database of curated influenza sequences**

The allele frequencies of alleles located over a set of codon positions are required to calculate posterior probability of the focal strain. During detailed process of calculation, we only consider the alleles carrying higher allele frequency in egg strains when compared to other strains (*p* value from chi-square test <0.0001).

*For the H3N2 subtype, please refer to "**/data/H3N2/H3N2_allele_freq_pvalue**".*

Glossary of terms

- **Adaptive Distance (AD)**

  The physical distance between major clustering of background strains and the target vaccine candidate strain in the principle component plot. It is used to quantify the strength of egg passage adaptation. The higher AD, and the stronger egg passage adaptation.

- **Enrichment Score (ES)**

The ratio between the frequency of an allele in egg passage isolates (Pegg) and the frequency in the total isolates (Ptotal). The higher ES, and the more enriched for the allele in egg passage isolates.

- **Vaccine Efficacy (VE)**

  The percentage reduction of disease in a vaccinated group of people compared to an unvaccinated group. The higher VE, and the more effective of the vaccine treatment.