Genin2

Genin2 is a lightining-fast bioinformatic tool to predict genotypes for clade 2.3.4.4b H5Nx viruses collected in Europe since October 2020. Genotypes are assigned using the methods described in this article. Genin2 identifies only epidemiologically relevant European genotypes, i.e., detected in at least 3 viruses collected from at least 2 countries. You can inspect the up-to-date list of supported genotypes in this file.

Table of contents:

Features

- 🐧 **Cross-platform**: Genin2 can be run on any platform that supports the Python interpreter. Including, but not limited to: Windows, Linux, MacOS.
- 🎈 **Extremely lightweight**: the prediction models weight less than 1 MB
- 🌸 **Easy on the resources**: genin2 can be run on any laptop; 1 CPU and 200 MB of RAM is all it takes
- ⚡ **Lightning-fast**: on a single 2.30 GHz core, Genin2 can process more than 1'200 sequences per minute

Installation

Genin2 is compatible with Windows, Linux, and macOS. It can be installed in two ways:

- Using Python's package manager (PIP)
- Using the Conda package management system

Method 1: PIP

Before proceeding, please ensure you have already installed Python and Pip (the latter is usually already included with the Python installation). Then, open a terminal and run:

```
pip install genin2
```

Method 2: Conda

Genin2 is available on Conda from the bioconda channel. Ensure you have installed Conda and run:

```
conda install -c bioconda genin2
```

Usage

Launching **Genin2** is as easy as:

```
genin2 input.fa output.tsv
```

Input guidelines

Genin2 expects the input to be a nucleotidic, IUPAC-encoded, FASTA file. Please ensure that each sequence name starts with the `>` character and ends with an undersore (_) followed by the name of the segment, e.g.:

```
>any_text|any_string/seq_name_PB1
                             ^^^^
```

For additional deatils on the accepted input format, please see the FAQs section.

Output Format and Interpretation

The results of the analysis are saved to disk as Tab-Separated Values (TSV). This format allows for quick and easy handling as they can be opened as tables with MS Excel, but also for simple and efficient processing by other scripts if you are setting up **Genin2** to work inside of a larger pipeline.

The results table consists of 10 columns:

- **Column 1**: Sample Name

  The sample name, as read from the input FASTA

- **Column 2**: Genotype

  The assigned genotype. Note that a value is only written here when it is certain; in all other cases the genotype is set as `[unassigned]` and the *Notes* column will provide additional information (see below).

- **Columns 3 to 9**: PB2, PB1, PA, NP, NA MP, NS

  The version that each segment is classified as.

- If the confidence of the prediction is below a safety threshold, an asterisk (`*`) is appended to the number.
- If the confidence is also below an acceptance threshold, it is discarded. In this and all other cases where a version is not available, a `?` is displayed, with additional information in the *Notes* column.
- Note: HA is ignored, as all samples are assumend to bellong to the 2.3.4.4b H5 clade.
- Note: MP is always assumed to be version "20", as it is the only version present in Genin2's genotypes list.

- **Column 10**: Notes

  Details on failed or discarded predictions and assigments. This column contains information about these events:

  - Genotypes might be `[unassigned]` because of an unknown composition (*"unknown composition"*), or because accepted versions are too few and the composition matches more than a single genotype (*"insufficient data"*). In the latter case however, if the set of matches is small they are listed as "*compatible with*".
  - Segment versions might be `?` if the segment was not present in the input file (*"missing*"), the sequence had poor quality or many `N`s or gaps (*"low quality"*), if the prediction reported insufficient confidence (*"low confidence*"), or the classification failed in general (*"unassigned"*).

FAQs

- General
  - Which genotypes are recognized by Genin2?
- About input data
  - Do I need to use a particular format for the FASTA headers?
  - Can the input file contain more than a single sample?
  - Are my sequences required to have all segments?
  - Do sequences need to be complete?

Q: Which genotypes are recognized by Genin2?

Answer:

Genin2's prediction models are regularely updated to include relevant new genotypes. You can inspect the table on which predictions are based upon by opening the file src/genin2/compositions.tsv. Generally speaking, we aim to support all epidemiologically relevant European genotypes, i.e., those observed in at least 3 occurences in at least 2 different coutnries.

Q: Do I need to use a particular format for the FASTA headers?

Answer:

Yes. The header should follow this format:

- Start with the `>` character
- Contain a sample identifier, such as `A/species/nation/XYZ`. This part can contain any text you wish, and it will be used to group segments together. Ensure it is the same for all segments belonging to the same sample, and that there are no duplicates across different samples.
- End with the unsercsore character (`_`) and one of the following segment names: `PB2`, `PB1`, `PA`, `HA`, `NP`, `NA`, `MP`, `NS`. The correct association between sequence and segment is essential for the correct choice of the prediction parameters. A valid header might look like this: `>A/chicken/Italy/ID_XXYYZZ/1997_PA`

Q: Can the input file contain more than a single sample?

Answer:

Yes, you can use how many samples you wish.

Q: Are my sequences required to have all segments?

Answer:

No, any number of available segments is accepted by the program. Clearly, missing genes might prevent the unique assignment of a genotype, but you will nonetheless gain knowledge on the versions of the processed segments. Moreover, HA and MP are ignored regardless: the former is assumed from the clade, while the latter, as of now, is only present in the dataset with the version "20".

Q: Do sequences need to be complete?

Answer:

No, not necessarily. Partial sequences are accepted, but the prediction will be based solely on the available data. Sometimes a chunk of sequence is enough for a confident discrimination, and some other times is not.

Cite Genin2

We are currently writing the paper. Until the publication please cite the GitHub repository:

https://github.com/izsvenezie-virology/genin2

License

Genin2 is licensed under the GNU Affero v3 license (see LICENSE).

Fundings

This work was supported by KAPPA-FLU HORIZON-CL6-2022-FARM2FORK-02-03 (grant agreement No 101084171) and by the NextGeneration EU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (Project no. PE00000007, INF-ACT).

> Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HEDEA). Neither the European Union nor the granting authority can be held responsible for them