

# K-shuff User Guide

---

**NOTE:** This version of K-shuff is still being developed, and we thank you for agreeing to test it. Please let us know what you think and if any problems come up. Please **DO NOT** distribute this program without our permission. If you know someone who would like to use it, please ask them to contact Dr. Kamlesh Jangid ([jangidk@uga.edu](mailto:jangidk@uga.edu), [jangidk@gmail.com](mailto:jangidk@gmail.com)) or Dr. Ming-Hung (Jason) Kao ([mhkao@math.asu.edu](mailto:mhkao@math.asu.edu), [jason.mhk@gmail.com](mailto:jason.mhk@gmail.com)).

## Contents

- [Introduction](#)
- [The theory behind K-shuff](#)
- [K-functions](#)
- [Test Statistics](#)
- [Permutations Tests](#)
- [Example Dataset](#)
- [Installation instructions](#)
- [How to run K-shuff](#)
- [Output Files](#)
- [Interpretation](#)
- [FAQs](#)
- [References](#)

## Introduction

K-shuff is a powerful computer program designed to identify spatial clustering in a given dataset based on the reduced second moment measure, or  $K$ -function (Diggle et al., 2007; Diggle and Chetwynd, 1991). In essence, K-shuff can be adapted for comparing any data from two (or more) samples to understand their relationship with each other. As an example, we adapt this technique to compare 16S rRNA gene sequence libraries from different environmental samples by treating gene sequences as points in space with hundreds of dimensions. To do the analysis, the program requires a PHYLIP formatted (<http://evolution.genetics.washington.edu/phylip.html>) distance matrix and a control file which contains the parameter settings for the run. K-shuff allows the user to determine the presence of both structural as well as compositional differences between the libraries (as explained below). In addition, K-shuff is capable of performing multiple comparisons in a single run.



## The theory behind K-shuff

The  $K$ -function is defined on a distance measure. In our context, it is set to the evolutionary distances among gene sequences. Denoting the evolutionary distance between sequence  $i$  and  $j$  by  $d_{ij}$ , the  $K$ -function for a population of size  $N$  is defined as

$$K(r) = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{i \neq j}^N I(d_{ij} \leq r)$$

Where  $I(E)$  is an indicator function (=1, when  $E$  is true; =0, otherwise), and  $r$  is any positive number on the real line. In essence, the  $K$ -function is the empirical cumulative distribution function of the evolutionary distances. In other words, for every value of evolutionary distance  $r$  in the distance matrix,  $K(r)$  is the fraction of  $d_{ij}$  values less than or equal to  $r$ .

## K-functions

We define two types of  $K$ -functions, namely intra- $K$ -functions (or IKFs) and cross- $K$ -functions (or CKFs), for testing two different aspects regarding the homogeneity among compared libraries. In essence, both the IKF and the CKF measure the genetic diversity. The only difference lies in the composition of the underlying library(ies) considered by these  $K$ -functions (Figure 1). While the IKF describes the richness or evenness of genetic diversity within each library, the CKF provides a natural measurement for the dissimilarity or homogeneity in the membership or composition between paired libraries. In addition, the difference in the IKFs among libraries signifies the difference in the genetic diversity (also referred to as the structural difference hereinafter).

Library	1	1	1	1	1	1	2	2	2	2	2	2	.....
1	--	11	11	11	11	11	12	12	12	12	12	12	.....
1	11	--	11	11	11	11	12	12	12	12	12	12	.....
1	11	11	--	11	11	11	12	12	12	12	12	12	.....
1	11	11	11	--	11	11	12	12	12	12	12	12	.....
1	11	11	11	11	--	11	12	12	12	12	12	12	.....
1	11	11	11	11	11	--	12	12	12	12	12	12	.....
2	21	21	21	21	21	21	--	22	22	22	22	22	.....
2	21	21	21	21	21	21	22	--	22	22	22	22	.....
2	21	21	21	21	21	21	22	22	--	22	22	22	.....
2	21	21	21	21	21	21	22	22	22	--	22	22	.....
2	21	21	21	21	21	21	22	22	22	22	--	22	.....
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 1: Hypothetical distance matrices illustrating the calculation of intra- $K$ -function (IKF) and cross- $K$ -function (CKF). The IKF is calculated from the comparison within each library, i.e., library 1 to itself (yellow cells), library 2 to itself (blue cells) and so on. The CKF is calculated from comparisons between libraries, i.e., library 1 to library 2 (green cells). In this case, the comparison of library 1 to 2 (green cells) is the same as the comparison of library 2 to 1 (orange cells).

In the figure below, two illustrative intra- $K$ -functions are presented. The Sea Water (SW) population represented by the left-most  $K$ -function is of the smallest diversity since a large proportion of the gene sequences are closely related to each other as represented by the steep increase in the proportion of sequences within a relatively small evolutionary distance. In contrast, a relatively large diversity exists in the Marsh Sediment (MS) populations as represented by the right-most  $K$ -function. In

$K$ -shuff, the area above the curve is a measurement of the diversity of the library. Hence, SW (Blue Area) has smaller diversity compared to MS (Blue+Green Area). Moreover, the green area in Figure 2 represents the difference in the genetic diversity between SW and MS.

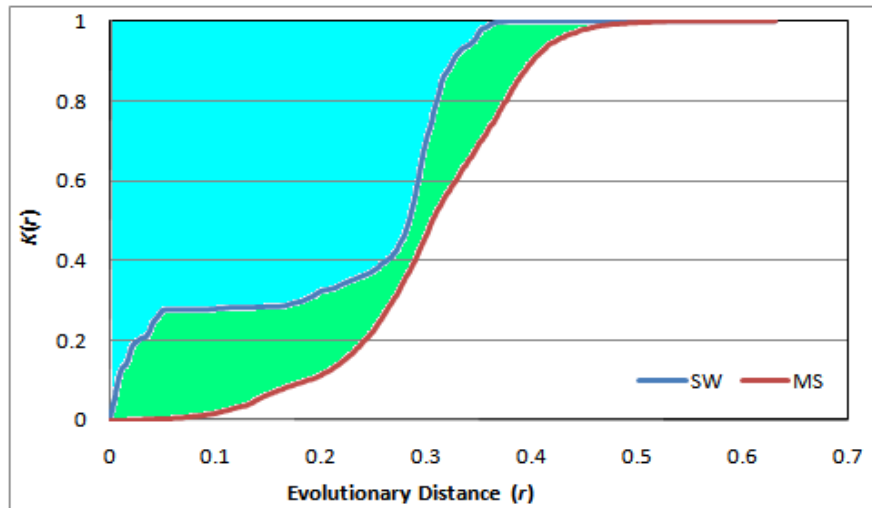


Figure 2: Two examples of intra- $K$ -functions: the left-most Sea Water (SW) population indicates low diversity, whereas the right-most Marsh Sediment (MS) population has high diversity. As illustrated, when the IKF is ‘steep’ over a small distance, the genetic diversity is small within the corresponding library.

In contrast, the CKFs signify the compositional difference between libraries; i.e. the difference in the membership. When the genetic diversity measured between libraries (by CKFs) is ‘larger’ than that measured within libraries (by IKFs), we tend to believe that the compositions of the libraries are different. As illustrated in Figure 3, the sum of the area between each IKF curve and the CKF is a measure of the compositional differences between the libraries. In other words, the area between  $IKF_{SW}$  (blue curve) and CKF (green curve), plus that between  $IKF_{MS}$  (red curve) and CKF (green curve) represents the compositional difference between these two communities.

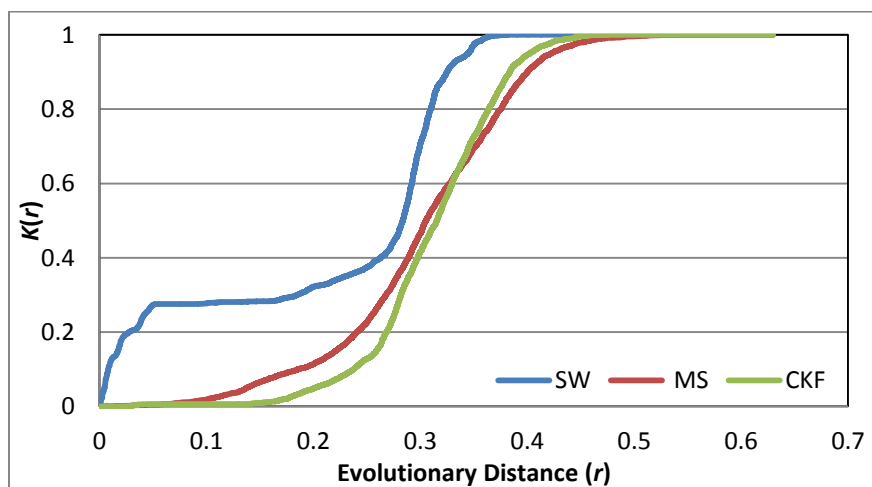


Figure 3: Measurement of structural and compositional differences between communities in  $K$ -shuff using the two  $K$ -functions, namely intra- $K$ -functions (or IKFs) and cross- $K$ -functions (or CKFs), respectively. A CKF, which is ‘flat’ at the beginning, indicates the dissimilarity in the composition between the two corresponding libraries.

## Test Statistics

We propose two test statistics for detecting 1) the disparities in the genetic diversity within clone libraries (structural difference) and 2) the diversity in the membership between each pair of libraries (compositional difference). With no plausible underlying statistical distribution, inferences are made based on the Monte Carlo testing procedure.

For detecting the structural difference, we consider the test statistic,  $T_s$ . For libraries of the same genetic diversity, their IKFs are similar and the  $T_s$  is small. On the other hand, large  $T_s$  is an indication for structural differences.

For testing the compositional difference, we propose  $T_c$ . This statistics compares the CKF with the averaged IKF, which amounts to comparing the genetic diversity “between” libraries to that “within” libraries. When the composition between the two libraries is different, the diversity between would be larger than that within. In such a case  $T_c$  would be large and the null hypothesis of no compositional difference tends to be rejected.

## Permutation Tests

The probability distribution of the  $K$ -function is given by randomly choosing gene sequences from the population. Working under this method, the moments of the  $K$ -function can be derived, and hence, the moments of the test statistics. However, the complicity of these moments hinders their use. In contrast, the permutation method, which renders exact inferences and is easy to implement, is more suitable for this application. We thus resort to this powerful tool.

With no plausible distributions of our test statistics, permutation method is a natural choice. The homogeneity among libraries ensures the exchangeability of the gene sequences. As the exchangeability holds, the distributions of our test statistics are invariant under the permutations of the sequences among libraries. Inferences can thus be made based on these distributions stemmed from permutations.

## Example Dataset

The WSCF-dataset, originally reported by Jangid et al. (2008) and Lasher et al. (2009) is available for download for learning purposes from here (<http://whitman.myweb.uga.edu/K-shuff/WSCF-dataset.zip>).

The dataset contains the following files:

1. **WSCF.dist**: A PHYLIP-formatted distance matrix for 328 sequences comprising of 82 sequences for each of the following four libraries:
  - Sea Water (Library Code= SW, from Lasher et al., 2009)
  - Marsh Sediments (Library Code= MS, from Lasher et al., 2009)
  - Soil from conventionally tilled Cropland (Library Code= SC, from Jangid et al., 2008)
  - Soil from Forest (Library Code= SF, from Jangid et al., 2008).

These libraries were chosen to be of the same size by random selection from larger libraries.

2. **ControlFile.txt**: The control file containing the parameter settings to be used to perform  $K$ -shuff analysis.



## Installation Instructions

At this time, *K*-shuff is only available as a Windows executable and as a FORTRAN based source code for Linux. You can download these files from the *K*-shuff homepage at <http://whitman.myweb.uga.edu/K-shuff.html>

**For Windows PC**, simply download, save and double click on the executable file to run the program.

**For Linux users**, follow the steps given below:

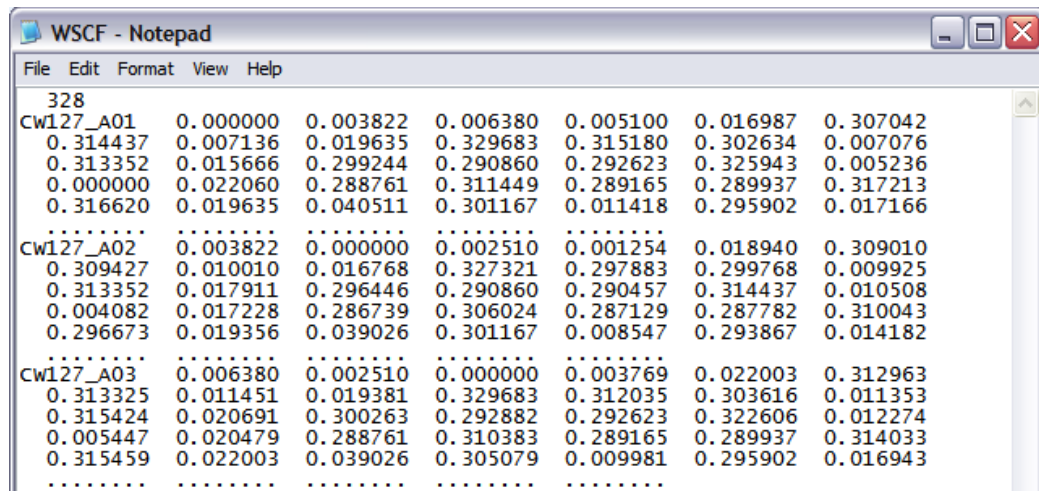
1. Compile *K*shuffMar2010.f90 by typing the following in your terminal window  
gfortran KshuffMar2010.f90 -o kshuff
2. Run *K*-shuff by typing  
./kshuff



## How to run *K*-shuff

Before you begin, we strongly recommend that you download the example dataset as explained above. Now, make sure that both the *K*-shuff program file and your input files are in the same folder. In order to run *K*-shuff, you will need two input files:

1. **PHYLIP-formatted distance matrix.** At this time, we know that PHYLIP distance matrices of upto 4000 sequences generated by both DNADist (part of the PHYLIP package) and DistEx (part of the data extraction tools available at <http://whitman.myweb.uga.edu/detools.html>) work successfully with *K*-shuff. In addition, we have tested that sequence identifiers of unequal character length also work with *K*-shuff without any problem. The WSCF.distfile from the trial dataset should look like this:



```
WSCF - Notepad
File Edit Format View Help
328
Cw127_A01  0.000000  0.003822  0.006380  0.005100  0.016987  0.307042
0.314437  0.007136  0.019635  0.329683  0.315180  0.302634  0.007076
0.313352  0.015666  0.299244  0.290860  0.292623  0.325943  0.005236
0.000000  0.022060  0.288761  0.311449  0.289165  0.289937  0.317213
0.316620  0.019635  0.040511  0.301167  0.011418  0.295902  0.017166
.....
Cw127_A02  0.003822  0.000000  0.002510  0.001254  0.018940  0.309010
0.309427  0.010010  0.016768  0.327321  0.297883  0.299768  0.009925
0.313352  0.017911  0.296446  0.290860  0.290457  0.314437  0.010508
0.004082  0.017228  0.286739  0.306024  0.287129  0.287782  0.310043
0.296673  0.019356  0.039026  0.301167  0.008547  0.293867  0.014182
.....
Cw127_A03  0.006380  0.002510  0.000000  0.003769  0.022003  0.312963
0.313325  0.011451  0.019381  0.329683  0.312035  0.303616  0.011353
0.315424  0.020691  0.300263  0.292882  0.292623  0.322606  0.012274
0.005447  0.020479  0.288761  0.310383  0.289165  0.289937  0.314033
0.315459  0.022003  0.039026  0.305079  0.009981  0.295902  0.016943
.....
```

Figure 4: Screenshot of input distance matrix from the WSCF-dataset.

2. **CONTROL file.** The control file is a simple text file that contains parameter settings for the run, such as the location of the distance matrix, the number of libraries in the matrix, size of each of the libraries, etc. The required parameters and the allowed options in the control file are listed below:

1. Name of the data file containing the distance matrix:  
Enter the location of the distance matrix file name.
2. Total number of libraries:  
Enter the total number of libraries that comprise the distance matrix.
3. Size of each library:  
Enter the number of sequences for each library within the input distance matrix in the order of their appearance. Each number should be separated by a TAB.
4. Name of each library (no longer than 10 characteristics for each name):  
For easy identification, what do you want to call your libraries? Each library name should be separated by a TAB. Each library name cannot be longer than 10 characters.
5. Number of permutations:  
The total number of iterations (usually 1000) you wish to run for the test of significance. Enter 1 if you do not wish to carry out any permutations.
6. Random Seed:  
Enter the number of random seeds. Enter 0 (zero) if you want the program to randomly assign the number.
7. Output test statistics:  
Do you want an output file to be created for the test statistics? Enter 0 (zero) for no, or 1 for yes.
8. Output areas of  $K$ -functions:  
Do you want an output file to be created for the sum of areas of  $K$ -functions? Enter 0 (zero) for no, or 1 for yes.
9. Output  $K$ -functions:  
Do you want output files to be created for the  $K$ -functions, IKF and CKF? Enter 0 (zero) for no, or 1 for yes. This allows plotting of the  $K$ -functions.
10. Output difference between  $K$ -functions:  
Do you want output files to be created for the difference between  $K$ -functions for the compared libraries? Enter 0 (zero) for no, or 1 for yes.

The ControlFile.txt file from the trial dataset should look like this:

```

ControlFile - Notepad
File Edit Format View Help
1. name of the data set containing the distance matrix:
WSCF.dist
2. total number of libraries:
4
3. size of each library:
82      82      82      82
4. name of each library (no longer than 10 characteristics for each name):
SW      MS      CS      FS
5. number of permutations (=1 if no permutation test):
1000
6. Random Seed (=0 if randomly assigned):
0
7. output test statistics (0=No, 1=yes):
1
8. output areas of kfunctions (0=No, 1=yes):
1
9. output K functions (0=No, 1=yes):
1
10. output difference between K functions (0=No, 1=yes):
1
Ln 1, Col 1

```

Figure 5: Screenshot of input Control File from the WSCF-dataset.

Once you are sure of the input file formats and the parameter settings, simply double click on the *K-shuff* executable in Windows OS to run the program. Linux users, please type the following in your terminal window to run *K-shuff*:

```
./kshuff
```

The following analysis window will pop-up on your screen and will be automatically updated as *K-shuff* progresses to completion:

```

C:\KshuffMar2010.exe
*****
* K-shuff v1.0
*
* Jangid, Kamlesh, University of Georgia
* Kao, Ming-Hung, Arizona State University
* Rathbun, Stephen, University of Georgia
* Whitman, William, University of Georgia
*
* Website: http://whitman.myweb.uga.edu/
*****
Intra K functions are in: intraK.txt
Cross K functions for SW are in: crossKSW.txt
Cross K functions for MS are in: crossKMS.txt
Cross K functions for CS are in: crossKCS.txt
Difference of IKFs between SW and others are in:
diff_IKF_SW.txt
Difference of IKFs between MS and others are in:
diff_IKF_MS.txt
Difference of IKFs between CS and others are in:
diff_IKF_CS.txt
Difference of CKFs between SW and others are in:
diff_CKF_SW.txt
Difference of CKFs between MS and others are in:
diff_CKF_MS.txt
Difference of CKFs between CS and others are in:
diff_CKF_CS.txt
Areas of K functions are in AreaK.txt
test statistics are presented in: teststat.txt
initiating random permutation... this might take a few minutes
random permutation started... 0%
random permutation completed... 20%
random permutation completed... 40%
random permutation completed... 60%
random permutation completed... 80%

```

Figure 6: Progress of *K-shuff* analysis for the WSCF-dataset as displayed on a Windows PC.

Upon completion, all the desired output files should have been created in the working folder and the analysis window will close automatically.



## Output files

*K-shuff* generates seven different types of output files, provided the parameters 7 through 10 in the control file are set to 1. Below, we explain the contents of each of the files and what the values mean.

### teststat.txt

This file is generated when parameter 7 in the control file is set to 1. This file gives the output for the test statistics,  $T_s$  (column 1) and  $T_c$  (column 2) to make conclusions about the structural and compositional differences between the compared libraries.

```

teststat (pairwise) = SW      Lib1      Lib2      IKF      CKF
teststat (pairwise) = SW      MS        CS        0.7973   0.6709
teststat (pairwise) = SW      CS        CS        0.6257   1.5151
teststat (pairwise) = SW      FS        FS        0.5316   2.2149
teststat (pairwise) = MS      CS        CS        0.0436   0.2335
teststat (pairwise) = MS      FS        FS        0.0649   0.4959
teststat (pairwise) = CS      FS        FS        0.0053   0.0245
teststat (overall) =          1.0341

```

Figure 7: Screenshot of output file teststat.txt generated after *K*-shuff analysis of the WSCF-dataset.

### AreaK.txt

This file is generated when parameter 8 in the control file is set to 1. This file gives differences between the communities represented as the sum of areas between the CKF and the IKF of each community (Column labeled  $|CKF-IKF|$ ). Remember, a larger sum of areas between communities indicates that the two compared communities are more distant and they share fewer members between them. The remaining two columns represent the area above the IKFs. The area between the paired IKFs can be calculated from the difference between the last two columns for each comparison.

```

Area of (1-ICKF) and (|CKF - IKF1| + |CKF - IKF2|)
|CKF-ICKF|   (1-ICKF1)   (1-ICKF2)
SW vs. MS    0.1169      0.2140     0.3018
SW vs. CS    0.1585      0.2140     0.2895
SW vs. FS    0.1897      0.2140     0.2825
MS vs. CS    0.0595      0.3018     0.2895
MS vs. FS    0.0873      0.3018     0.2825
CS vs. FS    0.0200      0.2895     0.2825

```

Figure 8: Screenshot of output file AreaK.txt generated after *K*-shuff analysis of the WSCF-dataset.

### pvalue.txt

This file is generated when parameter 5 in the control file is set to anything larger than 1. This file gives the p values for both IKF and CKF values as outputted in the AreaK.txt file above.

```

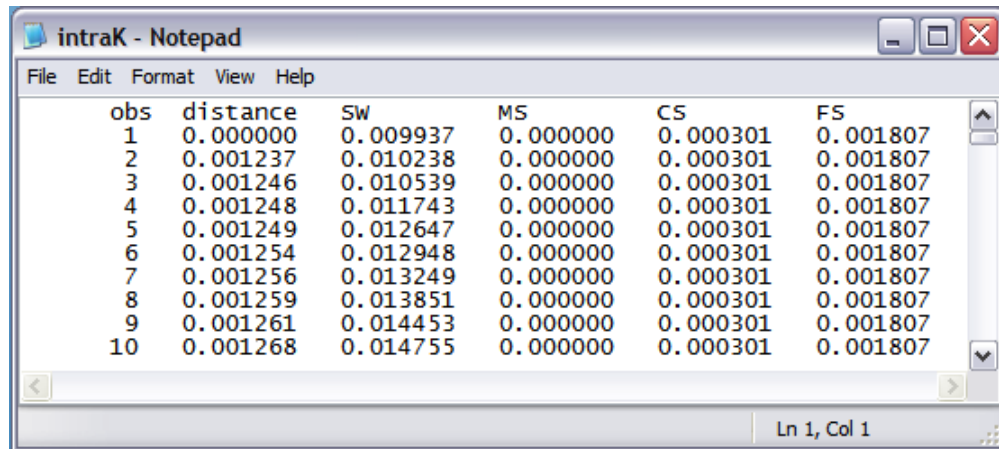
pvalue (pairwise) = Lib1      Lib2      p-value: IKF  p-value: CKF
pvalue (pairwise) = SW      MS        0.0010        0.0010
pvalue (pairwise) = SW      CS        0.0010        0.0010
pvalue (pairwise) = SW      FS        0.0010        0.0010
pvalue (pairwise) = MS      CS        0.2050        0.0010
pvalue (pairwise) = MS      FS        0.1060        0.0010
pvalue (pairwise) = CS      FS        0.8110        0.0010
pvalue (overall) =          0.0010

```

Figure 9: Screenshot of output file pvalue.txt generated after *K*-shuff analysis of the WSCF-dataset.

### IntraK.txt

This file is generated when parameter 9 in the control file is set to 1. A single output file will be generated for the IKFs for all the libraries in the data matrix file. The data in this file can be used to generate the IKF coverage curves as shown in Figure 2.

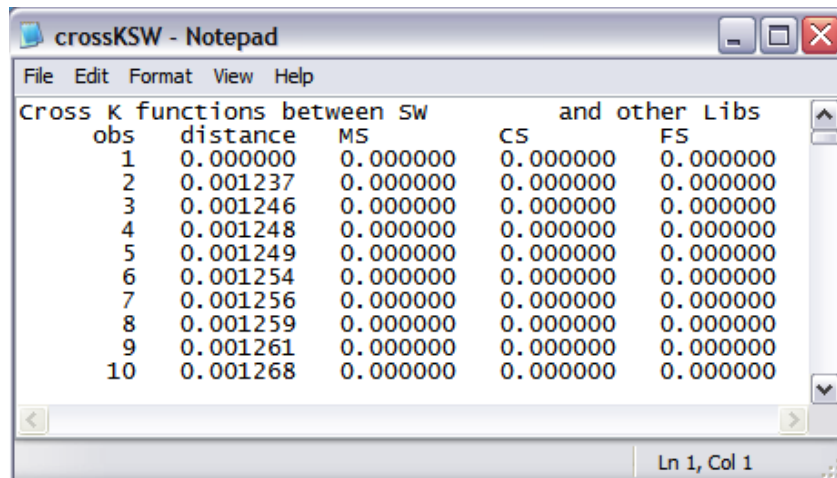


obs	distance	SW	MS	CS	FS
1	0.000000	0.009937	0.000000	0.000301	0.001807
2	0.001237	0.010238	0.000000	0.000301	0.001807
3	0.001246	0.010539	0.000000	0.000301	0.001807
4	0.001248	0.011743	0.000000	0.000301	0.001807
5	0.001249	0.012647	0.000000	0.000301	0.001807
6	0.001254	0.012948	0.000000	0.000301	0.001807
7	0.001256	0.013249	0.000000	0.000301	0.001807
8	0.001259	0.013851	0.000000	0.000301	0.001807
9	0.001261	0.014453	0.000000	0.000301	0.001807
10	0.001268	0.014755	0.000000	0.000301	0.001807

Figure 10: Screenshot of output file IntraK.txt generated after K-shuff analysis of the WSCF-dataset.

### crossKn.txt

This file is generated when parameter 9 in the control file is set to 1. A total of N-1 files (where N= total number of libraries) will be created for the CKFs and these will be named as “crossKn.txt” where n= library name. The data in this file can be used to generate the CKF coverage curve as shown in Figure 3.



Cross K functions between SW and other Libs					
obs	distance	MS	CS	FS	
1	0.000000	0.000000	0.000000	0.000000	
2	0.001237	0.000000	0.000000	0.000000	
3	0.001246	0.000000	0.000000	0.000000	
4	0.001248	0.000000	0.000000	0.000000	
5	0.001249	0.000000	0.000000	0.000000	
6	0.001254	0.000000	0.000000	0.000000	
7	0.001256	0.000000	0.000000	0.000000	
8	0.001259	0.000000	0.000000	0.000000	
9	0.001261	0.000000	0.000000	0.000000	
10	0.001268	0.000000	0.000000	0.000000	

Figure 11: Screenshot of output file crossKSW.txt generated after K-shuff analysis of the WSCF-dataset.

### diff\_IKF\_n.txt

This file is generated when parameter 10 in the control file is set to 1. A total of N-1 files (where N= total number of libraries) will be created for the difference in the IKFs of the compared libraries and these will be named as “diff\_IKF\_n.txt” where n= library name.

Difference of IKFs between SW		and other Libs		
obs	distance	MS	CS	FS
1	0.000000	0.009937	0.009636	0.008130
2	0.001237	0.010238	0.009937	0.008431
3	0.001246	0.010539	0.010238	0.008732
4	0.001248	0.011743	0.011442	0.009937
5	0.001249	0.012647	0.012346	0.010840
6	0.001254	0.012948	0.012647	0.011141
7	0.001256	0.013249	0.012948	0.011442
8	0.001259	0.013851	0.013550	0.012045
9	0.001261	0.014453	0.014152	0.012647
10	0.001268	0.014755	0.014453	0.012948

Figure 12: Screenshot of output file diff\_IKF\_SW.txt generated after K-shuff analysis of the WSCF-dataset.

### diff\_CKF\_n.txt

This file is generated when parameter 10 in the control file is set to 1. A total of N-1 files (where N= total number of libraries) will be created for the difference in CKFs between the compared libraries and these will be named as “diff\_CKF\_n.txt” where n= library name.

Difference of CKFs between SW		and other Libs		
obs	distance	MS	CS	FS
1	0.000000	0.009937	0.009636	0.008130
2	0.001237	0.010238	0.009937	0.008431
3	0.001246	0.010539	0.010238	0.008732
4	0.001248	0.011743	0.011442	0.009937
5	0.001249	0.012647	0.012346	0.010840
6	0.001254	0.012948	0.012647	0.011141
7	0.001256	0.013249	0.012948	0.011442
8	0.001259	0.013851	0.013550	0.012045
9	0.001261	0.014453	0.014152	0.012647
10	0.001268	0.014755	0.014453	0.012948

Figure 13: Screenshot of output file diff\_CKF\_SW.txt generated after K-shuff analysis of the WSCF-dataset.



## Interpretation

Based on the K-shuff analysis we carried out for the WSCF dataset, the IKF which is a measure of the genetic diversity within each library compared well with other measures of bacterial diversity (Table 1) calculated using DOTUR (Schloss & Handelsman, 2005) at an evolutionary distance of 0.03. Regardless of the diversity measure, the communities in SW are less diverse as compared to other three libraries MS, SC and SF.

Table 1. Comparison of IKF with conventional diversity indices of estuarine seawater (SW), salt marsh sediments (MS), and crop (SC) and forest soil (SF).

Diversity Index	Source of library			
	SW	MS	SC	SF
IKF	0.2140	0.3018	0.2895	0.2825
S <sup>a</sup>	21	73	70	62
N <sup>b</sup>	82	82	82	82
Shannon (H) <sup>c</sup>	2.26	4.24	4.16	3.97
H/H <sub>max</sub>	0.51	0.96	0.94	0.90
Evenness <sup>d</sup>	1.71	2.28	2.25	2.21
Simpson (1/D) <sup>e</sup>	5	332	175	87
Chao 1 <sup>f</sup>	25	375	566	530
95 % lci	21	191	231	195
95 % hci	36	645	1018	956

<sup>a</sup>Number of OTUs formed at 97 % sequence similarity using DOTUR (Schloss & Handelsman, 2005).

<sup>b</sup>Number of clones in the library

<sup>c</sup>Shannon diversity index,  $H = -\sum[(n/N)\ln(n/N)]$ ; at  $H_{max}$ ,  $n = N$ .

<sup>d</sup>Evenness =  $H/\log(S)$ . Minimum and maximum evenness values were 0 & 2.3, respectively.

<sup>e</sup>Simpson's index,  $D = \sum n(n-1)/N(N-1)$

<sup>f</sup>Chao1 =  $S + n_1^2/2n_2$ , where  $n_2$  is the number of clones that occur twice. 'lci' is the 95% lower confidence interval and 'hci' is the 95% higher confidence interval for Chao1 estimator.

In contrast, the CKF which represents the compositional differences between the four communities indicated that all four libraries had significantly different communities (Table 2). Moreover, the communities in SW were most different from those in the other three libraries, especially when compared to SF.

Table 2. Differences between the bacterial communities in seawater, salt marsh sediments and soils<sup>a</sup>.

Bacterial community	SW	MS	SC	SF
Seawater (SW)	-	0.0010	0.0010	0.0010
Salt marsh sediment (MS)	0.1169	-	0.0010	0.0010
Soil cropland (SC)	0.1585	0.0594	-	0.0010
Soil forest (SF)	0.1897	0.0873	0.0200	-

<sup>a</sup>P-values for the CKF comparisons are above the diagonal. The  $p$ -value for the overall test = 0.0010. Below the diagonal are the differences between the communities represented as the sum of the area between the CKF and IKF for each community, i.e.,  $|CKF-IKF|$ .

In addition, the CKF values below the diagonal in table 2 can be used in multi-dimensional scaling analysis to represent the extent of differences between the communities as represented in Figure 14:

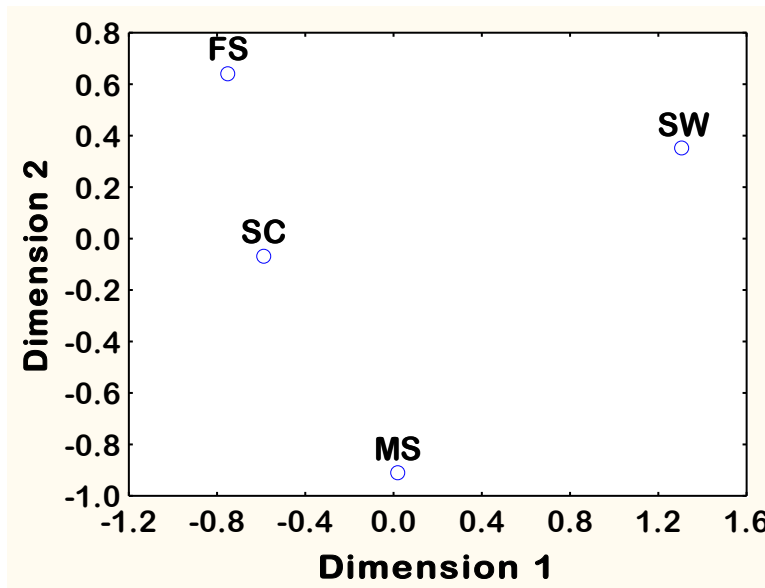


Figure 14. Multidimensional scaling analysis of the bacterial communities in WSCF-dataset based on the CKF values listed in Table 2.

From this we conclude that the composition of all four libraries is different from each other ( $p \geq 0.003$ ). Moreover, the composition of the SC and SF libraries are similar to each other than to the MS and SW libraries. Finally, the SW library is the most unlike the other libraries.



## Frequently asked questions

### Why do we ask your e-mail during the download page?

In order to notify all users of any future updates, we request that you enter your e-mail contacts on the download page. In addition, this will allow us to keep track of the number of *K*-shuff users and will help us for any future grant opportunities.

### What if I have problem running *K*-shuff?

In our experience, most errors in running *K*-shuff occur due to an incorrect input format of the distance matrix or if the sum of the number of sequences listed for each library in your control file do not match with those in the distance matrix. However, if you think the error is due to some other issue, please contact Ming-Hung (Jason) Kao ([mhkao@math.asu.edu](mailto:mhkao@math.asu.edu), [jason.mhk@gmail.com](mailto:jason.mhk@gmail.com)) and/or Kamlesh Jangid ([jangidk@uga.edu](mailto:jangidk@uga.edu), [jangidk@gmail.com](mailto:jangidk@gmail.com)).

### How do I cite *K*-shuff for my research?

We are in the process of submitting the manuscript for publication. As soon as we have the acceptance, we will notify all users about the citation information.

### Please be Kind.

We are still testing this program, so if things don't work out let us know.



## References

1. Diggle, P.J., Chetwynd, A.G., 1991. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics* 47: 1155-1163.
2. Diggle, P.J., Gomez-Rubio, V., Brown, P.E., Chetwynd, A.G., Gooding, S., 2007. Second-order analysis of inhomogeneous spatial point processes using case-control data. *Biometrics* 63: 550-557.
3. Jangid, K., Williams, M.A., Franzluebbers, A.J., Sanderlin, J.S., Reeves, J.H., Jenkins, M.B., Endale, D.M., Coleman, D.C., Whitman, W.B. 2008. Relative impacts of land-use, management intensity and fertilization upon soil microbial community structure in agricultural systems. *Soil Biology & Biochemistry* 40: 2843-2853.
4. Lasher, C., Dyszynski, G., Everett, K., Edmonds, J., Ye, W., Sheldon, W., Wang, S., Joye, S.B., Moran, M.A., Whitman, W.B. 2009. The diverse bacterial community in intertidal, anaerobic sediments at Sapelo Island, Georgia. *Microbial Ecology* 58: 244-261.
5. Schloss, P.D., Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* 71: 1501-1506.

