Bioinformatics, 37(11), 2021, 1595–1597
doi: 10.1093/bioinformatics/btaa951
Advance Access Publication Date: 3 November 2020
Applications Note

Genome analysis

MiRKAT: kernel machine regression-based global association tests for the microbiome

Nehemiah Wilson[1], Ni Zhao[2], Xiang Zhan[3], Hyunwook Koh[4], Weijia Fu[5], Jun Chen[6], Hongzhe Li[7], Michael C. Wu[8] and Anna M. Plantinga[1,*]

[1]Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, USA, [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA, [3]Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA, [4]Department of Applied Mathematics and Statistics, The State University of New York, Korea (SUNY Korea), Incheon 21985, South Korea, [5]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA 98121, USA, [6]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA, [7]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA and [8]Public Health Sciences Division, Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on August 1, 2020; revised on October 13, 2020; editorial decision on October 27, 2020; accepted on October 28, 2020

Abstract

Summary: Distance-based tests of microbiome beta diversity are an integral part of many microbiome analyses. MiRKAT enables distance-based association testing with a wide variety of outcome types, including continuous, binary, censored time-to-event, multivariate, correlated and high-dimensional outcomes. Omnibus tests allow simultaneous consideration of multiple distance and dissimilarity measures, providing higher power across a range of simulation scenarios. Two measures of effect size, a modified R-squared coefficient and a kernel RV coefficient, are incorporated to allow comparison of effect sizes across multiple kernels.

Availability and implementation: MiRKAT is available on CRAN as an R package.

Contact: amp9@williams.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Distance-based analysis of microbiome beta diversity is a powerful approach for detecting global associations between microbial community composition and a wide variety of phenotypes or experimental conditions, such as obesity or type 2 diabetes (Qin *et al.*, 2012; Turnbaugh *et al.*, 2009). High power is attained by avoiding stringent multiple comparison corrections, aggregating modest effect sizes and incorporating specialized features of microbiome associations such as presence/absence of rare taxa and phylogenetic relationships among taxa. This last benefit is operationalized by choosing a dissimilarity measure that encodes the desired features. Because the optimal dissimilarity is rarely known a priori and a poor choice of dissimilarity may result in drastic power loss, omnibus tests that consider multiple dissimilarities are vital. A second challenge of distance-based analysis is effect size estimation. PERMANOVA reports an R^2 statistic (Anderson, 2005), but no such measure is available for other, more flexible and computationally efficient distance-based tests.

Here, we present MiRKAT, an R package that includes distance-based tests of association for continuous, binary, censored time-to-event, multivariate, structured high-dimensional and correlated phenotypes in a kernel machine regression framework. The tests are computationally efficient due to analytical P-value calculation, and the regression framework allows flexible confounder adjustment. Omnibus tests are available for all supported outcome types. An R^2 statistic and the KRV test statistic are provided as measures of effect size; their utility and limitations are discussed below.

2 Software description and demonstration

MiRKAT comprises several kernel machine regression-based variance component score tests. Technical details are included in Supplementary Section S1. Table 1 lists all of the tests available in MiRKAT and summarizes key functionality components: whether P-values are calculated computationally using, for example, the Davies approach (Davies, 1980) or by permutation; whether an omnibus test is available; and whether measures of effect size R^2 and KRV are supported. All MiRKAT functions enable adjustment for confounders. Examples of function usage with real and simulated data are included in Supplementary Section S4.

Table 1. Tests available in MiRKAT and associated functionality

Name	Outcome type	Computational *P*-values	Omnibus test	R^2 and KRV	Reference
MiRKAT	Continuous, binary	Yes (Davies)	Yes (MinP)	Yes	Zhao *et al.* (2015)
MiRKAT-S	Time-to-event	Yes (Davies)	Yes (MinP)[a]	Yes	Plantinga *et al.* (2017)
MMiRKAT	Multivariate	Yes (Davies)	No; use KRV omnibus	Yes	Zhan *et al.* (2017b)
KRV	Structured high-dimensional	Yes (Moment matching)	Yes (Omnibus kernel)	Yes	Zhan *et al.* (2017a)
MiRKAT-R	Continuous; robust regression	Yes (Moment matching)	Yes (Omnibus kernel)	Yes	Unpublished
CSKAT	Correlated continuous	Yes (Davies)	Yes (MinP)	No	Zhan *et al.* (2018)
GLMM-MiRKAT	Correlated continuous, binary or Poisson	No	Yes (MinP)	No	Koh *et al.* (2019)

Note: Tests with computational *P*-value calculation often also provide permutation *P*-values, which may be preferred for small samples.
[a]Introduced in Koh *et al.* (2018).

2.1 Computation time

A major advantage of MiRKAT is computational efficiency. Comparing MiRKAT computation times with continuous outcomes to PERMANOVA shows that MiRKAT with Davies *P*-values is over ten times faster than PERMANOVA (Supplementary Fig. S1 and Supplementary Section S2). MiRKAT with permutation *P*-values is slightly slower than PERMANOVA for small sample sizes with a single kernel ($n \leq 100$), but much more efficient for large samples or when multiple kernels are considered due to sharing of the permutation-based null distribution across kernels.

2.2 Omnibus tests

Like other distance-based methods, MiRKAT requires the choice of a measure of dissimilarity for comparing two microbial communities. Common ecological dissimilarities include UniFrac distances, which incorporate phylogenetic relationships among taxa and may emphasize rare or common taxa, and the Bray-Curtis dissimilarity, which summarizes differences in taxon abundance without regard for phylogeny. Power is highest when the characteristics captured by the dissimilarity match those that drive the true microbiome association. Omnibus tests increase robustness by considering multiple dissimilarities simultaneously.

MiRKAT permits omnibus testing via the Cauchy combination test (Liu and Xie, 2020), a MinP procedure that uses residual permutation or the construction of a combination/omnibus kernel via a weighted linear combination of all candidate kernels as described in Zhan *et al.* (2017a).

2.3 Effect size estimation

Effect size estimation enables the researcher to evaluate the scientific importance of a result separately from its statistical significance. Among existing distance-based tests, only PERMANOVA provides a version of R^2 for effect size estimation.

MiRKAT provides an R^2 statistic and the KRV test statistic for quantification of effect sizes. For continuous outcomes, the coefficient of determination (R^2) may be calculated as $R_M^2 = \mathrm{Corr}^2(\mathbf{L}^{\mathrm{vec}}, \mathbf{K}^{\mathrm{vec}})$ where $\mathbf{L} = (\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)'$ is the cross product of the residuals under the null model, $\mathbf{K}$ is the kernel matrix for the microbiome and the superscript *vec* denotes vectorization, i.e. $\mathbf{L}^{vec} = (L_{11}, \ldots, L_{n1}, \ldots, L_{1n}, \ldots, L_{nn})$. This R^2 statistic is proportional to the MiRKAT score statistic (Zhan, 2019) and may be generalized to other univariate outcomes by using the appropriate set of residuals, or to multivariate outcomes using the outcome kernel $\mathbf{L}$ constructed in the KRV test. Effect sizes may also be quantified using the KRV test statistic $KRV(\mathbf{Y}, \mathbf{Z}) = tr(\mathbf{LK})/\{\sqrt{tr(\mathbf{LL})}\sqrt{tr(\mathbf{KK})}\}$ where $\mathbf{L}$ is a Gower-centered kernel associated with the phenotype $\mathbf{Y}$ (possibly the cross product of the residuals) and $\mathbf{K}$ is a microbiome kernel.

Comparing MiRKAT R^2 (R_M^2), the KRV statistic and PERMANOVA R^2 (R_P^2) shows that even in the presence of very strong associations, all of the R^2 and KRV estimates are small, with maximum values of approximately 0.02–0.2 (Supplementary Fig. S2 and Supplementary Section S3). The association among estimates is strong and positive, though R_M^2 is non-linearly related to the other two (Supplementary Fig. S3). Of the three estimates, only R_M^2 consistently identifies the kernel best matching that form of association as having the largest effect size.

Measures of effect size rely on a particular kernel matrix and are not available for the omnibus tests. The omnibus *P*-value can be combined with effect size estimates from individual kernels to evaluate both the strength of evidence for an association and the likely form of association.

3 Conclusion

We have developed the R package MiRKAT to perform distance-based microbiome analyses with a wide variety of phenotypes and study designs, including binary, continuous, time-to-event, high-dimensional and correlated data. The tests are computationally efficient, and they provide natural confounder adjustment due to the regression framework. Omnibus tests are available for all outcome types to maximize power under unknown forms of association. R^2 and the KRV test statistic are provided as measures of effect size.

Funding

This work was supported by the National Institutes of Health [R21AI144765 to X.Z., R01GM129512 to M.C.W.] and the National Science Foundation [1953189 to X.Z.].

Conflict of Interest: none declared.

References

Anderson,M.J. (2005) *Permutational Multivariate Analysis of Variance*, Vol. **26**. Department of Statistics, University of Auckland, Auckland, pp. 32–46.

Davies,R.B. (1980) The distribution of a linear combination of chi-2 random variables. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, **29**, 323–333.

Koh,H. *et al.* (2018) A highly adaptive microbiome-based association test for survival traits. *BMC Genomics*, **19**, 210.

Koh,H. *et al.* (2019) A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. *Front. Genet.*, **10**, 458.

Liu,Y. and Xie,J. (2020) Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.*, **115**, 393–402.

Plantinga,A. *et al.* (2017) MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome*, **5**, 17.

Qin,J. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Zhan,X. (2019) Relationship between MiRKAT and coefficient of determination in similarity matrix regression. *Processes*, **7**, 79.

Zhan,X. *et al.* (2017a) A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, **73**, 1453–1463.

Zhan,X. *et al.* (2017b) A small-sample multivariate kernel machine test for microbiome association studies. *Genet. Epidemiol.*, **41**, 210–220.

Zhan,X. *et al.* (2018) A small-sample kernel association test for correlated data with application to microbiome association studies. *Genet. Epidemiol.*, **42**, 772–782.

Zhao,N. *et al.* (2015) Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.*, **96**, 797–807.