skater: An R package for SNP-based Kinship Analysis, Testing, and Evaluation

Stephen D. Turner[1], V. P. Nagraj[1], Matthew Scholz[1], Shakeel Jessa[1], Carlos Acevedo[1], Jianye Ge[2], August E. Woerner[2], and Bruce Budowle[2]

[1]**Signature Science, LLC., Austin, TX 78759, USA.**
[2]**Center for Human Identification, Department of Microbiology, Immunology, and Genetics, University of North Texas Health Science Center, Fort Worth, TX 76107, USA.**

Abstract

Motivation: SNP-based kinship analysis with genome-wide relationship estimation and IBD segment analysis methods produces results that often require further downstream processing and manipulation. A dedicated software package that consistently and intuitively implements this analysis functionality is needed.

Results: Here we present the skater R package for **S**NP-based **k**inship **a**nalysis, **t**esting, and **e**valuation with **R**. The skater package contains a suite of well-documented tools for importing, parsing, and analyzing pedigree data, performing relationship degree inference, benchmarking relationship degree classification, and summarizing IBD segment data.

Availability: The skater package is implemented as an R package and is released under the MIT license at https://github.com/signaturescience/skater. Documentation is available at https://signaturescience.github.io/skater.

Keywords

bioinformatics, kinship, R, genealogy, SNPs, single nucleotide polymorphisms, relatedness

R version: R version 4.0.4 (2021-02-15)

skater package version: 0.1.0

Introduction

Inferring familial relationships between individuals using genetic data is a common practice in population genetics, medical genetics, and forensics. There are multiple approaches to estimating relatedness between samples, including genome-wide measures, such as those implemented in Plink [1] or KING [2], and methods that rely on identity by descent (IBD) segment detection, such as GERMLINE [3], hap-IBD [4], and IBIS [5]. Recent efforts focusing on benchmarking these methods [6] have been aided by tools for simulating pedigrees and genome-wide SNP data [7]. Analyzing results from genome-wide SNP-based kinship analysis or comparing analyses to simulated data for benchmarking have to this point required writing one-off analysis functions or utility scripts that are seldom distributed with robust documentation, test suites, or narrative examples of usage. There is a need in the field for a well-documented software package with a consistent design and API that contains functions to assist with downstream manipulation, benchmarking, and analysis of SNP-based kinship assessment methods. Here we present the skater package for **S**NP-based **k**inship **a**nalysis, **t**esting, and **e**valuation with **R**.

Methods

Implementation

The skater package provides an intuitive collection of analysis and utility functions for SNP-based kinship analysis. Functions in the package include tools for importing, parsing, and analyzing pedigree data, performing relationship degree inference, benchmarking relationship degree classification, and summarizing IBD segment data, described in full in the *Use Cases* section below. The package adheres to "tidy" data analysis principles, and builds upon the tools released under the tidyverse R ecosystem [8].

The skater package is hosted in the Comprehensive R Archive Network (CRAN) which is the main repository for R packages: `http://CRAN.R-project.org/package=skater`. Users can install skater in R by executing the following code:

```
install.packages("skater")
```

Alternatively, the development version of skater is available on GitHub at `https://github.com/signaturescience/skater`. The development version may contain new features which are not yet available in the version hosted on CRAN. This version can be installed using the `install_github()` function in the devtools package:

```
install.packages("devtools")
devtools::install_github("signaturescience/skater", build_vignettes=TRUE)
```

When installing skater, other packages which skater depends on are automatically installed, including magritr, tibble, dplyr, tidyr, readr, purrr, kinship2, corrr, rlang, and others.

Operation

Minimal system requirements for installing and using skater include R (version 3.0.0 or higher) and several tidyverse packages [8] that many R users will already have installed. Use cases are demonstrated in detail below. In summary, the skater package has functions for:

- Reading in various output files produced by commonly used tools in SNP-based kinship analysis
- Pedigree parsing, manpulation, and analysis
- Relationship degree inference
- Benchmarking and assessing relationship classification accuracy
- IBD segment analysis post-processing

A comprehensive reference for all the functions in the skater package is available at `https://signaturescience.github.io/skater/`.

Use Cases

Pedigree parsing, manipulation, and analysis

The skater package has several functions for importing, parsing, and analyzing pedigree data. Pedigrees define familial relationships in a hierarchical structure. Many genomics tools for working with pedigrees start with a .fam file, which is a tabular format with one row per individual and columns for unique IDs of the mother, father, and the family unit. The skater package contains the function `read_fam()` to read in a PLINK-formatted .fam file and another function `fam2ped()` to convert the content into a pedigree object as a nested tibble with one row per family. All pedigree processing from skater internally leverages a data structure from the kinship2 package [9]. Further functions such as `plot_pedigree()` produce a multi-page PDF drawing a diagram of the pedigree for each family, while `ped2kinpair()` produces a pairwise list of relationships between all individuals in the data with the expected kinship coefficients for each pair (see skater package vignette).

Relationship degree inference and benchmarking

The skater package includes functions to translate kinship coefficients to relationship degrees. The kinship coefficients could come from `ped2kinpair()` or other kinship estimation software.

The `dibble()` function creates a **d**egree **i**nference t**ibble**, with degrees up to the specified maximum degree resolution, expected kinship coefficient, and lower and upper inference ranges as defined in Manichaikul et al. [2]. The `kin2degree()` function infers the relationship degree given a kinship coefficient and a maximum degree resolution (e.g., 7th-degree relatives) up to which anything more distant is classified as unrelated.

Once estimated kinship is converted to degree, it may be of interest to compare the inferred degree to known degrees of relatedness. When aggregated over many relationships and inferences, this can help benchmark performance of a particular kinship analysis method. The skater package adapts a `confusion_matrix()` function from Clark [10] to provide standard contingency table metrics (e.g. sensitivity, specificity, PPV, precision, recall, F1, etc.) with a new reciprocal RMSE (R-RMSE) metric. The R-RMSE metric is defined more thoroughly in the skater package vignette and may be a preferable measure of classification accuracy when benchmarking relationship degree estimation. In many kinship benchmarking analyses, classification error is treated in a categorical manner (exact match plus or minus one degree), neglecting the true amount of sharing as a real number. Taking the reciprocal of the target and predicted degree in a typical RMSE calculation results in larger penalties for more egregious misclassifications (e.g., classifying a first-degree relative pair as second-degree) than misclassifications at more distant relationships (e.g., classifying a fourth-degree relative pair as fifth-degree).

IBD segment analysis

Tools such as hap-IBD [4], and IBIS [5] detect shared IBD segments between individuals. The skater package includes functionality to take those IBD segments, compute shared genomic centimorgan (cM) length, and converts that shared cM to a kinship coefficient. In addition to inferred segments, these functions can estimate "truth" kinship from simulated IBD segments [7]. The `read_ibd()` function reads pairwise IBD segments from IBD inference tools and from simulated IBD segments. The `read_map()` function reads in genetic map in a standard format which is required to translate the total centimorgans shared IBD to a kinship coefficient using the `ibd2kin()` function.

Summary

The skater R package provides a robust software package for data import, manipulation, and analysis tasks typically encountered when working with SNP-based kinship analysis tools. All package functions are internally documented with examples, and the package contains a vignette demonstrating usage, inputs, outputs, and interpretation of all key functions. The package contains internal tests that are automatically run with continuous integration via GitHub Actions whenever the package code is updated. The skater package is permissively licensed (MIT) and is easily extensible to accommodate outputs from new genome-wide relatedness and IBD segment methods as they become available.

Software availability

1. Software available from: `http://CRAN.R-project.org/package=skater`.
2. Source code available from: `https://github.com/signaturescience/skater`.
3. Archived source code at time of publication: **FIXME Zenodo DOI to come here**.
4. Software license: MIT License.

Author information

SDT and VPN developed the R package.

All authors contributed to method development.

SDT wrote the first draft of the manuscript.

All authors assisted with manuscript revision.

All authors read and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported in part by award 2019-DU-BX-0046 (Dense DNA Data for Enhanced Missing Persons Identification) to B.B., awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice and by internal funds from the Center for Human Identification. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

References

[1] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007. ISSN 0002-9297. doi: 10.1086/519795.

[2] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, 26(22):2867–2873, November 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq559.

[3] Alexander Gusev, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, February 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.081398.108.

[4] Ying Zhou, Sharon R. Browning, and Brian L. Browning. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *The American Journal of Human Genetics*, 106(4):426–437, April 2020. ISSN 0002-9297. doi: 10.1016/j.ajhg.2020.02.010.

[5] Daniel N. Seidman, Sushila A. Shenoy, Minsoo Kim, Ramya Babu, Ian G. Woods, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, and Amy L. Williams. Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. *American Journal of Human Genetics*, 106 (4):453–466, April 2020. ISSN 1537-6605. doi: 10.1016/j.ajhg.2020.02.012.

[6] Monica D. Ramstetter, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, Jason G. Mezey, and Amy L. Williams. Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics*, 207(1):75–82, September 2017. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.117.1122.

[7] Madison Caballero, Daniel N. Seidman, Ying Qiao, Jens Sannerud, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, Shai Carmi, and Amy L. Williams. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLOS Genetics*, 15(12):e1007979, December 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007979.

[8] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686. URL `https://doi.org/10.21105/joss.01686`.

[9] Jason P. Sinnwell, Terry M. Therneau, and Daniel J. Schaid. The kinship2 R Package for Pedigree Data. *Human heredity*, 78(2):91–93, 2014. ISSN 0001-5652. doi: 10.1159/000363105.

[10] Michael Clark. https://github.com/m-clark/confusionmatrix, January 2021.