Package: a2bcovid (via r-universe)

August 28, 2024

Type Package

Title Inferring COVID-19 Transmission Events from Sequence and Location Data

Version 0.1.0

Maintainer Chris Illingworth <chris.illingworth@mrc-bsu.cam.ac.uk>

Description A tool which combines genome sequence and the locations of infected individuals, using a statistical and evolutionary model, to estimate the likelihood that transmission occurred between particular individuals, and then to identify clusters of infections. It is currently designed to apply to COVID-19 infection dynamics on hospital wards.

License GPL-3

Encoding UTF-8

LazyData true

Imports Rcpp (>= 1.0.5), shiny, shinyBS, rstudioapi, ggplot2, ggtext, scales, tidyr, tidyselect, dplyr, magrittr

Suggests testthat, knitr, rmarkdown

LinkingTo Rcpp

RoxygenNote 7.1.2

VignetteBuilder knitr

Depends R (>= 2.10)

Repository https://chjackson.r-universe.dev

RemoteUrl https://github.com/chjackson/a2bcovid

RemoteRef HEAD

RemoteSha bd1f41d86daf8292bb32214c0f5c62fc66a413d9

Contents

`a2bcovid` *Identifying clusters of COVID-19 infections from genome sequence and individual location data*

Description

A tool which combines genome sequence and the locations of infected individuals, using a statistical and evolutionary model, to estimate the likelihood that transmission occurred between particular individuals, and then to identify clusters of infections. It is currently designed to apply to COVID-19 infection dynamics on hospital wards.

Usage

```
a2bcovid(
  pat_file,
  hcw_loc_file = "",
  ali_file = "",
  pat_loc_file = "",
  strain = "default",
  ucta = 2.59321520957074,
  uctb = 3.77600606639754,
  ucto = 3.11208004146092,
  uct_mean = 6.67992,
  evo_rate = 8e-04,
  seq_noise = 0.41369,
  chat = 0.5,
  max_n = 10,
  min_qual = 0.8,
  diagnostic = FALSE,
  hcw_default = 0.5714286,
  pat_default = 1,
  use_all_seqs = 0,
  symptom_uncertainty_calc = 0
)
```

Arguments

`pat_file` (Required) A character string with the path to a file containing the basic data for each individual. This should be a comma separated (.csv) file with data in columns:

1. Individual ID (A code or identifier corresponding to the individual)

2. Onset date. The date at which the individual first experienced symptoms. Date format should be dd/mm/yyyy.

3. Onset date source : Equal to 1 if the date of onset is known. Equal to 2 if the infection was asymptomatic. In this case the onset date is the date on which the first positive swab was collected. Equal to 3 if data is missing or unknown. In this case the onset date is the date on which the first positive swab was collected. If the onset date is anything other than 1 the true onset date is estimated by the code using data collected from Cambridge University hospitals (argument `uct_mean`).

4. Infection type : Equal to 1 if the individual is a patient and a community case (i.e. who could not have been infected by others in the dataset but who could potentially transmit the virus to others). This was defined as being positive for the virus 48 hours before admission to hospital with no healthcare contact in the previous 14 days prior to admission). Equal to 2 if the individual is a patient and not a community case (i.e. who could potentially transmit and receive infection). Equal to 3 if the individual is a healthcare worker. Whether or not an individual is a healthcare worker is set by this parameter.

5. Sequence ID : A code used to link the individual to genome sequence information. This should match the header of the sequence corresponding to the individual in the accompanying .fasta file (arcument `ali_file`).

6. Date of sample collection : Used in evolutionary calculations. Date format should be dd/mm/yyyy.

7. Sample received date : Currently not used in the calculation but necessary. If this seventh column is missing A2B-Covid has been reported to crash R completely.

An example is given with the installed package. The path to the example file can be shown by the R command `system.file("extdata", "Example_genetic_temporal_data.csv", package="a2bcovid")`

`hcw_loc_file` A character string with the path to a file of data describing when specific health care workers were on the ward in question. If this argument is omitted or set to an empty string, then this kind of data is not used in the calculation.

The first line is a header line with column names. The first two of these are labels, while those from the third column onwards describe dates, specified in dd.mm.yyyy format. After the first line, the data is specified in columns as follows:

1. Individual ID (same as for `pat_file`).

2. Cluster ID e.g. the name of the ward in question.

3 onwards: Presence/absence data. A 'Y' indicates that the health care worker was on the ward on the date specified for that column in the first row. An 'N' indicates that the health care worker was not present on the ward on that date. Either 'Y' or 'N' should be specified for each date.

An example is given with the installed package. The path to the example file can be shown by the R command `system.file("extdata", "Example_movement_file.csv", package="a2bcovid")`

`ali_file` A character string with the path to a file in FASTA format containing genome sequence alignments. This file must contain all required sequences, specified

by the sequence ID in the data of `pat_file`. If this argument is omitted or set to an empty string, then genomic data are not used in the calculation. Note: Sequences must be properly aligned for the code to work. It is worth manually checking the sequence alignment before running the code.

An example is given with the installed package. The path to the example file can be shown by the R command `system.file("extdata","Example_sequences.fa", package="a2bcovid")`

`pat_loc_file` A character string with the path to a file containing the location of patients over time. If this argument is omitted or set to an empty string, then this kind of data is not used in the calculation.

This should be a comma separated (.csv) file. Two alternative formats are accepted, "wide" and "long" formats. These are based on the formats in use in the hospital setting where the package was developed.

The first line of the file should be a header with variable names. If there is a variable called `"start_date"` then the file is assumed to be in long format. Or if there is a variable called `"StartDate_0"` then the file is assumed to be in wide format. Otherwise the variable names are ignored. The columns should appear in the specified order.

The names don't matter, but the columns should appear in the specified order.

In wide format, each row represents a different patient. The file should have the following columns:

1. Individual ID (same as for `pat_file`).

2. Cluster ID e.g. the name of the ward being studied.

3. Infection type e.g. 'patient' or 'HCW' for health care worker.

4. Availability of data e.g. 'patient_moves_available'.

5 onwards. Data of the location of a patient, in sets of three columns. These specify in turn: i) The name of the location of the individual e.g. WARD_01. ii) The start date of the individual being in that location. iii) The end date of the individual being in that location. In practice only the first column, and columns from 5 onwards are used. An example is given with the installed package. The path to the example file can be shown by the R command `system.file("extdata", "Example_pat_loc_file_wide.csv", package="a2bcovid")`

In long format, each row represents a single stay on a specific ward for a specific patient. The file should have the following columns:

1. Individual ID

2. Cluster ID, typically the name of the ward

3. Start date/time for the ward stay, in d/m/Y format (optionally in d/m/Y H:M format, but the time is currently ignored)

4. Name of the ward the patient went to next (or "Discharge") if they were discharged

5. End date/time for the ward stay, in d/m/Y format (optionally in d/m/Y H:M format, but the time is currently ignored).

An example is given with the installed package. The path to the example file can be shown by the R command `system.file("extdata", "Example_pat_loc_file_wide.csv", package="a2bcovid")`

`strain` Specification of parameters describing transmission dynamics. `strain = "default"` uses parameter representing the original strain of the SARS-CoV-2 virus. This is the default. `strain = "delta"` uses parameters representing the Delta variant of SARS-CoV-2.

`ucta` Alpha parameter for a gamma distribution of the times bewterrn becomining symptomatic and testing positive. Currently not used.

`uctb` Beta parameter for a gamma distribution of the times bewterrn becomining symptomatic and testing positive. Currently not used.

`ucto` Offset parameter for a gamma distribution of the times bewterrn becomining symptomatic and testing positive. Currently not used.

`uct_mean` Mean time between an individual becoming symptomatic for coronavirus infection and testing positive. This value is used to estimate times of individuals becoming symptomatic in the case that no symptom dates are available

`evo_rate` Rate of evolution of the virus, specified in nucleotide substitutions per locus per year.

`seq_noise` An estimate of the number of mutations separating two genome sequences that arises from sequencing noise. The default parameter was estimated from data collected by Cambridge University Hospitals within single hosts, using the criteria that at least 90% of the reported nucleotides were unambiguous.

`chat` Prior estimate of the probability of any two individuals being in contact on any given day, conditional on transmission between the two individuals having taken place.

`max_n` Maximum number of ambiguous nucleotides tolerated in a sequence counted at positions in the sequence data for which there is a polymorphism. This parameter deals with a case of a sequence of generally high quality in which the missing coverage of the genome is all at critical sites

`min_qual` Minimum sequence quality for a sequence to be included, measured as a fraction of genome coverage (e.g. 0.8 would indicate that at least 80% of the genome must have been specified by a sequence

`diagnostic` Binary flag to enable extensive diagnostic output from the function.

`hcw_default` Default probability of a health care worker being present on the ward on a given day if no location information is specified for that individual. Default is 4/7.

`pat_default` Default probability of a patient being present on the ward on a given day if no location information is specified for that individual. Default 1.

`use_all_seqs` Binary flag to use multiple sequences from an individual, rather than simply the first collected. Reports the maximum likelihood calculated across all sequences from an individual.

`symptom_uncertainty_calc` Binary flag to use a complete offset gamma distribution, specified by the parameters ucta, uctb, and ucto, to model the uncertainty in the date of onset of symptom.

Value

A data frame with the following columns

from

to

hcw_from

hcw_to

ordered_i

ordered_j

likelihood

consistency

under_threshold

Author(s)

Chris Illingworth <chris.illingworth@mrc.bsu.cam.ac.uk>, Chris Jackson <chris.jackson@mrc.bsu.cam.ac.uk>.

References

"A2B-Covid: A method for evaluating potential Covid-19 transmission events". Illingworth C., Hamilton W., Jackson C. et al. Under preparation.

See Also

plot_a2bcovid

Examples

```
## Example data supplied with the package
pat_file <- system.file("extdata", "Example_genetic_temporal_data.csv", package="a2bcovid")
hcw_loc_file <- system.file("extdata", "Example_movement_file.csv", package="a2bcovid")
ali_file <- system.file("extdata", "Example_sequences.fa", package="a2bcovid")
pat_loc_file <- system.file("extdata", "Example_pat_loc_file.csv", package="a2bcovid")

res <- a2bcovid(pat_file = pat_file, hcw_loc_file = hcw_loc_file, ali_file = ali_file,
                pat_loc_file = pat_loc_file)
plot_a2bcovid(res, hi_from="from_hcw", hi_to="to_hcw")
```

---

a2bcovid_app *Web app interface to a2bcovid*

Description

Web app interface to a2bcovid

Usage

```
a2bcovid_app(rstudio = FALSE)
```

Arguments

rstudio	Set to TRUE to open the app in the RStudio Viewer. If FALSE (the default), an external web browser is launched.

---

long_to_wide *Convert patient location data for an a2bcovid analysis from long to wide format*

Description

Convert patient location data for an a2bcovid analysis from long to wide format

Usage

```
long_to_wide(long_file)
```

Arguments

long_file	A path name to a CSV file in long format.

Value

A path name to a temporary file containing the equivalent data in wide format. This can be read with read.csv.

The names of the columns in the wide and long formats are both documented in the a2bcovid help page, argument pat_loc_file.

---

plot_a2bcovid *Plot results of an a2bcovid analysis*

Description

Plots a grid of colours indicating likelihood of transmission paths between each pair of individuals.

Usage

```
plot_a2bcovid(
  x,
  cluster = TRUE,
  hi_from = "from_hcw",
  hi_to = "to_hcw",
  hi_col = "red",
  hi_lab = NULL,
  palette = NULL,
  continuous = FALSE,
  direction = 1
)
```

Arguments

x	Data frame returned by a2bcovid.
cluster	If TRUE (the default) then the individual IDs are rearranged using a heuristic clustering method so that apparent clusters of infections appear contiguously in the plot. If FALSE then individuals are arranged in their original order in the data.
hi_from	Character string, naming a variable in the dataframe indicating "from" individual IDs to be highlighted in the plot. If not supplied, then no IDs will be highlighted.
hi_to	Character string indicating "to" individual IDs to be highlighted, similarly.
hi_col	Colour to use to highlight individual IDs.
hi_lab	Legend to describe which individuals are highlighted. By default this is "Health-care workers".
palette	Colour palette, passed to ggplot2::scale_fill_brewer(). If omitted, a default will be chosen.
continuous	If TRUE then the p-values are plotted on a continuous colour scale. Currently only implemented with the default colour palette. If FALSE (the default), then the p-values are classified into ranges and plotted as a categorical variable.
direction	Direction of colours in the brewer palettes. Defaults to 1. Change to -1 to reverse the order of colours.

Value

A **ggplot2** plot object.

See Also

a2bcovid

---

wide_to_long *Convert patient location data for an a2bcovid analysis from wide to long format*

Description

Convert patient location data for an a2bcovid analysis from wide to long format

Usage

```
wide_to_long(wide_file)
```

Arguments

wide_file	A path name to a CSV file in wide format.

Value

A path name to a temporary file containing the equivalent data in long format. This can be read with `read.csv`.

The names of the columns in the wide and long formats are both documented in the `a2bcovid` help page, argument `pat_loc_file`.

Index