Package 'bumblebee'

May 8, 2026

Title Quantify Disease Transmission Within and Between Population Groups

Version 0.1.0

Description A simple tool to quantify the amount of transmission of an infectious disease of interest occurring within and between population groups. 'bumblebee' uses counts of observed directed transmission pairs, identified phylogenetically from deep-sequence data or from epidemiological contacts, to quantify transmission flows within and between population groups accounting for sampling heterogeneity. Population groups might include: geographical areas (e.g. communities, regions), demographic groups (e.g. age, gender) or arms of a randomized clinical trial. See the 'bumblebee' website for statistical theory, documentation and examples <https://magosil86.github.io/bumblebee/>.

License MIT + file LICENSE

URL https://magosil86.github.io/bumblebee/

BugReports https://github.com/magosil86/bumblebee/issues

LazyData true

Depends R (>= 2.10)

Imports dplyr (>= 1.0.2), gtools (>= 3.8.2), Hmisc (>= 4.4-2), magrittr (>= 2.0.1), rmarkdown (>= 2.6)

Suggests covr (>= 3.5.1), knitr (>= 1.30), markdown (>= 1.1), testthat (>= 3.0.1)

Encoding UTF-8

RoxygenNote 7.1.1

Config/testthat/edition 3

VignetteBuilder knitr

NeedsCompilation no

Author Lerato E Magosi [aut] (ORCID: <https://orcid.org/0000-0002-3388-9892>), Marc Lipsitch [aut], Lerato E Magosi [cre]

Maintainer Lerato E Magosi <magosil86@gmail.com>

Repository CRAN

Date/Publication 2021-05-11 09:42:19 UTC

Contents

`counts_hiv_transmission_pairs`

Observed HIV transmission pairs

Description

Counts of directed HIV transmission pairs observed within and between intervention and control communities in the 30-community BCPP/Ya Tsie HIV prevention trial in Botswana (2013-2018). The Botswana -Ya Tsie trial was a pair-matched community randomized trial that evaluated the effect of a universal HIV test and treat intervention in reducing population-level incidence. For further details see references and: https://magosil86.github.io/bumblebee/.

Usage

```
counts_hiv_transmission_pairs
```

Format

A data frame:

H1_group Name of population group 1

H2_group Name of population group 1

num_linked_pairs_observed Number of observed directed transmission pairs between samples from population groups 1 and 2

Source

https://magosil86.github.io/bumblebee/

References

Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.

`estimated_hiv_transmission_flows`

Estimated HIV transmission flows

Description

Estimated HIV transmissions within and betweeen intervention and control communities in the BCPP/Ya Tsie trial population adjusted for variability in sampling.

Usage

```
estimated_hiv_transmission_flows
```

Format

A data frame:

H1_group Name of population group 1

H2_group Name of population group 2

number_hosts_sampled_group_1 Number of individuals sampled from population group 1

number_hosts_sampled_group_2 Number of individuals sampled from population group 2

number_hosts_population_group_1 Estimated number of individuals in population group 1

number_hosts_population_group_2 Estimated number of individuals in population group 2

max_possible_pairs_in_sample Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2

max_possible_pairs_in_population Number of distinct possible transmission pairs between individuals in population groups 1 and 2

num_linked_pairs_observed Number of observed directed transmission pairs between samples from population groups 1 and 2

p_hat Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked

est_linkedpairs_in_population Estimated transmission pairs between population groups 1 and 2

theta_hat Estimated transmission flows or relative probability of transmission within and between population groups 1 and 2 adjusted for sampling heterogeneity. More precisely, the conditional probability that a pair of pathogen sequences is from a specific population group pairing given that the pair is linked.

obs_trm_pairs_est_goodman Point estimate, Goodman method Confidence intervals for observed transmission pairs

obs_trm_pairs_lwr_ci_goodman Lower bound of Goodman confidence interval

obs_trm_pairs_upr_ci_goodman Upper bound of Goodman confidence interval

est_goodman Point estimate, Goodman method Confidence intervals for estimated transmission flows

lwr_ci_goodman Lower bound of Goodman confidence interval

upr_ci_goodman Upper bound of Goodman confidence interval

prob_group_pairing_and_linked Probability that a pair of pathogen sequences is from a specific population group pairing and is linked

c_hat Probability that a randomly selected pathogen sequence in one population group links to at least one pathogen sequence in another population group i.e. probability of clustering

est_goodman_cc Point estimate, Goodman method Confidence intervals with continuity correction

lwr_ci_goodman_cc Lower bound of Goodman confidence interval

upr_ci_goodman_cc Upper bound of Goodman confidence interval

est_sisonglaz Point estimate, Sison-Glaz method Confidence intervals

lwr_ci_sisonglaz Lower bound of Sison-Glaz confidence interval

upr_ci_sisonglaz Upper bound of Sison-Glaz confidence interval

est_qhurst_acswr Point estimate, Queensbury-Hurst method Confidence intervals via ACSWR r package

lwr_ci_qhurst_acswr Lower bound of Queensbury-Hurst confidence interval

upr_ci_qhurst_acswr Upper bound of Queensbury-Hurst confidence interval

est_qhurst_coinmind Point estimate, Queensbury-Hurst method Confidence intervals via CoinMinD r package

lwr_ci_qhurst_coinmind Lower bound of Queensbury-Hurst confidence interval

upr_ci_qhurst_coinmind Upper bound of Queensbury-Hurst confidence interval

lwr_ci_qhurst_adj_coinmind Lower bound of Queensbury-Hurst confidence interval adjusted

upr_ci_qhurst_adj_coinmind Upper bound of Queensbury-Hurst confidence interval adjusted

Source

https://magosil86.github.io/bumblebee/

References

Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.

estimate_c_hat estimate_c_hat *Estimates probability of clustering*

Description

This function estimates c_hat, the probability that a randomly selected pathogen sequence in one population group links to at least one pathogen sequence in another population group.

Usage

```
estimate_c_hat(df_counts_and_p_hat, ...)

## Default S3 method:
estimate_c_hat(df_counts_and_p_hat, ...)
```

Arguments

df_counts_and_p_hat	A data.frame returned by the function: estimate_p_hat()
...	Further arguments.

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1
- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2
- p_hat, Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked
- c_hat, Probability that a randomly selected pathogen sequence in one population group links to at least one pathogen sequence in another population group i.e. probability of clustering

Methods (by class)

- default: Estimates probability of clustering

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Carnegie, N.B., et al., Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Computational Biology, 2014. 10(1): p. e1003430.

See Also

See `estimate_p_hat` to prepare input data to estimate `c_hat`

Examples

```
library(bumblebee)
library(dplyr)

# Estimate the probability of clustering between individuals from two population groups of interest

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs, ?sampling_frequency and ?estimated_hiv_transmission_flows


# Load and view data
#
# The input data comprises counts of observed directed HIV transmission pairs within and
# between intervention and control communities in the BCPP/Ya Tsie trial, sampling
# information and the probability of linkage between individuals sampled from
# intervention and control communities (i.e. \code{p_hat})
#
# See ?estimate_p_hat() for details on estimating p_hat
results_estimate_p_hat <- estimated_hiv_transmission_flows[, c(1:10)]

results_estimate_p_hat

# Estimate c_hat
results_estimate_c_hat <- estimate_c_hat(df_counts_and_p_hat = results_estimate_p_hat)

# View results
results_estimate_c_hat
```

`estimate_multinom_ci` `estimate_multinom_ci` *Estimates confidence intervals for transmission flows*

Description

This function computes simultaneous confidence intervals at the 5% significance level for estimated transmission flows. Available methods for computing confidence intervals are: Goodman, Goodman with a continuity correction, Sison-Glaz and Queensbury-Hurst.

Usage

```
estimate_multinom_ci(df_theta_hat, ...)

## Default S3 method:
estimate_multinom_ci(df_theta_hat, detailed_report = FALSE, ...)
```

Arguments

`df_theta_hat`	A data.frame returned by the function: `estimate_theta_hat()`
`...`	Further arguments.
`detailed_report`	A boolean value to produce detailed output of the analysis. (Default is `FALSE`)

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1
- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2
- p_hat, Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked
- est_linkedpairs_in_population, Estimated transmission pairs between population groups 1 and 2
- theta_hat, Estimated transmission flows or relative probability of transmission within and between population groups 1 and 2 adjusted for sampling heterogeneity. More precisely, the conditional probability that a pair of pathogen sequences is from a specific population group pairing given that the pair is linked.
- obs_trm_pairs_est_goodman, Point estimate, Goodman method Confidence intervals for observed transmission pairs

- obs_trm_pairs_lwr_ci_goodman, Lower bound of Goodman confidence interval
- obs_trm_pairs_upr_ci_goodman, Upper bound of Goodman confidence interval
- est_goodman, Point estimate, Goodman method Confidence intervals for estimated transmission flows
- lwr_ci_goodman, Lower bound of Goodman confidence interval
- upr_ci_goodman, Upper bound of Goodman confidence interval

The following additional fields are returned if the detailed_report flag is set

- est_goodman_cc, Point estimate, Goodman method Confidence intervals with continuity correction
- lwr_ci_goodman_cc, Lower bound of Goodman confidence interval
- upr_ci_goodman_cc, Upper bound of Goodman confidence interval
- est_sisonglaz, Point estimate, Sison-Glaz method Confidence intervals
- lwr_ci_sisonglaz, Lower bound of Sison-Glaz confidence interval
- upr_ci_sisonglaz, Upper bound of Sison-Glaz confidence interval
- est_qhurst_acswr, Point estimate, Queensbury-Hurst method Confidence intervals via ACSWR r package
- lwr_ci_qhurst_acswr, Lower bound of Queensbury-Hurst confidence interval
- upr_ci_qhurst_acswr, Upper bound of Queensbury-Hurst confidence interval
- est_qhurst_coinmind, Point estimate, Queensbury-Hurst method Confidence intervals via CoinMinD r package
- lwr_ci_qhurst_coinmind, Lower bound of Queensbury-Hurst confidence interval
- upr_ci_qhurst_coinmind, Upper bound of Queensbury-Hurst confidence interval
- lwr_ci_qhurst_adj_coinmind, Lower bound of Queensbury-Hurst confidence interval adjusted
- upr_ci_qhurst_adj_coinmind, Upper bound of Queensbury-Hurst confidence interval adjusted

Methods (by class)

- `default`: Estimates confidence intervals for transmission flows

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Goodman, L. A. On Simultaneous Confidence Intervals for Multinomial Proportions Technometrics, 1965. 7, 247-254.
3. Cherry, S., A Comparison of Confidence Interval Methods for Habitat Use-Availability Studies. The Journal of Wildlife Management, 1996. 60(3): p. 653-658.
4. Sison, C.P and Glaz, J. Simultaneous confidence intervals and sample size determination for multinomial proportions. Journal of the American Statistical Association, 1995. 90:366-369.

5. Glaz, J., Sison, C.P. Simultaneous confidence intervals for multinomial proportions. Journal of Statistical Planning and Inference, 1999. 82:251-262.
6. May, W.L., Johnson, W.D. Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells. Journal of Statistical Software, 2000. 5(6). Paper and code available at https://www.jstatsoft.org/v05/i06.
7. Carnegie, N.B., et al., Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Computational Biology, 2014. 10(1): p. e1003430.
8. Ratmann, O., et al., Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nature Communications, 2019. 10(1): p. 1411.
9. Wymant, C., et al., PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Molecular Biology and Evolution, 2017. 35(3): p. 719-733.

See Also

See estimate_theta_hat to prepare input data to estimate confidence intervals.

```
To learn more about the Goodman and Sison-Glaz confidence interval methods
see \code{\link[DescTools]{MultinomCI}}. For Queensbury-Hurst confidence
intervals see \code{\link[ACSWR]{QH_CI}} and \code{\link[CoinMinD]{QH}}
```

Examples

```
library(bumblebee)
library(dplyr)

# Compute confidence intervals for estimated transmission flows

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs and ?sampling_frequency


# Load and view data
#
# The data comprises counts of observed directed HIV transmission pairs between individuals
# sampled from intervention and control communities (i.e. num_linked_pairs_observed);
# and the estimated HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie trial population adjusted for sampling heterogneity
# (i.e. \code{est_linkedpairs_in_population}). See ?estimate_theta_hat() for details on
# computing \code{est_linkedpairs_in_population} and \code{theta_hat}.

results_estimate_theta_hat <- estimated_hiv_transmission_flows[, c(1:13)]

results_estimate_theta_hat

# Compute Goodman confidence intervals (Default)
results_estimate_multinom_ci <- estimate_multinom_ci(
    df_theta_hat = results_estimate_theta_hat,
```

```
    detailed_report = FALSE)

# View results
results_estimate_multinom_ci

# Compute Goodman, Sison-Glaz and Queensbury-Hurst confidence intervals
results_estimate_multinom_ci_detailed <- estimate_multinom_ci(
    df_theta_hat = results_estimate_theta_hat,
    detailed_report = TRUE)

# View results
results_estimate_multinom_ci_detailed
```

`estimate_prob_group_pairing_and_linked` `estimate_prob_group_pairing_and_linked` *Estimates joint probability of linkage*

Description

This function computes the joint probability that a pair of pathogen sequences is from a specific population group pairing and linked.

Usage

```
estimate_prob_group_pairing_and_linked(
  df_counts_and_p_hat,
  individuals_population_in,
  ...
)

## Default S3 method:
estimate_prob_group_pairing_and_linked(
  df_counts_and_p_hat,
  individuals_population_in,
  verbose_output = FALSE,
  ...
)
```

Arguments

`df_counts_and_p_hat`	A data.frame returned by function: `estimate_p_hat()`
`individuals_population_in`	A numeric vector of the estimated number of individuals per population group
`...`	Further arguments.
`verbose_output`	A boolean value to display intermediate output. (Default is `FALSE`)

Details

For a population group pairing (u, v), the joint probability that a pair is from groups (u, v) and is linked is computed as

$$(N_u v / N_c hoose_2) * p_h at_u v,$$

where,

- N_uv = N_u * N_v: maximum distinct possible (u, v) pairs in population
- p_hat_uv: probability of linkage between two individuals randomly sampled from groups u and v
- N choose 2 or (N * (N - 1))/2 : all distinct possible pairs in population.

See bumblebee website for more details https://magosil86.github.io/bumblebee/.

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1
- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2
- p_hat, Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked
- prob_group_pairing_and_linked, Probability that a pair of pathogen sequences is from a specific population group pairing and is linked

Methods (by class)

- `default`: Estimates joint probability of linkage

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Carnegie, N.B., et al., Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Computational Biology, 2014. 10(1): p. e1003430.

See Also

See estimate_p_hat to prepare input data to estimate prob_group_pairing_and_linked

Examples

```
library(bumblebee)
library(dplyr)

# Estimate joint probability that a pair is from a specific group pairing and linked

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs and ?sampling_frequency


# Load and view data
#
# The input data comprises counts of observed directed HIV transmission pairs
# within and between intervention and control communities in the BCPP/Ya Tsie
# trial, sampling information and the probability of linkage between individuals
# sampled from intervention and control communities (i.e. \code{p_hat})
#
# See ?estimate_p_hat() for details on estimating p_hat
results_estimate_p_hat <- estimated_hiv_transmission_flows[, c(1:10)]

results_estimate_p_hat

# Estimate prob_group_pairing_and_linked
results_prob_group_pairing_and_linked <- estimate_prob_group_pairing_and_linked(
    df_counts_and_p_hat = results_estimate_p_hat,
    individuals_population_in = sampling_frequency$number_population)

# View results
results_prob_group_pairing_and_linked
```

---

estimate_p_hat — estimate_p_hat *Estimates probability of linkage between two individuals*

Description

This function computes the probability that pathogen sequences from two individuals randomly sampled from their respective population groups (e.g. communities) are linked.

Usage

```
estimate_p_hat(df_counts, ...)

## Default S3 method:
estimate_p_hat(df_counts, ...)
```

Arguments

df_counts	A data.frame returned by the function: prep_p_hat()
...	Further arguments.

Details

For a population group pairing (u, v), p_hat is computed as the fraction of distinct possible pairs between samples from groups u and v that are linked. Note: The number of distinct possible (u, v) pairs in the sample is the product of sampled individuals in groups u and u. If $u = v$, then the distinct possible pairs is the number of individuals sampled in population group u choose 2. See bumblebee website for more details https://magosil86.github.io/bumblebee/.

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1
- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2
- p_hat, Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked

Methods (by class)

- default: Estimates probability of linkage between two individuals

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Carnegie, N.B., et al., Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Computational Biology, 2014. 10(1): p. e1003430.

See Also

See prep_p_hat to prepare input data to estimate p_hat

Examples

```
library(bumblebee)
library(dplyr)

# Estimate the probability of linkage between two individuals randomly sampled from
# two population groups of interest.

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs and ?sampling_frequency

# Prepare input to estimate p_hat

# View counts of observed directed HIV transmissions within and between intervention
# and control communities
counts_hiv_transmission_pairs

# View the estimated number of individuals with HIV in intervention and control
# communities and the number of individuals sampled from each
sampling_frequency

results_prep_p_hat <- prep_p_hat(group_in = sampling_frequency$population_group,
                                 individuals_sampled_in = sampling_frequency$number_sampled,
                           individuals_population_in = sampling_frequency$number_population,
                                    linkage_counts_in = counts_hiv_transmission_pairs,
                                    verbose_output = FALSE)

# View results
results_prep_p_hat

# Estimate p_hat
results_estimate_p_hat <- estimate_p_hat(df_counts = results_prep_p_hat)

# View results
results_estimate_p_hat
```

`estimate_theta_hat` `estimate_theta_hat` *Estimates conditional probability of linkage (transmission flows)*

Description

This function estimates `theta_hat`, the relative probability of transmission within and between population groups accounting for variable sampling rates among population groups. This relative probability is also refferred to as transmission flows.

Usage

```
estimate_theta_hat(df_counts_and_p_hat, ...)

## Default S3 method:
estimate_theta_hat(df_counts_and_p_hat, ...)
```

Arguments

df_counts_and_p_hat	A data.frame returned by the function: estimate_p_hat()
...	Further arguments.

Details

For a population group pairing (u, v), the estimated transmission flows within and between population groups u and v, are represented by the vector theta_hat,

$$\hat{\theta} = (\hat{\theta}_{uu}, \hat{\theta}_{uv}, \hat{\theta}_{vu}, \hat{\theta}_{vv}),$$

and are computed as

$$\hat{\theta_{ij}} = Pr(pairfromgroups(i,j)|pairislinked), wherei = u, vandj = u, v,$$

$$\hat{\theta_{ij}} = \frac{N_{ij}p_{ij}}{\sum_m \sum_{n \geq m} N_{mn}p_{mn}}, wherei = u, vandj = u, v,$$

See bumblebee website for more details https://magosil86.github.io/bumblebee/.

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1
- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2
- p_hat, Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked

- est_linkedpairs_in_population, Estimated transmission pairs between population groups 1 and 2
- theta_hat, Estimated transmission flows or relative probability of transmission within and between population groups 1 and 2 adjusted for sampling heterogeneity. More precisely, the conditional probability that a pair of pathogen sequences is from a specific population group pairing given that the pair is linked.

Methods (by class)

- `default`: Estimates conditional probability of linkage (transmission flows)

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Carnegie, N.B., et al., Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Computational Biology, 2014. 10(1): p. e1003430.

See Also

See `estimate_p_hat` to prepare input data to estimate `theta_hat`

Examples

```
library(bumblebee)
library(dplyr)

# Estimate transmission flows within and between population groups accounting for variable
# sampling among population groups

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs and ?sampling_frequency


# Load and view data
#
# The input data comprises counts of observed directed HIV transmission pairs within
# and between intervention and control communities in the BCPP/Ya Tsie trial,
# sampling information and the probability of linkage between individuals sampled
# from intervention and control communities (i.e. \code{p_hat})
#
# See ?estimate_p_hat() for details on estimating p_hat
results_estimate_p_hat <- estimated_hiv_transmission_flows[, c(1:10)]

results_estimate_p_hat

# Estimate theta_hat
results_estimate_theta_hat <- estimate_theta_hat(df_counts_and_p_hat = results_estimate_p_hat)
```

```
# View results
results_estimate_theta_hat
```

`estimate_transmission_flows_and_ci` `estimate_transmission_flows_and_ci` *Estimates transmission flows and corresponding confidence intervals*

Description

This function estimates transmission flows or the relative probability of transmission within and between population groups accounting for variable sampling among population groups.

Corresponding confidence intervals are provided with the following methods: Goodman, Goodman with a continuity correction, Sison-Glaz and Queensbury-Hurst.

Usage

```
estimate_transmission_flows_and_ci(
  group_in,
  individuals_sampled_in,
  individuals_population_in,
  linkage_counts_in,
  ...
)

## Default S3 method:
estimate_transmission_flows_and_ci(
  group_in,
  individuals_sampled_in,
  individuals_population_in,
  linkage_counts_in,
  detailed_report = FALSE,
  verbose_output = FALSE,
  ...
)
```

Arguments

`group_in`	A character vector indicating population groups/strata (e.g. communities, age-groups, genders or trial arms) between which transmission flows will be evaluated,
`individuals_sampled_in`	A numeric vector indicating the number of individuals sampled per population group,

`individuals_population_in`
: A numeric vector of the estimated number of individuals per population group,

`linkage_counts_in`
: A data.frame of counts of linked pairs identified between samples of each population group pairing of interest.
The data.frame should contain the following three fields:
  - H1_group (character) Name of population group 1
  - H2_group (character) Name of population group 2
  - number_linked_pairs_observed (numeric) Number of observed directed transmission pairs between samples from population groups 1 and 2

`...`
: Further arguments.

`detailed_report`
: A boolean value to produce detailed output of the analysis

`verbose_output`
: A boolean value to display intermediate output (Default is `FALSE`)

Details

Counts of observed directed transmission pairs can be obtained from deep-sequence phylogenetic data (via phyloscanner) or from known epidemiological contacts. Note: Deep-sequence data is also commonly referred to as high-throughput or next-generation sequence data. See references to learn more about phyloscanner.

The `estimate_transmission_flows_and_ci()` function is a wrapper function that calls the following functions:

1. The `prep_p_hat()` function to determine all possible combinations of the population groups/strata provided by the user. Type `?prep_p_hat()` at R prompt to learn more.
2. The `estimate_p_hat()` function to compute the probability of linkage between pathogen sequences from two individuals randomly sampled from their respective population groups. Type `?estimate_p_hat()` at R prompt to learn more.
3. The `estimate_theta_hat()` function that uses `p_hat` estimates to compute the conditional probability of linkage that a pair of pathogen sequences is from a specific population group pairing given that the pair is linked. The conditional probability, `theta_hat` represents transmission flows or the relative probability of transmission within and between population groups adjusted for variable sampling among population groups. Type `?estimate_theta_hat()` at R prompt to learn more.
4. The `estimate_multinom_ci()` function to estimate corresponding confidence intervals for the computed transmission flows.

Further to estimating transmission flows and corresponding confidence intervals the `estimate_transmission_flows_and_c` function provides estimates for:

1. `prob_group_pairing_and_linked`, the joint probability that a pair of pathogen sequences is from a specific population group pairing and linked. Type `?estimate_prob_group_pairing_and_linked()` at R prompt to learn more.
2. `c_hat`, the probability of clustering that a pathogen sequence from a population group of interest is linked to one or more pathogen sequences in another population group of interest. Type `?estimate_c_hat()` at R prompt to learn more.

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1
- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2
- p_hat, Probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked
- est_linkedpairs_in_population, Estimated transmission pairs between population groups 1 and 2
- theta_hat, Estimated transmission flows or relative probability of transmission within and between population groups 1 and 2 adjusted for sampling heterogeneity. More precisely, the conditional probability that a pair of pathogen sequences is from a specific population group pairing given that the pair is linked.
- obs_trm_pairs_est_goodman, Point estimate, Goodman method Confidence intervals for observed transmission pairs
- obs_trm_pairs_lwr_ci_goodman, Lower bound of Goodman confidence interval
- obs_trm_pairs_upr_ci_goodman, Upper bound of Goodman confidence interval
- est_goodman, Point estimate, Goodman method Confidence intervals for estimated transmission flows
- lwr_ci_goodman, Lower bound of Goodman confidence interval
- upr_ci_goodman, Upper bound of Goodman confidence interval

The following additional fields are returned if the detailed_report flag is set

- prob_group_pairing_and_linked, Probability that a pair of pathogen sequences is from a specific population group pairing and is linked
- c_hat, Probability that a randomly selected pathogen sequence in one population group links to at least one pathogen sequence in another population group i.e. probability of clustering
- est_goodman_cc, Point estimate, Goodman method Confidence intervals with continuity correction
- lwr_ci_goodman_cc, Lower bound of Goodman confidence interval
- upr_ci_goodman_cc, Upper bound of Goodman confidence interval

- est_sisonglaz, Point estimate, Sison-Glaz method Confidence intervals
- lwr_ci_sisonglaz, Lower bound of Sison-Glaz confidence interval
- upr_ci_sisonglaz, Upper bound of Sison-Glaz confidence interval
- est_qhurst_acswr, Point estimate, Queensbury-Hurst method Confidence intervals via ACSWR r package
- lwr_ci_qhurst_acswr, Lower bound of Queensbury-Hurst confidence interval
- upr_ci_qhurst_acswr, Upper bound of Queensbury-Hurst confidence interval
- est_qhurst_coinmind, Point estimate, Queensbury-Hurst method Confidence intervals via CoinMinD r package
- lwr_ci_qhurst_coinmind, Lower bound of Queensbury-Hurst confidence interval
- upr_ci_qhurst_coinmind, Upper bound of Queensbury-Hurst confidence interval
- lwr_ci_qhurst_adj_coinmind, Lower bound of Queensbury-Hurst confidence interval adjusted
- upr_ci_qhurst_adj_coinmind, Upper bound of Queensbury-Hurst confidence interval adjusted

Methods (by class)

- `default`: Estimates transmission flows and accompanying confidence intervals

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Carnegie, N.B., et al., Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Computational Biology, 2014. 10(1): p. e1003430.
3. Cherry, S., A Comparison of Confidence Interval Methods for Habitat Use-Availability Studies. The Journal of Wildlife Management, 1996. 60(3): p. 653-658.
4. Ratmann, O., et al., Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nature Communications, 2019. 10(1): p. 1411.
5. Wymant, C., et al., PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Molecular Biology and Evolution, 2017. 35(3): p. 719-733.
6. Goodman, L. A. On Simultaneous Confidence Intervals for Multinomial Proportions Technometrics, 1965. 7, 247-254.
7. Sison, C.P and Glaz, J. Simultaneous confidence intervals and sample size determination for multinomial proportions. Journal of the American Statistical Association, 1995. 90:366-369.
8. Glaz, J., Sison, C.P. Simultaneous confidence intervals for multinomial proportions. Journal of Statistical Planning and Inference, 1999. 82:251-262.
9. May, W.L., Johnson, W.D. Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells. Journal of Statistical Software, 2000. 5(6). Paper and code available at https://www.jstatsoft.org/v05/i06.

See Also

estimate_theta_hat and estimate_multinom_ci to learn more about estimation of transmission flows and confidence intervals.

Examples

```
library(bumblebee)
library(dplyr)

# Estimate transmission flows and confidence intervals

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs and ?sampling_frequency

# View counts of observed directed HIV transmissions within and between intervention
# and control communities
counts_hiv_transmission_pairs

# View the estimated number of individuals with HIV in intervention and control
# communities and the number of individuals sampled from each
sampling_frequency

# Estimate transmission flows within and between intervention and control communities
# accounting for variable sampling among population groups.

# Basic output
results_estimate_transmission_flows_and_ci <- estimate_transmission_flows_and_ci(
    group_in = sampling_frequency$population_group,
    individuals_sampled_in = sampling_frequency$number_sampled,
    individuals_population_in = sampling_frequency$number_population,
    linkage_counts_in = counts_hiv_transmission_pairs)

# View results
results_estimate_transmission_flows_and_ci

# Retrieve dataset of estimated transmission flows
dframe <- results_estimate_transmission_flows_and_ci$flows_dataset

# Detailed output
results_estimate_transmission_flows_and_ci_detailed <- estimate_transmission_flows_and_ci(
    group_in = sampling_frequency$population_group,
   individuals_sampled_in = sampling_frequency$number_sampled,
   individuals_population_in = sampling_frequency$number_population,
   linkage_counts_in = counts_hiv_transmission_pairs,
   detailed_report = TRUE)

# View results
results_estimate_transmission_flows_and_ci_detailed

# Retrieve dataset of estimated transmission flows
dframe <- results_estimate_transmission_flows_and_ci_detailed$flows_dataset
```

```
# Options:
# To show intermediate output set verbose_output = TRUE

# Basic output
results_estimate_transmission_flows_and_ci <- estimate_transmission_flows_and_ci(
    group_in = sampling_frequency$population_group,
   individuals_sampled_in = sampling_frequency$number_sampled,
   individuals_population_in = sampling_frequency$number_population,
   linkage_counts_in = counts_hiv_transmission_pairs,
   verbose_output = TRUE)

# View results
results_estimate_transmission_flows_and_ci

# Detailed output
results_estimate_transmission_flows_and_ci_detailed <- estimate_transmission_flows_and_ci(
    group_in = sampling_frequency$population_group,
   individuals_sampled_in = sampling_frequency$number_sampled,
   individuals_population_in = sampling_frequency$number_population,
   linkage_counts_in = counts_hiv_transmission_pairs,
   detailed_report = TRUE,
   verbose_output = TRUE)

# View results
results_estimate_transmission_flows_and_ci_detailed
```

`prep_p_hat` `prep_p_hat` *Prepares input data to estimate* `p_hat`

Description

This function generates variables required for estimating `p_hat`, the probability that pathogen sequences from two individuals randomly sampled from their respective population groups are linked. For a population group pairing (u, v), `prep_p_hat` determines all possible group pairings i.e. (uu, uv, vu, vv).

Usage

```
prep_p_hat(
  group_in,
  individuals_sampled_in,
  individuals_population_in,
  linkage_counts_in,
  ...
)

## Default S3 method:
```

```
prep_p_hat(
  group_in,
  individuals_sampled_in,
  individuals_population_in,
  linkage_counts_in,
  verbose_output = FALSE,
  ...
)
```

Arguments

group_in	A character vector indicating population groups/strata (e.g. communities, age-groups, genders or trial arms) between which transmission flows will be evaluated,
individuals_sampled_in	A numeric vector indicating the number of individuals sampled per population group,
individuals_population_in	A numeric vector of the estimated number of individuals per population group,
linkage_counts_in	A data.frame of counts of linked pairs identified between samples of each population group pairing of interest. The data.frame should contain the following three fields: • H1_group (character) Name of population group 1 • H2_group (character) Name of population group 2 • number_linked_pairs_observed (numeric) Number of observed directed transmission pairs between samples from population groups 1 and 2
...	Further arguments.
verbose_output	A boolean value to display intermediate output. (Default is FALSE)

Details

Counts of observed directed transmission pairs can be obtained from deep-sequence phylogenetic data (via phyloscanner) or from known epidemiological contacts. Note: Deep-sequence data is also commonly referred to as high-throughput or next-generation sequence data. See references to learn more about phyloscanner.

Value

Returns a data.frame containing:

- H1_group, Name of population group 1
- H2_group, Name of population group 2
- number_hosts_sampled_group_1, Number of individuals sampled from population group 1
- number_hosts_sampled_group_2, Number of individuals sampled from population group 2
- number_hosts_population_group_1, Estimated number of individuals in population group 1

- number_hosts_population_group_2, Estimated number of individuals in population group 2
- max_possible_pairs_in_sample, Number of distinct possible transmission pairs between individuals sampled from population groups 1 and 2
- max_possible_pairs_in_population, Number of distinct possible transmission pairs between individuals in population groups 1 and 2
- num_linked_pairs_observed, Number of observed directed transmission pairs between samples from population groups 1 and 2

Methods (by class)

- `default`: Prepares input data to estimate `p_hat`

References

1. Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.
2. Ratmann, O., et al., Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nature Communications, 2019. 10(1): p. 1411.
3. Wymant, C., et al., PHYLOSCANNER: Inferring Transmission from Within and Between-Host Pathogen Genetic Diversity. Molecular Biology and Evolution, 2017. 35(3): p. 719-733.

See Also

estimate_p_hat

Examples

```
library(bumblebee)
library(dplyr)

# Prepare input to estimate p_hat

# We shall use the data of HIV transmissions within and between intervention and control
# communities in the BCPP/Ya Tsie HIV prevention trial. To learn more about the data
# ?counts_hiv_transmission_pairs and ?sampling_frequency

# View counts of observed directed HIV transmissions within and between intervention
# and control communities
counts_hiv_transmission_pairs

# View the estimated number of individuals with HIV in intervention and control
# communities and the number of individuals sampled from each
sampling_frequency

results_prep_p_hat <- prep_p_hat(group_in = sampling_frequency$population_group,
                                 individuals_sampled_in = sampling_frequency$number_sampled,
                           individuals_population_in = sampling_frequency$number_population,
                                    linkage_counts_in = counts_hiv_transmission_pairs,
```

```
                                    verbose_output = TRUE)

# View results
results_prep_p_hat
```

`sampling_frequency` *Sampling fequency*

Description

Estimated number of individuals with HIV in intervention and control communities of the BCPP/Ya Tsie trial, and the number of individuals sampled from each for HIV viral phylogenetic analysis.

Usage

```
sampling_frequency
```

Format

A data frame:

population_group Population group

number_sampled Number of individuals sampled per population group

number_population Estimated number of individuals in each population group

Source

https://magosil86.github.io/bumblebee/

References

Magosi LE, et al., Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial. To submit for publication, 2021.

Index