Package: paneljudge (via r-universe)

June 9, 2024

Title Judge the performance of a panel of genetic markers using simulated data

Version 0.0.0.9000

Description An R package to judge the performance of a panel of genetic markers using data simulated for pairs of haploid genotypes. The data are simulated under a hidden Markov model of relatedness (described in Taylor, A.R., Jacob, P.E., Neafsey, D.E. and Buckee, C.O., 2019. Estimating relatedness between malaria parasites. Genetics, 212(4), pp.1337-1351) using allele frequency estimates provided by the user and inter-marker distances. The markers are treated as categorical random variables whose realisations (alleles) are unordered. The effective cardinalities and diversities of the markers can be computed using the input allele frequency estimates. Panel performance can be judged in terms of the root mean square error (RMSE) and confidence interval width of estimated relatedness, where relatedness is estimated under the same model used to simulate the data. At present, the examples we provide do not consider model misspecification; do not account for uncertainty around input allele frequency estimates; do not consider relatedness between pairs of haploid genotypes simulated using different allele frequencies; do not account for marker drop-out (markers that fail to produce useful data, e.g. because they a monomorphic). Otherwise stated, in the examples provided, the performance of a panel is judged in its most favourable light; it will likely perform less well in reality.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

LinkingTo Rcpp

Imports Rcpp, doParallel, doRNG, foreach

Suggests knitr, rmarkdown

VignetteBuilder knitr

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

RemoteUrl https://github.com/aimeertaylor/paneljudge

RemoteRef HEAD

RemoteSha b489f6ab4669d0d3c6ff6e3bdfa9666375889a31

Contents

`chr_lengths` *Data on chromosome lengths.*

Description

Lengths in base pairs of chromosomes Pf3D7_01_v3 to Pf3D7_14_v3 of the 3D7 Plasmodium falciparum reference genome listed on PlasmoDB (see url below).

Usage

```
chr_lengths
```

Format

A numeric vector named by the chromosome number.

Source

https://plasmodb.org/plasmo/showApplication.do

`compute_diversities` *Function to compute marker diversities*

Description

Given a matrix of marker allele frequencies, `compute_diversities` returns the diversities of $t = 1, ..., m$ markers, where m is the marker count. Each diversity is calculated as described in [1], i.e. without correcting for finite sample sizes or considering uncertainty.

Usage

```
compute_diversities(fs, warn_fs = TRUE)
```

Arguments

`fs`	Matrix of marker allele frequencies, i.e. the fts in [1]. Specifically, a m by $Kmax$ matrix, where m is the marker count and $Kmax$ is the maximum cardinality (per-marker allele count) observed over all m markers. If, for any $t = 1, ..., m$, the maximum cardinality exceeds that of the t-th marker (i.e. if $Kmax > Kt$), then all `fs[t,1:Kt]` are in (0,1] and all `fs[t,(Kt+1):Kmax]` are zero. For example, if $Kt = 2$ and $Kmax = 4$ then `fs[t,]` might look like `[0.3, 0.7, 0, 0]`.
`warn_fs`	Logical indicating if the function should return warnings following allele frequency checks.

Value

Diversities for $t = 1, \ldots, m$ markers.

References

1. Taylor, A.R., Jacob, P.E., Neafsey, D.E. and Buckee, C.O., 2019. Estimating relatedness between malaria parasites. Genetics, 212(4), pp.1337-1351.

Examples

```
compute_diversities(fs = frequencies$Colombia)
```

`compute_eff_cardinalities` *Function to compute marker effective cardinalities*

Description

Given a matrix of marker allele frequencies, `compute_eff_cardinalities` returns the effective cardinalites of $t = 1, ..., m$ markers, where m is the marker count. Effective cardinalities are per-marker allele counts that account for inequifrequenct alleles. Each effective cardinality is calculated as described in [1], i.e. without correcting for finite sample sizes or considering uncertainty.

Usage

```
compute_eff_cardinalities(fs, warn_fs = TRUE)
```

Arguments

`fs`	Matrix of marker allele frequencies, i.e. the fts in [1]. Specifically, a m by $Kmax$ matrix, where m is the marker count and $Kmax$ is the maximum cardinality (per-marker allele count) observed over all m markers. If, for any $t = 1, ..., m$, the maximum cardinality exceeds that of the t-th marker (i.e. if $Kmax > Kt$), then all `fs[t,1:Kt]` are in (0,1] and all `fs[t,(Kt+1):Kmax]` are zero. For example, if $Kt = 2$ and $Kmax = 4$ then `fs[t,]` might look like `[0.3, 0.7, 0, 0]`.
`warn_fs`	Logical indicating if the function should return warnings following allele frequency checks.

Value

Effective cardinalities for $t = 1, \ldots, m$ markers.

References

1. Taylor, A.R., Jacob, P.E., Neafsey, D.E. and Buckee, C.O., 2019. Estimating relatedness between malaria parasites. Genetics, 212(4), pp.1337-1351.

Examples

```
compute_eff_cardinalities(fs = frequencies$Colombia)
```

`compute_r_and_k_CIs` *Function to compute confidence intervals for relatedness and switch rate parameters*

Description

Given a matrix of marker allele frequencies, a vector of inter-marker distances, and estimates of the relatedness and switch rate parameters, `compute_r_and_k_CIs` returns confidence intervals around the parameter estimates. The default confidence is 95%. The intervals are approximate. They are generated using parametric bootstrap draws of the parameter estimates based on genotype calls for haploid genotype pairs simulated under the HMM described in [1] using the input parameter estimates. The quality of the approximation and compute time increases with the number of parametric bootstrap draws, which are generated in parallel using a specified number of cores.

Usage

```
compute_r_and_k_CIs(
  fs,
  ds,
  khat,
  rhat,
  confidence = 95,
  nboot = 100,
  core_count = parallel::detectCores() - 1,
  warn_fs = TRUE,
  ...
)
```

Arguments

`fs`	Matrix of marker allele frequencies, i.e. the fts in [1]. Specifically, a m by $Kmax$ matrix, where m is the marker count and $Kmax$ is the maximum cardinality (per-marker allele count) observed over all m markers. If, for any $t = 1, ..., m$, the maximum cardinality exceeds that of the t-th marker (i.e. if $Kmax > Kt$), then all `fs[t,1:Kt]` are in (0,1] and all `fs[t,(Kt+1):Kmax]` are zero. For example, if $Kt = 2$ and $Kmax = 4$ then `fs[t,]` might look like `[0.3, 0.7, 0, 0]`.
`ds`	Vector of m inter-marker distances, i.e. the dts in [1]. The t-th element of the inter-marker distance vector, `ds[t]`, contains the distance between marker t and $t+1$ such that `ds[m] = Inf`, where m is the marker count. (Note that this differs slightly from [1], where `ds[t]` contains the distance between marker $t-1$ and t). Distances between markers on different chromosomes are also considered infinite, i.e. if the chromosome of marker $t+1$ is not equal to the chromosome of the t-th marker, `ds[t] = Inf`.
`khat`	Estimate of the switch rate parameter, i.e. estimate of k in [1].
`rhat`	Estimate of the relatedness parameter, i.e. estimate of r in [1].

confidence	Confidence level (percentage) of the confidence interval (default 95%).
nboot	Number of parametric bootstrap draws from which to compute the confidence interval. Larger values provide a better approximation but prolong computation.
core_count	Number of cores to use to do computation. Set to 2 or more for parallel computation. Defaults to the number detected on the machine minus one.
warn_fs	Logical indicating if the function should return warnings following allele frequency checks.
...	Arguments to be passed to simulate_Ys and estimate_r_and_k.

Value

Confidence intervals around input switch rate parameter, k, and relatedness parameter, r.

References

1. Taylor, A.R., Jacob, P.E., Neafsey, D.E. and Buckee, C.O., 2019. Estimating relatedness between malaria parasites. Genetics, 212(4), pp.1337-1351.

Examples

```
# First, stimulate some data
simulated_Ys <- simulate_Ys(fs = frequencies$Colombia, ds = markers$distances, k = 5, r = 0.25)

# Second, estimate the switch rate parameter, k, and relatedness parameter, r
krhat <- estimate_r_and_k(fs = frequencies$Colombia, ds = markers$distances, Ys = simulated_Ys)

# Third, compute confidence intervals (CIs)
compute_r_and_k_CIs(fs = frequencies$Colombia, ds = markers$distances, khat = krhat['khat'], rhat = krhat['rhat'])
```

---

estimate_r_and_k *Function to estimate relatedness and switch rate parameters*

Description

Given a matrix of marker allele frequencies, a vector of inter-marker distances, and a matrix of genotype calls for a pair of haploid genotypes, estimate_r_and_k returns the maximum likelihood estimates of the relatedness parameter, r, and the switch rate parameter, k, under the HMM described in [1].

Usage

```
estimate_r_and_k(
  fs,
  ds,
  Ys,
  epsilon = 0.001,
```

```
  rho = 7.4 * 10^(-7),
  kinit = 50,
  rinit = 0.5,
  warn_fs = TRUE
)
```

Arguments

fs	Matrix of marker allele frequencies, i.e. the fts in [1]. Specifically, a m by $Kmax$ matrix, where m is the marker count and $Kmax$ is the maximum cardinality (per-marker allele count) observed over all m markers. If, for any $t = 1, ..., m$, the maximum cardinality exceeds that of the t-th marker (i.e. if $Kmax > Kt$), then all `fs[t,1:Kt]` are in (0,1] and all `fs[t,(Kt+1):Kmax]` are zero. For example, if $Kt = 2$ and $Kmax = 4$ then `fs[t,]` might look like `[0.3, 0.7, 0, 0]`.
ds	Vector of m inter-marker distances, i.e. the dts in [1]. The t-th element of the inter-marker distance vector, `ds[t]`, contains the distance between marker t and $t+1$ such that `ds[m] = Inf`, where m is the marker count. (Note that this differs slightly from [1], where `ds[t]` contains the distance between marker $t-1$ and t). Distances between markers on different chromosomes are also considered infinite, i.e. if the chromosome of marker $t+1$ is not equal to the chromosome of the t-th marker, `ds[t] = Inf`.
Ys	Matrix of genotypes calls for a pair of simulated haploid genotypes, i.e. the Yts of the i-th and j-th haploid genotypes in [1]. Specifically, a m by 2 matrix, where m is the marker count and each column contains a haploid genotype. For all $t = 1, ..., m$ markers, alleles are enumerated 0 to $Kt - 1$, where Kt is the cardinality (per-marker allele count) of the t-th marker. For example, if $Kt = 2$, both `Ys[t,1]` and `Ys[t,2]` are either 0 or 1.
epsilon	Genotyping error, i.e. ϵ in [1]. The genotyping error is the probability of miscalling one specific allele for another. As such, the error rate for the t-th marker, $(Kt - 1)\epsilon$, scales with Kt (the per-marker allele count, cardinality).
rho	Recombination rate, i.e. ρ in [1]. The recombination rate corresponds to the probability of a crossover per base pair. It is assumed constant across the genome under the HMM of [1]. Its default value corresponds to an average rate estimated for *Plasmodium falciparum* [2].
kinit	Switch rate parameter value used to initialise optimization of the negative log-likelihood.
rinit	Relatedness parameter value used to initialise optimization of the negative log-likelihood.
warn_fs	Logical indicating if the function should return warnings following allele frequency checks.

Value

Maximum likelihood estimates of the switch rate parameter, k, and relatedness parameter, r.

References

1. Taylor, A.R., Jacob, P.E., Neafsey, D.E. and Buckee, C.O., 2019. Estimating relatedness between malaria parasites. Genetics, 212(4), pp.1337-1351.
2. Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., Drury, E., O'Brien, J. and Rubio, V.R., 2016. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome research, 26(9), pp.1288-1299.

Examples

```
# First stimulate some data
simulated_Ys <- simulate_Ys(fs = frequencies$Colombia, ds = markers$distances, k = 5, r = 0.25)

# Second estimate the switch rate parameter, k, and relatedness parameter, r
estimate_r_and_k(fs = frequencies$Colombia, ds = markers$distances, Ys = simulated_Ys)
```

---

frequencies *Data on allele frequencies of the example GTseq panel.*

Description

A data set of allele frequencies for four countries: Colombia, French Guiana, Mali and Sengal.

Usage

```
frequencies
```

Format

Each entry of the list is a matrix, `fs` say, with $m = 126$ rows and $Kmax = 44$ variables, where m is the marker count and $Kmax$ is the maximum cardinality (per-marker allele count) observed over all m markers. If, for any $t = 1, ..., m$, the maximum cardinality exceeds that of the t-th marker (i.e. if $Kmax > Kt$), then all `fs[t,1:Kt]` are in (0,1] and all `fs[t,(Kt+1):Kmax]` are zero. For example, for PF3D7_0103600 in Colombia, $Kt = 2$ and `frequencies$Colombia["PF3D7_0103600",] = (0.687075, 0.312925, 0, ..., 0)`.

Allele.1 Frequency (numeric) of the first allele ...

Allele.44 Frequency (numeric) of the $Kmax$ allele

Source

see https://github.com/artaylor85/paneljudge/blob/master/data_raw/Process_GTseq.R

markers *Data on markers of the example GTseq panel.*

Description

A data set of marker attributes for markers pertaining to the example GTseq panel.

Usage

```
markers
```

Format

A data frame with 126 rows and 8 variables:

Amplicon_name Name (character) of the microhaplotype marker ("Amplicon" because typed using an amplicon)

Chr Chromosome (character) of the microhaplotype marker

Start First base pair (integer) of the microhaplotype marker

Stop Last base pair (integer) of the microhaplotype marker

length Length (integer) of the microhaplotype marker in base pairs

pos Mid-point (numeric) of the microhaplotype marker

chrom Chromosome (numeric) of the microhaplotype marker

distance Inter mid-point distance (numeric) between the microhaplotype marker and its subsequent neighbour

Source

see https://github.com/artaylor85/paneljudge/blob/master/data_raw/Process_GTseq.R

simulate_Ys *Function to simulate genotype calls for a pair of haploid genotypes*

Description

Given a matrix of marker allele frequencies, a vector of inter-marker distances, a relatedness parameter, and a switch rate parameter, for a pair of haploid genotypes `simulate_Ys` returns genotype calls simulated under the HMM described in [1].

Usage

```
simulate_Ys(fs, ds, k, r, epsilon = 0.001, rho = 7.4 * 10^(-7), warn_fs = TRUE)
```

Arguments

fs	Matrix of marker allele frequencies, i.e. the fts in [1]. Specifically, a m by $Kmax$ matrix, where m is the marker count and $Kmax$ is the maximum cardinality (per-marker allele count) observed over all m markers. If, for any $t = 1, ..., m$, the maximum cardinality exceeds that of the t-th marker (i.e. if $Kmax > Kt$), then all `fs[t,1:Kt]` are in (0,1] and all `fs[t,(Kt+1):Kmax]` are zero. For example, if $Kt = 2$ and $Kmax = 4$ then `fs[t,]` might look like `[0.3, 0.7, 0, 0]`.
ds	Vector of m inter-marker distances, i.e. the dts in [1]. The t-th element of the inter-marker distance vector, `ds[t]`, contains the distance between marker t and $t+1$ such that `ds[m] = Inf`, where m is the marker count. (Note that this differs slightly from [1], where `ds[t]` contains the distance between marker $t-1$ and t). Distances between markers on different chromosomes are also considered infinite, i.e. if the chromosome of marker $t+1$ is not equal to the chromosome of the t-th marker, `ds[t] = Inf`.
k	Data-generating switch rate parameter, i.e. k in [1].
r	Data-generating relatedness parameter, i.e. r in [1].
epsilon	Genotyping error, i.e. ϵ in [1]. The genotyping error is the probability of miscalling one specific allele for another. As such, the error rate for the t-th marker, $(Kt-1)\epsilon$, scales with Kt (the per-marker allele count, cardinality).
rho	Recombination rate, i.e. ρ in [1]. The recombination rate corresponds to the probability of a crossover per base pair. It is assumed constant across the genome under the HMM of [1]. Its default value corresponds to an average rate estimated for *Plasmodium falciparum* [2].
warn_fs	Logical indicating if the function should return warnings following allele frequency checks.

Value

Simulated genotype calls for a pair of haploid genotypes, i.e. the Yts of the i-th and j-th haploid genotypes in [1]. Specifically, a m by 2 matrix, where m is the marker count and each column contains a haploid genotype. For all $t = 1, ..., m$ markers, alleles are enumerated 0 to $Kt - 1$, where Kt is the cardinality (per-marker allele count) of the t-th marker. For example, if $Kt = 2$, both `Ys[t,1]` and `Ys[t,2]` are either 0 or 1.

References

1. Taylor, A.R., Jacob, P.E., Neafsey, D.E. and Buckee, C.O., 2019. Estimating relatedness between malaria parasites. Genetics, 212(4), pp.1337-1351.
2. Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., Drury, E., O'Brien, J. and Rubio, V.R., 2016. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome research, 26(9), pp.1288-1299.

Examples

```
simulate_Ys(fs = frequencies$Colombia, ds = markers$distances, k = 10, r = 0.5)
```

Index