

Genetics and Population Analysis

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

Stephen D. Turner¹, V.P. Nagraj¹, Matthew Scholz¹, Shakeel Jessa¹, Carlos Acevedo¹, Jianye Ge², August Woener², Bruce Budowle²

¹Signature Science, LLC., Charlottesville, Virginia, 22911, USA.

²University of North Texas Health Science Center, Fort Worth, Texas, 76107, USA.

To whom correspondence should be addressed. E-mail: sturner@signaturescience.com

Associate Editor: XXX

Received on XXX; revised on XXX; accepted on XXX

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 SNP-based kinship analysis, testing, and evaluation 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>.

Contact: sturner@signaturescience.com

Supplementary information: Supplementary data are available at Bioinformatics Online.

1 Introduction

Inferring familial relationships between individuals using genetic data is a common problem 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 (Purcell *et al.*, 2007) or KING (Manichaikul *et al.*, 2010), and methods that rely on identity by descent (IBD) segment detection such as GERMLINE (Gusev *et al.*, 2009), hap-IBD (Zhou *et al.*, 2020), and IBIS (Seidman *et al.*, 2020). Recent efforts focusing on benchmarking these methods (Ramstetter *et al.*, 2017; de Vries *et al.*, 2021) have been aided by tools for simulating pedigrees and genome-wide SNP data (Caballero *et al.*, 2019). 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 SNP-based kinship analysis, testing, and evaluation with R.

2 The skater package

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. The package is designed to adhere to “tidy” data analysis principles, and builds upon the tools released under the tidyverse R ecosystem (Wickham *et al.*, 2019).

2.1 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 (Sinnwell and Therneau, 2020). 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.

2.2 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 degree inference tibble, with degrees up to the specified maximum degree resolution, expected kinship coefficient, and lower and upper inference ranges as defined in Manichaikul et al. (2010). 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 treated as unrelated.

Once estimated kinship is converted to degree, it may be of interest to compare the inferred degree to truth. 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 (2021) 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 (see Supplementary Material), and may be a preferable measure of classification accuracy when benchmarking relationship degree estimation. 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).

2.3 IBD segment analysis

Tools such as hap-IBD (Zhou et al., 2020), and IBIS (Seidman et al., 2020) detect shared IBD segments between individuals. The skater package includes functionality to take those IBD segments, compute shared genomic centimorgan (cM) length, and convert that shared cM to a kinship coefficient. In addition to inferred segments, these functions can estimate "truth" kinship from simulated IBD segments (Caballero et al., 2019). 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.

3 Conclusion

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 (see Supplementary Material). 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.

Funding

UNT FOLKS: NEED YOUR INPUT HERE! The following rules should be followed:

- The sentence should begin: 'This work was supported by ...' -
- The full official funding agency name should be given, i.e. 'National Institutes of Health', not 'NIH' (full RIN-approved list of UK funding agencies)
- Grant numbers should be given in brackets as follows: '[grant number xxxx]'
- Multiple grant numbers should be separated by a comma as follows: '[grant numbers xxxx, yyyy]'
- Agencies should be separated by a semi-colon (plus 'and' before the last funding agency)
- Where individuals need to be specified for certain sources of funding the following text should be added after the relevant agency or grant number 'to [author initials]'.

An example is given here: 'This work was supported by the National Institutes of Health [AA123456 to C.S., BB765432 to M.H.]; and the Alcohol & Education Research Council [hfygr667789].'

Oxford Journals will deposit all NIH-funded articles in PubMed Central. See Depositing articles in repositories – information for authors for details. Authors must ensure that manuscripts are clearly indicated as NIH-funded using the guidelines above.

References

- Caballero, M., Seidman, D. N., Qiao, Y., Sannerud, J., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Carmi, S., and Williams, A. L. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLOS Genetics*, **15**(12), e1007979.
- Clark, M. (2021). <https://github.com/m-clark/confusionmatrix>.
- de Vries, J. H., Kling, D., Vidaki, A., Arp, P., Kalamara, V., Verbiest, M. M. P. J., Piniewska-Róg, D., Parsons, T. J., Uitterlinden, A. G., and Kayser, M. (2021). Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy. *bioRxiv*, page 2021.06.25.449870.
- Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, **19**(2), 318–326.
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, **26**(22), 2867–2873.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.
- Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Mezey, J. G., and Williams, A. L. (2017).

- Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics*, **207**(1), 75–82.
- Seidman, D. N., Shenoy, S. A., Kim, M., Babu, R., Woods, I. G., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., and Williams, A. L. (2020). Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. *American Journal of Human Genetics*, **106**(4), 453–466.
- Sinnwell, J. and Therneau, T. (2020). *kinship2: Pedigree Functions*. R package version 1.8.5.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686.
- Zhou, Y., Browning, S. R., and Browning, B. L. (2020). 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.