MiRKAT Package

Ni Zhao

March 13, 2015

1 Overview

MiRKAT package has functions to test an association between a microbiome community and continuous/binary phenotypes via a kernel metric, where the kernel can be constructed using phylogenetic or non-phylogenetic distance metrics.

2 Required packages

CompQuadForm and a modified version of BiasedUrn are required for MiRKAT. The source file for CompQuadForm can be downloaded from http://cran.r-project.org/web/packages/CompQuadForm/index.html and the source file for modified BiasedUrn can be downloaded from the same website as MiRKAT as http://research.fhcrc.org/wu/en/software.html. In this vignette, we used the throat data from R package GUniFrac for demonstration. The kernels are constructed using the family of UniFrac distance. However, kernels can be constructed from other distances or directly (such as linear kernel) as well.

3 Testing an association between microbiome composition and phynotypes

3.1 Example Dataset

The throat data in GUniFrc contains 60 subjects with 28 smokers and 32 non-smokers. Microbiome data were collected from right and left nasopharynx and oropharynx region to form an OTU table with 856 OTUs. We want to evaluate whether smoking can affect the microbiome composition in the upper respiratory tract, taking into consideration additional covariates including gender and antibiotic use within 3 months.

```
> library(MiRKAT)
> library(GUniFrac)
> data(throat.tree)
```

```
> data(throat.otu.tab)
> data(throat.meta)
> attach(throat.meta)
```

3.2 Prepare the data

```
> set.seed(123)
> Male = (Sex == "Male")**2
> Smoker =(SmokingStatus == "Smoker") **2
> anti =  (AntibioticUsePast3Months_TimeFromAntibioticUsage != "None")^2
> cova = cbind(Male,  anti)
```

3.3 Create the UniFrac Distances

```
> otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff
> unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs
> D.weighted = unifracs[,,"d_1"]
> D.unweighted = unifracs[,,"d_UW"]
> D.BC= as.matrix(vegdist(otu.tab.rff , method="bray"))
```

3.4 Convert Distances to kernel matrices

```
> K.weighted = D2K(D.weighted)
> K.unweighted = D2K(D.unweighted)
> K.BC = D2K(D.BC)
```

3.5 Testing using a single kernel

```
> MiRKAT(y = Smoker, Ks = K.weighted, X = cbind(Male, anti),out_type = "D",
+        method = "davies")

[1] 0.00475982
```

"Method" indicates which method to use to compute kenel specific p-value. "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq. We adopt an exact variance component tests because most of the studies concerning microbiome compositions have modest sample size. "permutation" represents a residual permutation approach with nperm number of permutations. "moment" represents an approximation method that matches the first two moments. When out_type = "C" (continuous outcome y), the "moment" method is the Satterthwaite approaximation. When out_type = "D" (dichotomous outcome), the "moment" method is the small sample adjustment in Lee et al (2012). When sample size is modest($n < 100$ for continuous or $n < 200$ for dichotomous outcome), the "moment" method can be inflated at very small size (such as $\alpha = 0.001$), although the type I error at $\alpha = 0.05$ is usually sustained. Therefore, we suggest using "davies" or permutation approach for such situations.

One thing to note is that the "method" only concerns with the way that a kernel specific p-value is produced.

3.6 Testing using multiple kernels

```
> Ks = list(K.weighted, K.unweighted, K.BC)
> MiRKAT(y = Smoker, Ks = Ks, X = cbind(Male, anti),out_type = "D" ,
+        nperm = 10000, method = "davies")

$indivP
[1] 0.004759820 0.014192003 0.002044791

$omnibus_p
[1] 0.0031
```

This function outputs p-values for associatiion using each single kernel and an omnibus p-value considering all three kernels. The omnibus p-value is obtained through residual permutation where the minimum p-values from each of the individual tests are used as test statistics.

4 Reference

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). under revision.

Davies R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables, *Journal of the Royal Statistical Society. Series C* , 29, 323-333.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biom. Bull. 2, 110-114.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (2012) Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91, 224-237.

Zhou., J. J. and Zhou, H.(2015) Powerful Exact Variance Component Tests for the Small Sample Next Generation Sequencing Studies (eVCTest), in submission.